
Unsupervised Detection of Contextualized Embedding Bias with Application to Ideology

Valentin Hofmann^{1,2} Janet B. Pierrehumbert^{3,1} Hinrich Schütze²

Abstract

We propose a fully unsupervised method to detect bias in contextualized embeddings. The method leverages the assortative information latently encoded by social networks and combines orthogonality regularization, structured sparsity learning, and graph neural networks to find the embedding subspace capturing this information. As a concrete example, we focus on the phenomenon of ideological bias: we introduce the concept of an ideological subspace, show how it can be found by applying our method to online discussion forums, and present techniques to probe it. Our experiments suggest that the ideological subspace encodes abstract evaluative semantics and reflects changes in the political left-right spectrum during the presidency of Donald Trump.

1. Introduction

What kinds of biases are implicitly encoded by word embeddings? This question has attracted considerable attention recently, with a focus on gender (Bolukbasi et al., 2016; Caliskan et al., 2017; Zhao et al., 2019) and race (Tan & Celis, 2019; Jiang & Fellbaum, 2020; Guo & Caliskan, 2021). There has also been work on **ideological bias**, i.e., the association of word embeddings with political ideology resulting from framing (Webson et al., 2020; Bianchi et al., 2021; Rozado & al Gharbi, 2021).

Geometrically, bias can be represented as a **linear subspace** in embedding space that captures most of the relevant semantic information (Vargas & Cotterell, 2020). Prior studies have typically taken a supervised approach to detect this subspace, drawing upon external resources (e.g., word lists of gender-specific job titles). In the case of ideological bias,

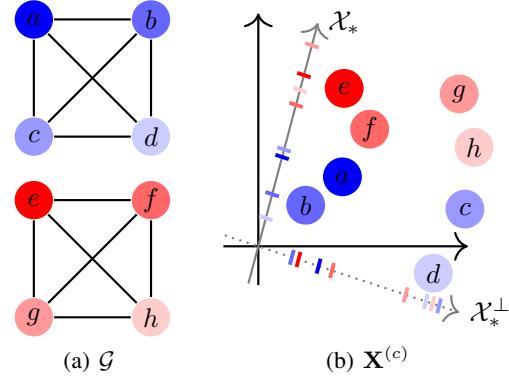


Figure 1: Framework. The example graph \mathcal{G} consists of two components reflecting different ideologies as indicated by node color. By projecting the corresponding embeddings $\mathbf{X}^{(c)}$ into the subspace capturing \mathcal{G} 's polarization (\mathcal{X}_*), we obtain representations of lower dimensionality that still allow for perfect predictions of \mathcal{G} 's edges.

additional supervision is required to distinguish texts stemming from different ideologies (e.g., manual labels). The heavy need for supervision limits the scalability and wider applicability of research on embedding bias.

In this paper, we propose a fully unsupervised method to detect bias in contextualized word embeddings. We draw upon the fact that **social networks** (a ubiquitous data structure) tend to be homophilous, i.e., neighboring nodes often have similar characteristics (McPherson et al., 2001). As a result, the structure of social networks latently encodes assortative information about variables relevant to bias, including gender (Psylia et al., 2017), race (DiPrete et al., 2011), and ideology (Conover et al., 2011). Our method exploits this to detect bias by rotating and shrinking the embedding space such that the resulting subspace is maximally informative about the social network topology (Figure 1). Algorithmically, we combine graph neural networks with orthogonality regularization and structured sparsity learning.

As a concrete application, we focus on online discussion forums, specifically Reddit, which can be modeled as networks with subforums as nodes and edges based on user overlap (Olson & Neal, 2015; Kumar et al., 2018). We lever-

¹Faculty of Linguistics, University of Oxford ²Center for Information and Language Processing, LMU Munich ³Department of Engineering Science, University of Oxford. Correspondence to: Valentin Hofmann <valentin.hofmann@ling-phil.ox.ac.uk>.

age the ideological information encoded by such networks to identify the ideological bias subspace. The unsupervised nature of our method makes it challenging to interpret the found subspace. We present two methods for remedy: **semantic probing**, which analyzes lexical-semantic regularities, and **indexical probing**, which aims to uncover the hidden ideological topology of the subspace.

Our **contributions** are as follows. We propose a fully unsupervised method that exploits the structure of social networks to detect bias in contextualized embeddings, focusing on the use case of ideological bias in online discussion forums. Our method combines orthogonality regularization, structured sparsity learning, and graph neural networks. We also present techniques to probe the found subspace. Our experiments show that the ideological subspace encodes abstract evaluative semantics and reflects changes in the left-right spectrum during the presidency of Donald Trump.¹

2. Related Work

A lot of research on **bias** in NLP (see Blodgett et al. (2020) and Cao et al. (2022) for reviews) has focused on linear subspaces in word embedding space that contain information about categories such as gender (Bolukbasi et al., 2016; Caliskan et al., 2017; Basta et al., 2019; Gonen & Goldberg, 2019) and race (Tan & Celis, 2019; Jiang & Fellbaum, 2020; Guo & Caliskan, 2021). There are also studies that measure word embedding associations to analyze differences between ideologies (Knoche et al., 2019; Tripodi et al., 2019; Xie et al., 2019; Webson et al., 2020; Bianchi et al., 2021; Rozado & al Gharbi, 2021; Walter et al., 2021). Our work differs from these studies in various ways: (i) it is fully unsupervised, i.e., it does not need a key word list to locate the subspace, nor does it need information about the ideological orientation of texts; (ii) it does not make assumptions about the number of ideologies and can handle multidimensional ideological spaces (e.g., different ideologies on all nodes), which is theoretically more sound (Heckman & Snyder, 1997); (iii) it uses contextualized embeddings, thus obviating the need to fit separate embeddings for each ideology (which is computationally infeasible in our setup).

Research on **ideological polarization** in the computational social sciences (Adamic & Glance, 2005; Yardi & Boyd, 2010; Conover et al., 2011; Guerra et al., 2013; Himelboim et al., 2013; Weber et al., 2013; Mejova et al., 2014; Bakshy et al., 2015; Garcia et al., 2015; Sylvester & Purver, 2015; Garimella et al., 2018; Green et al., 2020; Cann et al., 2021; Waller & Anderson, 2021) and NLP (Sagi et al., 2013; Iyyer et al., 2014; Preotiuc-Pietro et al., 2017; An et al., 2018; Kulkarni et al., 2018; An et al., 2019; Demszky et al., 2019;

¹We make our code available at https://github.com/valentinhofmann/unsupervised_bias.

