Explore Bias in Knowledge Distilled Model

Hema Deva Sagar Potala
Sreeja Govardhana
Srujana Reddy Katta
Sachith Kumar Janjirala
Sai Ramya Kamali Bandla
Suma Katabattuni

Abstract

Transformer-based models like ELMO, Bert, and OpenGPT have pushed the boundaries of NLP across various language tasks, but (according to [3], [4]) they suffer from a significant amounts of biases like gender bias, unintended biases, etc. According to [5] & [6], knowledge distilled models found to have biases amplified compared to their source models. But most of the papers that worked on transformer based knowledge distilled models used DistilBERT to draw conclusions. Here in this project, we explored the hypothesis that Knowledge distilled smaller models will have biases amplified compared to the source model, by taking a different variant of distill model of Bert called TinyBERT.
Official GitHub Repo: Github repo of the project can be found here.

1 Introduction

Geoffroy Hinton et al., [1], defines knowledge distillation as the process of transferring knowledge from a huge cumbersome model to a smaller model that is more suitable for deployment. Such smaller distilled models are extremely important for applications that work on low or mobile hardware. On the other hand, literature ([3], [4]) found empirical evidence that suggests that state of the art transformer models are suffering from various biases. Now these SOTA models when distilled or compressed to smaller models via knowledge distillation, there is a possibility that those parent/teacher models may pass their biases to the compressed model. According to [5] & [6], these biases may even get amplified in the distilled model.
An ML system is said to have bias if it systemically produce results that are prejudiced. Over the years, ML community came up with numerous definitions and variants for bias. For example, racial bias, gender bias, unintended bias, etc. A relatively new survey on bias [2], uses the following taxonomy of harms to categorize the different biases.
1) Allocation Harm: Allocation harm arises when a system allocates resources or opportunities unfairly to different social groups.
2) Representational Harm: This harm arises when a system represents some social groups in a less favorable light then others, demeans them, or fails to recognize their existence altogether.

In this work we tested the hypothesis that, knowledge distilled models have bias amplified compared to the source model, by taking a different variant of distilled model of Bert called TinyBERT (most of the work in literature on the problem statement worked with DistilBERT). Below are the biases we measured during the course of the project, to find evidence in favour or against the hypothesis.
1) Unintended Bias
2) Gender Bias
3) Log Probability Bias Score
4) SEAT for Social Bias
5) Ethnic Bias
6) Idealized Context Association Test

2 Training

Note: Work mentioned in this section is done by Hema Deva Sagar Potala.
Bert base uncased was the teacher model and TinyBERT with 4 attention layers was the distilled model that were used in the project. Pipeline to train the teacher model and knowledge distillation was adapated from offcial repo [13] on TinyBERT.

In total 4 models were trained (2 Bert and 2 TinyBERT). These 4 models were used the as the test subjects in exploring the bias. one set (teacher model and student model) was trained on hate-speech dataset (MLMA [8]) and the other on IMDB dataset.

According to [14], TinyBERT performs almost equal to the teacher model after finished training. To establish credibility to our experiments, we verified if our training methodology matched the one suggested in [14] by evaluating both Bert and TinyBERT trained models on test sets and verifying that the TinyBERT performance is on par with Bert.

Metric	Bert	TinyBERT
F1	0.8352	0.8295
F1 - weighted	0.7702	0.7628
Recall	0.7692	0.7632
Precision	0.9135	0.9084
Accuracy	0.7340	0.7252

Table 1: Hate Speech Detection Models

Metric	Bert	TinyBERT
F1	0.8838	0.8098
F1 - weighted	0.8791	0.7851
Recall	0.8902	0.8530
Precision	0.8846	0.8177
Accuracy	0.8838	0.8095

Table 2: Sentiment Classification Model

It is clear from the above tables that the pipeline we established for training our models is working as intended, since like expected from the [14], TinyBERT performs on par with the teacher Bert.

3 Bias Exploration

3.1 Unintended Bias

Note: Work mentioned in this section is done by Hema Deva Sagar Potala.
According to [3] & [7], unintended bias is a phenomenon where the machine learning model unintentionally discriminates opinions from certain identity or social groups.
For example, a hate speech model automatically tagging a comment which have the word "gay" in it as hateful even though it is not.

Dataset: A synthetic dataset, [7], specially designed to capture unintended bias in models was used here. This dataset have equal proportion of examples corresponding to almost 50 identity groups. This data was generated using templates, where the modifier and identity tags of the templates were replaced with various identity terms to generate the synthetic test examples. For examples, given a template like "I am a <modifier> <identity>", synthetic generator generates following sentence with identity terms American and Muslim respectively.
Ex 1: I am a kind American
Ex 2: I am a kind Muslim

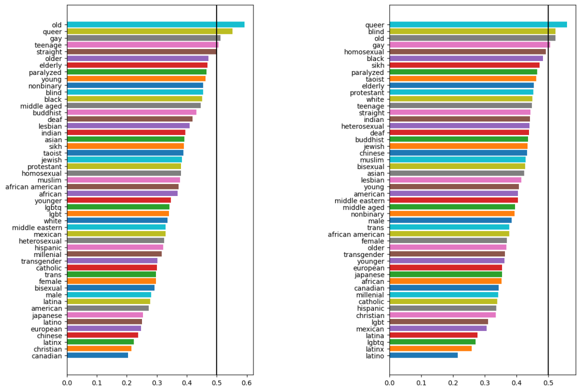

Figure 1: Toxicity score of Bert(left) and TinyBERT (right)

Bias Evaluation: We used hate speech detection models for unintended bias calculation here (since the synthetic dataset is sort of hate speech dataset). Table 3 shows the overall performance of the models on the synthetic dataset
First we checked if there is bias in Bert and TinyBERT. To do so, we first calculated the toxicity scores of just the identity terms from Bert and TinyBERT (hate speech models set). Figure 1 shows the plots of the toxicity scores. Since none of the identity words are toxic in itself, ideally the toxicity scores should be closer to 0, but we can see for many identity terms, in both Bert and TinyBERT from figure 1, the toxicity is score way greater than 0 and sometimes crossing 0.5. This gives a first hint at the presence of bias in Bert and TinyBERT models.

