BERT WEARS GLOVES: DISTILLING STATIC EMBEDDINGS FROM PRETRAINED CONTEXTUAL REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Contextualized word representations such as ELMo and BERT have become the de facto starting point for incorporating pretrained representations for downstream NLP tasks. In these settings, contextual representations have largely made obsolete their static embedding predecessors such as Word2Vec and GloVe. However, static embeddings do have their advantages in that they are straightforward to understand and faster to use. Additionally, embedding analysis methods for static embeddings are far more diverse and mature than those available for their dynamic counterparts. In this work, we introduce simple methods for generating static lookup table embeddings from existing pretrained contextual representations and demonstrate they outperform Word2Vec and GloVe embeddings on a variety of word similarity and word relatedness tasks. In doing so, our results also reveal insights that may be useful for subsequent downstream tasks using our embeddings or the original contextual models. Further, we demonstrate the increased potential for analysis by applying existing approaches for estimating social bias in word embeddings. Our analysis constitutes the most comprehensive study of social bias in contextual word representations (via the proxy of our distilled embeddings) and reveals a number of inconsistencies in current techniques for quantifying social bias in word embeddings. We publicly release our code and distilled word embeddings to support reproducible research and the broader NLP community.

1 INTRODUCTION

Word embeddings (Bengio et al., 2003; Collobert & Weston, 2008; Collobert et al., 2011) have been a hallmark of modern natural language processing (NLP) for several years. Pretrained embeddings in particular have seen widespread use and have experienced parallel and complementary innovations alongside neural networks for NLP. Advances in embedding quality in part have come from integrating additional information such as syntax (Levy & Goldberg, 2014b; Li et al., 2017), morphology (Cotterell & Schütze, 2015), subwords (Bojanowski et al., 2017), subcharacters (Stratos, 2017; Yu et al., 2017) and, most recently, context (Peters et al., 2018; Devlin et al., 2019). As a consequence of their representational potential, pretrained word representations have seen widespread adoption across almost every task in NLP and reflect one of the greatest successes of both representation learning and transfer learning for NLP (Ruder, 2019b).

The space of pretrained word representations can be partitioned into *static* vs. *dynamic* embeddings methods. Static methods such as Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and FastText (Bojanowski et al., 2017) yield representations that are fixed after training and generally associate a single vector with a given word in the style of a lookup table. While subsequent work addressed the fact that words may have multiple senses and should have different representations for different senses (Pilehvar & Collier, 2016; Lee & Chen, 2017; Pilehvar et al., 2017; Athiwaratkun & Wilson, 2017; Camacho-Collados & Pilehvar, 2018), fundamentally these methods cannot easily adapt to the inference time context in which they are applied. This contrasts with contextual, or dynamic, methods such as CoVe (McCann et al., 2017), ELMo (Peters et al., 2018), and BERT (Devlin et al., 2019), which produce vector representations for a word conditional on the inference time context in which it appears. Given that dynamic representations are arguably more linguistically valid,

more expressive (static embeddings are a special-case of dynamic embeddings that are optimally ineffective at being dynamic), and have yielded significant empirical improvements (Wang et al., 2019b;a; Ruder, 2019a), it would seem that static embeddings are outdated.

Static embeddings, however, have significant advantages over dynamic embeddings with regard to speed, computational resources, and ease of use. These benefits have important implications for time-sensitive systems, resource-constrained settings (Shen et al., 2019) or environmental concerns (Strubell et al., 2019), and broader accessibility of NLP technologies[1]. As a consequence of this dichotomy between static and dynamic representations and their disparate benefits, we propose in this work a simple yet effective mechanism for converting from dynamic representations to static representations. We begin by demonstrating that our method when applied to pretrained contextual models (BERT, GPT-2, RoBERTa, XLNet, DistilBERT) yields higher quality static embeddings than Word2Vec and GloVe when evaluated intrinsically on four word similarity and word relatedness datasets. Further, since our procedure does not rely on specific properties of the pretrained contextual model, it can be applied as needed to generate ever-improving static embeddings that will track advances in pretrained contextual word representations. Our approach offers the hope that high-quality embeddings can be maintained in both settings given their unique advantages and appropriateness in different settings.

At the same time, we show that by distilling static embeddings from their dynamic counterparts, we can then employ the more comprehensive arsenal of embedding analysis tools that have been developed in the static embedding setting to better understand the original contextual embeddings. As an example, we employ methods for identifying gender, racial, and religious bias (Bolukbasi et al., 2016; Garg et al., 2018; Manzini et al., 2019) to our distilled representations and find that these experiments not only shed light on the properties of our distilled embeddings for downstream use but can also serve as a proxy for understanding existing biases in the original pretrained contextual representations. Our large-scale and exhaustive evaluation of bias further reveals dramatic inconsistencies in existing measures of social bias and highlights sizeable discrepancies in the bias estimates obtained for distilled embeddings drawn from different pretrained models and individual model layers.

2 BACKGROUND

In this work, we study pretrained word embeddings, primarily of the static variety. As such, we focus on comparing our embeddings against existing pretrained static embeddings that have seen widespread adoption. We identify Word2Vec and GloVe as being the most prominent static embeddings currently in use and posit that these embeddings have been frequently chosen not only because of their high quality representations but also because lookup tables pretrained on large corpora are publicly accessible and easy to use. Similarly, in considering contextual models to distill from, we begin with BERT as it has been the most prominent in downstream use among the growing number of alternatives (e.g. ELMo (Peters et al., 2018), GPT (Radford et al., 2018), BERT (Devlin et al., 2019), Transformer-XL (Dai et al., 2019), GPT-2 (Radford et al., 2019), XLNet (Yang et al., 2019), RoBERTa (Liu et al., 2019b), and DistilBERT(Sanh, 2019)) though we provide similar analyses for several of the other models (GPT-2, XLNet, RoBERTa, DistilBERT) and more comprehensively address them in the appendices. We primarily report results for the `bert-base-uncased` model and include complete results for the `bert-large-uncased` model in the appendices as well.

3 METHODS

In order to use a contextual model like BERT to compute a single context-agnostic representation for a given word w, we define two operations. The first is *subword pooling*: the application of a pooling mechanism over the subword representations generated for w in context c to compute a single representation for w in c, i.e. $\{\mathbf{w}_c^1, \ldots, \mathbf{w}_c^k\} \mapsto \mathbf{w}_c$. Beyond this, we define *context combination* to be the mapping from representations $\mathbf{w}_{c_1}, \ldots, \mathbf{w}_{c_n}$ of w in different contexts $c_1, \ldots, c_n$ to a single static embedding $\mathbf{w}$ that is agnostic of context.

[1] A recent account from the perspective of a humanist about the (in)accessibility of BERT: `https://tedunderwood.com/2019/07/15/do-humanists-need-bert/`

3.1 SUBWORD POOLING

The tokenization procedure for BERT can be decomposed into two steps: performing a simple word-level tokenization and then potentially deconstructing a word into multiple subwords, yielding $w^1, \ldots, w^k$ such that $cat(w^1, \ldots, w^k) = w$ where $cat(\cdot)$ indicates concatenation. In English, the subword tokenization algorithm is WordPiece (Wu et al., 2016). As a consequence, the decomposition of a word into subwords is the same across contexts and the subwords can be unambiguously associated with their source word. Therefore, any given layer of the model outputs vectors $\mathbf{w}_c^1, \ldots, \mathbf{w}_c^k$. We consider four potential pooling mechanisms to compute $\mathbf{w}_c$ given these vectors:

$$\mathbf{w}_c = f(\mathbf{w}_c^1, \ldots, \mathbf{w}_c^k);\ f \in \{\min, \max, \text{mean}, \text{last}\} \tag{1}$$

$\min(\cdot)$ and $\max(\cdot)$ are element-wise min and max pooling, $\text{mean}(\cdot)$ indicates mean pooling, i.e. $\underset{x \in \mathcal{X}}{\text{mean}}\, g(x) = \frac{\sum_{x \in \mathcal{X}} g(x)}{|\mathcal{X}|}$ and $\text{last}(\cdot)$ indicates selecting the last vector, $\mathbf{w}_c^k$.

3.2 CONTEXT COMBINATION

In order to convert contextual representations into static ones, we describe two methods of specifying contexts $c_1, \ldots, c_n$ and then combining the resulting representations $\mathbf{w}_{c_1}, \ldots, \mathbf{w}_{c_n}$.

Decontextualized - For a word w, we use a single context where $c_1 = w$. That is, we feed the single word w by itself into the pretrained contextual model and consider the resulting vector to be the representation (applying subword pooling if the word is split into multiple subwords).

Aggregated - Observing that the **Decontextualized** strategy may be presenting an unnatural input to the pretrained encoder which may have never encountered w by itself without a surrounding phrase or sentence, we instead consider ways of combining the representations for w in multiple contexts. In particular, we sample n sentences from a large corpus $\mathcal{D}$, each of which contains the word w, and compute the vectors $\mathbf{w}_{c_1}, \ldots, \mathbf{w}_{c_n}$. Then, we apply a pooling strategy to yield a single representation that aggregates the representations across the n contexts as is shown in Equation 2.

$$\mathbf{w} = g(\mathbf{w}_{c_1}, \ldots, \mathbf{w}_{c_n});\ g \in \{\min, \max, \text{mean}\} \tag{2}$$

4 REPRESENTATION QUALITY

To assess the representational quality of our static embeddings, we evaluate on several word similarity and word relatedness datasets (see §A.2 for additional commentary). We consider 4 such datasets: RG65 (Rubenstein & Goodenough, 1965), WS353 (Agirre et al., 2009), SIMLEX999 (Hill et al., 2015) and SIMVERB3500 (Gerz et al., 2016). Taken together, these datasets contain 4917 examples and contain a vocabulary $\mathcal{V}$ of 2005 unique words. Each example is a pair of words (w_1, w_2) with a gold-standard annotation (provided by one or more humans depending on the dataset) of how semantically similar or how semantically related w_1 and w_2 are. A word embedding is evaluated by the relative correctness of its ranking of the similarity/relatedness of all examples in a dataset with respect to the gold-standard ranking using the Spearman ρ coefficient. Embedding predictions are computed using cosine similarity as in Equation 3:

$$\cos(w_1, w_2) = \frac{\mathbf{w_1} \cdot \mathbf{w_2}}{\|\mathbf{w_1}\| \, \|\mathbf{w_2}\|} \tag{3}$$

4.1 INTRINSIC EVALUATION

We begin by studying how the choices of f and g[2] impact the performance of embeddings distilled from `bert-base-uncased`. In Figure 1, we show the performance on all four datasets of the resulting static embeddings where embeddings computed using the **Aggregated** strategy are pooled over $N = 100000$ sentences. Here, N is the number of total contexts for all words (see §A.4). Across all four datasets, we see that $g = \text{mean}$ is the best performing pooling mechanism within the **Aggregated** strategy and also outperforms the **Decontexualized** strategy by a substantial margin.

[2]For brevity, we treat **Decontextualized** as a choice for g and denote it as *decont* in the figures. Additional shorthand is described in Appendix H.

Fixing $g = \text{mean}$, we further observe that mean pooling at the subword level also performs best. We further find that this trend that $f = \text{mean}, g = \text{mean}$ is optimal among the 16 possible pairs consistently holds for almost all pretrained contextual models we considered.

If we further consider the impacts of N as shown in Table 1, we see that performance for both `bert-base-uncased` and `bert-large-uncased` tends to steadily increase for all datasets with increasing N (and this trend holds for the 7 other pretrained models). In particular, in the largest setting with $N = 1000000$, the `bert-large-uncased` embeddings distilled from the best performing layer for each dataset dramatically outperform both Word2Vec and GloVe. However, this can be seen as an unfair comparison given that we are selecting the layer for specific datasets. As the middle band of table shows, we can fix a layer and still outperform both Word2Vec and Glove.

Beyond the benefits of using a larger N, Table 1 reveals an interesting relationship between N and the best-performing layer. In Figure 1, there is a clear preference towards the first quarter of the model's layers (layers 0-3) with a sharp drop-off in performance immediately thereafter (we see a similar preference for the first quarter in models with a different number of layers, e.g. Figure 3, Figure 10) . Given that our intrinsic evaluation is centered on lexical semantic understanding, this appears to be largely consistent with the findings of Liu et al. (2019a); Tenney et al. (2019). However, as we pool over a larger number of contexts, we see that the best-performing layer monotonically (with a single exception) shifts to be later and later within the pretrained model. What this indicates is that since the later layers did not perform better for smaller values of N, these layers demonstrate greater variance with respect to the layer-wise distributional mean and reducing this variance helps in our evaluation[3]. This may have implications for downstream use, given that later layers of the model are generally preferred by downstream practitioners (Zhang et al., 2019) and it is precisely these layers where we see the greatest variance. Accordingly, combining our stable static embeddings from layer ℓ with the contextual example-specific embeddings also from layer ℓ of the pretrained model as was suggested in Peters et al. (2018) may be a potent strategy in downstream settings. In general, we find these results suggest there may be merits towards further work studying the unification of static and dynamic methods.

Along with a trend towards later layers for larger values of N, we see a similar preference towards later layers as we consider each column of results from left to right. In particular, while the datasets are ordered chronologically[4], each dataset was explicitly introduced as an improvement over its predecessors (perhaps transitively, see §A.3). While it is unclear from our evaluation as to what differences in the examples in each dataset may cause this behavior, we find this correlation with dataset difficulty and layer-wise optimality to be intriguing. In particular, we see that SIMVERB3500 which contains verbs primarily (as opposed to nouns or adjectives which dominate the other datasets) tends to yield the best performance for embeddings distilled from the intermediary layers of the model (most clear for `bert-large-uncased`).