Shen & Rosé, 2019; Davoodi et al., 2020; Mokhberian et al., 2020; Roy & Goldwasser, 2020; Tyagi et al., 2020; Vorakitsphan et al., 2020; He et al., 2021; Mendelsohn et al., 2021) has shown that polarization can manifest itself on the level of social networks by a polarized network structure and on the level of political discourse by a range of linguistic phenomena including framing. Most closely related, Hofmann et al. (2022a) leverage the structure of social networks to detect polarized issues. Our work differs in both its topic and its methodology: (i) it focuses on ideological bias and the embedding subspace containing it; (ii) it is fully unsupervised, which increases its applicability.

3. Framing and Ideological Bias

Framing describes the mechanism by which proponents of different ideologies highlight different aspects of the same issue during political communication, thereby lending greater perceived importance to them (Entman, 1993; Nelson et al., 1997; Druckman, 2001; Chong & Druckman, 2007). In the US, e.g., liberals tend to frame immigrants as victims, underscoring their vulnerability, whereas conservatives often frame them as criminals, portraying them as threats to the public (Benson, 2013; Mendelsohn et al., 2021).

What is the relationship between framing and ideological bias in contextualized word embeddings? Framing results in language-internal and language-external associations that interact in creating ideological bias. Linguistically, framing is realized by bringing certain words into syntactic contiguity with each other, impacting cooccurrence statistics and leading to (**first-order**) **semantic associations** (e.g., between *immigrants* and *criminals*). Contextualized word embeddings encode such semantic associations by mapping words to vectors that vary with the context (Coenen et al., 2019; Field & Tsvetkov, 2019; Wiedemann et al., 2019), placing, e.g., the embedding of *immigrants* close to the embedding of *criminals*. Extralinguistically, certain frames are preferentially employed by proponents of certain ideologies, creating (**second-order**) **indexical associations** (Silverstein, 2003; Nguyen et al., 2021) between the linguistic manifestations of framing and ideologies (e.g., between the semantic association of *immigrants* with *criminals* on the one hand and conservatives on the other).² Such indexical associations are reflected by systematic covariation in the region occupied by the embeddings of a word and the ideological orientation of the text on which the embeddings are computed, making it possible (in the extreme case) to predict the ideology from the word embedding (e.g., predict that a text is conservative based on the fact that the embeddings of *immigrants* are close to the embeddings of *criminals*).

²Notice that while political ideology is not typically viewed as a sociolinguistic variable (Eckert, 2012; 2019), it impacts social identity construction in a similarly crucial way.

Our conceptualization of ideological bias is inherently neutral: rather than examining its potentially harmful effects (a topic in its own right), we aim to present ideological bias as a little-investigated property of contextualized word embeddings that can be used as an analytical lens to draw inferences about political reasoning. This focus sets our work apart from many other studies on bias in NLP (Blodgett et al., 2020; Cao et al., 2022) and puts it more in line with research on the linguistic manifestations of political slant (Gentzkow & Shapiro, 2010; Fulgoni et al., 2016; Fan et al., 2019; Baly et al., 2020), which so far has not touched on the topology of contextualized word embeddings.

It is important to notice that the connection between social groups, the typical word cooccurrence patterns they employ, and the resulting associations in contextualized embedding space apply to many other types of bias as well (e.g., bias in the way words are used by people of different gender or bias in the way concepts are used by different scientific fields). In all these cases, bias is only one factor besides many others such as syntax (Goldberg, 2019; Hewitt & Manning, 2019) impacting variation in the contextualized embeddings of a word. One of the key goals of this paper is to devise a method that overcomes this challenge and separates, for a set of words, the variation in embedding space caused by bias from the variation caused by other factors.

4. Ideological Subspace

Let $\mathcal{X} \subset \mathbb{R}^d$ be a d -dimensional embedding space. We want to find the d_* -dimensional orthogonal subspace $\mathcal{X}_* \subset \mathbb{R}^{d_*}$, with $d_* \ll d$, that contains all and only information relevant to ideological framing, and whose orthogonal complement \mathcal{X}_*^\perp contains information irrelevant to ideological framing. We call \mathcal{X}_* the **ideological subspace** of \mathcal{X} .

In this paper, we show how \mathcal{X}_* can be found for discussion forums that are divided into smaller subforums. While there is typically no explicit information about the subforum ideologies that would allow us to perform supervised learning, we argue that the network structure of the subforums is sufficient to determine \mathcal{X}_* . More formally, let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph consisting of a set of subforums \mathcal{V} and a set of edges between the subforums \mathcal{E} representing homophilous (McPherson et al., 2001) relations such as user overlap. Let C be a set of political concepts to be analyzed and $X = \{\mathbf{X}^{(c)}\}_{c \in C}$ a set of matrices with

$$\mathbf{X}^{(c)} = [\mathbf{x}_1^{(c)}, \dots, \mathbf{x}_{|\mathcal{V}|}^{(c)}]^\top, \quad (1)$$

i.e., each row in $\mathbf{X}^{(c)} \in \mathbb{R}^{|\mathcal{V}| \times d}$ contains the embedding $\mathbf{x}_i^{(c)} \in \mathcal{X}$ of concept c for subforum i . The embeddings capture the ideological framing of concept c in subforum i (which we want to be in \mathcal{X}_*) as well as other information irrelevant to ideology (which we want to be in \mathcal{X}_*^\perp).

Our key idea is that due to the homophily of \mathcal{G} , subforums close to (far from) each other in \mathcal{G} are expected to be ideologically similar (dissimilar), which should be reflected by similar (dissimilar) patterns of ideological bias while having little effect on other semantic characteristics. Put differently, for $\mathbf{X}^{(c)}$ representing concept c , its projection to \mathcal{X}_* should be informative about the proximity of two subforums in \mathcal{G} , but its projection to \mathcal{X}_*^\perp should not. We formalize this idea as the task of predicting links in \mathcal{G} using the embedding matrices in X as features while at the same time shrinking \mathcal{X} to \mathcal{X}_* , i.e., we leverage the training signal from link prediction to remove the task-irrelevant information in \mathcal{X}_*^\perp .

To make this more concrete, Figure 1 shows an example graph of eight nodes that fall into two components reflecting distinct ideologies as well as the corresponding embeddings for one example concept. By projecting the embeddings into the subspace capturing the network polarization (\mathcal{X}_*), we obtain representations that are of lower dimensionality while still allowing for perfect predictions of the links in \mathcal{G} . Projecting into the orthogonal complement (\mathcal{X}_*^\perp), on the other hand, does not allow for perfect predictions.