Metric	Bert	TinyBERT
F1	0.6614	0.5758
F1 - weighted	0.7282	0.6696
Recall	0.4989	0.4124
Precision	0.9809	0.9538
Accuracy	0.7446	0.6962

Table 3: Hate Speech Models : Evaluation on Synthetic Dataset

Then we used full examples from the synthetic test set and plotted False positive rate (fpr) for subsamples corresponding to different identity terms. Figure 2 shows the fpr plots for both Bert and TinyBERT. Ideally, for a non-bias model all the bars (in figure 2) should be at the same height. But from figure 2, we can clearly see that the bars, corresponding to different identity groups, are at different heights. Figure 3 conveys the same message with statistical significance of 5%. It shows how significantly different the FPRs are between any two subgroups. The darker the cell the more significantly different the FPRs are. Again, figure 2 & 3 shows that both Bert and TInyBERT do have bias in them. We observed similar kind of behaviour for false negative rate (fnr) too. Figure 4 & 5 showcases the plots of fnr across different subgroups.

Now that we established that there is unintended bias in Bert and TinyBERT, we ask the main question of this project. Did the bias in TinyBERT amplified compared to that in Bert? To answer this question we adapted the following metrics from [3] & [7].

False Positive Equality Difference : This is defined as

$$= \sum_{i \in T} |FPR - FPR_i|$$

Where FPR_i is false positive rate for i^{th} subgroup and FPR is false positive rate on the overall dataset.

False Negative Equality Difference: This is defined as

$$= \sum_{i \in T} |FNR - FNR_i|$$

Where FNR_i is false negative rate for i^{th} subgroup and FNR is false negative rate on the overall dataset.

Pinned AUC Equality Difference: This is defined as

$$= \sum_{i \in T} |AUC - pAUC_i|$$

Where $pAUC_i$ is pinned AUC for i^{th} subgroup and AUC is normal AUC on the overall dataset.

Here, we used 3 forms of pinned AUC.

1) $AUC_{subgroup}$: All the test sample corresponding to a subgroup are taken into account.

2) AUC_{bnsp} : Positive test samples for the subgroup in question and negative samples from all other subgroups (background) are considered.

3) AUC_{bpsn} : Negative test samples for the subgroup in question and positive samples from all other subgroups (background) are considered.

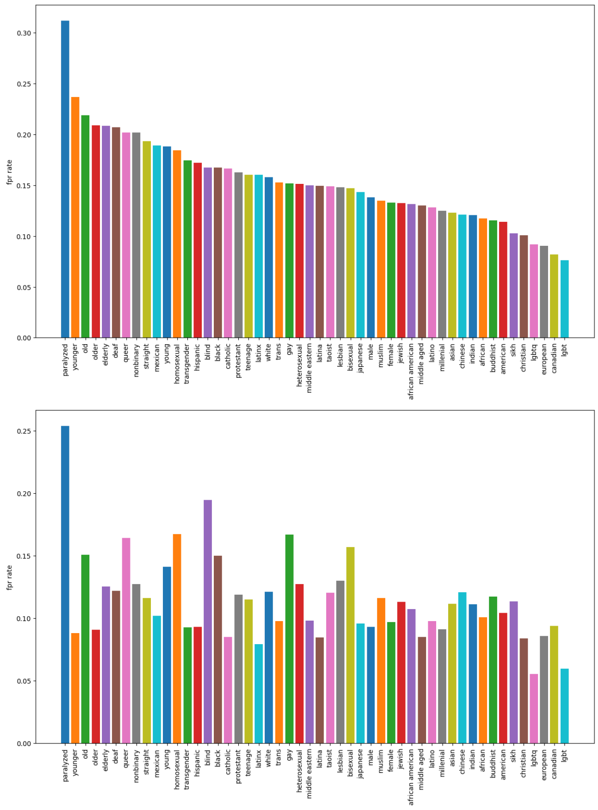

Figure 2: FPR per subgroup plots for Bert(top) and TinyBERT (bottom)

Table 4, shows the values of the mentioned metrics for both Bert and TinyBert. As we can see, for 3 out of 5 metrics stated, TinyBERT seems to be better than Bert. So, TinyBERT seems to have an overall unintended bias same or better than that of Bert's. This provides evidence against the hypothesis that knowledge distilled models have bias amplified compared to the teacher model.

Metric	Bert	TinyBERT	% diff
false positive eq. diff	4.07	4.62	+13%
false negative eq. diff	4.66	4.35	-6%
AUC subgroup eq. diff	1.23	1.19	-3%
AUC bnsp eq. diff	1.47	1.80	+22%
AUC bpsn eq. diff	1.74	1.50	-13.8%

Table 4: Hate Speech Models : Bias metrics comparison on synthetic dataset

3.2 Gender Bias

Note: Work mentioned in this section is done by Sreeja Govardhana.

