Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings

Rishi Bommasani
Cornell University
rb724@cornell.edu

Kelly Davis
Mozilla Corporation
kdavis@mozilla.com

Claire Cardie
Cornell University
cardie@cs.cornell.edu

Abstract

Contextualized representations (e.g. ELMo, BERT) have become the default pretrained representations for downstream NLP applications. In some settings, this transition has rendered their static embedding predecessors (e.g. Word2Vec, GloVe) obsolete. As a side-effect, we observe that older interpretability methods for static embeddings — while more mature than those available for their dynamic counterparts — are underutilized in studying newer contextualized representations. Consequently, we introduce simple and fully general methods for converting from contextualized representations to static lookup-table embeddings which we apply to 5 popular pretrained models and 9 sets of pretrained weights. Our analysis of the resulting static embeddings notably reveals that pooling over many contexts significantly improves representational quality under intrinsic evaluation. Complementary to analyzing representational quality, we consider social biases encoded in pretrained representations with respect to gender, race/ethnicity, and religion and find that bias is encoded disparately across pretrained models and internal layers even for models that share the same training data. Concerningly, we find dramatic inconsistencies between social bias estimators for word embeddings.

1 Introduction

Word embeddings (Bengio et al., 2003; Collobert and Weston, 2008; Collobert et al., 2011) have been a hallmark of modern natural language processing (NLP) for many years. Embedding methods have been broadly applied and have experienced parallel and complementary innovations alongside neural network methods for NLP. Advances in embedding quality in part have come from integrating additional information such as syntax (Levy and Goldberg, 2014a; Li et al., 2017), morphology (Cotterell and 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). Due to their tremendous representational power, pretrained contextualized representations, in particular, have seen widespread adoption across myriad subareas of NLP.

The recent dominance of pretrained contextualized representations such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) has served as the impetus for exciting and diverse interpretability research: Liu et al. (2019a); Tenney et al. (2019a) study what is learned across the layers of these models, Tenney et al. (2019b); Ethayarajh (2019) consider what is learned from context, Clark et al. (2019); Michel et al. (2019) look at specific attention heads, Hewitt and Manning (2019); Ettinger (2020) address linguistic understanding such as syntax and negation, and Wallace et al. (2019); Tan and Celis (2019) address ethical concerns such as security (adversarial robustness) and social bias. In fact, the neologism *BERTology* was coined specifically to describe this flurry of interpretability research.[1] While these works have provided nuanced fine-grained analyses by creating new interpretability schema/techniques, we instead take an alternate approach of trying to re-purpose methods developed for analyzing static word embeddings.

In order to employ static embedding interpretability methods to contextualized representations, we begin by proposing a simple strategy for converting from contextualized representations to static embeddings. Crucially, our method is fully general and assumes only that the contextualized model maps word sequences to vector sequences. Given this generality, we apply our method to 9 popular pretrained contextualized representations. The resulting static embeddings serve as *proxies* for the original contextualized model.

[1] We direct interested readers to a more complete survey of this work from Rogers et al. (2020).

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4758–4781
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

We initially examine the representational quality of these embeddings under intrinsic evaluation. Our evaluation produces several insights regarding layer-wise lexical semantic understanding and representational variation in contextualized representations. Importantly, our analyses suggest constructive improvements to potentially improve downstream practices in using contextualized models. Simultaneously, we find that our static embeddings substantially outperform Word2Vec and GloVe and therefore suggests our method serves the dual purpose of being a lightweight mechanism for generating static embeddings that track with advances in contextualized representations. Since static embeddings have significant advantages with respect to speed, computational resources, and ease of use, these results have important implications for resource-constrained settings (Shen et al., 2019), environmental concerns (Strubell et al., 2019), and the broader accessibility of NLP technologies.[2]

Alongside more developed methods for embedding analysis, the static embedding setting is also equipped with a richer body of work regarding social bias. In this sense, we view understanding the encoded social bias in representations as a societally critical special-case of interpretability research. We employ methods for identifying and quantifying gender, racial/ethnic, and religious bias (Bolukbasi et al., 2016; Garg et al., 2018; Manzini et al., 2019) to our static embeddings. These experiments not only shed light on the properties of our static embeddings for downstream use but can also serve as a proxy for understanding latent biases in the original pretrained contextual representations. We find that biases in different models and across different layers are quite disparate; this has important consequences on model and layer selection for downstream use. Further, for two sets of pretrained weights learned on the same training data, we find that bias patterns still remain fairly distinct. Most surprisingly, our large-scale evaluation makes clear that existing bias estimators are dramatically inconsistent with each other.

2 Methods

In order to use a contextualized 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 k subword representations generated for w in context c in order to compute a single representation for w in c, i.e. $\{\mathbf{w}_c^1, \dots, \mathbf{w}_c^k\} \mapsto \mathbf{w}_c$. Beyond this, we define *context combination* to be the mapping from representations $\mathbf{w}_{c_1}, \dots, \mathbf{w}_{c_n}$ of w in different contexts $c_1, \dots, c_n$ to a single static embedding $\mathbf{w}$ that is agnostic of context.

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 $w^1, \dots, w^k$ such that $cat(w^1, \dots, w^k) = w$ where $cat(\cdot)$ indicates concatenation. Then, every layer of the model computes vectors $\mathbf{w}_c^1, \dots, \mathbf{w}_c^k$. Given these vectors, we consider four pooling mechanisms to compute $\mathbf{w}_c$:

$$\mathbf{w}_c = f(\mathbf{w}_c^1, \dots, \mathbf{w}_c^k)$$
$$f \in \{\min, \max, \mathrm{mean}, \mathrm{last}\}$$

$\min(\cdot)$, $\max(\cdot)$ are element-wise min/max pooling, $\mathrm{mean}(\cdot)$ is the arithmetic mean and $\mathrm{last}(\cdot)$ indicates selecting the last vector, $\mathbf{w}_c^k$.

Context Combination. Next, we describe two approaches for specifying contexts $c_1, \dots, c_n$ and combining the associated representations $\mathbf{w}_{c_1}, \dots, \mathbf{w}_{c_n}$.

- **Decontextualized:** For a word w, we use a single context $c_1 = w$. That is, we feed the single word w into the pretrained model and use the outputted vector as the representation of w (applying subword pooling if the word is split into multiple subwords).

- **Aggregated:** Since the **Decontextualized** strategy presents an unnatural input to the pretrained encoder, which likely never encountered w in isolation, we instead aggregate representations of w across multiple contexts. In particular, we sample n sentences from a text corpus $\mathcal{D}$ (see §A.2) each of which contains the word w, and compute the vectors $\mathbf{w}_{c_1}, \dots, \mathbf{w}_{c_n}$. Then, we apply a pooling strategy to yield a single representation that aggregates representations across contexts:

$$\mathbf{w} = g(\mathbf{w}_{c_1}, \dots, \mathbf{w}_{c_n});\ g \in \{\min, \max, \mathrm{mean}\}$$

3 Setup

We begin by verifying that the resulting static embeddings that we derive retain their representational

[2] A humanist's outlook on the (in)accessibility of BERT: https://tedunderwood.com/2019/07/15/do-humanists-need-bert/

strength, to some extent. We take this step to ensure that properties we observe of the static embeddings can be attributed to, and are consistent with, the original contextualized representations. Inspired by concerns with probing methods/diagnostic classifiers (Liu et al., 2019a; Hewitt and Liang, 2019) regarding whether learning can be attributed to the classifier and not the underlying representation, we employ an exceptionally simple parameter-free method for converting from contextualized to static representations to ensure that any properties observed in the latter are not introduced via this process.

When evaluating static embedding performance, we consider Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) embeddings as baselines since they have been the most prominent pretrained static embeddings for several years. Similarly, we begin with BERT as the contextualized model as it is currently the most prominent in downstream use among the growing number of alternatives. We provide identical analyses for 4 other contextualized model architectures (GPT-2 (Radford et al., 2019), XLNet (Yang et al., 2019), RoBERTa (Liu et al., 2019b), DistilBERT (Sanh et al., 2019)) and, in total, 9 sets of pretrained weights. All models, weights, and naming conventions used are enumerated in Appendix C and Table 9. Additional representation quality results appear in Tables 4–7 and Figures 4–10. We primarily report results for `bert-base-uncased`; further results for `bert-large-uncased` appear in Figure 3.