As a result of the two-level structure of ideological bias, \mathcal{X}_* encodes both semantic and indexical information. For the running example of immigrants, \mathcal{X}_* might encode the semantic category of agency to capture the different framing as victims or criminals, and it might exhibit regions indexically linked to liberals and conservatives. This makes it possible to analyze \mathcal{X}_* from two complementary perspectives.

5. Model

We use pretrained (base, uncased) BERT (Devlin et al., 2019) to obtain subforum-specific representations for the concepts (d is 768).³ Specifically, for all concepts $c \in C$ and subforums $i \in \mathcal{V}$, we compute average contextualized embeddings $\mathbf{x}_i^{(c)}$ and use them as node features in a graph auto-encoder to perform link prediction on \mathcal{G} (i.e., we predict edges between subforums). Based on the assumption that the ideological subspace contains the information needed for this task, we simultaneously rotate and shrink the space to find the ideological subspace \mathcal{X}_* .

The first part of the model rotates $\mathbf{X}^{(c)}$ such that the information relevant to ideological bias corresponds to a small number of dimensions in the rotated space \mathcal{X}_r , i.e.,

$$\mathbf{X}_r^{(c)} = \mathbf{X}^{(c)} \mathbf{R}. \quad (2)$$

Here, $\mathbf{R} \in \mathbb{R}^{d \times d}$ is an orthogonal matrix that transforms \mathcal{X} into \mathcal{X}_r . By choosing \mathbf{R} to be orthogonal, we do not add or remove any information from the original space. \mathbf{R} is optimized as part of the training. To enforce the orthogonality

³We take the mean-pooled embedding if a concept is split into multiple WordPiece tokens.

Table 1: Dataset statistics. $|\mathcal{D}|$: number of comments; $|\mathcal{V}|$: number of nodes (subreddits); $|\mathcal{E}|$: number of edges; μ_d : average node degree; μ_π : average shortest path length; ρ : density; Q : maximum modularity.

Year	$ \mathcal{D} $	$ \mathcal{V} $	$ \mathcal{E} $	μ_d	μ_π	ρ	Q
2013	6,306,458	108	324	6.00	3.08	.056	.560
2014	6,664,567	132	335	5.08	3.86	.039	.663
2015	9,230,022	168	493	5.87	3.87	.035	.672
2016	34,801,075	255	1,318	10.34	3.14	.041	.603
2017	38,278,685	295	1,572	10.66	3.14	.036	.585
2018	40,222,627	316	1,604	10.15	3.17	.032	.584
2019	46,590,000	412	2,536	12.31	3.20	.030	.603

of \mathbf{R} , we use orthogonality regularization (Bousmalis et al., 2016; Brock et al., 2017; Vorontsov et al., 2017), i.e., we compute an orthogonality penalty,

$$\mathcal{L}_o = \|\mathbf{R}\mathbf{R}^\top - \mathbf{I}\|_F^2, \quad (3)$$

where $\mathbf{I} \in \mathbb{R}^{d \times d}$ is the identity matrix, and $\|\cdot\|_F^2$ is the squared Frobenius norm.

To perform link prediction, we use a graph auto-encoder with two convolutional layers (Kipf & Welling, 2016; 2017) that takes as input the rotated embeddings $\mathbf{X}_r^{(c)}$ as well as \mathcal{G} 's adjacency matrix $\mathbf{A} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$ and in each layer updates the embeddings according to

$$\mathbf{H}^{(l+1)} = \sigma \left(\tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}} \mathbf{H}^{(l)} \mathbf{W}^{(l)} \right), \quad (4)$$

where $\mathbf{H}^{(l)}$ is the embedding matrix after layer l , $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ is \mathcal{G} 's adjacency matrix with added self-loops, $\tilde{\mathbf{D}}$ is the degree matrix of $\tilde{\mathbf{A}}$, and $\mathbf{W}^{(l)}$ is the weight matrix of layer l . σ is the activation function, for which we use ReLU after the first and no non-linearity after the second layer. We set $\mathbf{H}^{(0)} = \mathbf{X}_r^{(c)}$. We reconstruct \mathbf{A} by means of a dot-product decoder (Kipf & Welling, 2016) and compute a prediction loss \mathcal{L}_p using binary cross-entropy.

To shrink \mathcal{X}_r , we combine the graph auto-encoder with structured sparsity learning (Yuan & Lin, 2006; Liu et al., 2015; Lebedev & Lempitsky, 2016; Wen et al., 2016; Yoon & Hwang, 2017), which has the effect that entire rows of the weight matrix $\mathbf{W}^{(0)}$ are set to zero during training. Writing $\mathbf{W}^{(0)} = [\mathbf{w}_1^{(0)}, \dots, \mathbf{w}_d^{(0)}]^\top$ as a series of row vectors, we define the sparsity penalty as

$$\mathcal{L}_s = \sum_{j=1}^d \|\mathbf{w}_j^{(0)}\|_2. \quad (5)$$

This is a mixed ℓ_1/ℓ_2 regularization (the ℓ_1 norm of the row ℓ_2 norms) that leads to sparsity on the level of rows. When all entries in a row $\mathbf{w}_j^{(0)}$ are zero, this has the effect of essentially removing dimension j from \mathcal{X}_r . The rows with

non-zero weights after training determine the d_* dimensions of the ideological subspace \mathcal{X}_* .

The final loss is $\mathcal{L} = \mathcal{L}_p + \lambda_o \mathcal{L}_o + \lambda_s \mathcal{L}_s$, where $\lambda_o, \lambda_s > 0$ are hyperparameters. Since \mathcal{L}_s is non-differentiable, we use proximal gradient descent (Bach et al., 2011; Parikh & Boyd, 2013) for optimization. We approximate the weighted proximal operator of the ℓ_1/ℓ_2 norm using the Newton-Raphson algorithm (Deleu & Bengio, 2021).

6. Experiments

6.1. Data

We base our study on the Reddit Politosphere (Hofmann et al., 2022b), a dataset covering the political discourse on the social media platform Reddit from 2008 to 2019. For each year, the Reddit Politosphere contains (i) a graph with political subforums (called *subreddits* in the context of Reddit) as nodes and edges computed by applying statistical backboning to the counts of users shared between subreddits, and (ii) all English comments posted in each of the political subreddits. To make training robust, we confine ourselves to years in which the graph has at least 100 nodes (2013 to 2019). Table 1 gives summary statistics. As indicated by the high modularity values, the graphs are polarized (nodes cluster by ideology). For C , we draw upon year-wise lists of 1,000 English political concepts (unigrams such as *abortion* and bigrams such as *social security*) determined by comparing the vocabulary of the political subreddits with the vocabulary of the default subreddits (i.e., subreddits users used to be subscribed to automatically) using mutual information (Hofmann et al., 2022a).

