INTERPRETING PRETRAINED CONTEXTUALIZED REPRESENTATIONS VIA REDUCTIONS TO STATIC EMBEDDINGS

RISHI BOMMASANI

KELLY DAVIS

CLAIRE CARDIE

BACKDROP

- Interpreting Pretrained Contextualized Representations
  - BERTology, ACL Interpretability Track, BlackBoxNLP
- New Interpretability Techniques
  - Probing Classifiers, Attention (Heads), Edge Probing

PREMISE

- Contextualized → Context-Agnostic

dog	-0.606	0.044	0.002
sesame	-0.222	-0.382	0.339
street	0.921	0.991	-0.922
natural	0.789	0.827	-0.648
language	0.669	-0.448	-0.685
processing	0.738	0.199	0.017

METHODOLOGY

REMOVING CONTEXT – SUBWORD POOLING

- Contextualized Subword Vector(s) → Contextualized Word Vector

$$context\ c,\ word\ w \in c,\ subwords\ w^1, \ldots, w^k$$

$$\textbf{Input:}\ \text{Contextual subword vectors}\ \mathbf{w}_c^1, \ldots, \mathbf{w}_c^k$$

$$\textbf{Output:}\ \text{Contextual word vector}\ \mathbf{w}_c \triangleq f\left(\mathbf{w}_c^1, \ldots, \mathbf{w}_c^k\right)$$

$$f \in \{min,\ max,\ mean,\ last\}$$

REMOVING CONTEXT – CONTEXT COMBINATION

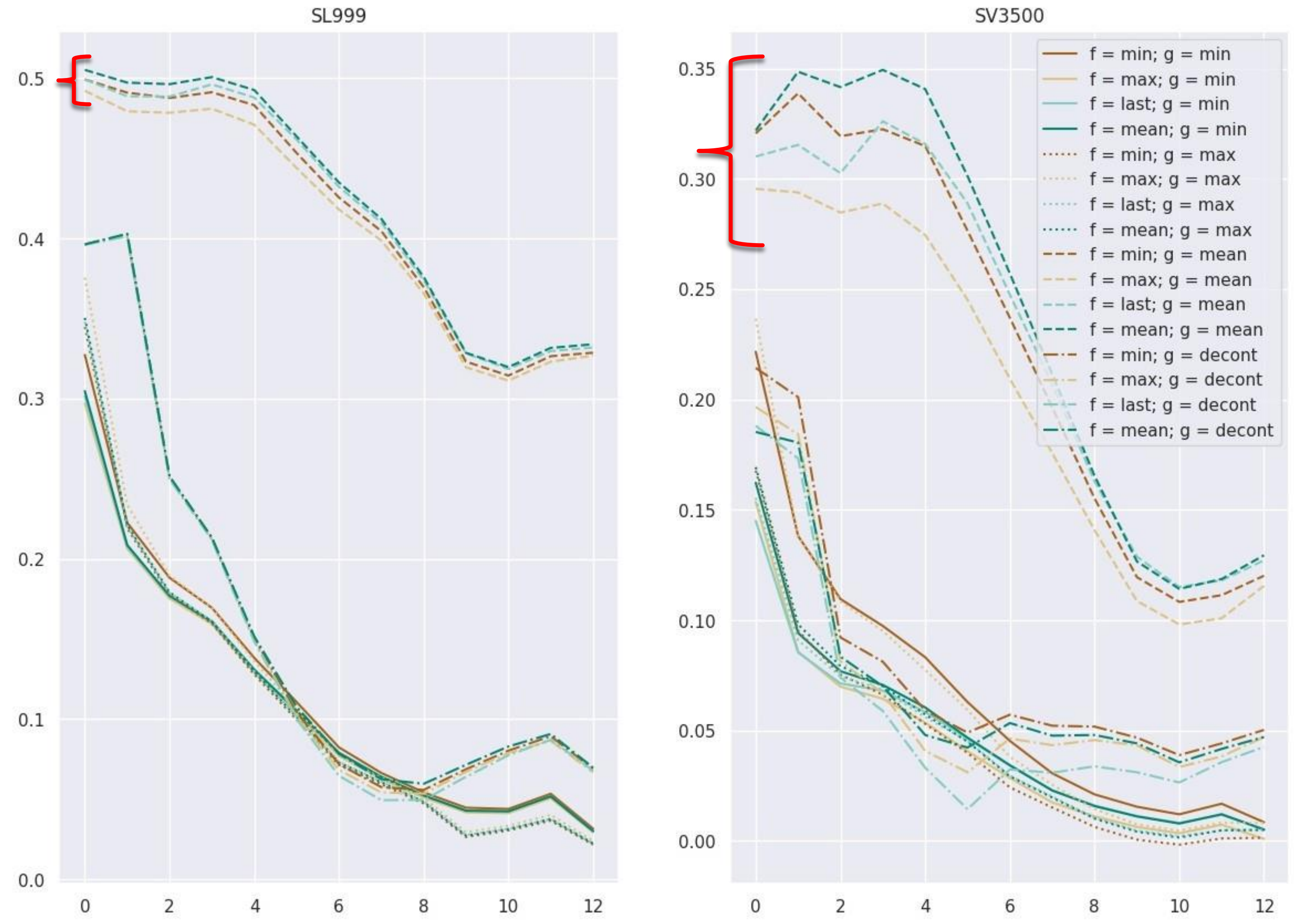
- Contextualized Word Vector(s) → Context-Agnostic Word Vector
  - **Decontextualize**: Specify a single context $c = w$
  - **Aggregate**: Pool across many contexts $c_1, \ldots, c_n; w \in c_i$
    - min, max, or mean pooling
    - Sample contexts from bot-filtered English Wikipedia