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)	500K	0.7206	0.7038	0.5019	0.3550
BERT-24 (1)	500K	0.7367	0.7074	0.5114	0.3687
BERT-24 (6)	500K	0.7494	0.7282	0.5116	0.4062
BERT-12	10K	0.5167 (1)	0.6833 (1)	0.4573 (1)	0.3043 (1)
BERT-12	100K	0.6980 (1)	0.7023 (1)	0.5007 (3)	0.3494 (3)
BERT-12	500K	0.7262 (2)	0.7038 (1)	0.5115 (3)	0.3853 (4)
BERT-12	1M	0.7242 (1)	0.7048 (1)	0.5134 (3)	0.3948 (4)
BERT-24	100K	0.7749 (2)	0.7179 (6)	0.5044 (1)	0.3686 (9)
BERT-24	500K	0.7643 (2)	0.7282 (6)	0.5116 (6)	0.4146 (10)
BERT-24	1M	**0.7768** (2)	**0.7301** (6)	**0.5244** (15)	**0.4280** (10)

Table 1: Performance of distilled BERT embeddings. f and g are set to mean and (#) indicates the layer the embeddings are distilled from. **Bold** indicates best performance for a given dataset of embeddings 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. f and g are set to mean, $N = 100\text{K}$, and (#) indicates the layer the embeddings are distilled from. **Bold** indicates best performance for a given dataset of embeddings depicted.

4 Representation Quality

4.1 Evaluation Details

To assess the representational quality of our static embeddings, we evaluate on several word similarity and word relatedness datasets.[3] We consider 4 such datasets: RG65 (Rubenstein and Goodenough, 1965), WS353 (Agirre et al., 2009), SimLex999 (Hill et al., 2015) and SimVerb3500 (Gerz et al., 2016) (see §A.4 for more details). Taken together, these datasets contain 4917 examples and specify 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) of the semantic similarity or relatedness between w_1 and w_2. 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.

4.2 Results

Pooling Strategy. In Figure 1, we show the performance on all 4 datasets for the resulting static embeddings. For embeddings computed using the **Aggregated** strategy, representations are aggregated over $N = 100\text{K}$ sentences where N is the number of total contexts for all words (§A.5). Across all four datasets, we see that $g = \mathrm{mean}$ is the best-performing pooling mechanism within the **Aggregated** strategy and also outperforms the **Decontexualized** strategy by a substantial margin. Fixing $g = \mathrm{mean}$, we further observe that mean pooling at the subword level also performs best (the dark green dashed line in all plots). We further find that this trend consistently holds across pretrained models.

Number of Contexts. In Table 1, we see that performance for both BERT-12 and BERT-24 steadily increases across all datasets with increas-

[3] Concerns with this decision are discussed in §A.3.

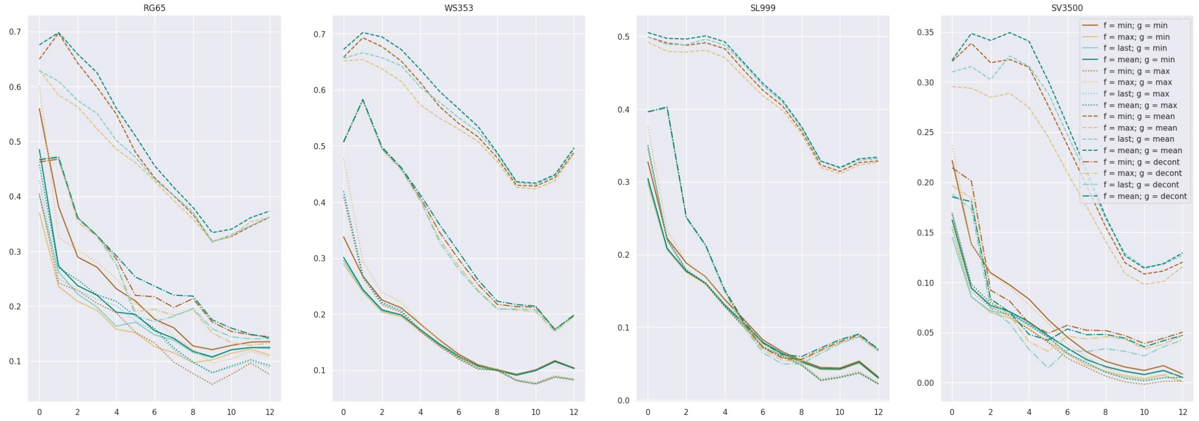

Figure 1: Layer-wise performance of distilled BERT-12 embeddings for all pairs (f, g) with $N = 100\text{K}$.

ing N; this trend holds for the other 7 pretrained models. In particular, in the largest setting with $N = 1\text{M}$, the BERT-24 embeddings distilled from the best-performing layer for each dataset drastically outperform both Word2Vec and GloVe. However, this can be seen as an unfair comparison given that we are selecting specific layers for specific datasets. As the middle band of Table 1 shows, we can fix a particular layer for all datasets and still outperform both Word2Vec and GloVe on all datasets.

Relationship between N and model 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. A similar preference for the first quarter of the model is observed in models with a different number of layers (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. (2019a) regarding where lexical semantic information is best encoded in pretrained contextualized models. However, as we pool over a larger number of contexts, Table 1 reveals an interesting relationship between N and the best-performing layer. The best-performing layer monotonically (with a single exception) shifts to be later and later within the pretrained model. 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 improves performance.[4] Since later layers of the model are generally preferred by downstream practitioners (Zhang et al., 2020), our findings suggest that downstream performance could be further improved by considering variance reduction as we suggest; Ethayarajh (2019) also provides concrete evidence of the tremendous variance in the later layers of deep pretrained contextualized models.

Cross-Model Results. 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). This is particularly noteworthy given that several works have found that different contextualized models pattern substantially differently (Liu et al., 2019a; Ethayarajh, 2019).

In Table 2, we summarize the performance of all models we studied. All of the models considered were introduced during a similar time period and have comparable properties in terms of downstream performance. In spite of this, we observe that their static analogues perform radically differently. For example, several do not reliably outperform Word2Vec and GloVe despite outperforming Word2vec and GloVe reliably in downstream evaluation. Future work may consider whether the reduction to static embeddings affects different models differently and whether this is reflective of the quality of context-agnostic lexical semantics from other types of linguistic knowledge (e.g. context modelling, syntactic understanding, and semantic composition). In general, these results

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

provide further evidence to suggest that linguistic understanding captured by different pretrained weights may be substantially different, even for models with near-identical Transformer (Vaswani et al., 2017) architectures.

Somewhat surprisingly, in Table 2, DistilBert-6 outperforms BERT-12 on three out of the four datasets despite being distilled (Ba and 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.

5 Bias

Bias is a complex and highly relevant topic in developing representations and models in NLP and ML. In this context, we study the social bias encoded within our static word representations as a proxy for understanding biases of the source contextualized 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 is starkly incomplete in terms of social biases of interest). Further, as we will show, different computational specifications of bias that evaluate the same underlying social phenomena yield markedly different results. As a direct consequence, we strongly caution that the results **must** 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. Instead, it satisfies the significantly weaker condition 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 measure of 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 $\mathbf{g}$:

$$\mathbf{g} = \texttt{PCA}\big([\mathbf{m}_1 - \mathbf{f}_1, \ldots, \mathbf{m}_n - \mathbf{f}_n]\big)[0]$$

where $[0]$ indicates 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:

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

This definition is only inherently applicable to binary bias settings, i.e. where there are exactly two *protected classes*. Multi-class generalizations are difficult to realize since constructing $\mathcal{P}$ requires aligned k-tuples whose entries only differ in the underlying social attribute and this becomes increasingly challenging for increasing k. Further, this definition assumes the first principal component explains a large fraction of the observed variance.

Garg et al. (2018) introduced a different definition that is not restricted to gender and 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, $\mu_i = \underset{w\in\mathcal{A}_i}{\text{mean}}\, \mathbf{w}$ is computed. Garg et al. (2018) computes the bias in two ways:

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

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

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). To permit comparison, we insert absolute values around 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 of representative words $\mathcal{A}_1, \ldots, \mathcal{A}_k$[5]:

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

5.2 Results

Inspired by the results of Nissim et al. (2020), in this work we transparently report social bias in existing static embeddings as well as the embeddings we produce. In particular, we exhaustively report the measured bias for all 3542 valid *(pretrained model, layer, social attribute, bias definition, target word list)* 5-tuples — all possible combinations of static embeddings and bias measures considered. The results for models beyond BERT appear in Figures 11–18.