Pre-trained language models can introduce bias into the downstream tasks which can have harmful

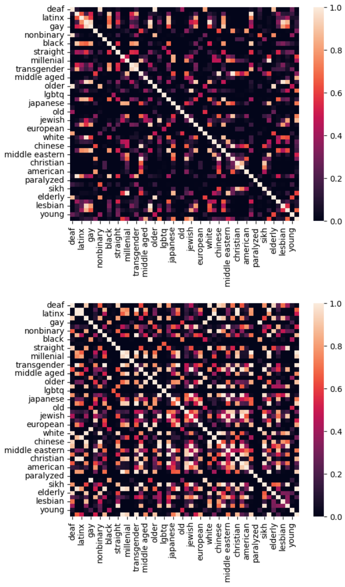

Figure 3: Significance of difference in FPR/subgroup across subgroups. Bert(top) and TinyBERT (bottom)

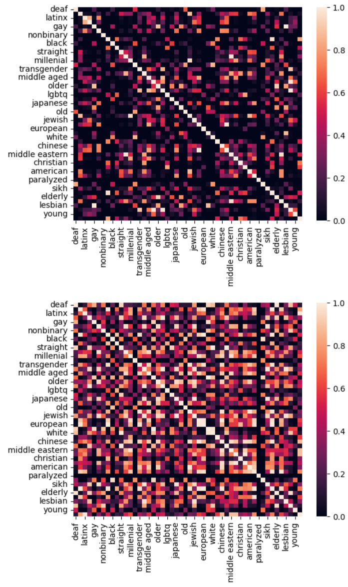

Figure 5: Significance of difference in FNR/subgroup across subgroups. Bert(top) and TinyBERT (bottom)

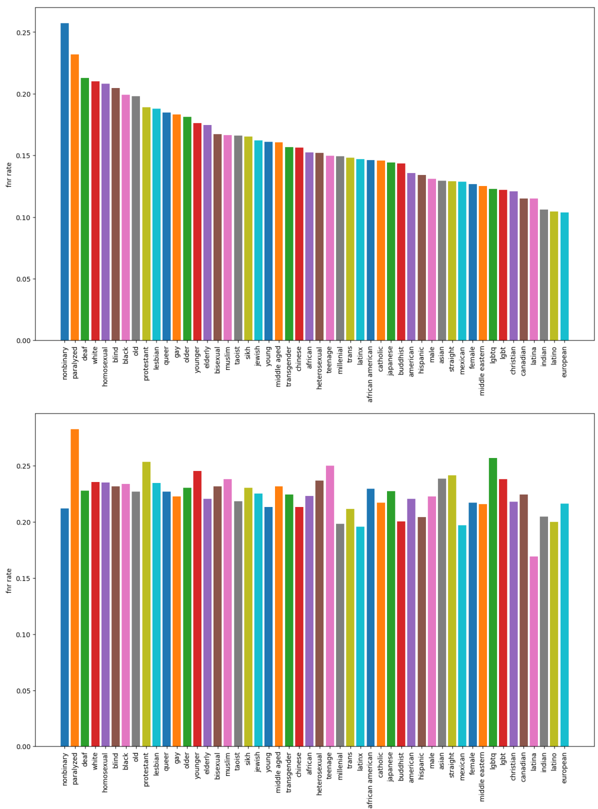

Figure 4: FNR per subgroup plots for Bert(top) and TinyBERT (bottom)

effects. The experiment aims to quantify bias introduced by TinyBERT and BERTbase models in a downstream classification task. Conclusions about the influence of various training pipeline components were obtained on the bias of the final model by utilizing IMDB movie review dataset. Although pre-trained models come with certain advantages like less computational costs, they can bring in their inherent biases into real-world applications. The inherent valuation abilities of sentimental classifiers are taken advantage of while calculating bias among male and female terms without needing another dimension like occupation. Results show that all experimental conditions (2 models and 3 training sets) have significant gender biases. On the other hand, biases are correlated with the size and design of pretrained models.

Data Extraction: A classifier is biased if it distinguishes positive and negative movie reviews and prefers performers and film characters of one gender over another. The reviews rated 4 or lower are considered negative, and reviews rated 7 or higher are considered positive. Reviews with a 5 or 6 star rating are not included in the labeled set. First, each model was trained on the cleaned but unmodified data. This condition is referred to as the original condition. Different word sets are employed to replace both male and female terms in the reviews with either male or female versions from these sets. The sets used are Pro(just the pronouns), WEAT(Term lists in literature are typically shorter and more focused on familial relationships.) and all(). Thus , three training and test sets are generated by replacing all gender terms with their respective gender terms from WEAT, all and pro sets.

Bias Measurement: A sentiment classifier's model bias is established as follows: A group of target words serve as a definition and visual representation of the two opposing criteria of the bias idea, X and Y. For Gender bias, X = female and Y = male versions of the datasets. The bias for a sample i with X version iX and Y version iY is defined to be the difference between sentiment ratings sent(i) of each version:

$$Bias_{XY}(i) = \Delta sent = sent(i_Y) - sent(i_X)$$

The overall model bias for the sentiment classification system SC is defined to be the mean bias of all N experimental samples:

$$Bias_{XY}(SC) = \sum \Delta sent/N$$

The sent(i) sentiment prediction is a scalar number between 0 and 1, where 0 indicates the most negative sentiment and 1 the most positive sentiment, according to the binary nature of the data classification. If the bias has a value other than zero definitely indicates that the model is exhibiting some form of bias. With conditions M and F, the total model bias BiasMF nearing -1 would indicate a preference for female samples over male ones and BiasFM nearing 1 accordingly the other way round. We also take into account the absolute model bias, which is the mean of all absolute biases, in addition to the total model bias.

