Unsupervised Detection of Contextualized Embedding Bias with Application to Ideology

Valentin Hofmann[*‡], Janet Pierrehumbert[†*], Hinrich Schütze[‡]

[*]Faculty of Linguistics, University of Oxford
[†]Department of Engineering Science, University of Oxford
[‡]Center for Information and Language Processing, LMU Munich

valentin.hofmann@ling-phil.ox.ac.uk

ICML 2022

Ideological Bias

- Framing: proponents of different ideologies highlight different aspects of the same issue during political discussion
- Example: left-wing/right-wing ideologies tend to frame immigrants as victims/criminals (Mendelsohn et al., 2021)
- Contextualized embeddings: covariation in the region occupied by the embeddings of a word and the ideology of a text
  - Left-wing text: immigrants $\leftrightarrow$ victims
  - Right-wing text: immigrants $\leftrightarrow$ criminals

Ideological Subspace

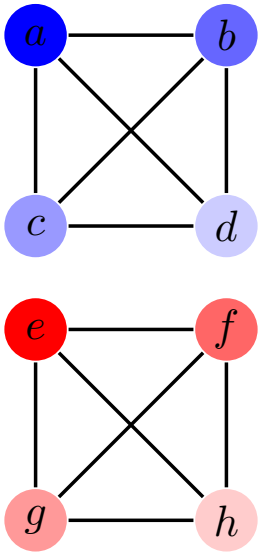
- $\mathcal{X} \subset \mathbb{R}^d$: contextualized embedding space
- $\mathcal{X}_* \subset \mathbb{R}^{d_*}$ ($d_* \ll d$): subspace of $\mathcal{X}$ that contains all and only information relevant to framing and ideological bias
- $\mathcal{X}_*^{\perp} \subset \mathbb{R}^{d-d_*}$: orthogonal complement of $\mathcal{X}_*$ that contains information irrelevant to framing and ideological bias
- Prior work: find $\mathcal{X}_*$ with supervision (e.g., Webson et al., 2020)
- This study: determine $\mathcal{X}_*$ in an unsupervised way

Social Networks and Ideology

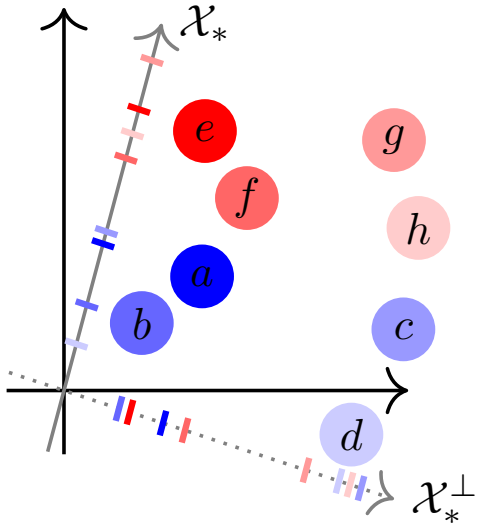
- Social networks reflect ideology: nodes close to each other tend to be ideologically similar (Adamic & Glance, 2005)
- The Reddit Politosphere (Hofmann et al., 2022)
  - Comments from political subreddits between 2013 and 2019
  - Year-wise social networks with political subreddits as nodes and edges based on user overlap
  - Year-wise sets of 1,000 political concepts, i.e., unigrams and bigrams (e.g., *immigrants*, *tax reform*)
- Key idea: leverage ideological information latently encoded by structure of social networks to determine $\mathcal{X}_*$