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). We study bias with respect to target word lists of professions $\mathcal{N}_{prof}$ and adjectives $\mathcal{N}_{adj}$. These results are by no means intended to be comprehensive with regards to the breadth of bias socially and only address a restricted subset of social biases which notably does not include 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 definitions of bias are grounded in the norms of the United States. All word lists are provided in Appendix B and are sourced in §A.6.

Layer-wise Bias Trends. In Figure 2, we report layer-wise bias across all (*attribute, definition*) pairs. We clearly observe that for every social attribute, there is a great deal of variation across the layers in the quantified amount of bias for a fixed bias estimator. Further, while we are not surprised that different bias measures for the same social attribute and the same layer 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 $\underset{\text{GARG-EUC}}{\text{bias}}$ we see a distribution with two peaks corresponding to layers at the start or end of the pretrained model with less bias within the intermediary layers. For estimating the same quantity, $\underset{\text{GARG-COS}}{\text{bias}}$ is mostly uniform across the layers. Similarly, in looking at the religious bias, we see similar inconsistencies with the bias increasing monotonically from layers 2 through 8 under $\underset{\text{MANZINI}}{\text{bias}}$, decreasing monotonically under $\underset{\text{GARG-EUC}}{\text{bias}}$, and remaining roughly constant under $\underset{\text{GARG-COS}}{\text{bias}}$. In general, while the choice of $\mathcal{N}$ (and the choice of $\mathcal{A}_i$ for gender) does affect the absolute bias estimates, the relative trends across layers are fairly robust to these choices for a specific definition.

Consequences. 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 and Goldberg, 2019).

Cross-Model Bias Trends. 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 $\mathcal{N}_{prof}$ due to the aforementioned stability (and for visual clarity) with results for $\mathcal{N}_{adj}$ appearing in Table 8. Since we do not preprocess or normalize embeddings, the scores using $\underset{\text{GARG-EUC}}{\text{bias}}$ are incomparable (and may be improper to compare in the layer-wise case) as

[5] We slightly modify the definition of Manzini et al. (2019) by (a) using cosine similarity where they use cosine distance and (b) inserting absolute values around each term in the mean over $\mathcal{N}$. We make these changes to introduce consistency with the other definitions and to permit comparison.

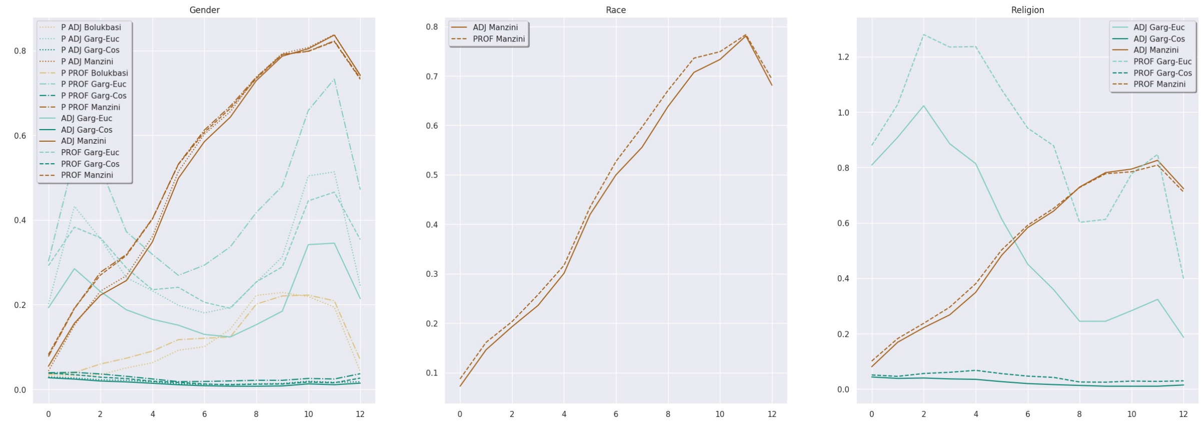

Figure 2: Layer-wise bias of distilled BERT-12 embeddings for $f = \text{mean}, g = \text{mean}, N = 100\text{K}$.

	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 encoded within different pretrained models with respect to a set of professions $\mathcal{N}_{prof}$. Parameters are discussed in the supplement. Lowest bias in a particular column is denoted in **bold**.

they are sensitive to the absolute norms of the embeddings.[6] Further, we note that $\text{bias}_{\text{BOLUKBASI}}$ may not be a reliable indicator since the first principal component explains less than 35% of the variance for the majority of distilled embedding (Zhao et al. (2019a) show similar findings for ELMo). For $\text{bias}_{\text{MANZINI}}$ and $\text{bias}_{\text{GARG-COS}}$, 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 embeddings consistently get high scores when compared to other distilled embeddings but under $\text{bias}_{\text{GARG-COS}}$ they are deemed the least biased.

Data alone does not determine bias. Comparing the results for BERT-12 and BERT-24 (full layer-wise results for BERT-24 appear in Figure 11) reveals that bias trends for BERT-12 and BERT-24 are starkly different for any fixed bias measure. What this indicates is the bias observed in contextualized models is not strictly determined by the training data (as these models share the same training data as do all other 12 and 24 model pairs) and must also be a function of the architecture, training procedure, and/or random initialization.

Takeaways. Ultimately, given the aforementioned issues regarding the reliability of bias measures, it is difficult to arrive at clear consensus of the how the bias encoded compares between our distilled representations 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 estimates.

6 Related Work

Contextualized → Static. Recently, Akbik et al. (2019) introduced an approach that gradually aggregates representations during training to accumulate global information and demonstrated improvements over only contextualized representations for NER. May et al. (2019) instead syntheti-

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

cally construct a single *semantically-bleached* sentence which is fed into a sentence encoder to yield a static representation. In doing so, they introduce SEAT as a means for studying biases in sentence encoders by applying WEAT (Caliskan et al., 2017) to the resulting static representations. This approach appears inappropriate for quantifying bias in sentence encoders[7] as sentence encoders are trained on semantically-meaningful sentences and semantically-bleached constructions are not representative of this distribution and their templates heavily rely on *deictic expressions* which are difficult to adapt for certain syntactic categories such as verbs (as required for SimVerb3500 especially). Given these concerns, our reduction method may be preferable for use in estimation of bias in contextualized representations. Due to the fact that we use mean-pooling, our approach may lend itself to interpretations of the bias in a model on average across contexts.

Ethayarajh (2019) considers a similar method to ours where pooling is replaced by PCA. While this work demonstrated contextualized representations are highly contextual, our work naturally explores the complementary problem of what value can be extracted from the static analogue of these representations.

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 downstream behavior (Zhao et al., 2017, 2018a, 2019a; Stanovsky et al., 2019).[8] Our bias evaluation is in the style of (a) and we consider multi-class social bias in the lens of gender, race, and religion whereas prior work has centered on binary gender. Additionally, while most prior work has discussed the static embedding setting, recent work has considered sentence encoders and contextualized models. Zhao et al. (2019a) consider gender bias in ELMo when applied to coreference systems and Kurita et al. (2019) extend these results by leveraging the masked language modeling objective of BERT. Similarly, Basta et al. (2019) considers intrinsic gender bias in ELMo via 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 many of 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 practitioners often do not use the last layer of deep Transformer models (Liu et al., 2019a; Zhang et al., 2020; Zhao et al., 2019b).[9]

7 Future Directions

Our work furnishes multiple insights about pretrained contextualized models that suggest changes (subword pooling, layer choice, beneficial variance reduction via averaging across contexts) to improve downstream performance. Recent models have combined static and dynamic embeddings (Peters et al., 2018; Bommasani et al., 2019; Akbik et al., 2019) and our representations may also support drop-in improvements in these settings.