PRETRAINED REPRESENTATIONS

- GPT-2, BERT, XLNet, RoBERTa
  - Small (12 layer); large (24 layer)
- DistilBERT
  - 6 layer
- Word2Vec, GloVe

EXPERIMENT 1: SETUP

- Task: Word Similarity / Relatedness
  - Method: Cosine similarity between representations
  - Metric: Spearman correlation coefficient
- Datasets:
  - RG65, WS353, SimLex999, SimVerb3500



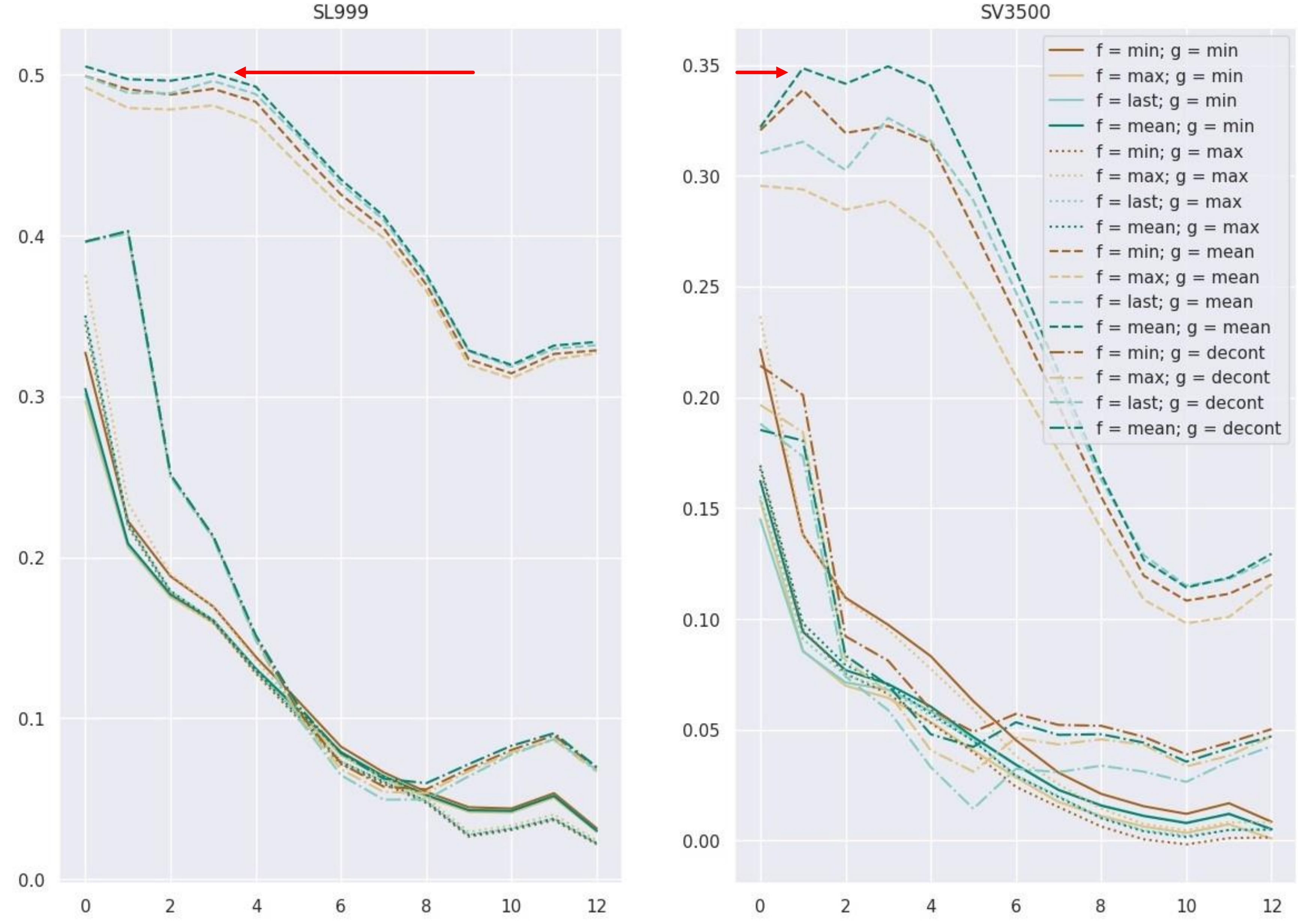
Finding:

Mean-pool across contexts

Decontextualized performs poorly

Visual Cue:

Red braces in figures



Finding:

Mean-pool across subwords

Visual Cue:

Red arrows in figures

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)

Finding:
More contexts is better

Visual Cue:
Red braces in figure

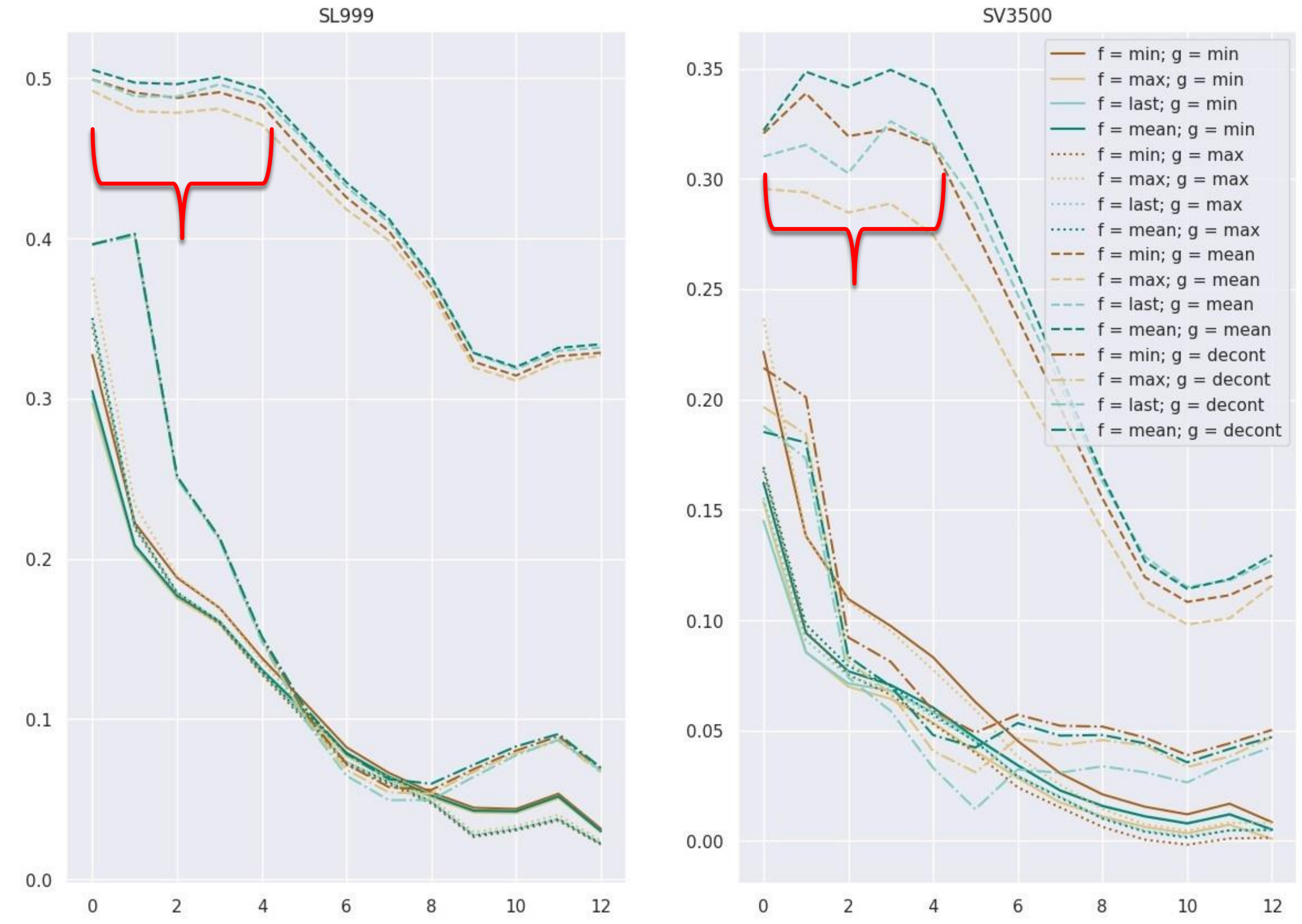
Model	N	RG65	WS353	SimLex999	SimVerb3500
Word2Vec	-	0.6787	0.6838	0.4420	0.3636
GloVe	-	0.6873	0.6073	0.3705	0.2271
BERT-12 (1)	500K	0.7206	0.7038	0.5019	0.3550
BERT-24 (1)	500K	0.7367	0.7074	0.5114	0.3687
BERT-24 (6)	500K	0.7494	0.7282	0.5116	0.4062
BERT-12	10K	0.5167 (1)	0.6833 (1)	0.4573 (1)	0.3043 (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)

Finding:

Significantly outperform Word2Vec, GloVe

Visual Cue:

Red, blue arrows in figure



Finding:

Early layers perform best*

Visual Cue:

Red braces in figures

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)

Finding:

Best layer is higher
as # contexts increases

Visual Cue:

Blue, red arrows in figure

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)

Finding:

Absolute performance is quite different across models

EXPERIMENT 1: INTERPRETABILITY / UNDERSTANDING

- ☑Many patterns generalize across all 9 weights, all 4 datasets
  - ☑ Absolute performance differences are not explained
- ☑Clarification on where lexical semantics is best encoded
  - ☑ Dependence on number of contexts
- ☑Evidence that representations are *over-contextualized*
  - ☑ Potentially related to anisotropy (Ethayarajh, 2019)
- ☑High quality word embedding can be easily extracted
  - ☑ Evaluation is restricted to intrinsic word similarity / relatedness tasks

EXPERIMENT 1: ENGINEERING / MODELLING

- ☑Mean-pooling across subwords, contexts
- ☑Variance reduction across contexts may be desirable
  - ☑Especially since later model layers are usually used
- ☑High quality word embeddings
  - ☑Easier to use
  - ☑On-device computation / resource-constrained settings
  - ☑Faster / environmental concerns