Remarkably, we find that most tendencies we observe generalize well to all other pretrained models we study (specifically the optimality of $f = \text{mean}, g = \text{mean}$, the improved performance for larger N, and the layer-wise tendencies with respect to N and dataset). In Table 2, we summarize the results of all models employing the **Aggregated** strategy with $f = \text{mean}, g = \text{mean}$ and $N = 100000$ contexts. Surprisingly, despite the fact that many of these models perform approximately equally on many downstream evaluations, we observe that their corresponding distilled embeddings perform radically differently even when the same distillation procedure is applied. These results can be interpreted as suggesting that some models learn better lexical semantic representations whereas others learn other behaviors such as context representation and semantic composition more accurately. More generally, we argue that these results warrant reconsideration of analyses performed on only one pretrained model as they may not generalize to other pretrained models even when the models considered have (nearly) identical Transformer architectures. A noteworthy result in Table 2 is that of DistilBert-6 which outperforms BERT-12 on three out of the four datasets despite being distilled using knowledge distillation (Ba & Caruana, 2014; Hinton et al., 2015) from BERT-12. Analogously, RoBERTa, which was introduced as a direct improvement over BERT, does not reliably outperform the corresponding BERT models when comparing the derived static embeddings.

[3] Shi et al. (2019) concurrently proposes a different approach with similar motivations.

[4] Incidentally, they also are ordered by dataset size. However, we do not believe this explains the layer-wise trends.

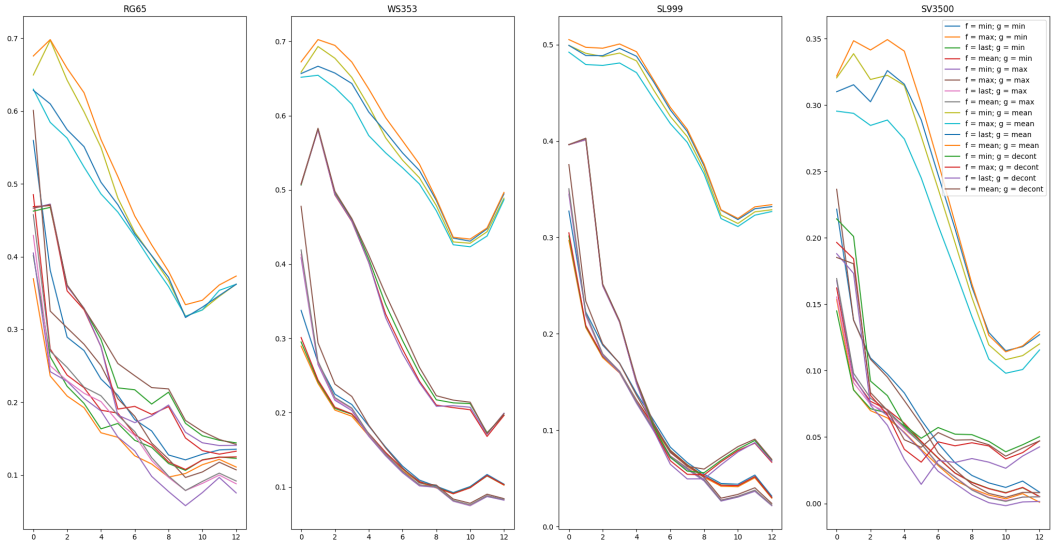

Figure 1: Layer-wise performance of distilled BERT-12 embeddings for all possible choices of f, g with $N = 100000$.

Model	N	RG65	WS353	SIMLEX999	SIMVERB3500
Word2Vec	-	0.6787	0.6838	0.4420	0.3636
GloVe	-	0.6873	0.6073	0.3705	0.2271
BERT-12 (1)	500000	0.7206	0.7038	0.5019	0.3550
BERT-24 (1)	500000	0.7367	0.7074	0.5114	0.3687
BERT-24 (6)	500000	0.7494	0.7282	0.5116	0.4062
BERT-12	10000	0.5167 (1)	0.6833 (1)	0.4573 (1)	0.3043 (1)
BERT-12	100000	0.6980 (1)	0.7023 (1)	0.5007 (3)	0.3494 (3)
BERT-12	500000	0.7262 (2)	0.7038 (1)	0.5115 (3)	0.3853 (4)
BERT-12	1000000	0.7242 (1)	0.7048 (1)	0.5134 (3)	0.3948 (4)
BERT-24	100000	0.7749 (2)	0.7179 (6)	0.5044 (1)	0.3686 (9)
BERT-24	500000	0.7643 (2)	0.7282 (6)	0.5116 (6)	0.4146 (10)
BERT-24	1000000	**0.7768** (2)	**0.7301** (6)	**0.5244** (15)	**0.4280** (10)

Table 1: Performance of distilled BERT embeddings on word similarity and word relatedness tasks. f and g are set to mean and (#) indicates the layer the embeddings are distilled from. **Bold** indicates best performing embeddings for a given dataset of those depicted.

Model	RG65	WS353	SIMLEX999	SIMVERB3500
BERT-12	0.6980 (1)	0.7023 (1)	0.5007 (3)	0.3494 (3)
BERT-24	**0.7749** (2)	**0.7179** (6)	0.5044 (1)	0.3686 (9)
GPT2-12	0.5156 (1)	0.6396 (0)	0.4547 (2)	0.3128 (6)
GPT2-24	0.5328 (1)	0.6830 (0)	0.4505 (3)	0.3056 (0)
RoBERTa-12	0.6597 (0)	0.6915 (0)	0.5098 (0)	0.4206 (0)
RoBERTa-24	0.7087 (7)	0.6563 (6)	0.4959 (0)	0.3802 (0)
XLNet-12	0.6239 (1)	0.6629 (0)	0.5185 (1)	0.4044 (3)
XLNet-24	0.6522 (3)	0.7021 (3)	**0.5503** (6)	**0.4545** (3)
DistilBERT-6	0.7245 (1)	0.7164 (1)	0.5077 (0)	0.3207 (1)

Table 2: Performance of static embeddings from different pretrained models on word similarity and word relatedness tasks. f and g are set to mean for all models, $N = 100000$, and (#) indicates the layer the embeddings are distilled from. **Bold** indicates best performing embeddings for a given dataset of those depicted.

5 BIAS

Bias is a complex and highly relevant topic in developing representations and models in machine learning and natural language processing. In this context, we study the social bias encoded within static word representations. As Kate Crawford argued for in her NIPS 2017 keynote, while studying individual models is important given that specific models may propagate, accentuate, or diminish biases in different ways, studying the representations that serve as the starting point and that are shared across models (which are used for possibly different tasks) allows for more generalizable understanding of bias (Barocas et al., 2017).

In this work, we simultaneously consider multiple axes of social bias (i.e. gender, race, and religion) and multiple proposed methods for computationally quantifying these biases. We do so precisely because we find that existing NLP literature has primarily prioritized gender (which may be a technically easier setting) and because we find that different computational specifications of bias that evaluate the same social phenomena yield different results. As a direct consequence, we strongly caution that the results should be taken with respect to the definitions of bias being applied. Further, we note that an embedding which receives low bias scores cannot be assumed to be (nearly) unbiased, rather that under existing definitions the embedding exhibits low bias and perhaps additional more nuanced definitions are needed.

5.1 DEFINITIONS

Bolukbasi et al. (2016) introduced a definition for computing gender bias which assumes access to a set $\mathcal{P} = \{(m_1, f_1), \ldots, (m_n, f_n)\}$ of (male, female) word pairs where m_i and f_i only differ in gender (e.g. 'men' and 'women'). They compute a gender direction $\vec{g}$:

$$\vec{g} = PCA([E(m_1) - E(f_1); \ldots; E(m_n) - E(f_n)][0] \tag{4}$$

where $E(\cdot)$ is the embedding function, ";" indicates horizontal concatenation/stacking and $[0]$ indicates taking the first principal component.

Then, given a set $\mathcal{N}$ of target words that we are interested in evaluating the bias with respect to, Bolukbasi et al. (2016) specifies the bias as:

$$\underset{\text{BOLUKBASI}}{\text{bias}}(\mathcal{N}) = \underset{w \in \mathcal{N}}{\text{mean}} \, |\cos(E(w), \vec{g})| \tag{5}$$

This definition is only inherently applicable to binary bias settings, i.e. where there are exactly two *protected classes*, but still is difficult to apply to binary settings beyond gender as constructing a set $\mathcal{P}$ can be challenging. Similarly, multi-class generalizations of this bias definition are also difficult to propose due to the issue of constructing k-tuples that only differ in the underlying social attribute. This definition also assumes the first principal component is capable of explaining a large fraction of the variance.

Garg et al. (2018) introduced a different definition for computing binary bias that is not restricted to gender, which assumes access to sets $\mathcal{A}_1 = \{m_1, \cdots, m_n\}$ and $\mathcal{A}_2 = \{f_1, \cdots, f_{n'}\}$ of representative words for each of the two protected classes. For each class, $\mu_i = \text{mean}_{w \in \mathcal{A}_i} E(w)$ is computed. Garg et al. (2018) computes the bias in the following ways:

$$\underset{\text{GARG-EUC}}{\text{bias}}(\mathcal{N}) = \underset{w \in \mathcal{N}}{\text{mean}} \, \|E(w) - \mu_1\|_2 - \|E(w) - \mu_2\|_2 \tag{6}$$

$$\underset{\text{GARG-COS}}{\text{bias}}(\mathcal{N}) = \underset{w \in \mathcal{N}}{\text{mean}} \, \cos(E(w), \mu_1) - \cos(E(w), \mu_2) \tag{7}$$

Compared to the definition of Bolukbasi et al. (2016), these definitions may be more general as constructing $\mathcal{P}$ is strictly more difficult than constructing $\mathcal{A}_1, \mathcal{A}_2$ (as $\mathcal{P}$ can always be split into two such sets but the reverse is not generally true) and Garg et al. (2018)'s definition does not rely on the first principal component explaining a large fraction of the variance. However, unlike the first definition, Garg et al. (2018) computes the bias in favor of/against a specific class (meaning if $\mathcal{N} = \{$'programmer', 'homemaker'$\}$ and 'programmer' was equally male-biased as 'homemaker' was female-biased, then under the definition of Garg et al. (2018), there would be no bias in aggregate). For the purposes of comparison, we adjust their definition by taking the absolute value of each term in the mean over $\mathcal{N}$.

Manzini et al. (2019) introduced a definition for quantifying multi-class bias which assumes access to sets $\mathcal{A}_1, \ldots, \mathcal{A}_k$ of representative words as in Garg et al. (2018). They quantify the bias as[5]:

$$\underset{\text{MANZINI}}{\text{bias}}\ (\mathcal{N}) = \underset{w \in \mathcal{N}}{\text{mean}}\ \underset{i \in \{1,\ldots,k\}}{\text{mean}}\ \underset{a \in \mathcal{A}_i}{\text{mean}}\ \cos(E(w), E(a)) \tag{8}$$

Similar to the adjustment made for the Garg et al. (2018) definition, we again take the absolute value of each term in the mean over $\mathcal{N}$.

5.2 Results

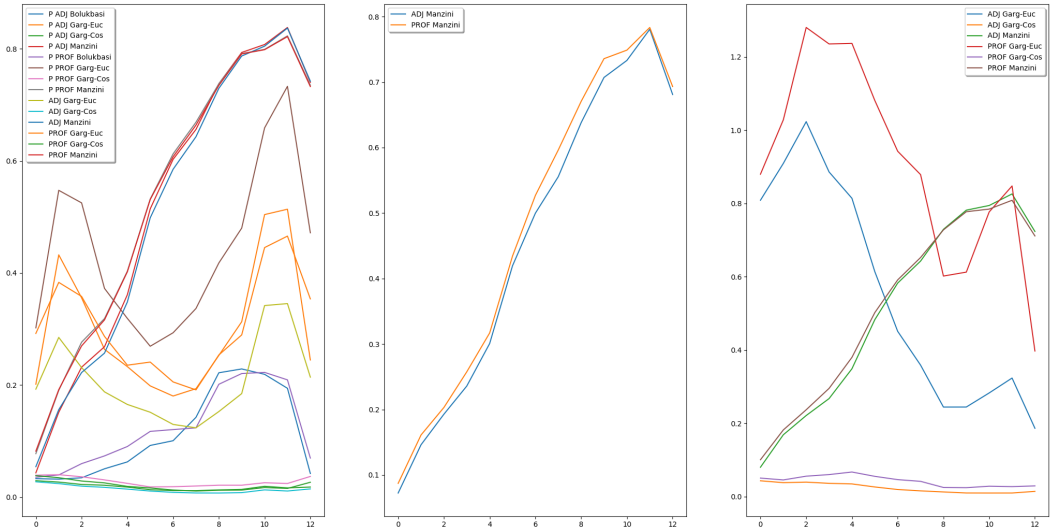

Figure 2: Layer-wise bias of distilled BERT-12 embeddings for $f = \text{mean}, g = \text{mean}, N = 100000$ **Left:** Gender, **Center:** Race, **Right:** Religion

	Gender							Race	Religion		
	B, $\mathcal{P}$	GE, $\mathcal{P}$	GC, $\mathcal{P}$	M, $\mathcal{P}$	GE	GC	M	M	GE	GC	M
Word2Vec	**0.0503**	**0.1758**	0.075	0.2403	**0.1569**	0.0677	0.2163	**0.0672**	**0.0907**	0.053	0.14
GloVe	0.0801	0.3534	0.0736	**0.1964**	0.357	0.0734	**0.1557**	0.1171	0.2699	0.0702	**0.0756**
BERT-12	0.0736	0.3725	0.0307	0.3186	0.2868	0.0254	0.3163	0.2575	1.2349	0.0604	0.2955
BERT-24	0.0515	0.6418	0.0462	0.234	0.4674	0.0379	0.2284	0.1956	0.6476	0.0379	0.2316
GPT2-12	0.4933	25.8743	0.0182	0.6464	2.0771	0.0062	0.7426	0.6532	4.5282	0.0153	0.776
GPT2-24	0.6871	40.1423	0.0141	0.8514	2.3244	**0.0026**	0.9019	0.8564	8.9528	**0.0075**	0.9081
RoBERTa-12	0.0412	0.2923	**0.0081**	0.8546	0.2077	0.0057	0.8551	0.8244	0.4356	0.0111	0.844
RoBERTa-24	0.0459	0.3771	0.0089	0.7879	0.2611	0.0064	0.783	0.7479	0.5905	0.0144	0.7636
XLNet-12	0.0838	1.0954	0.0608	0.3374	0.6661	0.042	0.34	0.2792	0.8537	0.0523	0.318
XLNet-24	0.0647	0.7644	0.0407	0.381	0.459	0.0268	0.373	0.328	0.8009	0.0505	0.368
DistilBERT-6	0.0504	0.5435	0.0375	0.3182	0.3343	0.0271	0.3185	0.2786	0.8128	0.0437	0.3106

Table 3: Social bias within static embeddings from different pretrained models with respect to a set of professions $\mathcal{N}_{prof}$. Parameters are set as $f = \text{mean}, g = \text{mean}, N = 100000$ and the layer of the pretrained model used in distillation is $\lfloor \frac{X}{4} \rfloor$. Lowest bias in a particular column is denoted in **bold**.