6.2. Experimental Setup

To estimate how much ideological information we lose by shrinking the space, we compare against a model that directly uses the concept embeddings as input to the graph auto-encoder, i.e., it neither rotates nor shrinks the space. Of course, the \mathcal{X}_* embeddings will not improve over the \mathcal{X} embeddings, but their performance should not be substantially worse, i.e., the comparison tells us how much task-relevant information we lose by removing \mathcal{X}_*^\perp .

We split concepts and edges for each year into train (60%), dev (20%), and test (20%). Models are trained separately for the years. Our training regime consists of *superepochs* in which we loop over all train concepts, and *epochs* in which we predict the train edges between the subreddits using the embeddings of a certain concept as node features. Put differently, on each epoch we perform one pass through the model as described in Section 5, with the node features changing on every epoch as we loop over concepts. This is repeated for a chosen number of superepochs (i.e., if we train for 100 superepochs, the embeddings of each train

Table 2: Performance on link prediction (MAUC). The embeddings in \mathcal{X}_* perform similarly to the embeddings in \mathcal{X} while being of much lower dimensionality. d_* for \mathcal{X}_* is 17 (2013, 2014, 2015), 13 (2016), 6 (2017), 19 (2018), and 6 (2019) versus 768 for \mathcal{X} . Differences between better (gray) and lower performance are significant (underlined) for only six columns as shown by two-tailed t -tests ($p < .01$), underscoring the similar performance of \mathcal{X} and \mathcal{X}_* .

Space	Dev								Test							
	2013	2014	2015	2016	2017	2018	2019	$\mu \pm \sigma$	2013	2014	2015	2016	2017	2018	2019	$\mu \pm \sigma$
\mathcal{X}_*	.671	.717	.656	.709	.700	.657	.728	.691±.028	.699	.735	.660	.669	.687	.695	.696	.692±.022
\mathcal{X}	.649	.699	.681	.705	.695	.673	.741	.692±.027	.683	.725	.699	.683	.687	.700	.707	.698±.014

concept are used 100 times as node features). For evaluation, we loop over the dev (or test) concepts and compute the area under the curve (AUC) that results from predicting the dev (or test) edges between the subreddits using the embeddings of a certain concept as node features. Finally, we compute the mean AUC (MAUC) for all dev (or test) concepts.

We use gradient accumulation to make training robust, i.e., weights are only updated after 10 concepts (epochs in our training regime). We perform grid search for the learning rate $r \in \{1 \times 10^{-4}, 3 \times 10^{-4}, 1 \times 10^{-3}\}$. For the model used to find \mathcal{X}_* , we further perform grid search for the orthogonality constant $\lambda_o \in \{1 \times 10^{-3}, 3 \times 10^{-3}, 1 \times 10^{-2}\}$ as well as the sparsity constant $\lambda_s \in \{1 \times 10^{-2}, 3 \times 10^{-2}, 1 \times 10^{-1}\}$. In total, there are 3 hyperparameter search trials for \mathcal{X} and 27 for \mathcal{X}_* per year. We use Adam (Kingma & Ba, 2015) as the optimizer. See Appendix A.1 for further details about hyperparameter tuning and runtime.

6.3. Performance

The performance on link prediction as a function of d_* exhibits a pronounced knee shape (Figure 2), i.e., we can shrink \mathcal{X} substantially without losing performance. To detect the knee, we use the algorithm proposed by Satopää et al. (2011). Table 2 shows that when comparing to the baseline without sparsity constraint, the embeddings in \mathcal{X}_* (with tuned d_*) and \mathcal{X} perform very similarly, i.e., the embeddings from the subspace allow to reconstruct the edges between the subreddits with nearly identical performance as the embeddings from the full space. The difference in performance is statistically insignificant for the majority of cases. The fact that between 97.5% (2018) and 99.2% (2017 and 2019) of \mathcal{X} can be discarded without major detrimental effects on performance suggests that most of the information encoded by \mathcal{X} does not bear relevance for framing. This is in line with prior work showing that contextualized embeddings represent various kinds of information such as syntax (Goldberg, 2019; Hewitt & Manning, 2019) that do not play a role in framing. Interestingly, even though training is performed separately for the years, it converges to subspaces of similar sizes ($5 < d_* < 20$), suggesting that the type of information encoded by \mathcal{X}_* might also be similar.

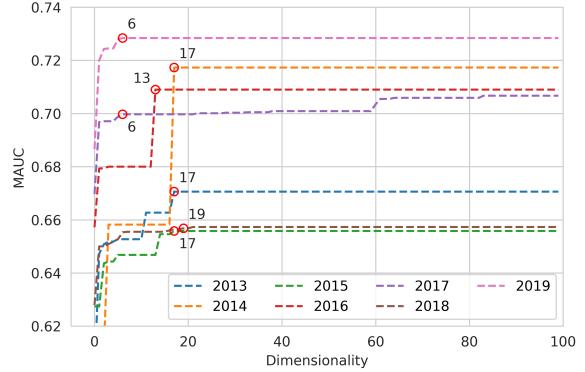


Figure 2: Performance on link prediction (MAUC). The figure shows how the performance varies as a function of d_* , the dimensionality of \mathcal{X}_* . We truncate at $d_* = 100$ since there is no further change for larger values. We highlight the found knees and corresponding values of d_* .

What information is encoded by \mathcal{X}_* ? Given the two-level structure of ideological bias, we examine this question with a view on \mathcal{X}_* 's semantic and indexical associations.

6.4. Semantic Probing of \mathcal{X}_*

To probe the semantic information encoded by \mathcal{X}_* , we draw upon AntSyn (Nguyen et al., 2016; 2017), a dataset of antonym (and synonym) pairs containing POS information, which we use to create semantic axes.⁴ More specifically, for each pair of antonyms $a = (p, q)$ (e.g., *small/big*), we compute year-wise average contextualized embeddings $\mathbf{x}^{(p)}$ and $\mathbf{x}^{(q)}$, using pretrained (base, uncased) BERT (Devlin et al., 2019) and pooling across subreddits. We discard antonym pairs unless both p and q occur at least 100 times in each year. AntSyn often contains several competing antonym pairs, some of which can be highly context-specific (e.g., *conventional/unconventional* and *conventional/nuclear*). As a simple method to determine the most general antonym pair in such cases (e.g., *conventional/unconventional*), we only keep an antonym pair a if p and q are nearest neighbors of each other, resulting in a final set of 972 antonym pairs. We

⁴We tried other datasets (Shwartz et al., 2017; An et al., 2018) but found AntSyn to work best for our use case.