While not central to our goals, we discovered that our static embeddings substantially outperform Word2Vec and GloVe under intrinsic evaluation. Future research may consider downstream gains as improved static embeddings are critical for resource-constrained settings and may help address environmental concerns in NLP (Strubell et al., 2019), machine learning (Canziani et al., 2016), and the broader AI community (Schwartz et al., 2019). Future research could explore weighting schema in the averaging process analogous to SIF (Arora et al., 2017) for sentence representations computed via averaging (Wieting et al., 2016).

The generality of the proxy analysis method implies that other interpretability methods for static embeddings can also be considered. Further, post-processing approaches beyond analysis/interpretability such as dimensionality reduction may be particularly intriguing given that this is often challenging to perform within large multi-layered networks like BERT (Sanh et al., 2019) but has been successfully demonstrated for static embeddings (Nunes and Antunes, 2018; Mu and Viswanath, 2018; Raunak et al., 2019).

Future work may revisit the choice of the corpus $\mathcal{D}$ from which contexts are drawn. For downstream use, setting $\mathcal{D}$ to be the target domain may serve as a lightweight domain adaptation strategy similar to findings for averaged word representations for

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

[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).

out-of-domain settings (Wieting et al., 2016).

8 Discussion and Open Problems

While our work demonstrates that contextualized representations retain substantial representational power even when reduced to be noncontextual, it is unclear what information is lost. After all, contextualized representations have been so effective precisely because they are tremendously contextual (Ethayarajh, 2019). As such, the validity of treating the resulting static embeddings as reliable proxies for the original contextualized model still remains open.

On the other hand, human language processing has often been conjectured to have both context-dependent and context-independent properties (Barsalou, 1982; Rubio-Fernández, 2008; Depraetere, 2014, 2019). Given this divide, our approach may provide an alternative mechanism for clarifying how these two properties interact in the computational setting from both an interpretability standpoint (i.e. comparing results for analyses on the static embeddings and the original contextualized representations) and a downstream standpoint (i.e. comparing downstream performance for models initialized using the static embeddings and the original contextualized representations). However, the precise relationship between the role of context in human language processing and computational language processing remains unclear.

Theoretical explanation for the behavior we observe in two settings is also needed. First, it is unclear why learning contextualized representations and then reducing them to static embeddings drastically outperforms directly learning static embeddings. In particular, the GloVe embeddings we use are learned using 6 billion tokens whereas the BERT representations were trained on roughly half as much data (3.3 billion tokens). Perhaps the behavior is reminiscent of the benefits of modelling in higher dimensional settings temporarily as is seen in other domains (e.g. the kernel trick and Mercer's theorem for learning non-linear classifiers using inner product methods): begin by recasting the problem in a more expressive space (contextualized representations) and then project/reduce to the original space (static embeddings). Second, the reason for the benefits of the variance reduction that we observe are unclear. Given that best-performing mechanism is to average over many contexts, it may be that approaching the asymptotic mean of the distribution across contexts is desirable/helps combat the anisotropy that exists in the original contextualized space (Ethayarajh, 2019).

9 Conclusion

In this work, we consider how methods developed for analyzing static embeddings can be re-purposed for understanding contextualized representations. We introduce simple and effective procedures for converting from contextualized representations to static word embeddings. When applied to pretrained models like BERT, we find the resulting embeddings are useful proxies that provide insights into the pretrained model while simultaneously outperforming Word2Vec and GloVe substantially under intrinsic evaluation. We further study the extent to which various social biases (gender, race, religion) are encoded, employing several different quantification schemas. Our large-scale analysis reveals that bias is encoded disparately across different popular pretrained models and different model layers. Our findings also have significant implications with respect to the reliability of existing protocols for estimating bias in word embeddings.

10 Reproducibility

All data, code and visualizations are made publicly available.[10] Further details are explictly and comprehensively reported in Appendix A.

Acknowledgments

We thank Ge Gao, Marty van Schijndel, Forrest Davis, and members of the Mozilla DeepSpeech and Cornell NLP groups for their valuable advice. We especially thank the reviewers and area chairs for their articulate and constructive feedback.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. 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*, pages 19–27, Boulder, Colorado. Association for Computational Linguistics.

Alan Akbik, Tanja Bergmann, and Roland Vollgraf. 2019. Pooled contextualized embeddings for named

[10] https://github.com/rishibommasani/Contextual2Static

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)*, pages 724–728, Minneapolis, Minnesota. Association for Computational Linguistics.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations*.

Jimmy Ba and Rich Caruana. 2014. Do deep nets really need to be deep? In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2654–2662. Curran Associates, Inc.

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

Lawrence W Barsalou. 1982. Context-independent and context-dependent information in concepts. *Memory & cognition*, 10(1):82–93.

Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy. Association for Computational Linguistics.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. 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, editors, *Advances in Neural Information Processing Systems 29*, pages 4349–4357. Curran Associates, Inc.

Rishi Bommasani, Arzoo Katiyar, and Claire Cardie. 2019. SPARSE: Structured prediction using argument-relative structured encoding. In *Proceedings of the Third Workshop on Structured Prediction for NLP*, pages 13–17, Minneapolis, Minnesota. Association for Computational Linguistics.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

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

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Ronan Collobert and Jason Weston. 2008. 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, pages 160–167, New York, NY, USA. ACM.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537.

Ryan Cotterell and Hinrich Schütze. 2015. Morphological word-embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1287–1292, Denver, Colorado. Association for Computational Linguistics.

Ilse Depraetere. 2014. Modals and lexically-regulated saturation. *Journal of Pragmatics*, 71:160–177.

Ilse Depraetere. 2019. Meaning in context and contextual meaning: A perspective on the semantics-pragmatics interface applied to modal verbs. *Anglophonia*, 28.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 30–35, Berlin, Germany. Association for Computational Linguistics.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. SimVerb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2173–2182, Austin, Texas. Association for Computational Linguistics.

Anna Gladkova and Aleksandr Drozd. 2016. Intrinsic evaluations of word embeddings: What can we do better? In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 36–42, Berlin, Germany. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Souleiman Hasan and Edward Curry. 2017. Word re-embedding via manifold dimensionality retention. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 321–326, Copenhagen, Denmark. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word 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)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014a. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014b. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, Ann Arbor, Michigan. Association for Computational Linguistics.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Bofang Li, Tao Liu, Zhe Zhao, Buzhou Tang, Aleksandr Drozd, Anna Rogers, and Xiaoyong Du. 2017. 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*, pages 2421–2431, Copenhagen, Denmark. Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. 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)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Edward Loper and Steven Bird. 2002. 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, pages 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. 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)*, pages 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. 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)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 14014–14024. Curran Associates, Inc.

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

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

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*.

Malvina Nissim, Rik van Noord, and Rob van der Goot. 2020. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, pages 1–17.

Davide Nunes and Luis Antunes. 2018. Neural random projections for language modelling. *CoRR*, abs/1807.00930.

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

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. 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)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

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

Vikas Raunak, Vivek Gupta, and Florian Metze. 2019. Effective dimensionality reduction for word embeddings. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 235–243, Florence, Italy. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *ArXiv*, abs/2002.12327.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Commun. ACM*, 8(10):627–633.

Paula Rubio-Fernández. 2008. Concept narrowing: The role of context-independent information. *Journal of semantics*, 25(4):381–409.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

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

Dinghan Shen, Pengyu Cheng, Dhanasekar Sundararaman, Xinyuan Zhang, Qian Yang, Meng Tang, Asli Celikyilmaz, and Lawrence Carin. 2019. Learning compressed sentence representations for on-device text processing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 107–116, Florence, Italy. Association for Computational Linguistics.

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

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Karl Stratos. 2017. A sub-character architecture for Korean language processing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 721–726, Copenhagen, Denmark. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

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

Yi Chern Tan and L. Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 13230–13241. Curran Associates, Inc.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong, China. Association for Computational Linguistics.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Towards universal paraphrastic sentence embeddings. In *International Conference on Learning Representations*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 33*. Curran Associates, Inc.

Jinxing Yu, Xun Jian, Hao Xin, and Yangqiu Song. 2017. Joint embeddings of Chinese words, characters, and fine-grained subcharacter components. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 286–291, Copenhagen, Denmark. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019a. 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)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. 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*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019b. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 563–578, Hong Kong, China. Association for Computational Linguistics.