Inspired by the results of Nissim et al. (2019), in this work we transparently report social bias in existing static embeddings as well as the embeddings we compute. In particular, we exhaustively report the bias for all 3542 valid *(pretrained model, layer, social attribute, bias definition)* 4-tuples which describe all combinations of static embeddings and bias measures referenced in this work.

[5]Manzini et al. (2019) describes their score using slightly different phrasing; their score can easily be verified to be equivalent to our rephrasing up to two differences: (a) we use cosine similarity where they use cosine distance and (b) we insert absolute values in the mean over $\mathcal{N}$. We make these changes to introduce consistency with the other definitions and to permit comparison.

We specifically report results for binary gender (male, female), two-class religion (Christianity, Islam) and three-class race (white, Hispanic, and Asian), directly following Garg et al. (2018). These results are by no means intended to be comprehensive with regards to the breadth of bias socially and only address a restricted class of social biases which notably does not include the important class of intersectional biases. The types of biases being evaluated for are taken with respect to specific word lists (which are sometimes subjective albeit being peer-reviewed) that serve as exemplars and with respect to definitions of bias grounded in the norms of the United States.

Beginning with `bert-base-uncased`, we report the layer-wise bias across all *(attribute, definition)* pairs in Figure 2. What we immediately observe is that for any given social attribute, there is a great deal of variation across the layers in the quantified amount of bias. Further, while we are unsurprised that different bias measures for the same social attribute assign different absolute scores, we observe that they also do not agree in relative judgments. For gender, we observe that the bias estimated by the definition of Manzini et al. (2019) steadily increases before peaking at the penultimate layer and slightly decreasing thereafter. In contrast, under $\text{bias}_{\text{GARG-EUC}}$ we see a distribution with two peaks corresponding to layers at the start or end of the pretrained contextual model with lower bias observed in the intermediary layers. For estimating the same quantity, $\text{bias}_{\text{GARG-COS}}$ is mostly uniform across the layers (though the scale of the axes visually lessens the variation displayed). Similarly, in looking at the religious bias, we see similar inconsistencies with the bias increasing monotonically from layers 2 through 8 under $\text{bias}_{\text{MANZINI}}$, decreasing monotonically under $\text{bias}_{\text{GARG-EUC}}$, and remaining roughly constant under $\text{bias}_{\text{GARG-COS}}$. In general, while the choice of $\mathcal{N}$ (and the choice of $\mathcal{A}_i$ in the gender bias case) does affect the absolute bias estimates under any given definition, we find that the general trends in the bias across layers are approximately invariant under these choices for a specific definition.

Taken together, our analysis suggests a concerning state of affairs regarding bias quantification measures for (static) word embeddings. In particular, while estimates are seemingly stable to some types of choices regarding word lists, bias scores for a particular word embedding are tightly related to the definition being used and existing bias measures are markedly inconsistent with each other. We find this has important consequences beyond understanding the social biases in our representations. Concretely, we argue that without certainty regarding the extent to which embeddings are biased, it is impossible to properly interpret the meaningfulness of debiasing procedures (Bolukbasi et al., 2016; Zhao et al., 2018a;b; Sun et al., 2019) as we cannot reliably estimate the bias in the embeddings both before and after the procedure. This is further compounded with the existing evidence that current intrinsic measures of social bias may not handle geometric behavior such as clustering (Gonen & Goldberg, 2019).

In light of the above, next we compare bias estimates across different pretrained models in Table 3. Given the conflicting scores assigned by different definitions, we retain all definitions along with all social attributes in this comparison. However, we only consider target words given by $\mathcal{N}_{prof}$ for visual clarity as well as due to the aforementioned stability to the choice of $\mathcal{N}$, with the results for adjectives provided in Table 8. We begin by noting that since we do not perform preprocessing to normalize embeddings, the scores using $\text{bias}_{\text{GARG-EUC}}$ are not comparable (and may not have been proper to compare in the layer-wise case either) as they are sensitive to the absolute norms of the embeddings which cannot be expected to be similar across models[6]. Further, we note that $\text{bias}_{\text{BOLUKBASI}}$ may not be a reliable indicator as similar to Zhao et al. (2019a), we find that the first principal component explains less than 35% of the variance in the majority of the static embeddings distilled from contextual models. Of the two bias definitions not mentioned thus far, we find that all distilled static embeddings have substantially higher scores under $\text{bias}_{\text{MANZINI}}$ but generally lower scores under $\text{bias}_{\text{GARG-COS}}$ when compared to Word2Vec and GloVe. Interestingly, we see that under $\text{bias}_{\text{MANZINI}}$ both GPT-2 and RoBERTa embedding consistently get high scores across social attributes when compared to other distilled embeddings but under $\text{bias}_{\text{GARG-COS}}$ they receive the lowest scores among distilled embeddings.

Ultimately, given the aforementioned issues regarding the reliability of bias measures, it is difficult to arrive at a clear consensus of the comparative bias between our distilled embeddings and prior static embeddings. What our analysis does resolutely reveal is a pronounced and likely problematic effect of existing bias definitions on the resulting bias scores.

[6]When we did normalize using the Euclidean norm, we found the relative results to reliably coincide with those for $\text{bias}_{\text{GARG-COS}}$ which is consistent with the findings of Garg et al. (2018).

6 RELATED WORK

Distilled Static Representations. Recently, Akbik et al. (2019) introduced an approach similar to our **Aggregated** strategy where representations are gradually aggregated across instances in a dataset during training to model global information. Between epochs, the *memory* of past instances is reset and during testing, inference-time instances are added into the memory. In that work, the computed static embeddings are an additional feature that is used to achieve the state-of-the-art on several NER datasets. Based on our results, we believe their approach could be further improved by different decisions in pretrained model and layer choice. Their results may be explained by the (desirable) variance reduction we observe in pooling over many contexts. Additionally, since they only pool over instances in an online fashion within an epoch, the number of contexts is relatively small in their approach as compared to ours which may help to explain why they find that $\min$ or $\max$ pooling perform slightly better than mean pooling as the choice for g.

May et al. (2019) proposes a different approach to convert representations from sentence encoders into static embeddings as a means for applying the WEAT (Caliskan et al., 2017) implicit bias tests to a sentence encoder. In their method, a single *semantically-bleached* sentence is *synthetically* constructed from a template and then fed into the encoder to compute a static embedding for the word of interest. We argue that this approach may inherently not be appropriate for quantifying bias in sentence encoders[7] in the general case as sentence encoders are trained on semantically-meaningful sentences and semantically-bleached constructions are not representative of this distribution. Moreover, the types of templated constructions presented heavily rely on *deictic expressions* and therefore are difficult to adapt for certain syntactic categories such as verbs (as would be required for the **SimVerb3500** dataset especially) without providing arguments for the verb. These concerns are further exacerbated by our findings given the poor representational behavior seen in our **Decontextualized** embeddings which have similar deficiencies with their static embeddings and the poor representational behavior when we pool over relatively few semantically-meaningful contexts using the **Aggregated** strategy (e.g. our results for $N = 10000$ which is still 50 instances per word on average and is much more than the single instance they consider). We believe our quantification of bias as a result can be taken as a more faithful estimator of bias in sentence encoders.

Concurrently, Hu et al. (2019) considers a similar approach towards diachronic sense modelling. In particular, given a word, they find its senses and example sentences of each sense in the Oxford English Dictionary and use these to compute static embeddings using the **Aggregated** strategy with the last layer of `bert-base-uncased` and n_i upper-bounded at 10. Given our results, their performance could likely be improved by pooling over more sentences, using `bert-large-uncased`, and considering layer choice as their task heavily relies on lexical understanding which seems to be better captured in earlier layers of the model than the last one. Since they require sense annotations for their setting (and the number of example sentences in a dictionary for a sense is inherently constrained), our findings also suggest that additional sense-annotated or weakly sense-annotated sentences would be beneficial.

Lightweight Pretrained Representations. Taken differently, our approach can be seen as a method for integrating pretraining in a more lightweight fashion. Model compression (LeCun et al., 1990; Frankle & Carbin, 2019) and knowledge distillation (Ba & Caruana, 2014; Hinton et al., 2015) are well-studied techniques in machine learning that have been recently applied for similar purposes. In particular, several concurrent approaches have been proposed to yield lighter pretrained sentence encoders and contextual word representations (Gururangan et al., 2019; Shen et al., 2019; Sanh, 2019; Tsai et al., 2019; Tang et al., 2019; Jiao et al., 2019). Our approach along with these recent approaches yield representations that are more appropriate for resource-constrained settings such as on-device models for mobile phones (Shen et al., 2019), for real-time settings where we require low-latency and short inference times, and for users that may not have access to GPU or TPU computational resources (Tsai et al., 2019). Additionally, this line of work is particularly timely given the emergent concerns of the environmental impact/harm of training and using increasingly large models in NLP (Strubell et al., 2019), machine learning (Li et al., 2016; Canziani et al., 2016), and the broader AI community (Schwartz et al., 2019).

Bias. Social bias in NLP has been primarily evaluated in three ways: (a) using geometric similarity between embeddings (Bolukbasi et al., 2016; Garg et al., 2018; Manzini et al., 2019), (b) adapting psychological association tests (Caliskan et al., 2017; May et al., 2019), and (c) considering down-

[7]The authors also identified several empirical concerns that draw the meaningfulness of this method into question.

stream behavior (Zhao et al., 2017; 2018a; 2019a; Stanovsky et al., 2019)[8]. In relation to this body of work, our bias evaluation is in the style of (a) as we are interested in intrinsic bias in embeddings and considers (potentially) multi-class social bias in the lens of gender, race, and religion whereas prior work has primarily focused on gender. Additionally, while most of the work on bias in embeddings has considered the static embedding setting, recent work has considered sentence encoders and contextual models. Zhao et al. (2019a) considers gender bias in ELMo when applied to NER and Kurita et al. (2019) extends these results by considering not only NER but also bias using WEAT by leveraging the masked language modeling objective of BERT. Similarly, Basta et al. (2019) considers intrinsic gender bias using ELMo by studying gender-swapped sentences. When compared to these approaches, we study a broader class of biases under more than one bias definition and consider more than one model. Further, while these approaches generally neglect reporting bias values for different layers of the model, we show this is crucial as bias is not uniformly distributed throughout model layers and downstream practitioners often do not use the last layer of deep Transformer models (Liu et al., 2019a; Tenney et al., 2019; Zhang et al., 2019; Zhao et al., 2019b) [9].

7 Future Directions

Pretrained contextual word representations have quickly gained traction in the NLP community, largely because of the flurry of empirical successes that have followed since their introduction. For downstream practitioners, our work suggests several simple (e.g. subword pooling mechanism choice) and more sophisticated (e.g. layer choice, benefits of variance reduction by using multiple contexts) strategies that may yield better downstream performance. Additionally, some recent models have combined static and dynamic embeddings (Peters et al., 2018; Bommasani et al., 2019; Akbik et al., 2019) and our representations may support drop-in improvements in these settings.

Beyond furthering efforts in representation learning, this work introduces a new approach towards the understanding of contextual word representations via proxy analysis. In particular, while in this work we choose to study social bias, similar analyses toward other forms of interpretability and understanding would be valuable. Additionally, post-processing approaches that go beyond analysis such as dimensionality reduction may be particularly intriguing given that this is often challenging to do within large multi-layered networks like BERT (Sanh, 2019) but has been successfully done for static embeddings (Nunes & Antunes, 2018; Mu & Viswanath, 2018; Raunak et al., 2019).

Future work may also consider the choice of the corpus $\mathcal{D}$ from which contexts are drawn. In particular, we believe choosing $\mathcal{D}$ to be drawn from the target domain for some downstream task may serve as an extremely lightweight domain adaptation strategy. Additionally, in this work we choose to provide contexts of sentence length in order to facilitate regularity in the comparison across models. But for some models, such as Transformer-XL or XLNet which are trained with memories to handle larger contexts, better performance may be achieved by using larger contexts.

8 Conclusion

In this work, we propose simple but effective procedures for converting contextual word representations into static word embeddings. When applied to pretrained models like BERT, we find the resulting embeddings outperform Word2Vec and GloVe substantially under intrinsic evaluation and provide insights into the pretrained model. We further demonstrate the resulting embeddings are more amenable to (existing) embedding analysis methods and report the extent of various social biases (gender, race, religion) across a number of measures. Our large-scale analysis furnishes several findings with respect to social bias encoded in popular pretrained contextual representations via the proxy of our embeddings and has implications towards the reliability of existing protocols for quantifying bias in word embeddings.

9 Reproducibility

All data, code, visualizations (and code to produce to them), and distilled word embeddings will be publicly released. Additional reproducibility details are provided in Appendix A.

[8] Sun et al. (2019) provides a taxonomy of the work towards understanding gender bias within NLP.

[9] This is the only layer studied in Kurita et al. (2019).

REFERENCES

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 19–27, Boulder, Colorado, June 2009. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N09-1003.

Alan Akbik, Tanja Bergmann, and Roland Vollgraf. Pooled contextualized embeddings for named entity recognition. 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)*, pp. 724–728, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1078. URL https://www.aclweb.org/anthology/N19-1078.

Ben Athiwaratkun and Andrew Wilson. Multimodal word distributions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1645–1656, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1151. URL https://www.aclweb.org/anthology/P17-1151.

Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 27*, pp. 2654–2662. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5484-do-deep-nets-really-need-to-be-deep.pdf.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. The problem with bias: from allocative to representational harms in machine learning. *Special Interest Group for Computing, Information and Society (SIGCIS)*, 2017.

Christine Basta, Marta R. Costa-jussà, and Noe Casas. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pp. 33–39, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3805. URL https://www.aclweb.org/anthology/W19-3805.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, March 2003. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=944919.944966.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. doi: 10.1162/tacl_a_00051. URL https://www.aclweb.org/anthology/Q17-1010.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 4349–4357. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddi
pdf.

Rishi Bommasani, Arzoo Katiyar, and Claire Cardie. SPARSE: Structured prediction using argument-relative structured encoding. In *Proceedings of the Third Workshop on Structured Prediction for NLP*, pp. 13–17, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1503. URL https://www.aclweb.org/anthology/W19-1503.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. ISSN 0036-8075. doi: 10.1126/science.aal4230. URL https://science.sciencemag.org/content/356/6334/183.

Jose Camacho-Collados and Mohammad Taher Pilehvar. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63:743–788, 2018.

Alfredo Canziani, Adam Paszke, and Eugenio Culurciello. An analysis of deep neural network models for practical applications. *CoRR*, abs/1605.07678, 2016. URL http://arxiv.org/abs/1605.07678.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pp. 160–167, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390177. URL http://doi.acm.org/10.1145/1390156.1390177.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537, November 2011. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1953048.2078186.

Ryan Cotterell and Hinrich Schütze. Morphological word-embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1287–1292, Denver, Colorado, May–June 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1140. URL https://www.aclweb.org/anthology/N15-1140.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–2988, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1285.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N19-1423.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pp. 30–35, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2506. URL https://www.aclweb.org/anthology/W16-2506.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJl-b3RcF7.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115 (16):E3635–E3644, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1720347115. URL https://www.pnas.org/content/115/16/E3635.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. SimVerb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2173–2182, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1235. URL https://www.aclweb.org/anthology/D16-1235.

Anna Gladkova and Aleksandr Drozd. Intrinsic evaluations of word embeddings: What can we do better? In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pp. 36–42, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2507. URL https://www.aclweb.org/anthology/W16-2507.

Hila Gonen and Yoav Goldberg. 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)*, pp. 609–614, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N19-1061.

Suchin Gururangan, Tam Dang, Dallas Card, and Noah A. Smith. Variational pretraining for semi-supervised text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5880–5894, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1590.

Souleiman Hasan and Edward Curry. Word re-embedding via manifold dimensionality retention. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 321–326, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1033. URL https://www.aclweb.org/anthology/D17-1033.

Felix Hill, Roi Reichart, and Anna Korhonen. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, December 2015. doi: 10.1162/COLI_a_00237. URL https://www.aclweb.org/anthology/J15-4004.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. URL http://arxiv.org/abs/1503.02531.

Renfen Hu, Shen Li, and Shichen Liang. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3899–3908, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1379.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. *CoRR*, abs/1909.10351, 2019. URL https://arxiv.org/abs/1909.10351.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pp. 166–172, Florence, Italy, August 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W19-3823.

Yann LeCun, John S. Denker, and Sara A. Solla. Optimal brain damage. In D. S. Touretzky (ed.), *Advances in Neural Information Processing Systems 2*, pp. 598–605. Morgan-Kaufmann, 1990. URL http://papers.nips.cc/paper/250-optimal-brain-damage.pdf.

Guang-He Lee and Yun-Nung Chen. MUSE: Modularizing unsupervised sense embeddings. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 327–337, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1034. URL https://www.aclweb.org/anthology/D17-1034.

Omer Levy and Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pp. 171–180, Ann Arbor, Michigan, June 2014a. Association for Computational Linguistics. doi: 10.3115/v1/W14-1618. URL https://www.aclweb.org/anthology/W14-1618.

Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 302–308, Baltimore, Maryland, June 2014b. Association for Computational Linguistics. doi: 10.3115/v1/P14-2050. URL https://www.aclweb.org/anthology/P14-2050.

Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015. doi: 10.1162/tacl_a_00134. URL https://www.aclweb.org/anthology/Q15-1016.

Bofang Li, Tao Liu, Zhe Zhao, Buzhou Tang, Aleksandr Drozd, Anna Rogers, and Xiaoyong Du. Investigating different syntactic context types and context representations for learning word embeddings. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2421–2431, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1257. URL https://www.aclweb.org/anthology/D17-1257.

Da Li, Xinbo Chen, Michela Becchi, and Ziliang Zong. Evaluating the energy efficiency of deep convolutional neural networks on cpus and gpus. In *2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom)(BDCloud-SocialCom-SustainCom)*, pp. 477–484. IEEE, 2016.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. Linguistic knowledge and transferability of contextual representations. 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)*, pp. 1073–1094, Minneapolis, Minnesota, June 2019a. Association for Computational Linguistics. doi: 10.18653/v1/N19-1112. URL https://www.aclweb.org/anthology/N19-1112.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019b. URL http://arxiv.org/abs/1907.11692.

Edward Loper and Steven Bird. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pp. 63–70, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1118108.1118117. URL https://doi.org/10.3115/1118108.1118117.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. Black is to criminal as caucasian is to police: Detecting and removing 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)*, pp. 615–621, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N19-1062.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. 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)*, pp. 622–628, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N19-1063.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pp. 6294–6305, 2017.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

George A Miller and Walter G Charles. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28, 1991.

Jiaqi Mu and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HkuGJ3kCb.

Malvina Nissim, Rik van Noord, and Rob van der Goot. Fair is better than sensational: Man is to doctor as woman is to doctor. *arXiv preprint arXiv:1905.09866*, 2019.

Davide Nunes and Luis Antunes. Neural random projections for language modelling. *CoRR*, abs/1807.00930, 2018. URL http://arxiv.org/abs/1807.00930.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://www.aclweb.org/anthology/D14-1162.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://www.aclweb.org/anthology/N18-1202.

Mohammad Taher Pilehvar and Nigel Collier. De-conflated semantic representations. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1680–1690, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1174. URL https://www.aclweb.org/anthology/D16-1174.

Mohammad Taher Pilehvar, Jose Camacho-Collados, Roberto Navigli, and Nigel Collier. Towards a seamless integration of word senses into downstream NLP applications. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1857–1869, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1170. URL https://www.aclweb.org/anthology/P17-1170.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Vikas Raunak, Vivek Gupta, and Florian Metze. Effective dimensionality reduction for word embeddings. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pp. 235–243, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4328. URL https://www.aclweb.org/anthology/W19-4328.

Herbert Rubenstein and John B. Goodenough. Contextual correlates of synonymy. *Commun. ACM*, 8(10):627–633, October 1965. ISSN 0001-0782. doi: 10.1145/365628.365657. URL http://doi.acm.org/10.1145/365628.365657.

Sebastian Ruder. Nlp-progress, 2019a. URL https://github.com/sebastianruder/NLP-progress.

Sebastian Ruder. *Neural Transfer Learning for Natural Language Processing*. PhD thesis, National University of Ireland, Galway, 2019b.

Victor Sanh. Smaller, faster, cheaper, lighter: Introducing distilbert, a distilled version of bert, Aug 2019. URL https://medium.com/huggingface/distilbert-8cf3380435b5.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green AI. *CoRR*, abs/1907.10597, 2019. URL http://arxiv.org/abs/1907.10597.

Dinghan Shen, Pengyu Cheng, Dhanasekar Sundararaman, Xinyuan Zhang, Qian Yang, Meng Tang, Asli Celikyilmaz, and Lawrence Carin. Learning compressed sentence representations for on-device text processing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 107–116, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1011.

Weijia Shi, Muhao Chen, Pei Zhou, and Kai-Wei Chang. Retrofitting contextualized word embeddings with paraphrases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Hong Kong, China, August 2019. Association for Computational Linguistics. URL https://arxiv.org/abs/1909.09700.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1679–1684, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1164.

Karl Stratos. A sub-character architecture for Korean language processing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 721–726, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1075. URL https://www.aclweb.org/anthology/D17-1075.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3645–3650, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1355.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1630–1640, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1159.

Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. Distilling task-specific knowledge from BERT into simple neural networks. *CoRR*, abs/1903.12136, 2019. URL http://arxiv.org/abs/1903.12136.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL https://www.aclweb.org/anthology/P19-1452.

Henry Tsai, Jason Riesa, Melvin Johnson, Naveen Arivazhagan, Xin Li, and Amelia Archer. Small and Practical BERT Models for Sequence Labeling. *arXiv e-prints*, art. arXiv:1909.00100, Aug 2019.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, NeurIPS'19, 2019a.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. 2019b. In the Proceedings of ICLR.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016. URL http://arxiv.org/abs/1609.08144.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019.

Jinxing Yu, Xun Jian, Hao Xin, and Yangqiu Song. Joint embeddings of Chinese words, characters, and fine-grained subcharacter components. In *Proceedings of the 2017 Conference on Empirical*

Methods in Natural Language Processing, pp. 286–291, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1027. URL https://www.aclweb.org/anthology/D17-1027.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav. Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2979–2989, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1323. URL https://www.aclweb.org/anthology/D17-1323.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 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)*, pp. 15–20, New Orleans, Louisiana, June 2018a. Association for Computational Linguistics. doi: 10.18653/v1/N18-2003. URL https://www.aclweb.org/anthology/N18-2003.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4847–4853, Brussels, Belgium, October-November 2018b. Association for Computational Linguistics. doi: 10.18653/v1/D18-1521. URL https://www.aclweb.org/anthology/D18-1521.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender bias in contextualized 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)*, pp. 629–634, Minneapolis, Minnesota, June 2019a. Association for Computational Linguistics. doi: 10.18653/v1/N19-1064. URL https://www.aclweb.org/anthology/N19-1064.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Hong Kong, China, August 2019b. Association for Computational Linguistics.

A Reproducibility Details

A.1 Data

We use English Wikipedia as the corpus $\mathcal{D}$ in context combination for the **Aggregated** strategy. The specific subset of English Wikipedia[10] used was lightly preprocessed with a simple heuristic to remove bot-generated content. Individual Wikipedia documents were split into sentences using NLTK (Loper & Bird, 2002). We chose to exclude sentences containing fewer than 7 sentences or greater than 75 tokens (token counts we computed using the NLTK word tokenizer) though we did not find this filtering decision to be particularly impactful in initial experiments.
The specific pretrained Word2Vec[11] and GloVe[12] embeddings used were both 300 dimensional. The Word2Vec embeddings were trained on approximately 100 billion words from Google News and the GloVe embeddings were trained on 6 billion tokens from Wikipedia 2014 and Gigaword 5. We chose the 300-dimensional embeddings in both cases as we believed they were the most frequently used and generally the best performing on both intrinsic evaluations (Hasan & Curry, 2017) and downstream tasks.

[10] https://blog.lateral.io/2015/06/the-unknown-perils-of-mining-wikipedia/
[11] https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit
[12] https://nlp.stanford.edu/projects/glove/

A.2 EVALUATION DECISIONS

In this work, we chose to conduct intrinsic evaluation experiments that focused on word similarity and word relatedness. We did not consider the related evaluation of lexical understanding via word analogies as they have been shown to decompose into word similarity subtasks (Levy & Goldberg, 2014a) and there are significant concerns about the validity of these analogies tests (Nissim et al., 2019). We acknowledge that word similarity and word relatedness tasks have also been heavily scrutinized (Faruqui et al., 2016; Gladkova & Drozd, 2016). A primary concern is that results are highly sensitive to (hyper)parameter selection (Levy et al., 2015). In our setting, where the parameters of the embeddings are largely fixed based on which pretrained models are publicly released and where we exhaustively report the impact of most remaining parameters, we find these concerns to still be valid but less relevant.

To this end, prior work has considered various preprocessing operations on static embeddings such as clipping embeddings on an elementwise basis (Hasan & Curry, 2017) when performing intrinsic evaluation. We chose not to study these preprocessing choices as they create discrepancies between the embeddings used in intrinsic evaluation and those used in downstream tasks (where this form of preprocessing is generally not considered) and would have added additional parameters implicitly. Instead, we directly used the computed embeddings from the pretrained model with no changes throughout this work.

A.3 REPRESENTATION QUALITY DATASET TRENDS

Rubenstein & Goodenough (1965) introduced a set of 65 noun-pairs and demonstrated strong correlation (exceeding 95%) between the scores in their dataset and additional human validation. Miller & Charles (1991) introduced a larger collection of pairs which they argued was an improvement over RG65 as it more faithfully addressed semantic similarity. Agirre et al. (2009) followed this work by introducing a even more pairs that included those of Miller & Charles (1991) as a subset and again demonstrated correlations with human scores exceeding 95%. Hill et al. (2015) argued that SIMLEX999 was an improvement in coverage over RG65 and more correctly quantified semantic similarity as opposed to semantic relatedness or association when compared to WS353. Beyond this, SIMVERB3500 was introduced by Gerz et al. (2016) to further increase coverage over all predecessors. Specifically, it shifted the focus towards verbs which had been heavily neglected in the prior datasets which centered on nouns and adjectives.

A.4 EXPERIMENTAL DETAILS

We used `PyTorch` (Paszke et al., 2017) throughout this work with the pretrained contextual word representations taken from the HuggingFace `pytorch-transformers` repository[13]. Tokenization for each model was conducted using its corresponding tokenizer, i.e. results for GPT2 use the GPT2Tokenizer in `pytorch-transformers`.

For simplicity, throughout this work, we introduce N as the total number of contexts used in distilling with the **Aggregated** strategy. Concretely, $N = \sum_{w_i \in \mathcal{V}} n_i$ where $\mathcal{V}$ is the vocabulary used (generally the 2005 words in the four datasets considered). As a result, in finding contexts, we filter for sentences in $\mathcal{D}$ that contain at least one word in $\mathcal{V}$. We choose to do this as this requires a number of candidate sentences upper bounded with respect to the most frequent word in $\mathcal{V}$ as opposed to filtering for a specific value for n which requires a number of sentences scaling in the frequency of the least frequent word in $\mathcal{V}$.