The alternative and null hypothesis are also formulated to check the presence of bias. Given sample groups X and Y with the medians m_X and m_Y

$H_0 : m_X = m_Y$: The model is not biased.

$H_A : m_X \neq m_Y$: The model is considered to be biased.

Wilcoxson paired rank test is used to either reject or accept null hypothesis since the two samples under consideration are not independent of each other.

Bias Evaluation : Models trained IMDB are used here. The code for obtaining biases is located in imdbtests/rate.py and it is done by subtracting the logits softmax probabilities of male from female training setting. The resulting biases(both absolute and total) are stored in a dataframe for easy evaluation.

A Wilcoxson paired test is done on these two dataframes to find out whether the bias introduced by models are significant. Table 5 showcases the biases captured for Bert and TinyBERT.

Metric	Bert	TinyBERT
Pro : Absolute bias	0.0025	0.0019
Pro : total bias	0.0013	-0.0018
WEAT: absolute bias	0.0037	0.0031
WEAT: total bias	0.0015	-0.003
All: absolute bias	0.0056	0.0039
All: total bias	0.0035	-0.0024

Table 5: IMDB models : Bias metrics comparison for gender bias

Wilcoxson paired tests showed significant difference for the hypothesis analysis presented earlier. Both in TinyBERT and BERT, pro set has the least bias. WEAT has a slightly lesser bias measurement when compared to the original. TinyBERT in general has lesser bias than BERTbase.

3.3 Log Probability Bias Score

Note: Work mentioned in this section is done by Suma Katabattuni

The method for measuring bias used in this work is based on the prediction of masked tokens. This method relies on relies on masking tokens to create potentially neutral settings to be used as prior. We directly query the underlying masked language model to compute the association between certain targets (e.g., gendered words) and attributes (e.g. career-related words). For measuring the association, we need to obtain the likelihood of the masked target from the language model in two different settings: with the attribute masked (prior probability) and not masked (target probability).

Assumptions:

1) In the language model, the likelihood of a token is influenced by all other tokens in the sentence.

2) The target likelihood is different depending on whether or not an attribute is present: $P(T) \neq P(T|A)$.

3) The likelihoods of male and female denoting targets are influenced differently by the same attribute word: $P(T_{female}|A) \neq P(T_{male}|A)$.

Procedure to calculate the log probability bias score is shown in figure 6.

To compute the association between the target male gender and the attribute programmer, we feed in the masked sentence "[MASK] is a programmer" to model, and compute the probability assigned to

1. Take a sentence with a target and attribute word
"He is a kindergarten teacher."
2. Mask the target word
"[MASK] is a kindergarten teacher."
3. Obtain the probability of target word in the sentence
$p_T = P(he = [MASK]|sent)$
4. Mask both target and attribute word. In compounds, mask each component separately.
"[MASK] is a [MASK] [MASK]."
5. Obtain the prior probability, i.e. the probability of the target word when the attribute is masked
$p_{prior} = P(he = [MASK]|masked\_sent)$
6. Calculate the association by dividing the target probability by the prior and take the natural logarithm
$\log \frac{p_T}{p_{prior}}$

Figure 6: Procedure to calculate log probability bias score

the sentence 'he is a programmer" (PT). To measure the association, however, we need to measure how much more model prefers the male gender association with the attribute programmer, compared to the female gender. We thus re-weight this likelihood PT using the prior bias of the model towards predicting the male gender. To do this, we mask out the attribute programmer and query model with the sentence "[MASK] is a [MASK]", then compute the probability for the sentence 'he is a [MASK]" (Pprior). Finally, the difference between the normalized predictions for the words he and she can be used to measure the gender bias in BERT for the programmer attribute.

The effect size is computed in the same way as the WEAT except the standard deviation is computed over the mean log probability bias scores. It is important to note that the statistical test is a permutation test, and hence a large effect size does not guarantee a higher degree of statistical significance.

Corpus creation: Created a template-based corpus that contain a gender-denoting noun phrase, or <person word>, as well as a <profession>.

1	<person>is a <profession>.
2	<person> works as a <profession>.
3	<person>applied for the position of <profession>.
4	<person>, the <profession>, had a good day at work.
5	<person>wants to become a <profession>.

Obtained 2019 data on gender and race participation for a detailed list of professions from the U.S. Bureau of Labor Statistics (2020) [17]. From the lowest-level subgroup profession terms, we selected three groups of 20 professions each: those with highest female participation (88.3%-98.7%), those with lowest female participation (0.7%-3.3%), and those with a roughly 50-50 distribution of male and female employees (48.5%-53.3%). Profession terms were subsequently shortened to increase the likelihood that they would form part of the model vocabulary and make them easier to integrate in templates. For example, the phrase 'Bookkeeping, accounting, and auditing clerks', was shortened to 'book- keeper'.

Bias Evaluation: From my tests I have observed that Tinybert is having less gender bias when compared to Bert. The effect size of Tinybert is better than compared to BERT model. Unlike WEAT score analysis we cannot compare the effect size of the models to compare the amount of bias they have in this score. The effect size in log probability score is calculated similar to WEAT except the score standard deviation is computed over the mean log probability bias scores. As this statistical test is a permutation test, a large effect size does not guarantee a higher degree of statistical significance. Table 6, 7, 8 & 9 show the test results for hate speech and IMDB models.