A Reproducibility Details

A.1 Additional Results

We provide layerwise model performance for all additional models in Figures 3-10 with corresponding tables for different N values (Tables 4-7). Similarly, we provide layerwise bias estimates for all additional models in Figures 11-18. Results for target words specified as adjectives are given in Table 8.

A.2 Data

We use English Wikipedia as the corpus $\mathcal{D}$ in context combination for the **Aggregated** strategy. The specific subset of English Wikipedia[11] used was lightly preprocessed with a simple heuristic to remove bot-generated content. Individual Wikipedia documents were split into sentences using `NLTK` (Loper and 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[12] and GloVe[13] 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 and Curry, 2017) and downstream tasks.

A.3 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 and Goldberg, 2014b) and there are significant concerns about the validity of these analogies tests (Nissim et al., 2020). We acknowledge that word similarity and word relatedness tasks have also been heavily scrutinized (Faruqui et al., 2016; Gladkova and 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 and 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.4 Representation Quality Dataset Trends

Rubenstein and 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 and 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 and 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

[11] `https://blog.lateral.io/2015/06/the-unknown-perils-of-mining-wikipedia/`

[12] `https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit`

[13] `https://nlp.stanford.edu/projects/glove/`

adjectives.

A.5 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[14]. 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 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.6 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 Appendix B 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 Word Lists

$\mathcal{N}_{\text{prof}}$ = {'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', 'geologist', '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',

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

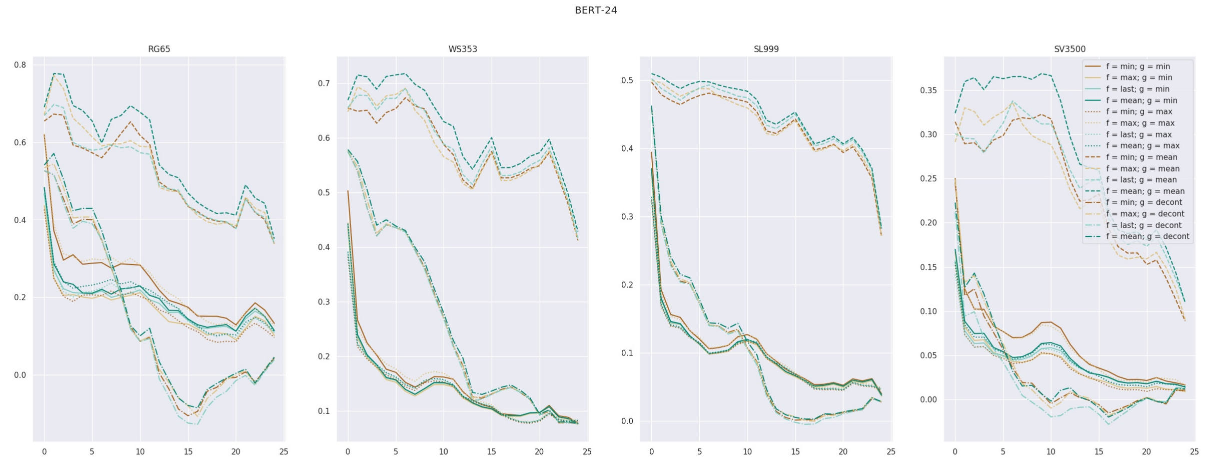

Figure 3: Layerwise performance of BERT-24 static embeddings for all possible choices of f, g

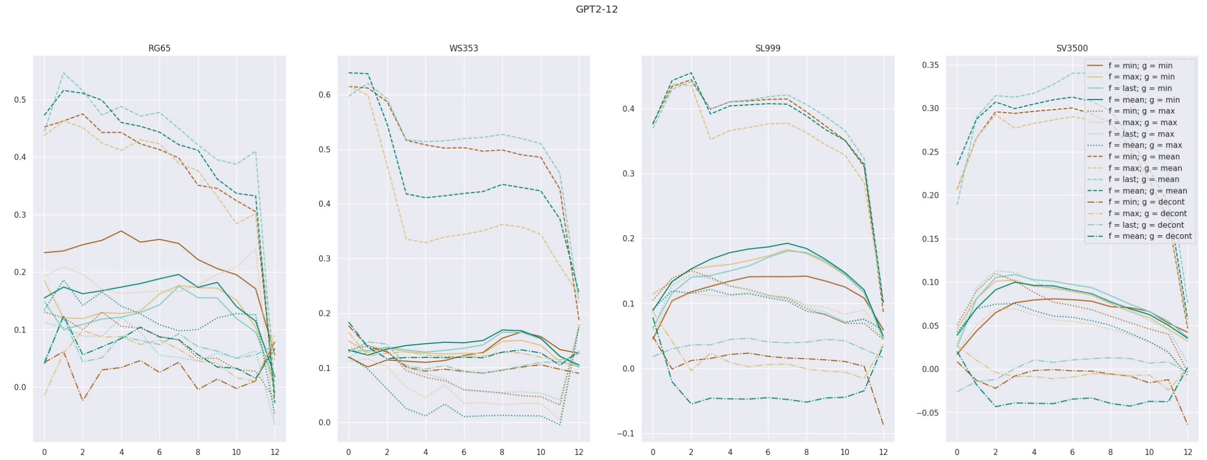

Figure 4: Layerwise performance of GPT2-12 static embeddings for all possible choices of f, g

'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',

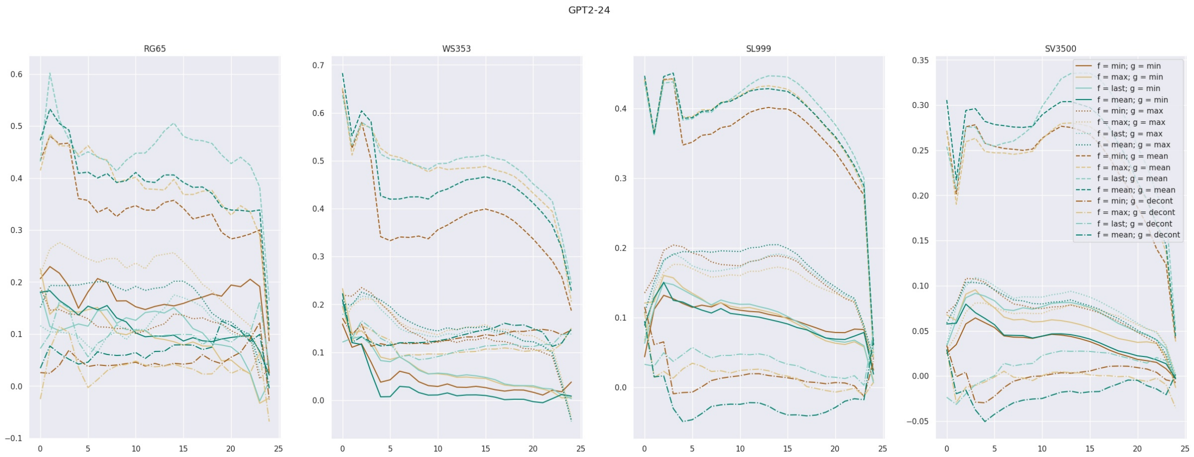

Figure 5: Layerwise performance of GPT-24 static embeddings for all possible choices of f, g

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.