Table 3: Concreteness and morality ratings for the 100 semantic axes with maximum and minimum s_a scores. Semantic axes that are strongly encoded by \mathcal{X}_* have consistently lower concreteness and higher morality scores than semantic axes that are weakly encoded by \mathcal{X}_* . The higher value per column (gray) is underlined if it is significantly ($p < .01$) higher than the lower value as shown by a two-tailed t -test ($p < .01$).

s_a	Concreteness							Morality								
	2013	2014	2015	2016	2017	2018	2019	$\mu \pm \sigma$	2013	2014	2015	2016	2017	2018	2019	$\mu \pm \sigma$
Max	.262	.307	.298	.301	.292	.283	.286	.290±.014	.116	.126	.128	.124	.118	.123	.122	.122±.004
Min	.423	.389	.443	.487	.383	.370	.405	.414±.037	.113	.111	.096	.100	.106	.100	.099	.104±.006

then map $\mathbf{x}^{(p)}$ (and analogously $\mathbf{x}^{(q)}$) into the ideological subspace \mathcal{X}_* by computing

$$\mathbf{x}_*^{(p)} = \mathbf{P}^\top \mathbf{R}^\top \mathbf{x}^{(p)}, \quad (6)$$

where \mathbf{R} is the learned (year-specific) rotation matrix, and $\mathbf{P} \in \mathbb{R}^{d \times d_*}$ is the projection matrix resulting from the learned (year-specific) sparsity pattern in $\mathbf{W}^{(0)}$. We can then define as

$$s_a = \|\mathbf{x}_*^{(p)}\|_2 + \|\mathbf{x}_*^{(q)}\|_2 \quad (7)$$

a score measuring the importance of the semantic axis imposed by a for the semantic information captured by the ideological subspace \mathcal{X}_* .⁵ Large (small) values of s_a indicate that \mathcal{X}_* captures much (little) of a 's semantics.

We first examine quantitatively whether there are systematic patterns regarding the lexical semantics of axes that are well captured by \mathcal{X}_* (s_a large) compared to axes where this is not the case (s_a small). Drawing upon crowdsourced datasets of affect (Warriner et al., 2013), concreteness (Brysbaert et al., 2014), and morality (Hopp et al., 2021) ratings for words (which all provide continuous human evaluation scores), we compute values for each a by averaging the ratings of p and q . Comparing the 100 pairs with top s_a (strongly encoded by \mathcal{X}_*) with the 100 pairs with bottom s_a (weakly encoded by \mathcal{X}_*), we find that they have consistently lower concreteness and higher morality values (Table 3), but there is no clear trend for affect (not shown). These results suggest that the ideologically-driven cooccurrence variations (i.e., differences in framing) that cause ideological bias are related to abstract semantics and moral-like reasoning. While the former is a property of political language in general, the latter can be related to recent insights about the central nature of moral judgments for political ideology (Haidt & Joseph, 2004; Haidt & Graham, 2007; Graham et al., 2009; 2013) and framing (Fulgoni et al., 2016; Mokhberian et al., 2020; Mendelsohn et al., 2021).

Furthermore, we qualitatively inspect pairs with highest and lowest values of s_a . Focusing on adjectives (Table 4), we find that many of them express either general or specific-

cally political and moral evaluative semantics such as *useful/useless*, *biased/impartial*, and *immoral/moral*. This is in line with our quantitative analysis, and it is also confirmed by the nominal and verbal axes (Appendix A.2), with pairs such as *ability/inability*, *patriot/traitor*, and *barbarism/culture* or *agree/disagree*, *overpay/underpay*, and *de-humanize/humanize*. We also notice that pairs with the lowest ranks (e.g., *north/south*, *husband/wife*, and *lock/unlock*) tend not to exhibit this evaluative character.

Thus, \mathcal{X}_* encodes abstract evaluative categories, specifically ones that are constitutive of political reasoning and framing. Furthermore, some pairs explicitly refer to opposing political concepts (e.g., *autocratic/democratic*), which begs the question whether \mathcal{X}_* represents ideology indexically.

6.5. Indexical Probing of \mathcal{X}_*

We are interested to see whether the topology of \mathcal{X}_* captures ideology, i.e., we want to find systematic patterns induced by the indexical associations of \mathcal{X}_* . To do so, we examine to what extent the embeddings in \mathcal{X}_* exhibit, for certain ideologies, a cluster structure. We focus on the left-right ideological spectrum due to its central importance for US politics (Heywood, 2017). We draw upon a left-wing (*socialism*) and a right-wing (*conservatives*) subreddit that are both among the largest subreddits.⁶ Using all concepts from train and plotting their embeddings for all years by means of PCA (Figure 3), we observe a strong clustering according to ideological groups in \mathcal{X}_* that becomes less pronounced after 2016. Furthermore, we do not observe a strong clustering for the PCA plots of \mathcal{X} and \mathcal{X}_*^\perp . Notice this ideological split is an *intrinsic property of the embeddings*: our method does not add any information to the embedding space but finds the subspace that already contains the ideological bias.

To test this more quantitatively, we split the concept embeddings for both ideologies into train (60%), dev (20%), and test (20%) and train year-wise logistic regression classifiers to predict the subreddit ideology (left-wing versus right-

⁵We tried other formulations and obtained similar results.

⁶Results are robust with respect to the selection of the subreddits and similar when using other left-wing and right-wing subreddits (e.g., *communism* and *Republican*).

Table 4: Top and bottom adjectival semantic axes. For each year, the table shows the four adjectival semantic axes with highest and lowest s_a scores. While the top axes tend to have abstract evaluative meanings, this is not the case for bottom axes. Corresponding tables for nominal and verbal semantic axes are provided in Appendix A.2.