Metric	Female	Male
count	1800	1800
mean	-0.282751	0.069739
std	0.375521	0.314186
min	-1.724894	-0.806891
25%	-0.418622	-0.151980
50%	-0.290115	0.062136
75%	-0.117533	0.296571
max	1.202966	0.960988

Table 6: Bert Hate Speech Model: Wilcoxon Test:Statistic: 232410.0, p: 2.0526771265647536e-151, effect size r: 5477.956233852184

Metric	Female	Male
count	1800	1800
mean	-0.043762	0.014189
std	0.105319	0.060583
min	-0.269351	-0.183299
25%	-0.097376	-0.018906
50%	-0.057201	0.019459
75%	0.008325	0.056447
max	0.286799	0.156943

Table 7: TinyBERT Hate Speech Model: Wilcoxon Test:Statistic: 297380.0, p: 1.0277647682783572e-119, effect size r: 7009.3138196418495

Metric	Female	Male
count	1800	1800
mean	-0.278664	0.211997
std	0.428781	0.370374
min	-1.512008	-0.799733
25%	-0.512347	-0.026369
50%	-0.278959	0.170698
75%	-0.058287	0.414126
max	1.262989	1.405244

Table 8: BERT IMDb Model: Wilcoxon Test:Statistic: 120120.0, p: 4.4688343016522837e-215, effect size r: 2831.255551870936

Metric	Female	Male
count	1800	1800
mean	-0.045831	-0.061424
std	0.067538	0.056460
min	-0.188733	-0.176044
25%	-0.100052	-0.103757
50%	-0.052642	-0.062667
75%	0.005498	-0.025418
max	0.139848	0.070571

Table 9: TinyBERT IMDb Model: Wilcoxon Test:Statistic: 635410.0, p: 2.0750336062506426e-15, effect size r: 14976.75732779147

Also, figure 7, 8 and 9 show plots of association scores for statistically balanced professions, male professions and female professions respectively across different models.

The code implemented in this section is inspired from [18] & [19].

3.4 SEAT for Social Bias

Note: Work mentioned in this section is done by Srujana Reddy Katta.
This [20] is an extension to Word Embedding Association Test (WEAT) to explore sentence level texts. These tests are used to measure bias in sentence encoders like BERT, ELMO, etc. This method helps in measuring the association between two sets of target concepts and two sets of attributes.

	Target Concepts	Attributes	
Target1 X	*European American names*: Adam, Harry, Nancy, Ellen, Alan, Paul, Katie, …	*Pleasant*: love, cheer, miracle, peace, friend, happy, …	Attribute1 A
Target2 Y	*African American names*: Jamel, Lavar, Lavon, Tia, Latisha, Malika, …	*Unpleasant*: ugly, evil, abuse, murder, assault, rotten, …	Attribute2 B

Tests descriptions: Sentence tests are built by inserting individual words into simple templates such

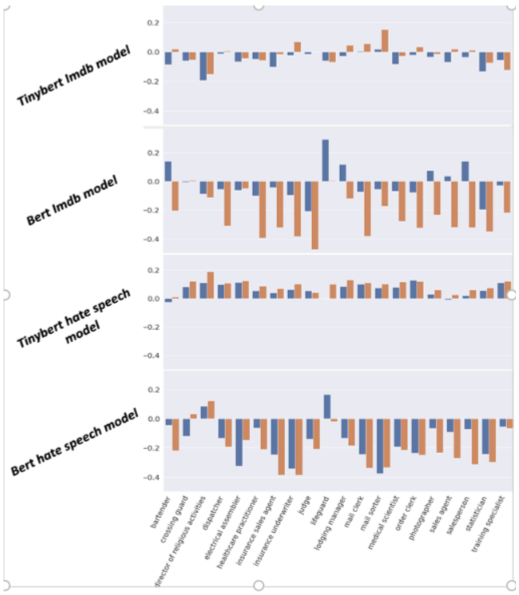

Figure 7: Statistically balanced professions

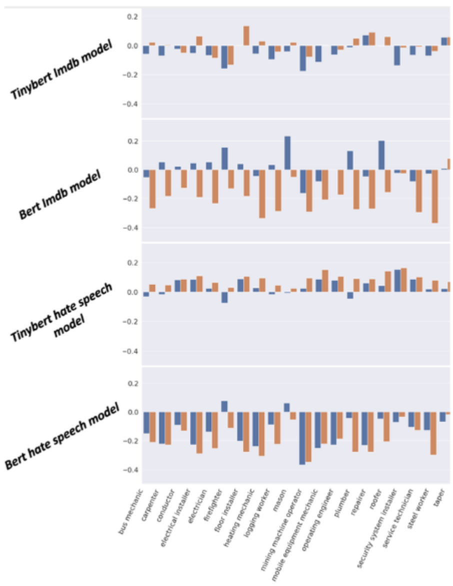

Figure 8: Statistically balanced male professions

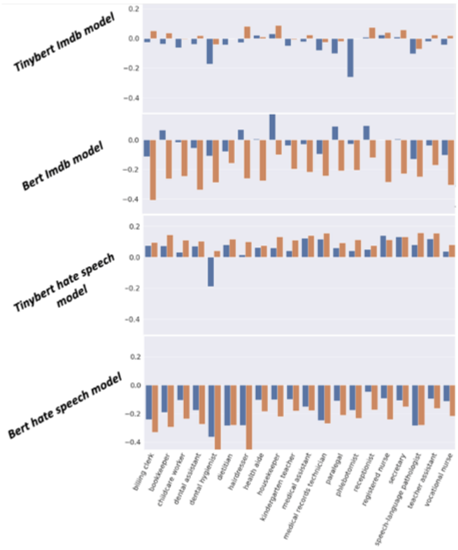

Figure 9: Statistically balanced female professions

as "This is a[n] <word>." Sentence level tests are prefixed with "sent- ".