The N samples from $\mathcal{D}$ for the **Aggregated** strategy were sampled uniformly at random. Accordingly, as the aforementioned discussion suggests, for word w_i, the number of examples n_i which contain w_i scales in the frequency of w_i in the vocabulary being used. As a consequence, for small values of N, it is possible that rare words would have no examples and computing a representation $\mathbf{w}$ using the **Aggregated** strategy would be impossible. In this case, we back-offed to using the **Decontextualized** representation for w_i.

Given this concern, in the bias evaluation, we fix $n_i = 20$ for every w_i. In initial experiments, we found the bias results to be fairly stable when choosing values $n_i \in \{20, 50, 100\}$. The choice of n_i would correspond to $N = 40100$ (as the vocabulary size was 2005) in the representation quality

[13] `https://github.com/huggingface/pytorch-transformers`

section in some sense (however this assumes a uniform distribution of word frequency as opposed to a Zipf distribution). The embeddings in the bias evaluation are drawn from layer $\lfloor \frac{X}{4} \rfloor$ using $f = \text{mean}, g = \text{mean}$ as we found these to be the best performing embeddings generally across pretrained models and datasets in the representational quality evaluation.

A.5 BIAS WORD LISTS

The set of gender-paired tuples $\mathcal{P}$ were taken from Bolukbasi et al. (2016). In the gender bias section, $\mathcal{P}$ for definitions involving sets $\mathcal{A}_i$ indicates that $\mathcal{P}$ was split into equal-sized sets of male and female work. For the remaining gender results, the sets described in §G.3 were used. The various attribute sets $\mathcal{A}_i$ and target sets $\mathcal{N}_j$ were taken from Garg et al. (2018) which can be further sourced to a number of prior works in studying social bias. We remove any multi-word terms from these lists.

B BERT-LARGE

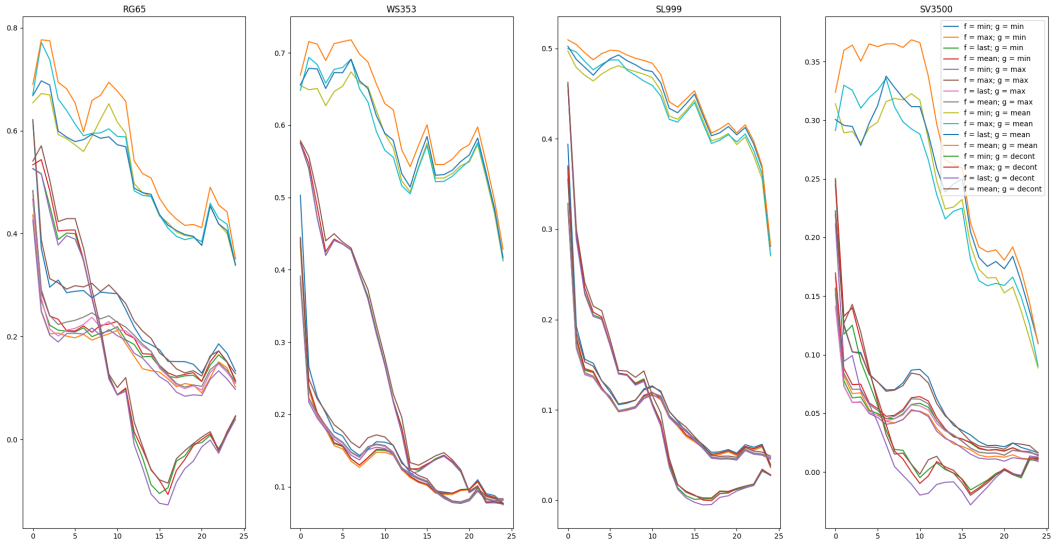

Figure 3: Layerwise performance of BERT-24 static embeddings for all possible choices of f, g

C GPT-2

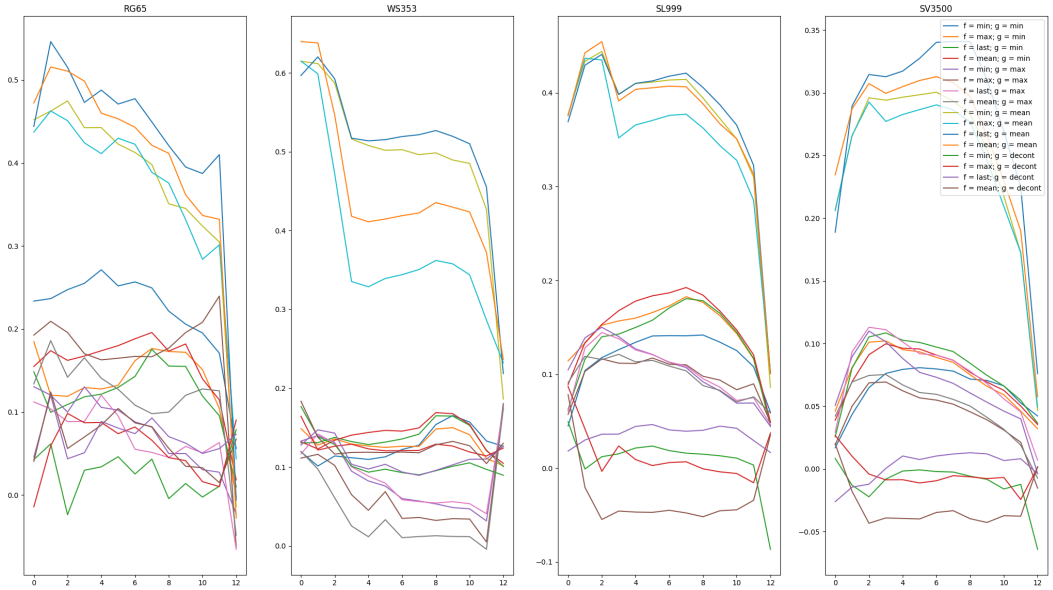

Figure 4: Layerwise performance of GPT2-12 static embeddings for all possible choices of f, g

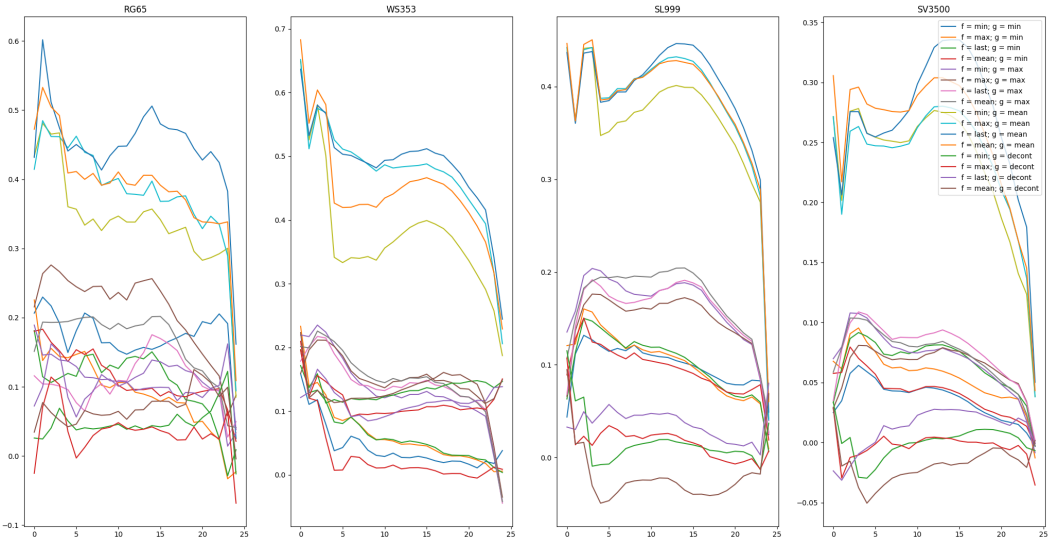

Figure 5: Layerwise performance of GPT-24 static embeddings for all possible choices of f, g

Model	N	RG65	WS353	SIMLEX999	SIMVERB3500
Word2Vec	-	0.6787	**0.6838**	0.4420	**0.3636**
GloVe	-	**0.6873**	0.6073	0.3705	0.2271
GPT2-12	10000	0.2843 (0)	0.4205 (1)	0.2613 (2)	0.1472 (6)
GPT2-12	50000	0.5000 (2)	0.5815 (1)	0.4378 (2)	0.2607 (2)
GPT2-12	100000	0.5156 (1)	0.6396 (0)	**0.4547** (2)	0.3128 (6)
GPT2-24	10000	0.3149 (0)	0.5209 (0)	0.2940 (0)	0.1697 (0)
GPT2-24	50000	0.5362 (2)	0.6486 (0)	0.4350 (0)	0.2721 (0)
GPT2-24	100000	0.5328 (1)	0.6830 (0)	0.4505 (3)	0.3056 (0)

Table 4: Performance of Static Embeddings on Word Similarity and Word Relatedness Tasks. f and g are set to mean for all GPT2-models and (#) indicates the layer the embeddings are distilled from. **Bold** indicates best performing embeddings for a given dataset.

D ROBERTA

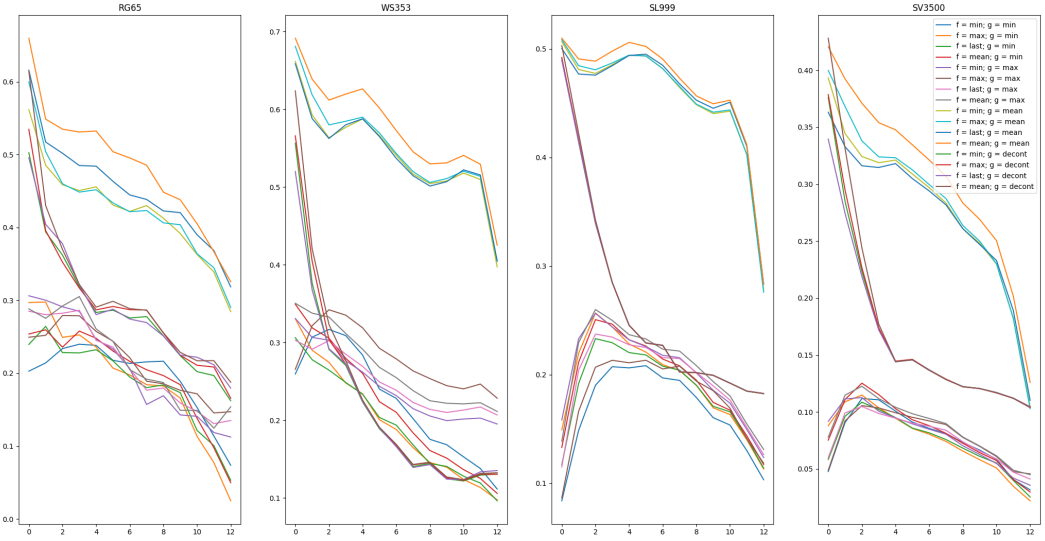

Figure 6: Layerwise performance of RoBERTa-12 static embeddings for all possible choices of f, g

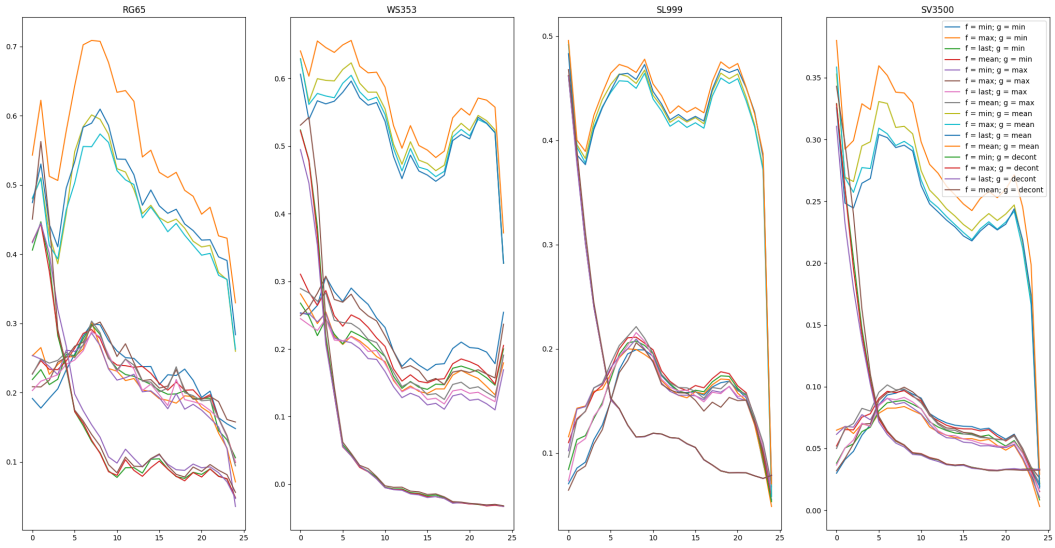

Figure 7: Layerwise performance of RoBERTa-24 static embeddings for all possible choices of f, g

Model	N	RG65	WS353	SIMLEX999	SIMVERB3500
Word2Vec	-	0.6787	0.6838	0.4420	0.3636
GloVe	-	0.6873	0.6073	0.3705	0.2271
RoBERTa-12	10000	0.5719 (0)	0.6618 (0)	0.4794 (0)	0.3968 (0)
RoBERTa-12	50000	0.6754 (0)	0.6867 (0)	0.501 (0)	0.4123 (0)
RoBERTa-12	100000	0.6597 (0)	0.6915 (0)	0.5098 (0)	0.4206 (0)
RoBERTa-12	500000	0.6675 (0)	0.6979 (0)	0.5268 (5)	0.4311 (0)
RoBERTa-12	1000000	0.6761 (0)	**0.7018** (0)	**0.5374** (5)	**0.4442** (4)
RoBERTa-24	10000	0.5469 (1)	0.6144 (0)	0.4499 (0)	0.3403 (0)
RoBERTa-24	50000	0.6837 (1)	0.6412 (0)	0.4855 (0)	0.371 (0)
RoBERTa-24	100000	0.7087 (7)	0.6563 (6)	0.4959 (0)	0.3802 (0)
RoBERTa-24	500000	**0.7557** (8)	0.663 (6)	0.5184 (18)	0.412 (6)
RoBERTa-24	1000000	0.739 (8)	0.6673 (6)	0.5318 (18)	0.4303 (9)

Table 5: Performance of Static Embeddings on Word Similarity and Word Relatedness Tasks. f and g are set to mean for all RoBERTa-models and (#) indicates the layer the embeddings are distilled from. **Bold** indicates best performing embeddings for a given dataset.