Year	Max s_a	Min s_a
2013	<i>executive/legislative</i>	<i>aware/unaware</i>
	<i>immoral/moral</i>	<i>adjacent/separate</i>
	<i>general/particular</i>	<i>happy/unhappy</i>
	<i>autocratic/democratic</i>	<i>cold/warm</i>
2014	<i>constitutional/unconstitutional</i>	<i>official/unofficial</i>
	<i>nonpartisan/partisan</i>	<i>less/more</i>
	<i>capable/incapable</i>	<i>first/second</i>
	<i>armed/unarmed</i>	<i>primary/secondary</i>
2015	<i>accurate/inaccurate</i>	<i>first/second</i>
	<i>boring/interesting</i>	<i>single/triple</i>
	<i>difficult/easy</i>	<i>primary/secondary</i>
	<i>biased/impartial</i>	<i>following/leading</i>
2016	<i>useful/useless</i>	<i>north/south</i>
	<i>ill/well</i>	<i>following/leading</i>
	<i>expensive/inexpensive</i>	<i>minus/plus</i>
	<i>common/uncommon</i>	<i>dark/light</i>
2017	<i>critical/uncritical</i>	<i>former/latter</i>
	<i>central/peripheral</i>	<i>happy/unhappy</i>
	<i>autocratic/democratic</i>	<i>likely/unlikely</i>
	<i>scientific/unscientific</i>	<i>aware/unaware</i>
2018	<i>autocratic/democratic</i>	<i>first/second</i>
	<i>critical/uncritical</i>	<i>former/latter</i>
	<i>biased/impartial</i>	<i>early/late</i>
	<i>armed/unarmed</i>	<i>likely/unlikely</i>
2019	<i>autocratic/democratic</i>	<i>likely/unlikely</i>
	<i>national/transnational</i>	<i>cold/warm</i>
	<i>biased/impartial</i>	<i>different/similar</i>
	<i>qualified/unqualified</i>	<i>former/latter</i>

wing) from the concept embeddings. Since we do not need to tune hyperparameters, we use the dev sets for additional testing. We compare the performance of the embeddings in \mathcal{X}_* to the ones in \mathcal{X} and \mathcal{X}_*^\perp .

We find that (i) the performance of the embeddings in \mathcal{X}_* drastically decreases after 2016, and (ii) that the embeddings in \mathcal{X}_* perform substantially better than the ones in \mathcal{X} and \mathcal{X}_*^\perp from 2013 to 2016, but similarly or worse from 2017 to 2019 (Table 5). This confirms the observation (Figure 3) that the left-right spectrum as reflected by \mathcal{X}_* becomes less pronounced after 2016. The embeddings in \mathcal{X}_*^\perp generally perform worse than the ones in \mathcal{X} , indicating that they contain less ideologically relevant information.

Why does the ideological left-right spectrum as reflected by \mathcal{X}_* become less pronounced over time? It is strik-

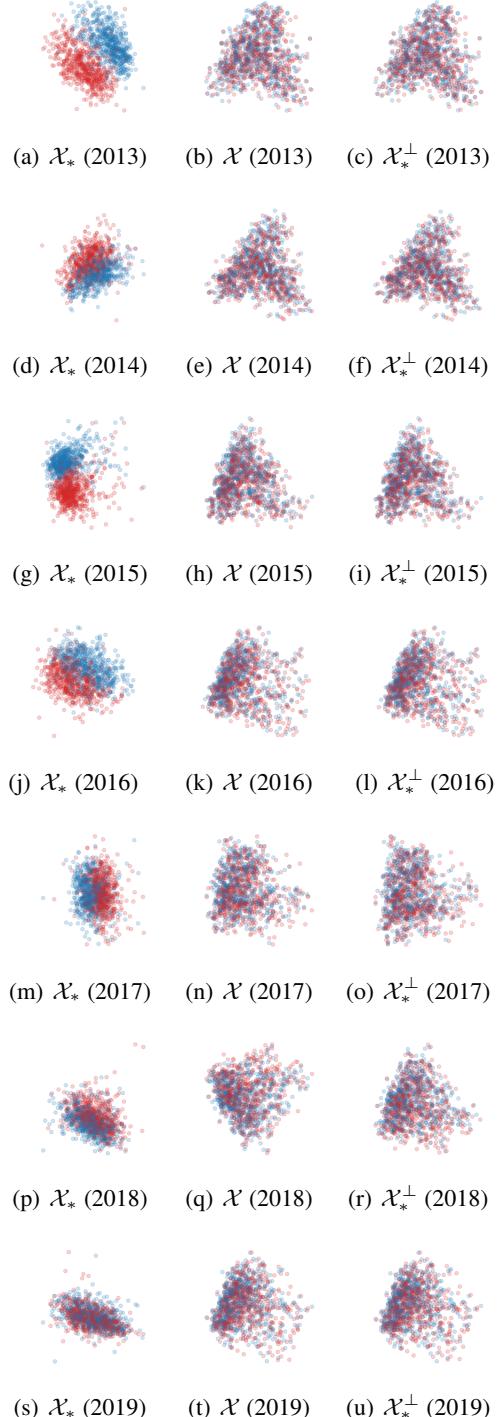


Figure 3: \mathcal{X} , \mathcal{X}_* , and \mathcal{X}_*^\perp . \mathcal{X}_* exhibits a clustering of embeddings into ideologically left (blue) and right (red). This is not the case for \mathcal{X} and \mathcal{X}_*^\perp . The clustering of \mathcal{X}_* into left and right becomes less pronounced after 2016.

ing to observe that the decreasing trend seems to be parallel with the inception of the presidency of Donald Trump. In fact, when we repeat the analysis for rep-

Table 5: Performance on ideology prediction (accuracy). The best performance per column (gray) is underlined if it is significantly ($p < .01$) better than the second-best performance as shown by a McNemar’s test (McNemar, 1947).

Space	Dev							Test							$\mu \pm \sigma$	
	2013	2014	2015	2016	2017	2018	2019	$\mu \pm \sigma$	2013	2014	2015	2016	2017	2018	2019	$\mu \pm \sigma$
\mathcal{X}_*	.938	<u>.954</u>	<u>.942</u>	<u>.879</u>	.754	.646	.504	.802±.161	<u>.946</u>	<u>.954</u>	<u>.925</u>	<u>.904</u>	<u>.783</u>	.625	.621	.823±.137
\mathcal{X}	.892	.887	.858	.767	<u>.804</u>	<u>.796</u>	.787	.827±.047	.887	.842	.850	.792	.779	<u>.817</u>	.838	.829±.034
\mathcal{X}^\perp	.754	.812	.692	.688	.779	.783	<u>.792</u>	.757±.046	.700	.750	.754	.717	.746	.775	<u>.842</u>	.755±.042

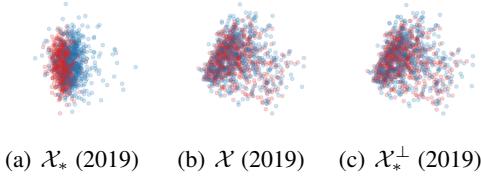


Figure 4: Embedding space topology of \mathcal{X} , \mathcal{X}_* , and \mathcal{X}_*^\perp . The first two principal components of \mathcal{X}_* exhibit a clustering of embeddings into pro-Trump (red) and anti-Trump (blue). This is not the case for \mathcal{X} and \mathcal{X}_*^\perp .