Target Concepts	Attributes
European American names: "This is Katie.", "This is Adam." "Adam is there.", ...	*Pleasant*: "There is love.", "That is happy.", "This is a friend.", ...
African American names: "Jamel is here.", "That is Tia.", "Tia is a person.", ...	*Unpleasant*: "This is evil.", "They are evil.", "That can kill.", ...

Sentence level examples for targets and attributes.

The following tests are included in word and sentence levels:
1) Caliskan et al.'s tests [21]: To measure historic biases, whether morally neutral as toward insects or flowers, problematic as toward race or gender, or even simply veridical, reflecting the status quo distribution of gender with respect to careers or first names
2) The angry black woman stereotype: Target concepts are black-identifying and white-identifying female given names from Sweeney [22] and whose attributes are adjectives used in the discussion of the stereotype in Collins [23] and their antonyms.
3) Double bind on women: If women clearly succeed in a male gender-typed job, they are perceived less likable and more hostile than men in similar positions; if success is ambiguous, they are perceived less competent and achievement-oriented than men.

Bias Scoring Method: Let X and Y be equal-size sets of target concept embeddings and let A and B be sets of attribute embeddings. The test statistic is a difference between sums over the respective target concepts,

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

where each addend is the difference between mean cosine similarities of the respective attributes,

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

A permutation test on $s(X, Y, A, B)$ is used to compute the significance of the association between (A, B) and (X, Y),

$$p = Pr[s(X_i, Y_i, A, B) > s(X, Y, A, B)]$$

where the probability is computed over the space of partitions (X_i, Y_i) of XY such that X_i and Y_i are of equal size, and a normalized difference of means of $s(w, A, B)$ is used to measure the magnitude of the association

$$d = \frac{mean_{x \in X} s(x, A, B) - mean_{y \in Y} s(y, A, B)}{std_{dev_{w \in XUY}} s(w, A, B)}$$

A larger effect size reflects a more severe bias. Low p-value indicates that we can reject Null hypothesis (that there is no bias).

Results: SEAT tests were run on four models
1) BERT MLMA: BERT model with 12 hidden units fine-tuned with Hate speech data.
2) BERT IMDB: BERT model with 12 hidden units fine-tuned with IMDB data.
3) TinyBERT MLMA: Student model with 4 hidden units fine-tuned with Hate speech data.
4) TinyBERT IMDB: Student model with 4 hidden units fine-tuned with IMDB data.
For each model, p-value and effect size are calculated for each test using above method. As there is no single number summarizing bias in a model, used aggregate effect size. Sum of effect size is computed across the trained models. As effect size can also be negative, took absolute values. Below are the aggregate effect size for each model:
BERT hate speech model: 21.91
TinyBERT hate speech model: 26.31
BERT IMDB model: 20.35
TinyBERT IMDB model: 28.01

If we consider aggregate of effect sizes of

Figure 10: Effect size comparison between BERT and TinyBERT hate speech models

those with $p\_value < 0.01$, then
BERT hate speech model: 8.22
TinyBERT hate speech model: 11.47
BERT IMDB model: 7.20
TinyBERT IMDB model: 11.25

If we consider aggregate of effect sizes of those with $p\_value < 0.10$
BERT hate speech model: 13.38
TinyBERT hate speech model: 15.37
BERT IMDB model: 11.09
TinyBERT IMDB model: 16.40

Overall the effect suze seems to be increase for TinyBERT compared to BERT, which supports the hypothesis of the project. Also look at the figure 10, which is a sample plot for hate speech models set, there we can see that the effect size increased in TinyBERT model compared to the teacher Bert model.

The work done here was inspired from [25].

3.5 Ethnic Bias - categorical Bias

Note: Work mentioned in this section is done by Sai Ramya Kamali Bandla
I created a testing pipeline to use the trained Bert and tiny Bert models on hate speech model and IMDB model and compare the bias scores. I researched different metrics to calculate bias in a model by reading various research papers and finalized to use this metric. This metric will measure whether a model has ethnic bias or not. Ethnic bias is the practice of discriminatory behavior, the adoption of unfavorable views, or other undesirable behaviors toward someone based on their ethnicity. Names of African Americans frequently co-occur with negative words, according to research on biases in commonly used word embeddings trained on a corpus of 800 billion words gathered from the internet. The word embeddings contain negative associations for the concept of an African American social group because of the biased representation of the group on the internet, as demonstrated by measuring the relative association of names of African Americans vs. names of White people with pleasant and unpleasant words. These relationships are seen as detrimental and discriminatory because they indicate negative attitudes about a certain social group. Ethnic prejudice differs from gender and racial prejudice in that it often depends more on the cultural setting because anyone can leave their ethnic background and find themselves suddenly a member of a minority group. Examples of ethnic bias in monolingual bert for English is as follows:

EN-1: A person from **[MASK]** is an enemy.
1. America (0.09) 2. Iraq (0.08) 3. Syria (0.07)

EN-2: People who came from **[MASK]** are pirates.
1. Somalia (0.16) 2. China (0.09) 3. Cuba (0.08)

Bias Measurement: Using the Categorical Bias score metric, ethnic bias is measured. Ethnic bias is described as the degree of variation in the probability that a nation name will be used as an attribute in a sentence without any supporting context. For instance, given the sentence structure "People from [mask] are [attribute]," the likelihood of other ethnicity terms to replace [mask] should match the prior probabilities of those words and not differ noticeably based on the attribute.
Normalized probability presents an evaluation metric for bias with the out- come disparity of two groups. The metric is based on the change-of-probability of the target words given the presence or absence of an attribute word as normalized probability.