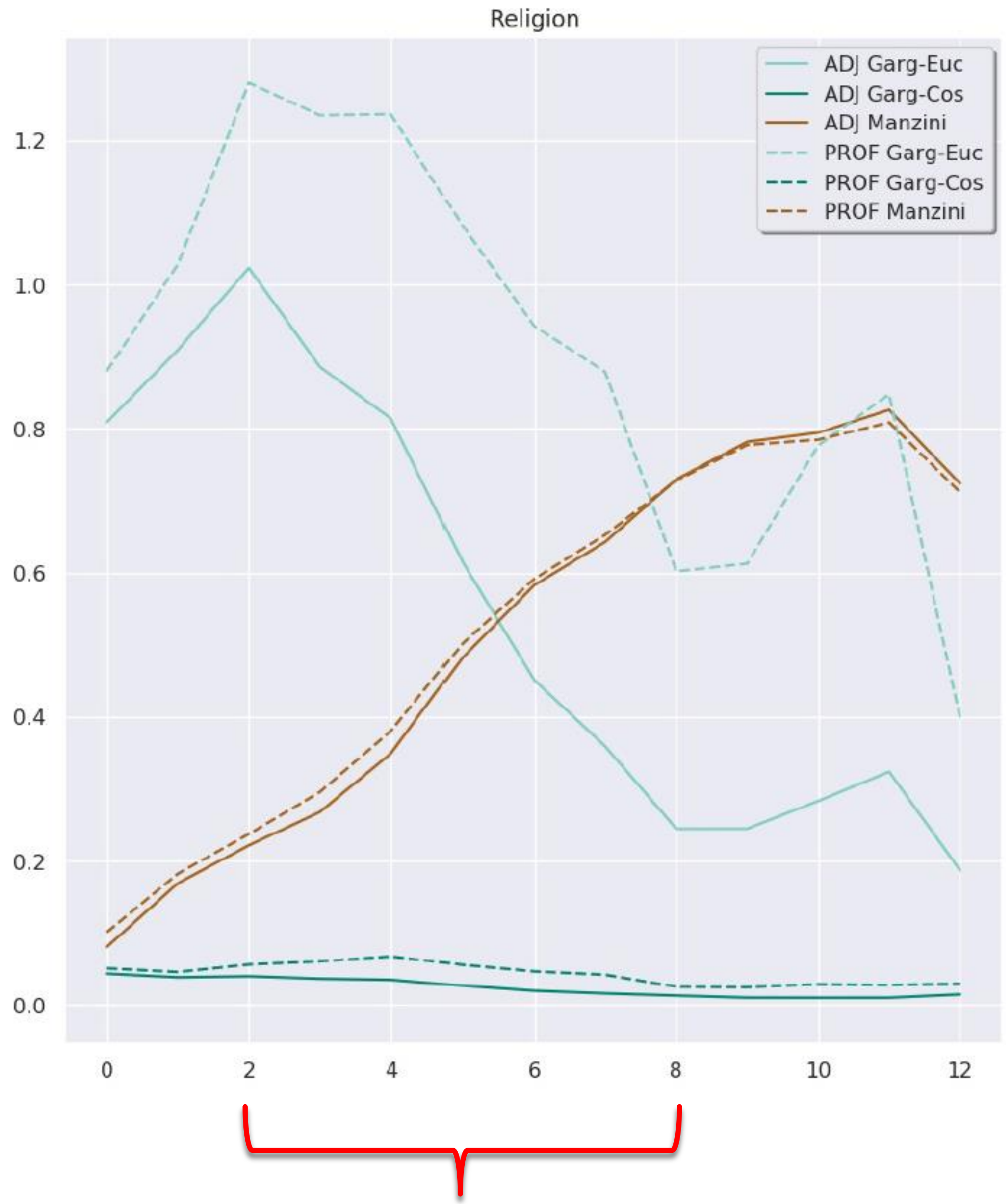
EXPERIMENT 2: SOCIAL BIAS

EXPERIMENT 2: SETUP

- Groups
  - Gender (male, female)
  - Race (white, Hispanic, Asian)
  - Religion (Christianity, Islam)
- Normative Considerations
  - Representational Harms
    - Measuring bias with respect to stereotypes within pretrained representations
  - Stereotypes pertaining to adjectives, professions may precipitate allocative harms

EXPERIMENT 2: METHODS

- Bolukbasi – Binary, PCA, aligned
- Garg-Cos – Binary, average, unaligned
- Garg-Cos – Binary, average, unaligned
- Manzini – Multiclass, average, unaligned