representative pro-Trump (The_Donald) and anti-Trump (AntiTrumpAlliance) subreddits, we see a separation for the years after 2016 (Figure 4).⁷ It is well known that Trump has profoundly impacted the political discourse on Reddit (Massachs et al., 2020). At the same time, Trump is notoriously hard to assign a position on the left-right spectrum (Carmines et al., 2016; Barber & Pope, 2019). Taken together, this suggests that the dominating ideological axis in the Reddit Politosphere after 2016 is not left versus right but rather pro-Trump versus anti-Trump, which is captured by the indexical structure of \mathcal{X}_* . This analysis is also supported by the observation that the performance is only decreasing for \mathcal{X}_* , but not for \mathcal{X} and \mathcal{X}_*^\perp .

The results underscore that \mathcal{X}_* contains ideological information from \mathcal{X} in distilled form and influenced by the dominant axes of antagonism in the data. If these axes change, so does the ideological information captured by \mathcal{X}_* .

6.6. Case Study: Dispersion in \mathcal{X}_*

Given our general framework, we expect concepts with more polarized ideological bias to (i) have embeddings with larger dispersion in \mathcal{X}_* and (ii) show better performance on link prediction.⁸ Measuring dispersion as the average ℓ_2 distance of a concept’s embeddings in \mathcal{X}_* to their centroid, we find a significant positive correlation between dispersion and AUC ($R^2 = .344$, $F(1, 598) = 314$, $p < .001$). Thus, large variance in the ideological subspace reflects the (polarized)

⁷Results are again robust with respect to the selection of the subreddits and similar when using other pro-Trump and anti-Trump subreddits (e.g., trump and Impeach_Trump).

⁸This section uses the 2016 data.

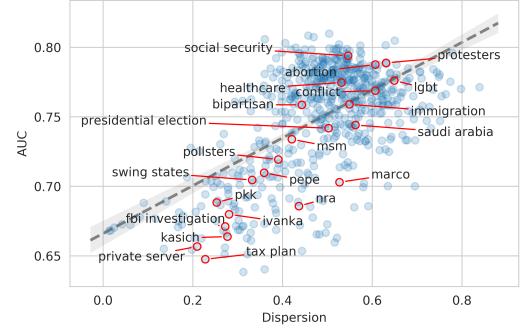
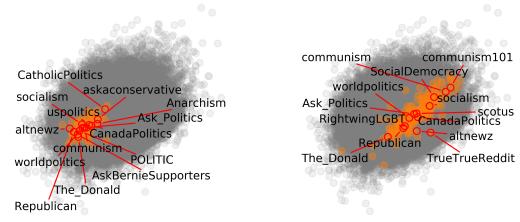


Figure 5: AUC as a function of dispersion. Concepts with higher dispersion in \mathcal{X}_* tend to result in better performance on link prediction. Concepts with large values for dispersion and AUC are most polarized. We provide annotations for selected concepts.



(a) *private server*

(b) *lgbt*

Figure 6: Dispersion of concepts in \mathcal{X}_* . The gray points represent all embeddings in \mathcal{X}_* . While the cluster for *private server* (unpolarized) is clumped, the cluster for *lgbt* (polarized) forms a skewed ellipse. We provide annotations for selected subreddits.

structure of the social network. Many of the concepts with large values for both dispersion and AUC such as *abortion* and *lgbt* are known to be polarized from previous research (Yardi & Boyd, 2010; Mendelsohn et al., 2020), but we also find concepts such as *protesters* and *social security* that have been studied less (Figure 5).

Taking *lgbt* (*private server*) as an example for a polarized (unpolarized) concept, we visualize the resulting clusters in \mathcal{X}_* by means of PCA (Figure 6). Whereas the cluster

for *private server* is clumped, the cluster for *lgbt* forms a skewed ellipse spread across the subspace. We further notice an ideological split for *lgbt*: left-wing and right-wing subreddits occupy opposite ends of the ellipse.

7. Limitations

Instead of direct supervision (e.g., in the form of word lists), our method finds the bias subspace by using the assortative information latently encoded in the structure of social networks. We believe that the ubiquity and variety of social networks online makes our method more scalable and more widely applicable than previous methods. However, there might be use cases for which high-quality external resources are readily available or social networks do not exist, and hence a supervised method might be preferable.

While we only apply our method to data from the Reddit Politosphere, the structure of Reddit as a forum divided into smaller subforums is very common on the web and shared by some of the most intensely researched online platforms (e.g., 4chan). Our method can also be applied to other types of social networks as long as (i) they are homophilous, and (ii) they have text attached to the nodes. For social networks whose nodes correspond to individual users (e.g., Twitter), careful preprocessing might be required to ensure enough data per node (e.g., graph clustering).

The success of our method depends on how accurately variables relevant for bias (in this study ideology) are reflected by the social network, which means that care must be taken during network selection (explicit networks) and construction (implicit networks). For example, user overlap on Reddit can also be due to conflict between subreddits (Datta et al., 2017; Kumar et al., 2018; Datta & Adar, 2019). While we do not find this to affect our results, it might be a limitation if the degree of homophily is too low.

8. Conclusion

We propose a fully unsupervised method that exploits the structure of social networks to detect bias in contextualized embeddings. The method combines orthogonality regularization, structured sparsity learning, and graph neural networks. While we focus on the use case of ideological bias in online discussion forums, our method can be easily applied to other types of bias (e.g., gender bias based on social networks encoding friendship relations or scientific bias based on citation networks). We also present semantic and indexical probing as two complementary techniques to probe the found subspace. Our experiments show that the ideological subspace encodes abstract evaluative semantics and reflects changes in the ideological left-right spectrum during the presidency of Donald Trump.

9. Acknowledgements

This work was funded by the European Research Council (#740516) and the Engineering and Physical Sciences Research Council (EP/T023333/1). The first author was also supported by the German Academic Scholarship Foundation and the Arts and Humanities Research Council. We thank the reviewers for their helpful comments.