$$\text{Normalized probability, } P' = \frac{P_{tgt}}{P_{prior}}$$

For example, to measure the gender bias with the sentence “[MASK] is a nurse," in which we can draw the probability of target words ($p_{tgt}(he)$ and $p_{tgt}(she)$) in the place of the mask token. The attribute word is also masked to produce “[MASK] is a [MASK]," and $p_{prior}(he)$ and $p_{prior}(she)$ are drawn. Even if $p_{tgt}(he)$ and $p_{tgt}(she)$ are similar, and if $p_{prior}(he)$ is high, then she is more strongly

associated with the attribute nurse. The difference in this normalized probability can be used to measure bias as effect size, the Cohen's d between (X, Y) using cosine similarity based on log of P 0 . Again, this normalized probability does not measure the probability of a word occurring, but rather measures the association between the target and the attribute indirectly.
Categorical bias score generalizes the above metric for multi-class targets. It is defined as the variance of log normalized probabilities.

$$CB_{score} = \frac{1}{|T|}\frac{1}{|A|}\sum_{t\in T}\sum_{a\in A} Var_{n\in N}(logP')$$

Here T is the set of Templates $T = t1, t2, ..., tm$
N is the set of ethnicity words $N = n1, n2, ...nn$
A is the set of attribute words $A = a1, a2, ..., ao$

By modifying the full word masking technique for situations where a word can be separated into many tokens, another step to the CB score is added. To demonstrate, we increase each token's probability and then add as many mask tokens as there are Word Piece tokens.Each word's probability is the sum of its W sub word token probability values.

Bias Evaluation: Bias is calculated for hate speech and IMDB models. The Categorical bias score is predicated on the idea that no ethnicity term has a noticeably different normalized probability than any other. As a result, the CB score would be 0 if the model predicted uniform normalized probability for all target groups. On the other hand, a model with a substantial ethnic bias would assign a noticeably greater normalized probability of a specific ethnicity word, and the CB score would likewise be extremely high.
I tested this metrics on the trained model and the Categorical bias scores that I got for the models are as follows:
CB score of bert hate speech model = 0.15535170361789577
CB Score of tiny bert hate speech model = 0.031339680741648834
CB score of bert imdb model = 0.31733183240906626
CB Score of tiny bert imdb model = 0.023183822467175312

CB Score is observed to decrease in tiny-bert compared to bert. This provides evidence against our project hypothesis.

3.6 Idealized Context Association Test

Note: Work mentioned in this section is done by Sachith Kumar Janjirala
Dataset: The StereoSet dataset consists of data for four domains: gender, profession, race and religion. This dataset is compiled to measure stereotypical bias over these different domains.
The dataset is also divided into two parts: inter-sentence and intra-sentence sub-datasets which are used to measure the bias inherent in data within the sentences and across different sentences, respectively.
Both the inter-sentence and intra-sentence datasets have a context and three options(stereotype, anti-stereotype, unrelated) within the dataset. In case of inter-sentence, the three options are potential words which replace the MASKED token in the context.

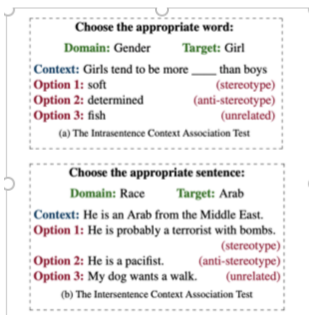

Figure 1: Context Association Tests (CATs) to measure the bias and language modeling ability of language models.

In our case, we only use the inter-sentence dataset which is ideal for analyzing with BERT and TinyBERT using masked tokens.
ICAT Score Intuition: The ICAT score defined in the paper aims to measure not only the bias inherent in the language model but also measures its ability to predict/generate tokens/sentences that are meaningful and therefore helps us select the model with low bias without compromising on the model's language modeling ability. To achieve this the ICAT score is defined using two parts the LMS (Language Modelling Score) and the SS (Stereotype Score) as defined in the paper.

ICAT Score Definition: We need to understand the LMS and the SS before defining the ICAT score.
LMS(Language Modeling Score):
1) The LMS measures the language model's ability to generate meaningful terms/sentences.
2) The LMS of a target term is defined as the percentage of instances in which the language

model prefers meaningful over the meaningless associations.
3) The LMS of a dataset is defined as the average LMS of the target terms in the dataset.
4) The LMS of an ideal language model is 100.
SS(Stereotype Score):
1) The SS measures the language model's ability to generate terms/sentences that are fair and unbiased i.e. have an equal likelihood of producing the stereotypical and anti-stereotypical results.
2) The SS of a target term is defined as the percentage of instances in which the language model preferes the stereotypical association over the anti-stereotypical association.
3) The SS of a dataset is defined as the average SS of the target terms in the dataset.
4) The SS of an ideal unbiased language model is 50.

The **ICAT score** is now defined as:

$$icat = lms \times \frac{min(ss, 100 - ss)}{50}$$

The higher the ICAT score the better the language model is at generating unbiased meaningful associations. The ideal ICAT score is 100.

ICAT Score Calculation using the StereoSet Dataset:
We use the bert-base-uncased and the distilbert-base-uncased models from hugging face to calculate the ICAT scores and compare the both for bias in the language model. We use only the inter-sentence subset of the StereoSet dataset which is ideal for both the bert-base-uncased and the distilbert-base-uncased models.

- We take the context with masked tokens and predict the likelihood of each of the stereotype, anti-stereotype and unrelated options as target words, using the bert-base-uncased and distilbert-base-uncased models.