‘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’, ‘ath-

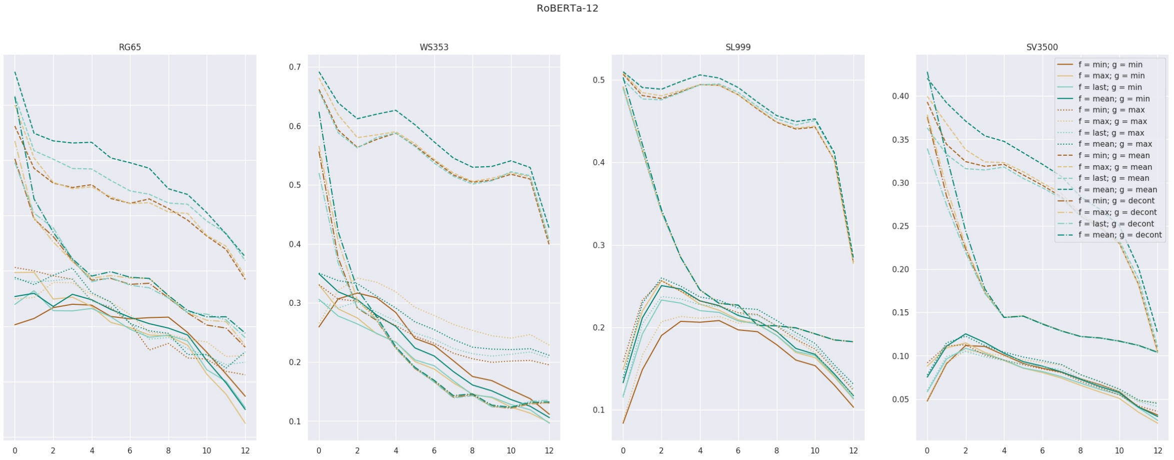

Figure 6: Layerwise performance of RoBERTa-12 static embeddings for all possible choices of f, g

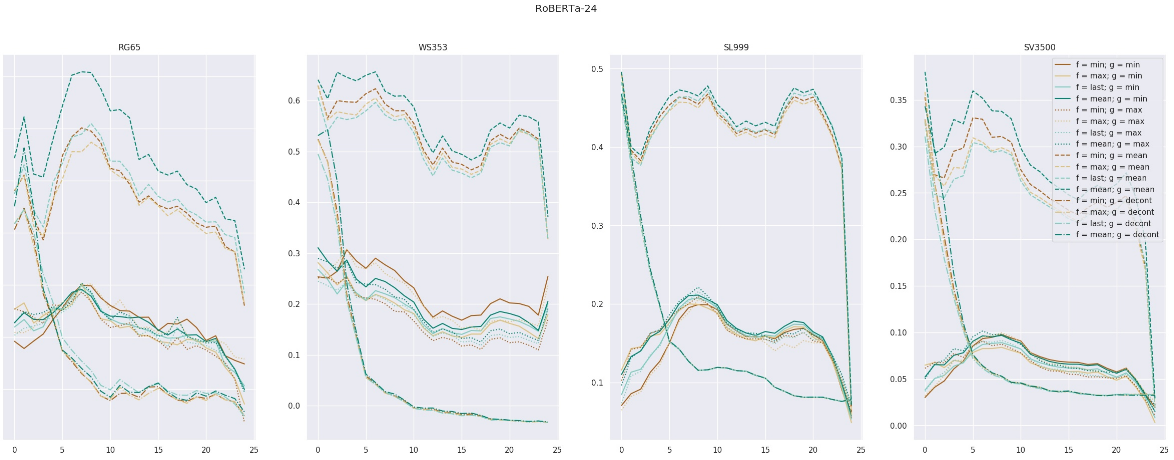

Figure 7: Layerwise performance of RoBERTa-24 static embeddings for all possible choices of f, g

letic', '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',

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.

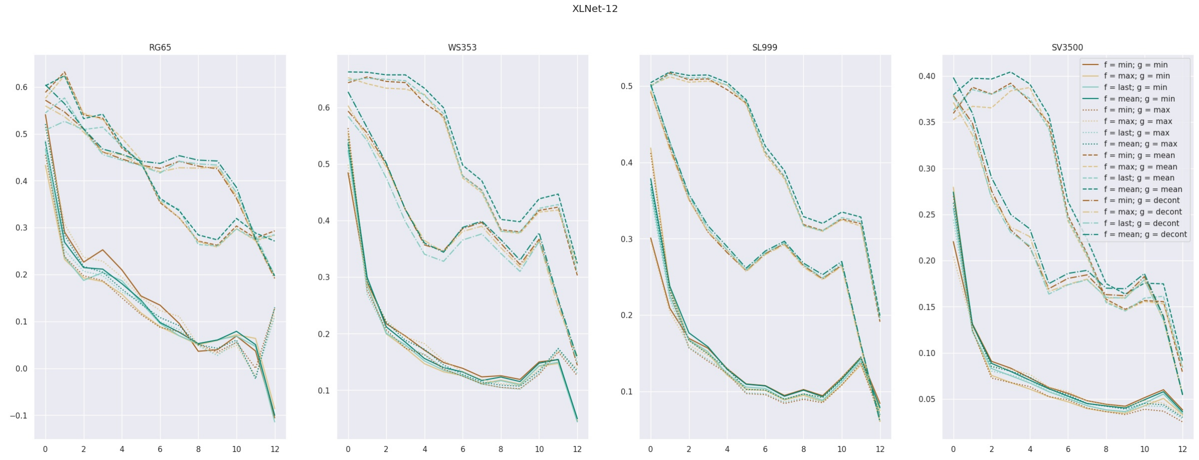

Figure 8: Layerwise performance of XLNet-12 static embeddings for all possible choices of f, g

‘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’[15], ‘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’, ‘gonza-

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

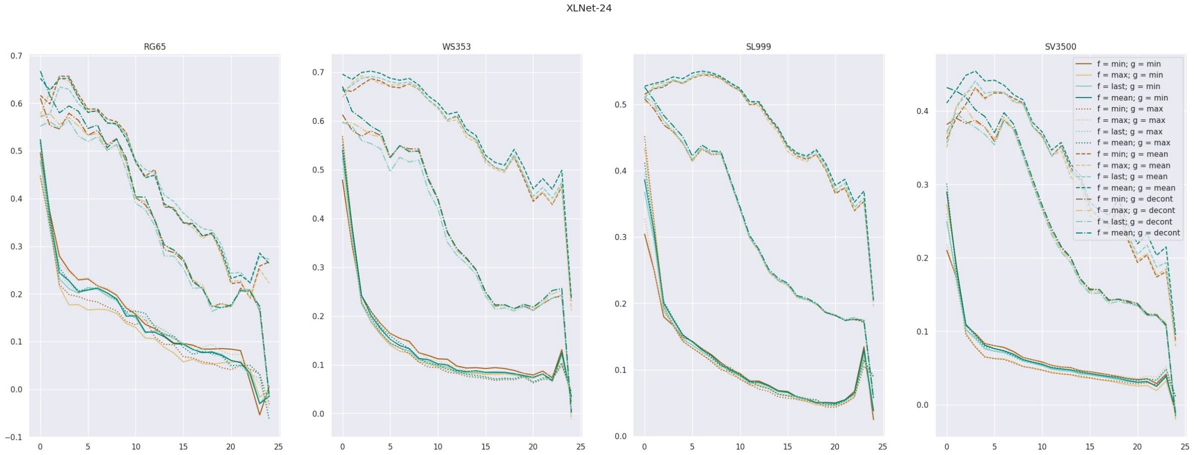

Figure 9: Layerwise performance of XLNet-24 static embeddings for all possible choices of f, g

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.

lez', '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'}

C 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.

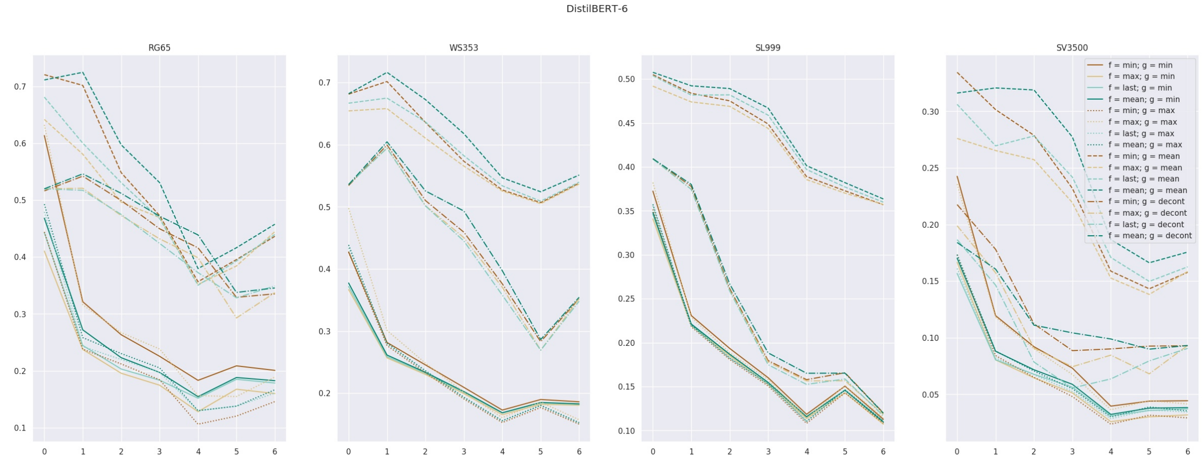

Figure 10: Layerwise performance of DistilBERT-6 static embeddings for all possible choices of f, g

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.