E XLNET

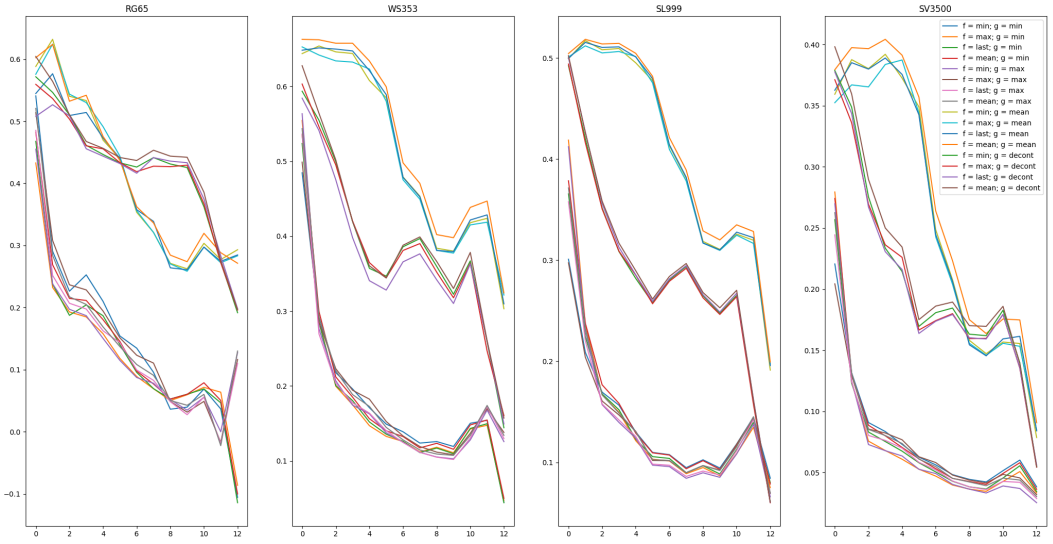

Figure 8: Layerwise performance of XLNet-12 static embeddings for all possible choices of f, g

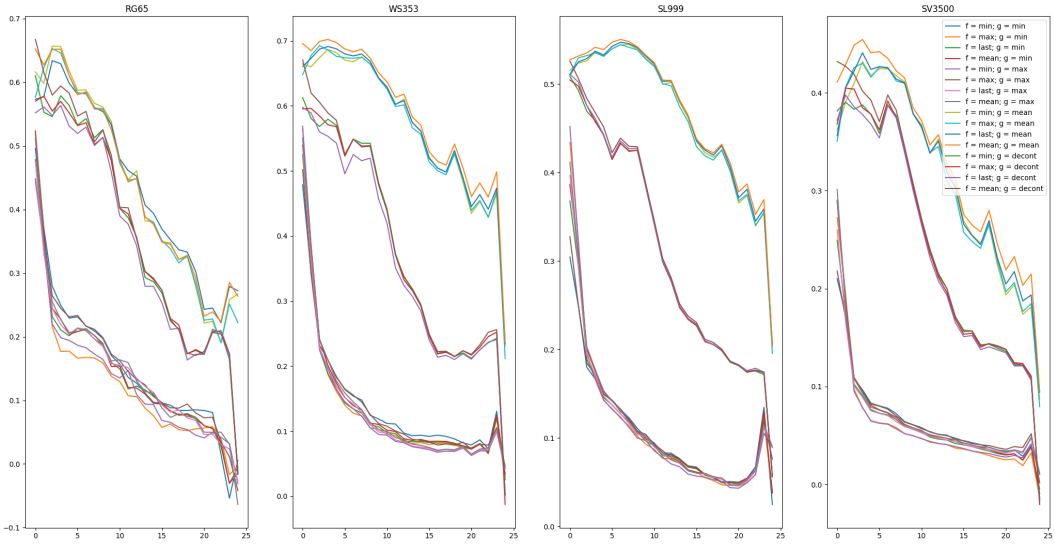

Figure 9: Layerwise performance of XLNet-24 static embeddings for all possible choices of f, g

Model	N	RG65	WS353	SIMLEX999	SIMVERB3500
Word2Vec	-	0.6787	0.6838	0.4420	0.3636
GloVe	-	0.6873	0.6073	0.3705	0.2271
XLNet-12	10000	0.604 (0)	0.6482 (0)	0.483 (0)	0.3916 (0)
XLNet-12	50000	0.6056 (1)	0.6571 (0)	0.5157 (1)	0.3973 (1)
XLNet-12	100000	0.6239 (1)	0.6629 (0)	0.5185 (1)	0.4044 (3)
XLNet-12	500000	0.6391 (3)	0.6937 (3)	0.5392 (3)	0.4747 (4)
XLNet-12	1000000	0.6728 (3)	0.7018 (3)	0.5447 (4)	0.4918 (4)
XLNet-24	10000	0.6525 (0)	0.6935 (0)	0.5054 (0)	0.4332 (1)
XLNet-24	50000	0.6556 (0)	0.6926 (0)	0.5377 (5)	0.4492 (3)
XLNet-24	100000	0.6522 (3)	0.7021 (3)	0.5503 (6)	0.4545 (3)
XLNet-24	500000	0.66 (0)	0.7378 (6)	0.581 (8)	0.5095 (6)
XLNet-24	1000000	**0.7119** (6)	**0.7446** (7)	**0.5868** (9)	**0.525** (6)

Table 6: Performance of Static Embeddings on Word Similarity and Word Relatedness Tasks. f and g are set to mean for all XLNet-models and (#) indicates the layer the embeddings are distilled from. **Bold** indicates best performing embeddings for a given dataset.

F DISTILBERT

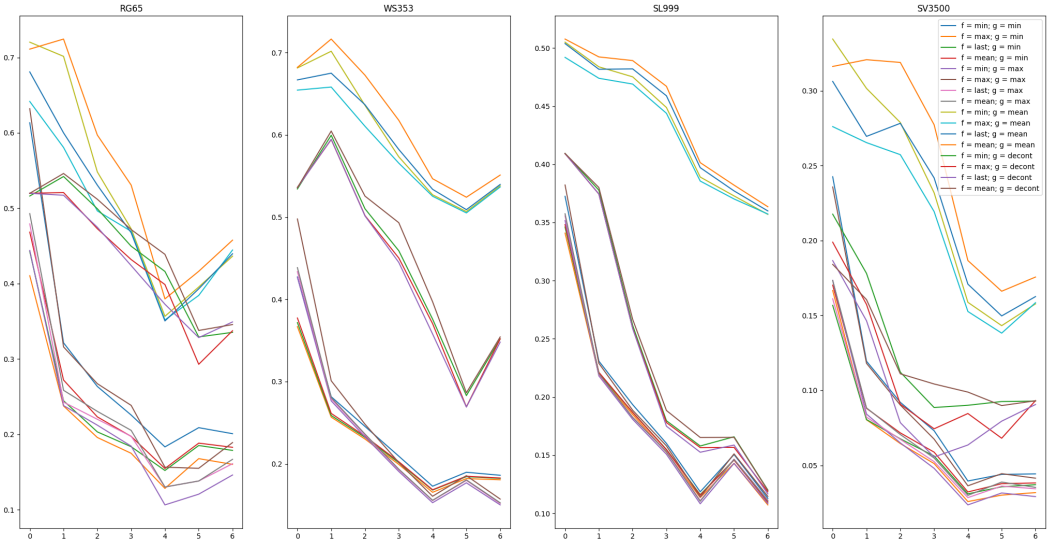

Figure 10: Layerwise performance of DistilBERT-6 static embeddings for all possible choices of f, g

Model	N	RG65	WS353	SIMLEX999	SIMVERB3500
Word2Vec	-	0.6787	0.6838	0.4420	**0.3636**
GloVe	-	0.6873	0.6073	0.3705	0.2271
DistilBERT-6	10000	0.57 (0)	0.6828 (1)	0.4705 (0)	0.2971 (0)
DistilBERT-6	50000	0.7257 (1)	0.6928 (1)	0.5043 (0)	0.3121 (0)
DistilBERT-6	100000	0.7245 (1)	0.7164 (1)	0.5077 (0)	0.3207 (1)
DistilBERT-6	500000	0.7363 (1)	0.7239 (1)	0.5093 (0)	0.3444 (2)
DistilBERT-6	1000000	**0.7443** (1)	**0.7256** (1)	**0.5095** (0)	0.3536 (3)

Table 7: Performance of Static Embeddings on Word Similarity and Word Relatedness Tasks. f and g are set to mean for all DistilBERT-models and (#) indicates the layer the embeddings are distilled from. **Bold** indicates best performing embeddings for a given dataset.

G BIAS

G.1 ADDITIONAL MODELS

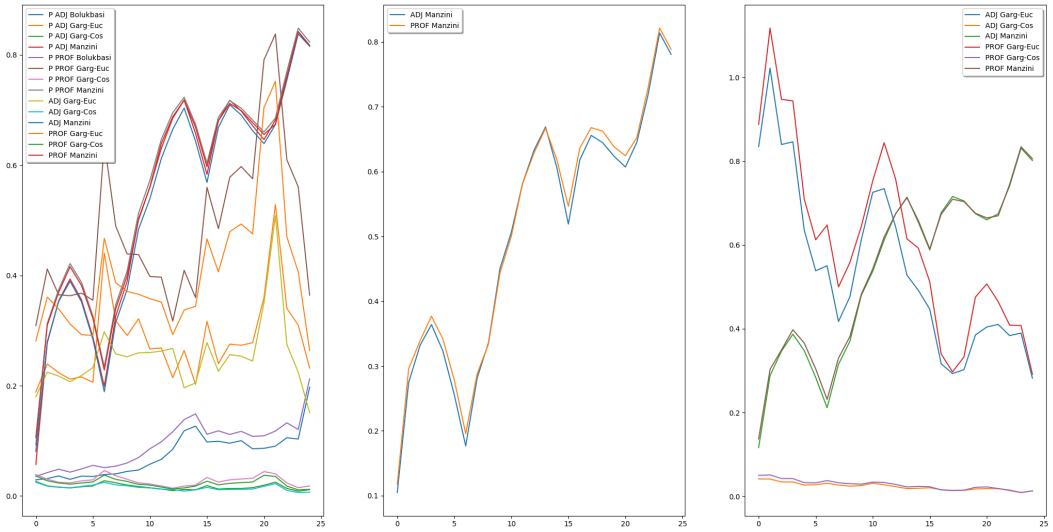

Figure 11: Layerwise bias of BERT-24 static embeddings for $f = \mathrm{mean}, g = \mathrm{mean}, N = 100000$
Left: Gender, **Center:** Race, **Right:** Religion

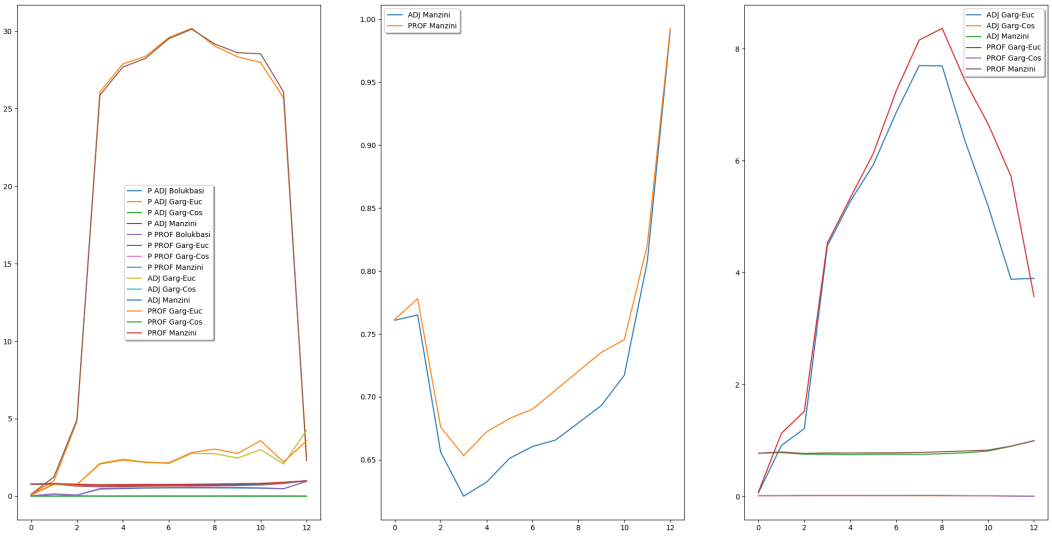

Figure 12: Layerwise bias of GPT2-12 static embeddings for $f = \text{mean}, g = \text{mean}, N = 100000$
Left: Gender, **Center:** Race, **Right:** Religion