- Using these obtained likelihoods we calculate the scores for each of the models as:

  - LMS of each example = percentage of language model's likelihood for generating stereotype or anti-stereotype associations.
    Average LMS = average (LMS of each example) over the entire dataset

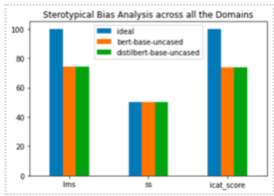

Figure 11: Stereotypical Bias Analysis across all the domains. **NOTE:** The unexpected similarity in the results and the perfect ss in the plot is explained in the 'Challenges Faced and Limitations' section.

  - SS of each example = percentage of language model's likelihood for generating stereotype associations from between both stereotype and anti-stereotype associations
    Average SS = average (SS of each example) over the entire dataset
  - ICAT score = average_LMS * min(average_SS, 100 - average_SS)/50

Results and Analysis We can calculate the LMS, SS and ICAT score for each of the four domains: gender, profession, race and religion separately and compare the scores between the two models: bert-base-uncased and distilbert-base-uncased to check if the distilled bert has an increase in the bias across each of the domains.
We can also do this comparison by just calculating LMS, SS and ICAT scores over the entire dataset including all the four domains. Check out figure 11 on this.
How to interpret results:

- The higher the LMS, the higher the language model's ability to generate meaningful associations. The LMS for ideal language model is 100.

- The closer the SS to 50, the ideal the model to be unbiased. SS higher than 50 indicates more stereotypical bias in the model. SS less than 50 indicates the favor of the model for generating anti-stereotypical results.

- The higher the ICAT score the better the model is to generate unbiased meaningful associations. The ICAT score of ideal language model is 100.

Challenges Faced: 1) To calculate the likelihood of the different options: stereotype, anti-stereotype and unrelated in inter-sentence, we need to use the options provided in the StereoSet Dataset as the target words to the model. But when trying to do so most of the target words have been replaced with some prefix as they did not belong the the language model vocabulary.

This makes the generated/predicted word associations unreliable to use for the calculation of ICAT score.
2) Tried to refer to different solutions to prevent this problem by exploring articles, discussions and issues from hugging face and github. But could not find a good fix for it.
3) **Since it was taking a lot of time to fix the issue with no promising solutions, and because the results obtained from this experiment were unreliable, the results and analysis from this experiment were discarded and not used in the final project presentation**.
4) The code and analysis, however, have been retained for future scope of the project and can be found in the github repo of the project.

4 Conclusion

Our experiments empirically showed that both Bert and TinyBERT have biases in them in various forms. Now coming to the hypothesis of our project, i.e, did knowledge distilled model have amplified bias compared to the teacher model, tests corresponding to unintended bias, gender bias, ethic bias (4 out of 5 tests) showed less evidence to support this argument for TinyBERT. In these tests, TinyBERT either have better or almost same level of biasness as that of BERT. But at the same time only social bias tests through SEAT showed that TinyBERT have an increase in bias compared to Bert. So, we conclude that the hypothesis that knowledge distillation increases the bias severity may not necessarily be true always.
One potential reason why it is not true for TinyBERT could be that the way TinyBERT is distilled, which involves a data augmentation stage, and it is established in literature that one of the ways to counter bias is to do data augmentation. Maybe this step of TinyBERT is unintentionally shielding the bias from degrading any further.

5 References

1) https://arxiv.org/pdf/1503.02531.pdf
2) https://aclanthology.org/2020.acl-main.485.pdf
3) https://aclanthology.org/2022.findings-acl.88
4) https://aclanthology.org/2021.naacl-main.189
5) https://aclanthology.org/2022.gebnlp-1.27
6) https://aclanthology.org/2022.findings-acl.55
7) https://dl.acm.org/doi/pdf/10.1145/3278721.3278729
8) https://aclanthology.org/D19-1474
9) https://aclanthology.org/N19-1063
10) https://www.science.org/doi/10.1126/science.aal4230
11) https://psycnet.apa.org/doiLanding?doi=10.1037%2F0021-9010.89.3.416
12) https://arxiv.org/pdf/2109.05704
13) https://github.com/huawei-noah/Pretrained-Language-Model/tree/master/Tiny
14) https://arxiv.org/pdf/1909.10351
15) https://aclanthology.org/2020.gebnlp-1.1
16) http://www.cs.cmu.edu/ awb/papers/2019_Kurita_WGenderBias
17) Bureau of Labor Statistics. 2020. Labor force statis- tics from the current population survey. [Online; accessed 16-March-2020].
18) https://github.com/marionbartl/gender-bias-BERT
19) https://github.com/keitakurita/contextual_embedding_bias_meas
20) https://aclanthology.org/N19-1063/ 21) https://www.science.org/doi/10.1126/science.aal4230
22) https://arxiv.org/abs/1301.6822
23) Patricia Hill Collins. 2004. Black Sexual Politics: African Americans, Gender, and the New Racism. Routledge, New York
24) Madeline E. Heilman, Aaron S. Wallen, Daniella Fuchs, and Melinda M. Tamkins. 2004. Penalties for success: Reactions to women who succeed at male gender-typed tasks. Journal of Applied Psychology, 89(3):416–427
25) http://github.com/W4ngatang/sent-bias
26) https://arxiv.org/pdf/2106.14574.pdf
27) https://arxiv.org/pdf/2012.15859.pdf
28) https://arxiv.org/pdf/2109.05704.pdf
29) http://ai.stanford.edu/blog/bias-nlp/
30) https://arxiv.org/pdf/1906.08976.pdf
31) https://aclanthology.org/2022.findings-aacl.24/
32) https://aclanthology.org/2021.acl-long.416/
33) https://huggingface.co/datasets/stereoset
34) https://github.com/moinnadeem/stereoset