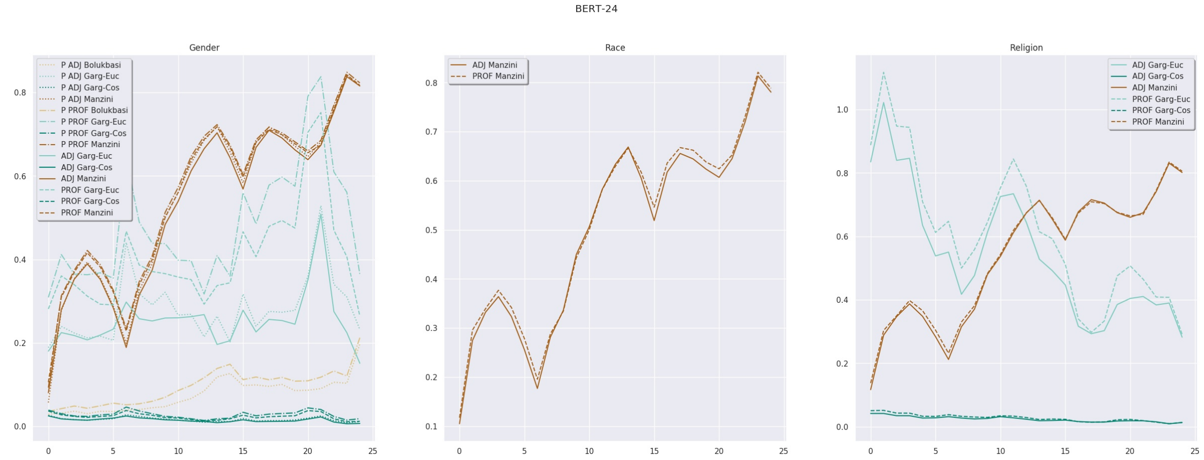

Figure 11: Layerwise bias of BERT-24 static embeddings for $f = \mathrm{mean}, g = \mathrm{mean}, N = 100000$
Left: Gender, **Center:** Race, **Right:** Religion

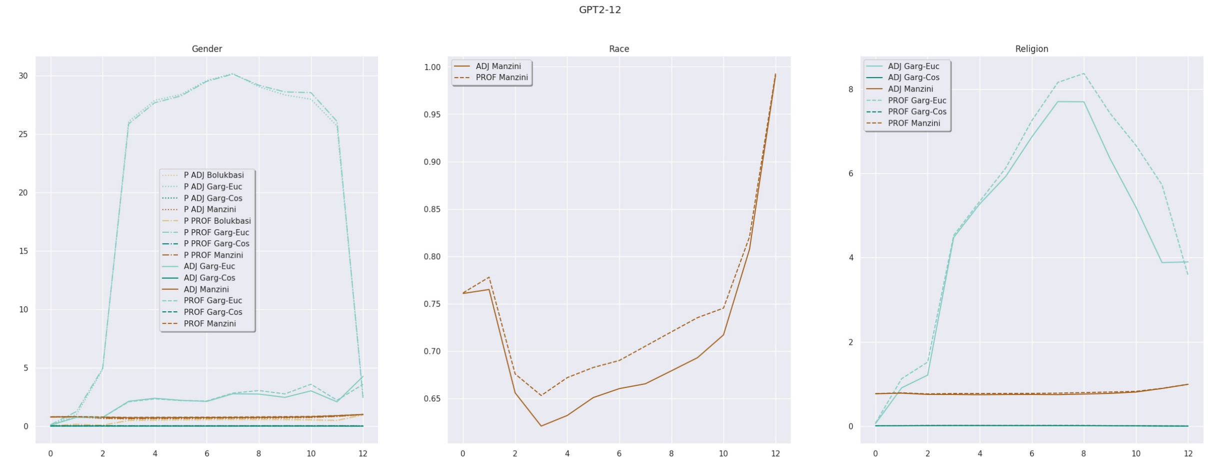

Figure 12: Layerwise bias of GPT2-12 static embeddings for $f = \text{mean}, g = \text{mean}, N = 100000$
Left: Gender, **Center:** Race, **Right:** Religion

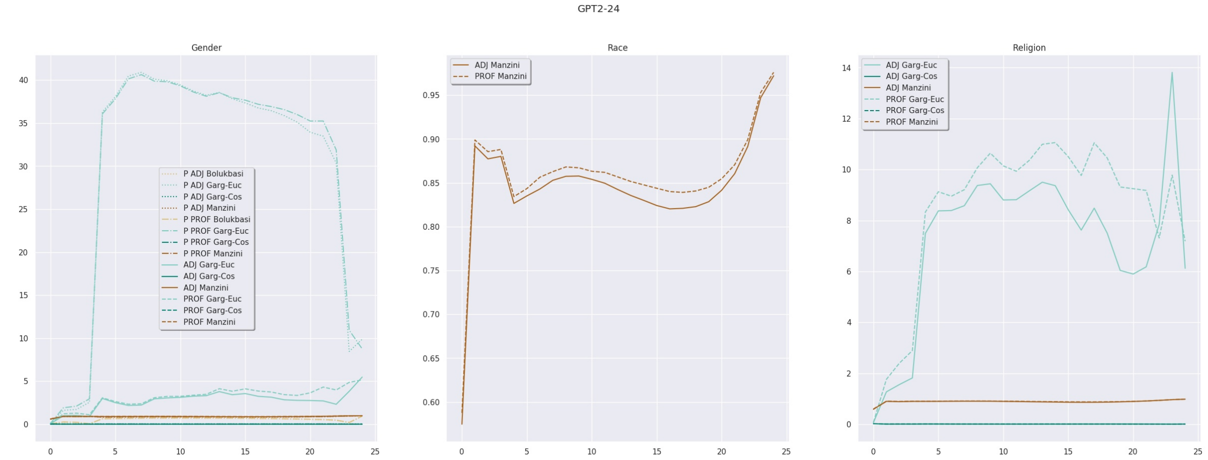

Figure 13: Layerwise bias of GPT2-24 static embeddings for $f = \text{mean}, g = \text{mean}, N = 100000$
Left: Gender, **Center:** Race, **Right:** Religion

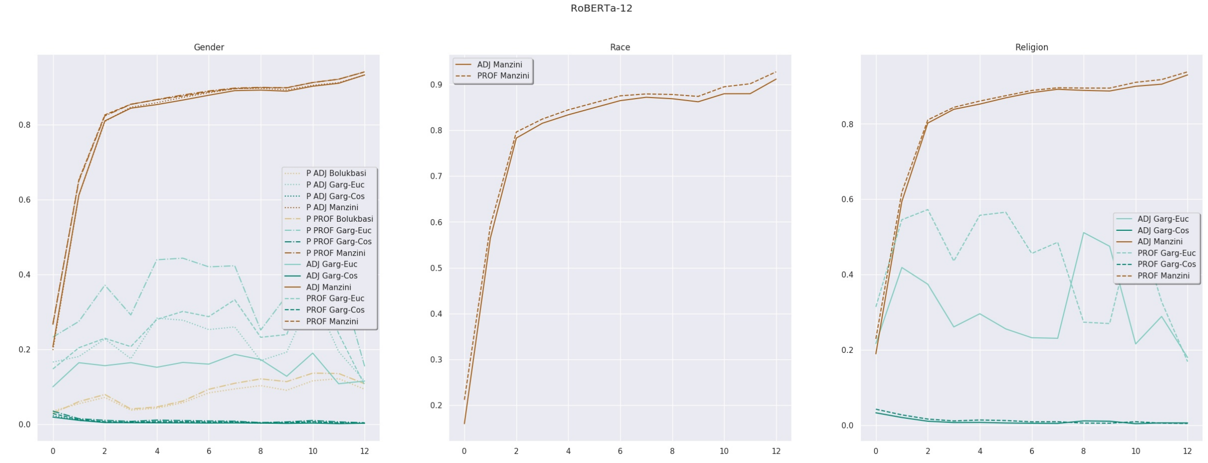

Figure 14: Layerwise bias of RoBERTa-12 static embeddings for $f = \text{mean}, g = \text{mean}, N = 100000$
Left: Gender, **Center:** Race, **Right:** Religion