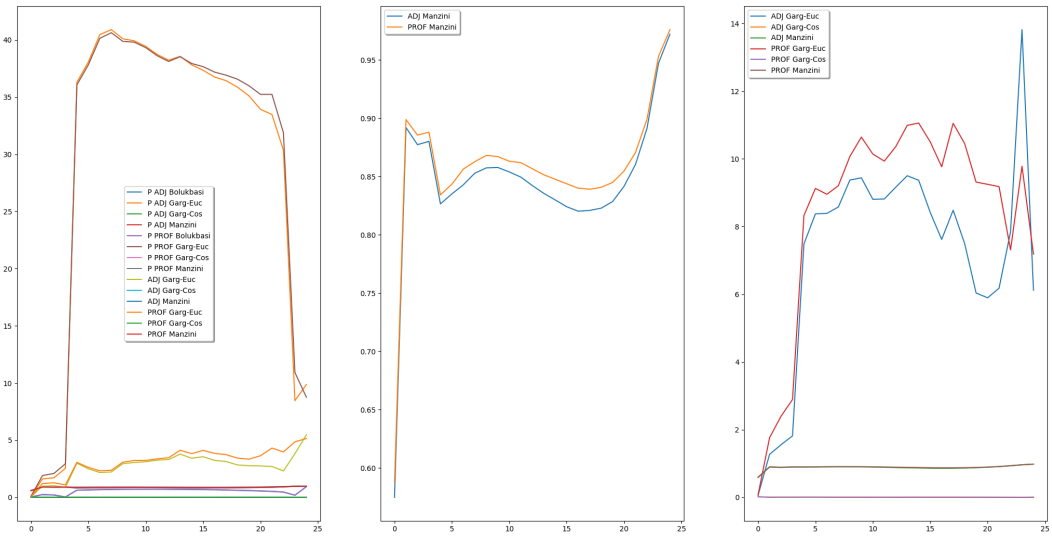

Figure 13: Layerwise bias of GPT2-24 static embeddings for $f = \text{mean}, g = \text{mean}, N = 100000$
Left: Gender, **Center:** Race, **Right:** Religion

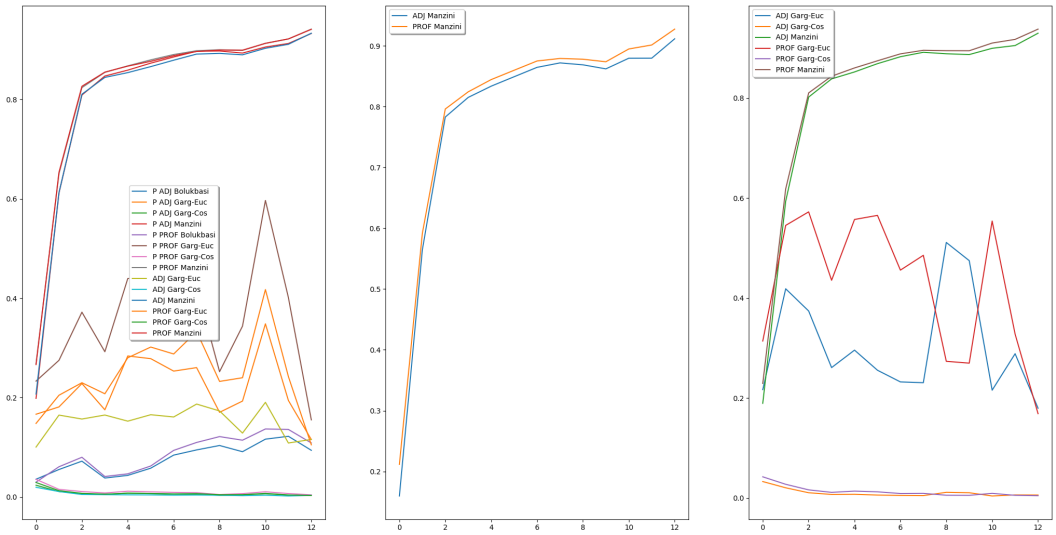

Figure 14: Layerwise bias of RoBERTa-12 static embeddings for $f = \text{mean}, g = \text{mean}, N = 100000$
Left: Gender, **Center:** Race, **Right:** Religion

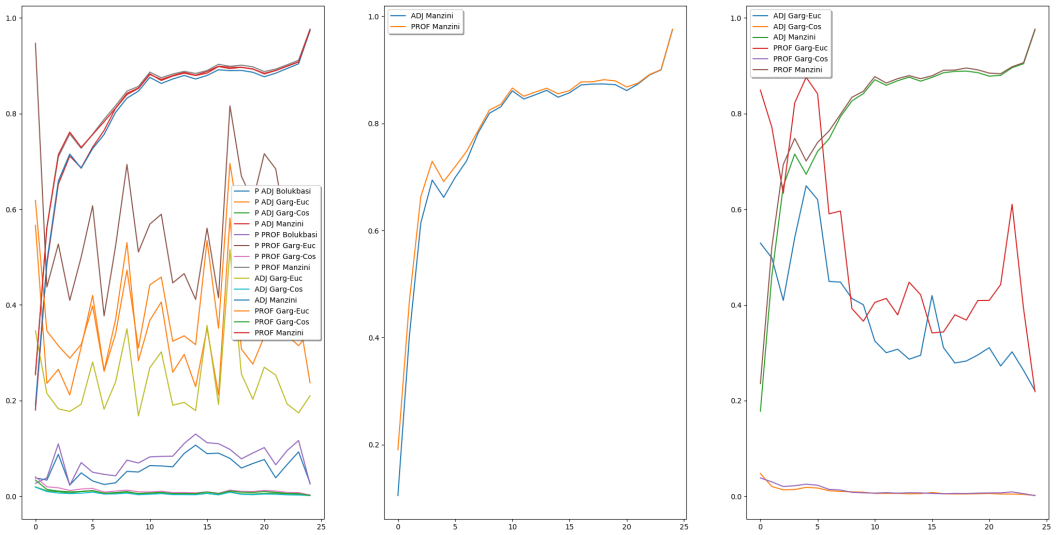

Figure 15: Layerwise bias of RoBERTa-24 static embeddings for $f = \text{mean}, g = \text{mean}, N = 100000$
Left: Gender, **Center:** Race, **Right:** Religion

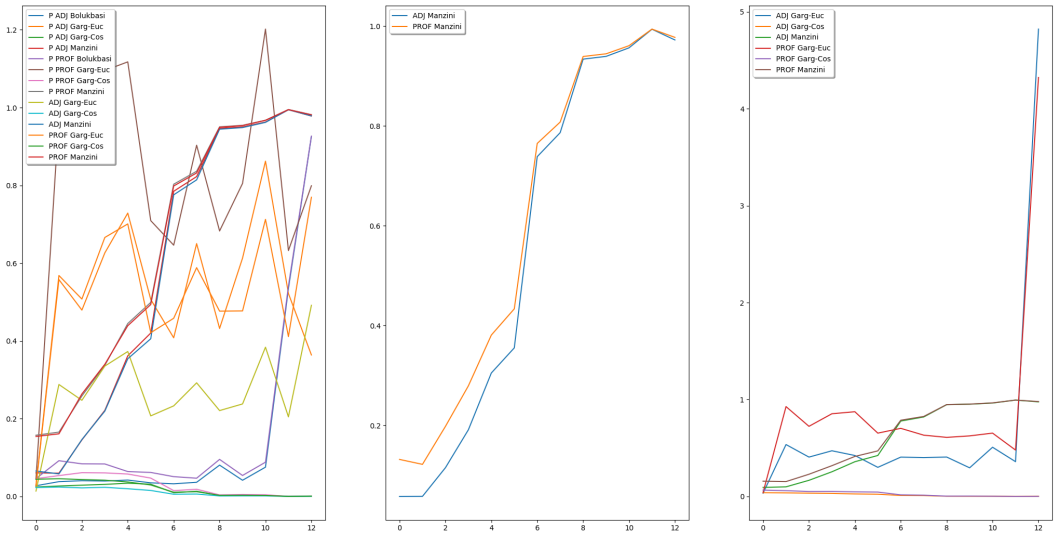

Figure 16: Layerwise bias of XLNet-12 static embeddings for $f = \text{mean}, g = \text{mean}, N = 100000$
Left: Gender, **Center:** Race, **Right:** Religion

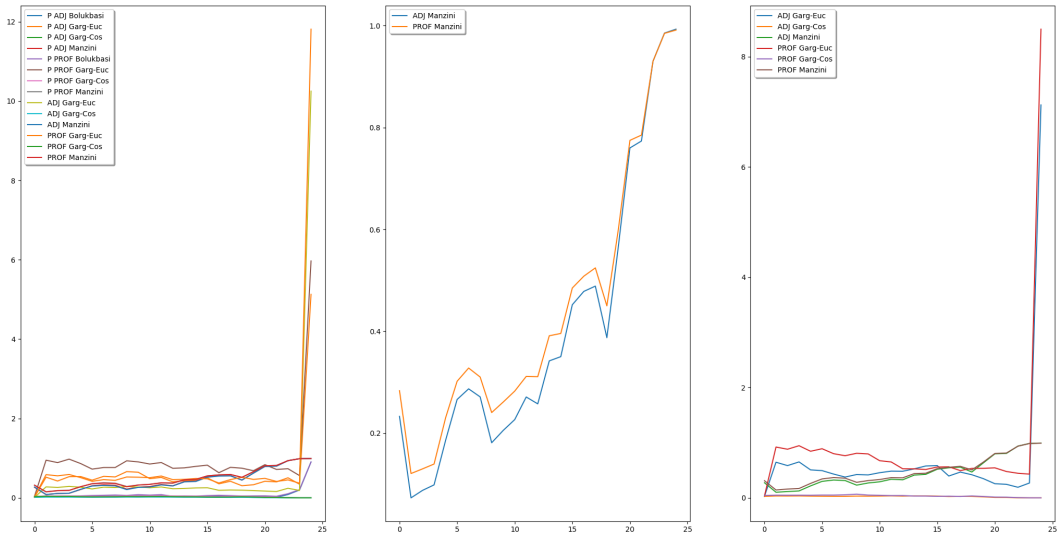

Figure 17: Layerwise bias of XLNet-24 static embeddings for $f = \text{mean}, g = \text{mean}, N = 100000$
Left: Gender, **Center:** Race, **Right:** Religion

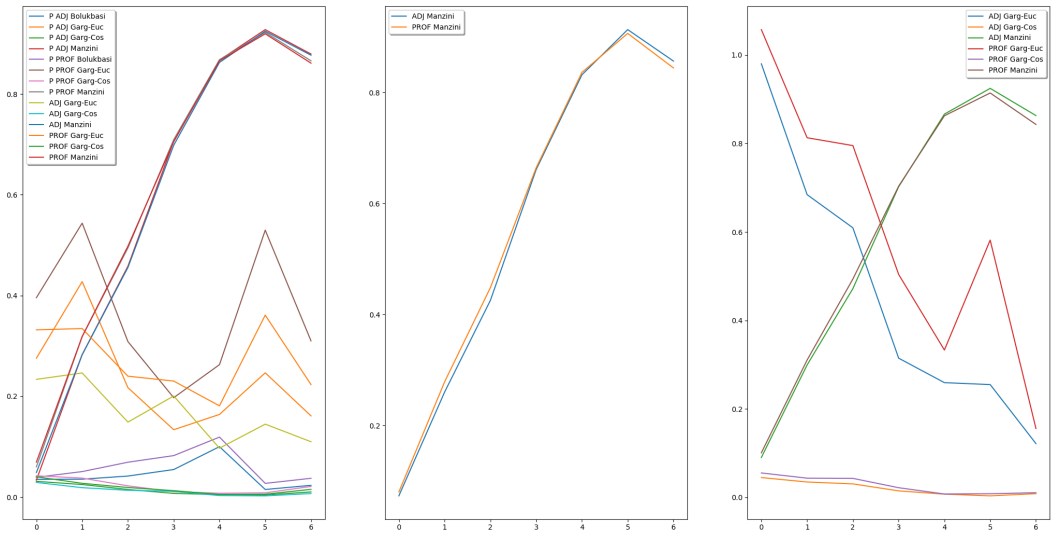

Figure 18: Layerwise bias of DistilBERT-6 static embeddings for $f = \text{mean}, g = \text{mean}, N = 100000$
Left: Gender, **Center:** Race, **Right:** Religion

G.2 Adjective Results

		Gender						Race	Religion		
	B, $\mathcal{P}$	GE, $\mathcal{P}$	GC, $\mathcal{P}$	M, $\mathcal{P}$	GE	GC	M	M	GE	GC	M
Word2Vec	0.0482	0.1656	0.0435	0.1347	0.1247	0.0343	0.1178	0.0661	0.13	0.0434	0.1264
GloVe	0.095	0.2206	0.0403	0.1289	0.2017	0.0355	0.1108	0.0714	0.2341	0.0606	0.0675
BERT-12	0.0506	0.2637	0.0213	0.2684	0.1879	0.0175	0.2569	0.2358	0.8858	0.0365	0.2677
BERT-24	0.0389	0.4405	0.0277	0.199	0.2978	0.0248	0.189	0.1768	0.5505	0.0316	0.212
GPT2-12	0.4631	26.0809	0.0176	0.6126	2.1238	0.0068	0.7101	0.621	4.4775	0.0152	0.7525
GPT2-24	0.6707	40.4664	0.0141	0.8367	2.1771	0.0023	0.89	0.843	8.3889	0.0064	0.9006
RoBERTa-12	0.0381	0.1754	0.005	0.8472	0.1649	0.0046	0.8444	0.8153	0.2608	0.0069	0.8387
RoBERTa-24	0.0248	0.2626	0.0064	0.7647	0.1821	0.0048	0.7562	0.73	0.4492	0.0117	0.7472
XLNet-12	0.0399	0.6265	0.0312	0.2214	0.3354	0.0237	0.2196	0.1911	0.4716	0.0321	0.2549
XLNet-24	0.0468	0.5423	0.025	0.3307	0.2697	0.0153	0.3144	0.2871	0.4318	0.0282	0.3235
DistilBERT-6	0.0353	0.4274	0.0247	0.2825	0.2461	0.0185	0.2824	0.2603	0.6842	0.035	0.2994

Table 8: Social bias within static embeddings from different pretrained models with respect to a set of adjectives, $\mathcal{N}_{adj}$. Parameters are set as $f = \text{mean}, g = \text{mean}, N = 100000$ and the layer of the pretrained model used in distillation is $\lfloor \frac{X}{4} \rfloor$.