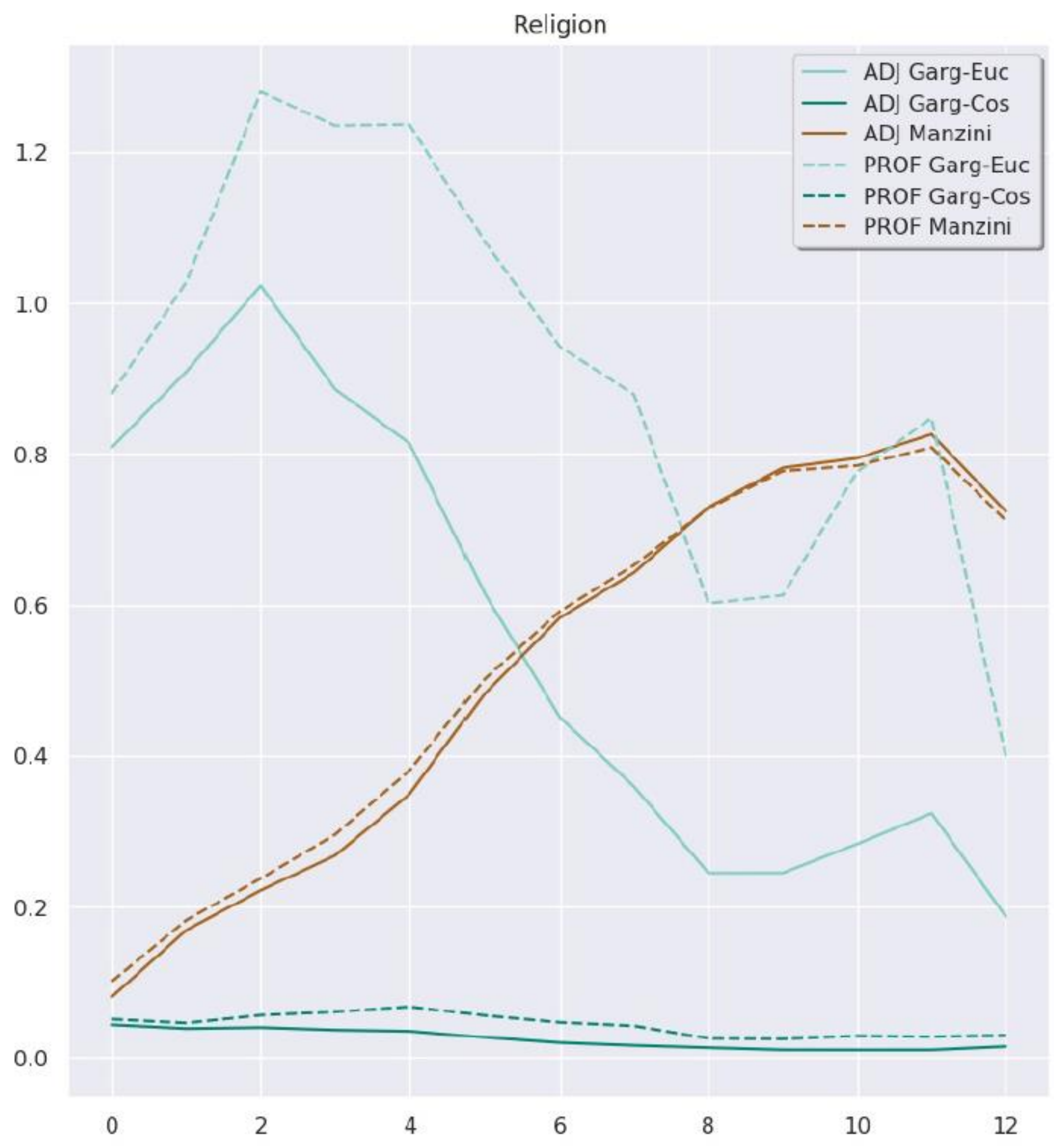


Finding:

Stark inconsistency in bias estimators

Visual Cue:

Red brace in figure

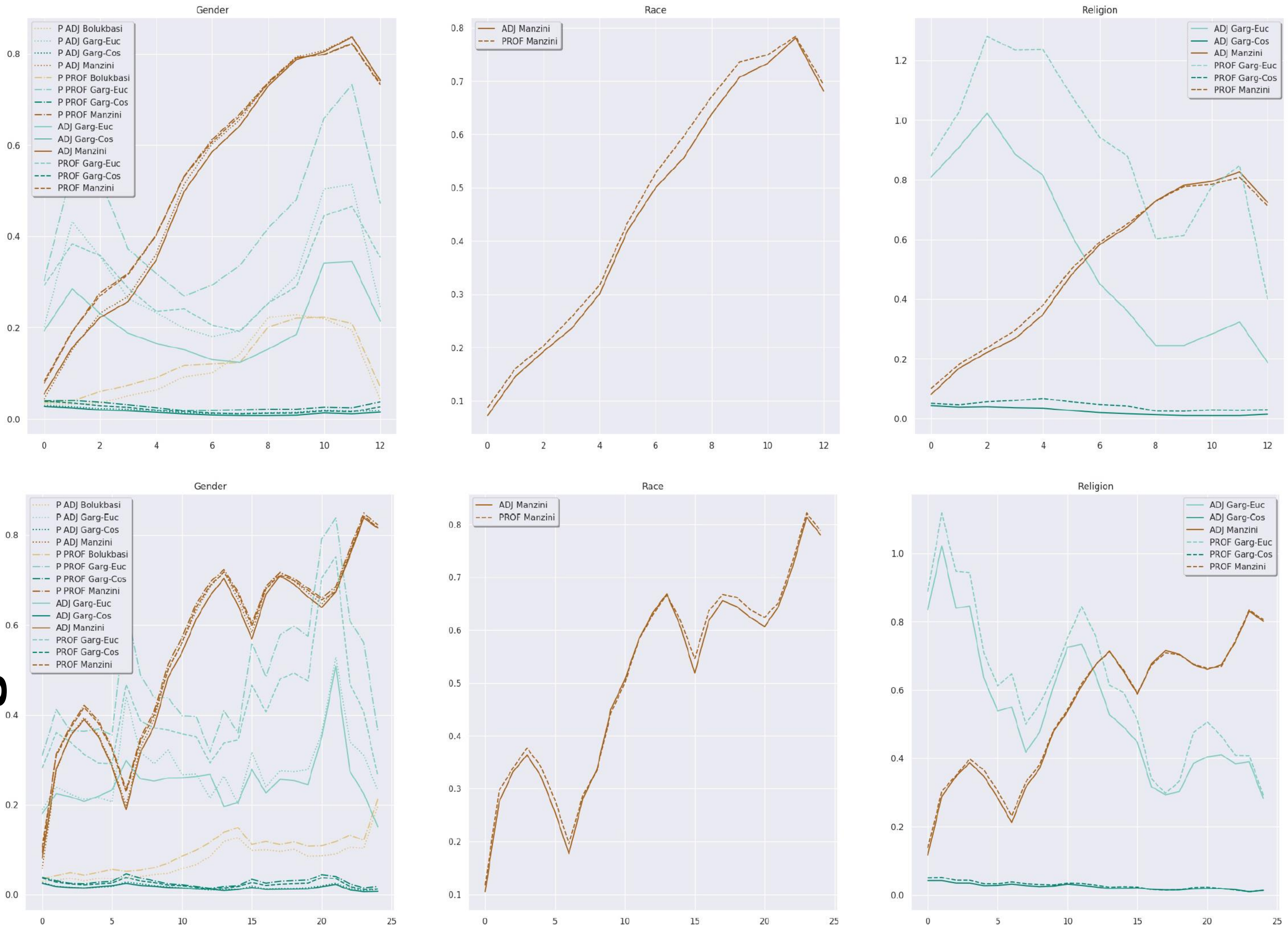


Finding:

Relative trends stable to target word list

Base

Large



Finding:

Data does *not* fully determine bias

CONCLUSIONS

WHAT TO REMEMBER ABOUT THIS WORK: 3 CONCRETE TAKEAWAYS + 2 LESSONS

- Reductions from Contextualized to Static
  - Interpretability, Bias, Dimensionality Reduction, Debiasing
- High quality word embeddings
- Social bias estimators are wildly inconsistent

- Breadth (many models, many layers, many estimators, many word lists, many social biases)
- Viewing aspects of social bias research as special cases of interpretability research

ACKNOWLEDGEMENTS

- Ge Gao, Marty van Schijndel, Forrest Davis
- Mozilla DeepSpeech
- Cornell NLP