Toy Example

Social network $\mathcal{G}$



Embedding space $\mathcal{X}$

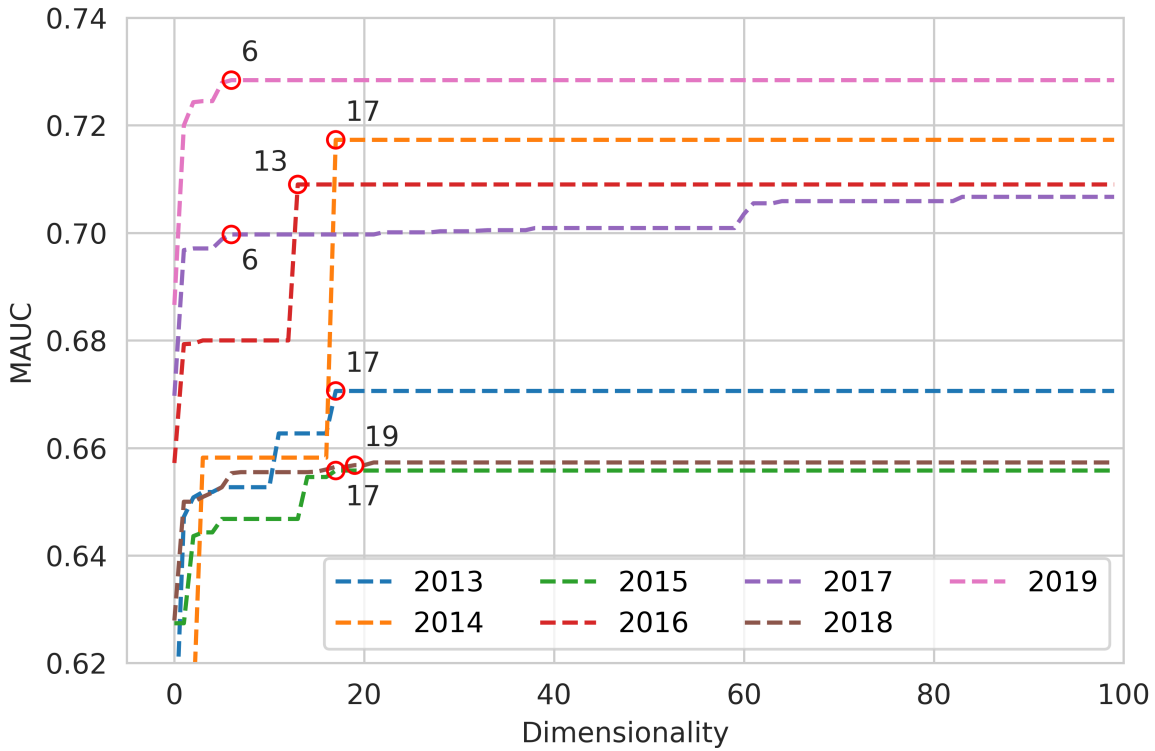


Model

- General setup: perform link prediction while rotating and shrinking the embedding space to find $\mathcal{X}_*$
- Input to model: $\mathbf{x}_i^{(c)} \in \mathcal{X} \subset \mathbb{R}^d$ (average contextualized BERT embedding of concept c for subreddit i)
- Rotation: $\tilde{\mathbf{x}}_i^{(c)} = \mathbf{x}_i^{(c)}\mathbf{R}$ with orthogonality penalty on $\mathbf{R}$
- $\tilde{\mathbf{x}}_i^{(c)}$ fed into graph auto-encoder (Kipf & Welling, 2016)
- Shrinkage: structured sparsity (mixed ℓ_1/ℓ_2 regularization) on first graph auto-encoder weight matrix
- Training regime in superepochs and epochs

Intrinsic Evaluation

- Performance (MAUC) measured for different values of d_*
- $\mathcal{X}$ can be shrunk substantially without loss in performance



Semantic Probing of $\mathcal{X}_*$

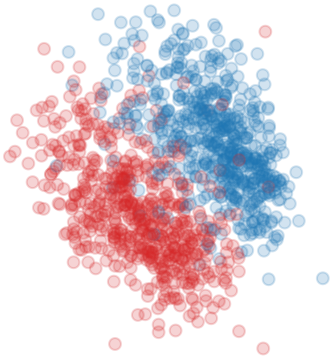
- Project AntSyn (Nguyen et al., 2016) antonym pairs a into $\mathcal{X}_*$ and compute importance scores s_a
- Top antonym pairs have abstract evaluative meanings

Year	Max s_a	Min s_a
2013	*executive/legislative* *immoral/moral* *general/particular* *autocratic/democratic*	*aware/unaware* *adjacent/separate* *happy/unhappy* *cold/warm*
2016	*useful/useless* *ill/well* *expensive/inexpensive* *common/uncommon*	*north/south* *following/leading* *minus/plus* *dark/light*
2019	*autocratic/democratic* *national/transnational* *biased/impartial* *qualified/unqualified*	*likely/unlikely* *cold/warm* *different/similar* *former/latter*

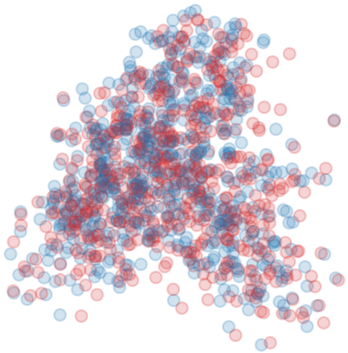
Indexical Probing of $\mathcal{X}_*$

- Train classifiers to predict ideological orientation of subreddits (left-wing vs. right-wing) based on concept embeddings
- $\mathcal{X}_*$ contains ideological information from $\mathcal{X}$ in distilled form

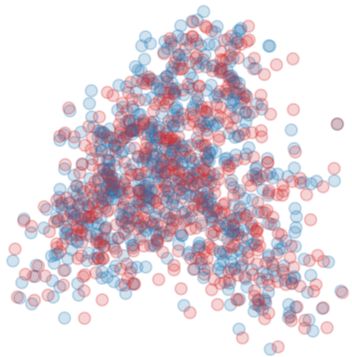
$\mathcal{X}_*$ (17): 94.6%



$\mathcal{X}$ (768): 88.7%



$\mathcal{X}_*^{\perp}$ (751): 70.0%



Take-away Points

- Ideological bias can be detected in an unsupervised way by leveraging information encoded by social networks
- Our method combines graph neural networks with structured sparsity learning and orthogonality regularization
- Ideological subspace can be probed semantically and indexically

Thank you!