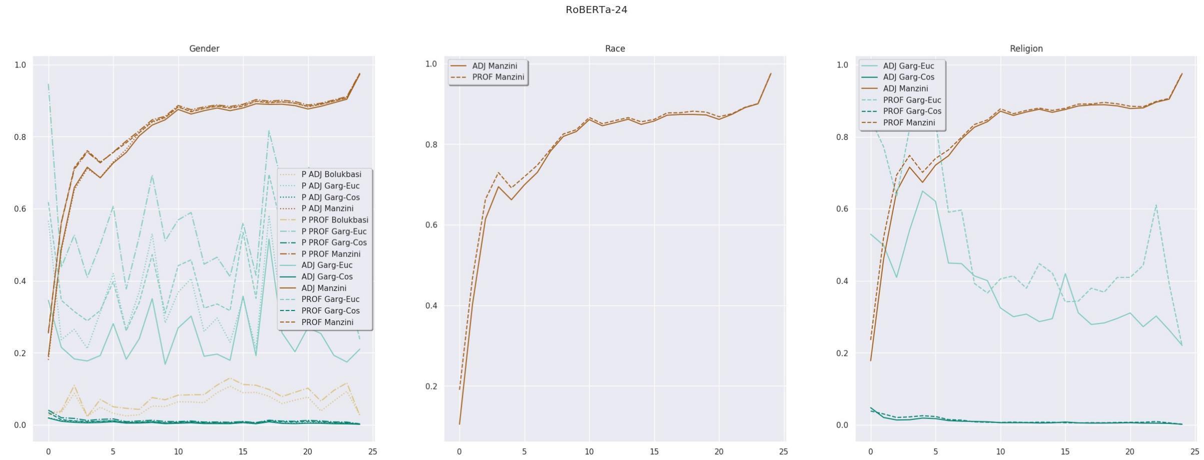

Figure 15: Layerwise bias of RoBERTa-24 static embeddings for f = mean, g = mean, N = 100000
Left: Gender, **Center:** Race, **Right:** Religion

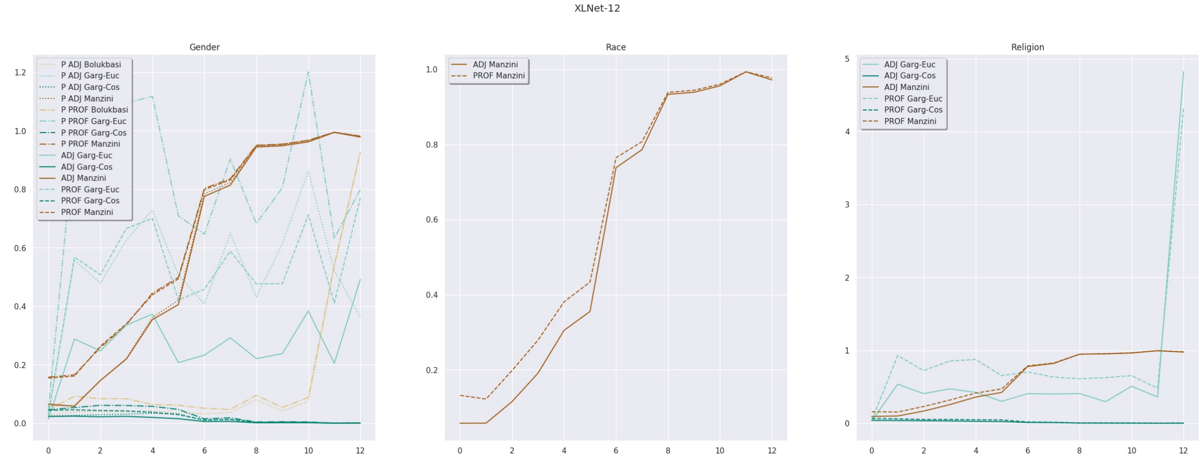

Figure 16: Layerwise bias of XLNet-12 static embeddings for f = mean, g = mean, N = 100000
Left: Gender, **Center:** Race, **Right:** Religion

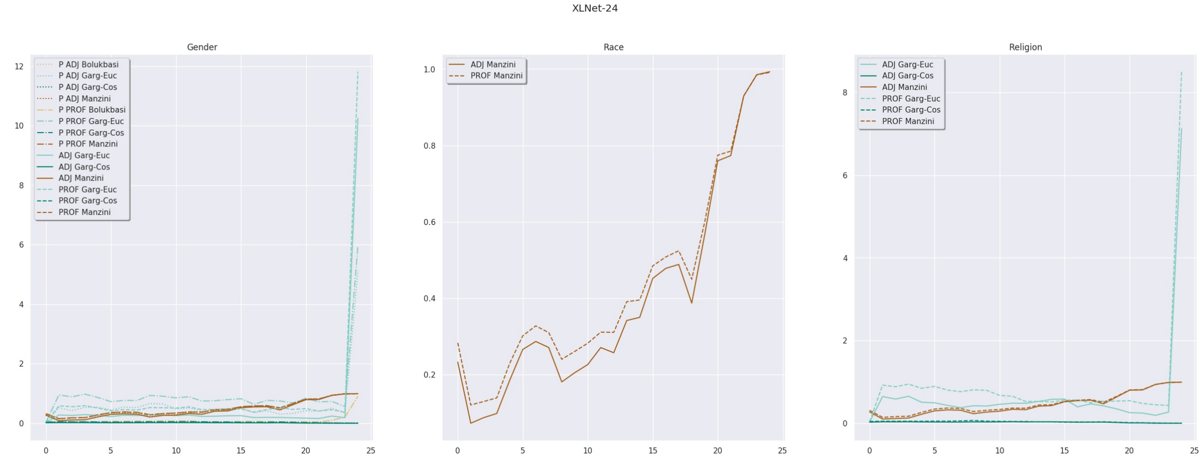

Figure 17: Layerwise bias of XLNet-24 static embeddings for f = mean, g = mean, N = 100000
Left: Gender, **Center:** Race, **Right:** Religion

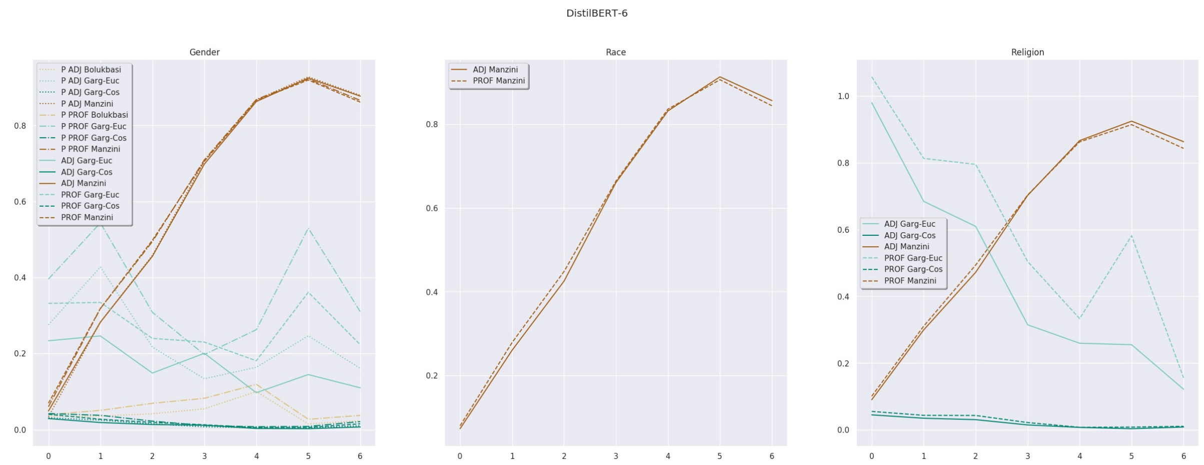

Figure 18: Layerwise bias of DistilBERT-6 static embeddings for $f = \text{mean}, g = \text{mean}, N = 100000$
Left: Gender, **Center:** Race, **Right:** Religion

	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 $\lfloor \frac{X}{4} \rfloor$.

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