G.3 Word Sets

$\mathcal{N}_{\text{prof}}$ = {'accountant', 'acquaintance', 'actor', 'actress', 'administrator', 'adventurer', 'advocate', 'aide', 'alderman', 'ambassador', 'analyst', 'anthropologist', 'archaeologist', 'archbishop', 'architect', 'artist', 'artiste', 'assassin', 'astronaut', 'astronomer', 'athlete', 'attorney', 'author', 'baker', 'ballerina', 'ballplayer', 'banker', 'barber', 'baron', 'barrister', 'bartender', 'biologist', 'bishop', 'bodyguard', 'bookkeeper', 'boss', 'boxer', 'broadcaster', 'broker', 'bureaucrat', 'businessman', 'businesswoman', 'butcher', 'cabbie', 'cameraman', 'campaigner', 'captain', 'cardiologist', 'caretaker', 'carpenter', 'cartoonist', 'cellist', 'chancellor', 'chaplain', 'character', 'chef', 'chemist', 'choreographer', 'cinematographer', 'citizen', 'cleric', 'clerk', 'coach', 'collector', 'colonel', 'columnist', 'comedian', 'comic', 'commander', 'commentator', 'commissioner', 'composer', 'conductor', 'confesses', 'congressman', 'constable', 'consultant', 'cop', 'correspondent', 'councilman', 'councilor', 'counselor', 'critic', 'crooner', 'crusader', 'curator', 'custodian', 'dad', 'dancer', 'dean', 'dentist', 'deputy', 'dermatologist', 'detective', 'diplomat', 'director', 'doctor', 'drummer', 'economist', 'editor', 'educator', 'electrician', 'employee', 'entertainer', 'entrepreneur', 'environmentalist', 'envoy', 'epidemiologist', 'evangelist', 'farmer', 'filmmaker', 'financier', 'firebrand', 'firefighter', 'fireman', 'fisherman', 'footballer', 'foreman', 'gangster', 'gardener', 'ge-

ologist', 'goalkeeper', 'guitarist', 'hairdresser', 'handyman', 'headmaster', 'historian', 'hitman', 'homemaker', 'hooker', 'housekeeper', 'housewife', 'illustrator', 'industrialist', 'infielder', 'inspector', 'instructor', 'inventor', 'investigator', 'janitor', 'jeweler', 'journalist', 'judge', 'jurist', 'laborer', 'landlord', 'lawmaker', 'lawyer', 'lecturer', 'legislator', 'librarian', 'lieutenant', 'lifeguard', 'lyricist', 'maestro', 'magician', 'magistrate', 'manager', 'marksman', 'marshal', 'mathematician', 'mechanic', 'mediator', 'medic', 'midfielder', 'minister', 'missionary', 'mobster', 'monk', 'musician', 'nanny', 'narrator', 'naturalist', 'negotiator', 'neurologist', 'neurosurgeon', 'novelist', 'nun', 'nurse', 'observer', 'officer', 'organist', 'painter', 'paralegal', 'parishioner', 'parliamentarian', 'pastor', 'pathologist', 'patrolman', 'pediatrician', 'performer', 'pharmacist', 'philanthropist', 'philosopher', 'photographer', 'photojournalist', 'physician', 'physicist', 'pianist', 'planner', 'playwright', 'plumber', 'poet', 'policeman', 'politician', 'pollster', 'preacher', 'president', 'priest', 'principal', 'prisoner', 'professor', 'programmer', 'promoter', 'proprietor', 'prosecutor', 'protagonist', 'protege', 'protester', 'provost', 'psychiatrist', 'psychologist', 'publicist', 'pundit', 'rabbi', 'radiologist', 'ranger', 'realtor', 'receptionist', 'researcher', 'restaurateur', 'sailor', 'saint', 'salesman', 'saxophonist', 'scholar', 'scientist', 'screenwriter', 'sculptor', 'secretary', 'senator', 'sergeant', 'servant', 'serviceman', 'shopkeeper', 'singer', 'skipper', 'socialite', 'sociologist', 'soldier', 'solicitor', 'soloist', 'sportsman', 'sportswriter', 'statesman', 'steward', 'stockbroker', 'strategist', 'student', 'stylist', 'substitute', 'superintendent', 'surgeon', 'surveyor', 'teacher', 'technician', 'teenager', 'therapist', 'trader', 'treasurer', 'trooper', 'trucker', 'trumpeter', 'tutor', 'tycoon', 'undersecretary', 'understudy', 'valedictorian', 'violinist', 'vocalist', 'waiter', 'waitress', 'warden', 'warrior', 'welder', 'worker', 'wrestler', 'writer'}

$\mathcal{N}_{\text{adj}}$ = {'disorganized', 'devious', 'impressionable', 'circumspect', 'impassive', 'aimless', 'effeminate', 'unfathomable', 'fickle', 'inoffensive', 'reactive', 'providential', 'resentful', 'bizarre', 'impractical', 'sarcastic', 'misguided', 'imitative', 'pedantic', 'venomous', 'erratic', 'insecure', 'resourceful', 'neurotic', 'forgiving', 'profligate', 'whimsical', 'assertive', 'incorruptible', 'individualistic', 'faithless', 'disconcerting', 'barbaric', 'hypnotic', 'vindictive', 'observant', 'dissolute', 'frightening', 'complacent', 'boisterous', 'pretentious', 'disobedient', 'tasteless', 'sedentary', 'sophisticated', 'regimental', 'mellow', 'deceitful', 'impulsive', 'playful', 'sociable', 'methodical', 'willful', 'idealistic', 'boyish', 'callous', 'pompous', 'unchanging', 'crafty', 'punctual', 'compassionate', 'intolerant', 'challenging', 'scornful', 'possessive', 'conceited', 'imprudent', 'dutiful', 'lovable', 'disloyal', 'dreamy', 'appreciative', 'forgetful', 'unrestrained', 'forceful', 'submissive', 'predatory', 'fanatical', 'illogical', 'tidy', 'aspiring', 'studious', 'adaptable', 'conciliatory', 'artful', 'thoughtless', 'deceptive', 'frugal', 'reflective', 'insulting', 'unreliable', 'stoic', 'hysterical', 'rustic', 'inhibited', 'outspoken', 'unhealthy', 'ascetic', 'skeptical', 'painstaking', 'contemplative', 'leisurely', 'sly', 'mannered', 'outrageous', 'lyrical', 'placid', 'cynical', 'irresponsible', 'vulnerable', 'arrogant', 'persuasive', 'perverse', 'steadfast', 'crisp', 'envious', 'naive', 'greedy', 'presumptuous', 'obnoxious', 'irritable', 'dishonest', 'discreet', 'sporting', 'hateful', 'ungrateful', 'frivolous', 'reactionary', 'skillful', 'cowardly', 'sordid', 'adventurous', 'dogmatic', 'intuitive', 'bland', 'indulgent', 'discontented', 'dominating', 'articulate', 'fanciful', 'discouraging', 'treacherous', 'repressed', 'moody', 'sensual', 'unfriendly', 'optimistic', 'clumsy', 'contemptible', 'focused', 'haughty', 'morbid', 'disorderly', 'considerate', 'humorous', 'preoccupied', 'airy', 'impersonal', 'cultured', 'trusting', 'respectful', 'scrupulous', 'scholarly', 'superstitious', 'tolerant', 'realistic', 'malicious', 'irrational', 'sane', 'colorless', 'masculine', 'witty', 'inert', 'prejudiced', 'fraudulent', 'blunt', 'childish', 'brittle', 'disciplined', 'responsive', 'courageous', 'bewildered', 'courteous', 'stubborn', 'aloof', 'sentimental', 'athletic', 'extravagant', 'brutal', 'manly', 'cooperative', 'unstable', 'youthful', 'timid', 'amiable', 'retiring', 'fiery', 'confidential', 'relaxed', 'imaginative', 'mystical', 'shrewd', 'conscientious', 'monstrous', 'grim', 'questioning', 'lazy', 'dynamic', 'gloomy', 'troublesome', 'abrupt', 'eloquent', 'dignified', 'hearty', 'gallant', 'benevolent', 'maternal', 'paternal', 'patriotic', 'aggressive', 'competitive', 'elegant', 'flexible', 'gracious', 'energetic', 'tough', 'contradictory', 'shy', 'careless', 'cautious', 'polished', 'sage', 'tense', 'caring', 'suspicious', 'sober', 'neat', 'transparent', 'disturbing', 'passionate', 'obedient', 'crazy', 'restrained', 'fearful', 'daring', 'prudent', 'demanding', 'impatient', 'cerebral', 'calculating', 'amusing', 'honorable', 'casual', 'sharing', 'selfish', 'ruined', 'spontaneous', 'admirable', 'conventional', 'cheerful', 'solitary', 'upright', 'stiff', 'enthusiastic', 'petty', 'dirty', 'subjective', 'heroic', 'stupid', 'modest', 'impressive', 'orderly', 'ambitious', 'protective', 'silly', 'alert', 'destructive', 'exciting', 'crude', 'ridiculous', 'subtle', 'mature', 'creative', 'coarse', 'passive', 'oppressed', 'accessible', 'charming', 'clever', 'decent', 'miserable', 'superficial', 'shallow', 'stern', 'winning', 'balanced', 'emotional', 'rigid', 'invisible', 'desperate', 'cruel', 'romantic', 'agreeable', 'hurried', 'sympathetic',

'solemn', 'systematic', 'vague', 'peaceful', 'humble', 'dull', 'expedient', 'loyal', 'decisive', 'arbitrary', 'earnest', 'confident', 'conservative', 'foolish', 'moderate', 'helpful', 'delicate', 'gentle', 'dedicated', 'hostile', 'generous', 'reliable', 'dramatic', 'precise', 'calm', 'healthy', 'attractive', 'artificial', 'progressive', 'odd', 'confused', 'rational', 'brilliant', 'intense', 'genuine', 'mistaken', 'driving', 'stable', 'objective', 'sensitive', 'neutral', 'strict', 'angry', 'profound', 'smooth', 'ignorant', 'thorough', 'logical', 'intelligent', 'extraordinary', 'experimental', 'steady', 'formal', 'faithful', 'curious', 'reserved', 'honest', 'busy', 'educated', 'liberal', 'friendly', 'efficient', 'sweet', 'surprising', 'mechanical', 'clean', 'critical', 'criminal', 'soft', 'proud', 'quiet', 'weak', 'anxious', 'solid', 'complex', 'grand', 'warm', 'slow', 'false', 'extreme', 'narrow', 'dependent', 'wise', 'organized', 'pure', 'directed', 'dry', 'obvious', 'popular', 'capable', 'secure', 'active', 'independent', 'ordinary', 'fixed', 'practical', 'serious', 'fair', 'understanding', 'constant', 'cold', 'responsible', 'deep', 'religious', 'private', 'simple', 'physical', 'original', 'working', 'strong', 'modern', 'determined', 'open', 'political', 'difficult', 'knowledge', 'kind'}

$\mathcal{P}$ = {('she', 'he'), ('her', 'his'), ('woman', 'man'), ('mary', 'john'), ('herself', 'himself'), ('daughter', 'son'), ('mother', 'father'), ('gal', 'guy'), ('girl', 'boy'), ('female', 'male')}

$\mathcal{A}_{\text{male}}$ = {'he', 'son', 'his', 'him', 'father', 'man', 'boy', 'himself', 'male', 'brother', 'sons', 'fathers', 'men', 'boys', 'males', 'brothers', 'uncle', 'uncles', 'nephew', 'nephews'}

$\mathcal{A}_{\text{female}}$ = {'she', 'daughter', 'hers', 'her', 'mother', 'woman', 'girl', 'herself', 'female', 'sister', 'daughters', 'mothers', 'women', 'girls', 'femen'[14], 'sisters', 'aunt', 'aunts', 'niece', 'nieces'}

$\mathcal{A}_{\text{white}}$ = {'harris', 'nelson', 'robinson', 'thompson', 'moore', 'wright', 'anderson', 'clark', 'jackson', 'taylor', 'scott', 'davis', 'allen', 'adams', 'lewis', 'williams', 'jones', 'wilson', 'martin', 'johnson'}

$\mathcal{A}_{\text{hispanic}}$ = {'castillo', 'gomez', 'soto', 'gonzalez', 'sanchez', 'rivera', 'martinez', 'torres', 'rodriguez', 'perez', 'lopez', 'medina', 'diaz', 'garcia', 'castro', 'cruz'}

$\mathcal{A}_{\text{asian}}$ = {'cho', 'wong', 'tang', 'huang', 'chu', 'chung', 'ng', 'wu', 'liu', 'chen', 'lin', 'yang', 'kim', 'chang', 'shah', 'wang', 'li', 'khan', 'singh', 'hong'}

$\mathcal{A}_{\text{islam}}$ = {'allah', 'ramadan', 'turban', 'emir', 'salaam', 'sunni', 'koran', 'imam', 'sultan', 'prophet', 'veil', 'ayatollah', 'shiite', 'mosque', 'islam', 'sheik', 'muslim', 'muhammad'}

$\mathcal{A}_{\text{christian}}$ = {'baptism', 'messiah', 'catholicism', 'resurrection', 'christianity', 'salvation', 'protestant', 'gospel', 'trinity', 'jesus', 'christ', 'christian', 'cross', 'catholic', 'church'}

H NAMING CONVENTIONS

Throughout this work, we make use of several naming conventions/substitutions. In the case of models, we use the form 'MODEL-X' where X indicates the number of layers in the model and consequently the model produces $X + 1$ representations for any given subword (including the initial layer 0 representation). Table 9 describes the complete correspondence of our shorthand and the full names. In the case of model names, the full form is the name assigned to the pretrained model (that was possibly reimplemented) released by HuggingFace.

[14] We remove 'femen' when using Word2Vec as it is not in the vocabulary of the pretrained embeddings we use.

Our Shorthand	Full Name
BERT-12	`bert-base-uncased`
BERT-24	`bert-large-uncased`
GPT2-12	`gpt2`
GPT2-24	`gpt2-medium`
RoBERTa-12	`roberta-base`
RoBERTa-24	`roberta-large`
XLNet-12	`xlnet-base-cased`
XLNet-24	`xlnet-base-cased`
DistilBERT-6	`distilbert-base-uncased`
SL999	SIMLEX999
SV3500	SIMVERB3500
B	$\text{bias}_{\text{BOLUKBASI}}$
GE	$\text{bias}_{\text{GARG-EUC}}$
GC	$\text{bias}_{\text{GARG-COS}}$
M	$\text{bias}_{\text{MANZINI}}$

Table 9: Naming conventions used throughout this work