References

- Adamic, L. A. and Glance, N. The political blogosphere and the 2004 U.S. Election: Divided they blog. In *International Workshop on Link Discovery (LinkKDD) 3*, 2005.
- An, J., Kwak, H., and Ahn, Y.-Y. SemAxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. In *Annual Meeting of the Association for Computational Linguistics (ACL) 56*, 2018.
- An, J., Kwak, H., Posegga, O., and Jungherr, A. Political discussions in homogeneous and cross-cutting communication spaces. In *International AAAI Conference on Web and Social Media (ICWSM) 13*, 2019.
- Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2011.
- Bakshy, E., Messing, S., and Adamic, L. A. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 384(6239):1130–1132, 2015.
- Baly, R., Da San Martino, G., Glass, J., and Nakov, P. We can detect your bias: Predicting the political ideology of news articles. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2020*, 2020.
- Barber, M. and Pope, J. C. Does party trump ideology? Disentangling party and ideology in America. *American Political Science Review*, 113(1):38–54, 2019.
- Basta, C., Costa-jussà, M. R., and Casas, N. Evaluating the underlying gender bias in contextualized word embeddings. In *Workshop on Gender Bias in Natural Language Processing (GeBNLP) 1*, 2019.
- Benson, R. *Shaping immigration news: A French-American comparison*. Cambridge University Press, Cambridge, UK, 2013.
- Bianchi, F., Marelli, M., Nicoli, P., and Palmonari, M. SWEAT: Scoring polarization of topics across different corpora. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2021*, 2021.

- Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. Language (technology) is power: A critical survey of “bias” in NLP. In *Annual Meeting of the Association for Computational Linguistics (ACL)* 58, 2020.
- Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems (NIPS)* 30, 2016.
- Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., and Erhan, D. Domain separation networks. In *Advances in Neural Information Processing Systems (NIPS)* 30, 2016.
- Brock, A., Lim, T., Ritchie, J. M., and Weston, N. Neural photo editing with introspective adversarial networks. In *International Conference on Learning Representations (ICLR)* 5, 2017.
- Brysbaert, M., Warriner, A. B., and Kuperman, V. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911, 2014.
- Caliskan, A., Bryson, J. J., and Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.
- Cann, T. J., Weaver, I. S., and Williams, H. T. Ideological biases in social sharing of online information about climate change. *PloS ONE*, 16(4):e0250656, 2021.
- Cao, Y. T., Pruksachatkun, Y., Chang, K.-W., Gupta, R., Kumar, V., Dhamala, J., and Galstyan, A. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In *Annual Meeting of the Association for Computational Linguistics (ACL)* 60, 2022.
- Carmines, E. G., Ensley, M. J., and Wagner, M. W. Ideological heterogeneity and the rise of Donald Trump. *The Forum*, 14(4):1631, 2016. ISSN 2194-6183.
- Chong, D. and Druckman, J. N. Framing theory. *Annual Review of Political Science*, 10:103–126, 2007.
- Coenen, A., Reif, E., Yuan, A., Kim, B., Pearce, A., Viégas, F., and Wattenberg, M. Visualizing and measuring the geometry of BERT. In *Advances in Neural Information Processing Systems (NeurIPS)* 33, 2019.
- Conover, M., Ratkiewicz, J., Francisco, M., Goncalves, B., Flammini, A., and Menczer, F. Political polarization on Twitter. In *International AAAI Conference on Web and Social Media (ICWSM)* 5, 2011.
- Datta, S. and Adar, E. Extracting inter-community conflicts in reddit. In *International AAAI Conference on Web and Social Media (ICWSM)* 13, 2019.
- Datta, S., Phelan, C., and Adar, E. Identifying misaligned inter-group links and communities. *Proceedings of the ACM on Human-Computer Interaction*, 1:1–23, 2017.
- Davoodi, M., Waltenburg, E., and Goldwasser, D. Understanding the language of political agreement and disagreement in legislative texts. In *Annual Meeting of the Association for Computational Linguistics (ACL)* 58, 2020.
- Deleu, T. and Bengio, Y. Structured sparsity inducing adaptive optimizers for deep learning. In *arXiv 2102.03869*, 2021.
- Demszky, D., Garg, N., Voigt, R., Zou, J., Gentzkow, M., Shapiro, J., and Jurafsky, D. Analyzing polarization in social media: Method and application to Tweets on 21 mass shootings. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HTL)* 2019, 2019.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HTL)* 2019, 2019.
- DiPrete, T. A., Gelman, A., McCormick, T., Teitler, J., and Zheng, T. Segregation in social networks based on acquaintanceship and trust. *American Journal of Sociology*, 116(4):1234–1283, 2011.
- Druckman, J. N. The implications of framing effects for citizen competence. *Political Behavior*, 23(2):225–256, 2001.
- Eckert, P. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology*, 41(1):87–100, 2012.
- Eckert, P. The limits of meaning: Social indexicality, variation, and the cline of interiority. *Language*, 95(4):751–776, 2019.
- Entman, R. M. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58, 1993.
- Fan, L., White, M., Sharma, E., Su, R., Choubey, P. K., Huang, R., and Wang, L. In plain sight: Media bias through the lens of factual reporting. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)* 2019, 2019.

- Field, A. and Tsvetkov, Y. Entity-centric contextual affective analysis. In *Annual Meeting of the Association for Computational Linguistics (ACL)* 57, 2019.
- Fulgoni, D., Carpenter, J., Ungar, L. H., and Preotiuc-Pietro, D. An empirical exploration of moral foundations theory in partisan news sources. In *International Conference on Language Resources and Evaluation (LREC)* 10, 2016.
- Garcia, D., Abisheva, A., Schweighofer, S., Serdült, U., and Schweitzer, F. Ideological and temporal components of network polarization in online political participatory media. *Policy and Internet*, 7(1):46–79, 2015.
- Garimella, K., Morales, G. D. F., Gionis, A., and Mathioudakis, M. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1):1–27, 2018.
- Gentzkow, M. and Shapiro, J. M. What drives media slant? evidence from U.S. daily newspapers. *Econometrica*, 78(1):35–71, 2010.
- Goldberg, Y. Assessing BERT’s syntactic abilities. In *arXiv 1901.05287*. 2019.
- Gonen, H. and Goldberg, Y. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HTL) 2019*, 2019.
- Graham, J., Haidt, J., and Nosek, B. A. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5):1029–1046, 2009.
- Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., and Ditto, P. H. Moral foundations theory. *Advances in Experimental Social Psychology*, 47:55–130, 2013.
- Green, J., Edgerton, J., Naftel, D., Shoub, K., and Cranmer, S. Elusive consensus: Polarization in elite communication on the COVID-19 pandemic. *Science Advances*, 6: eabc2717, 2020.
- Guerra, P. H., Meira, W., Cardie, C., and Kleinberg, R. A measure of polarization on social media networks based on community boundaries. In *International AAAI Conference on Web and Social Media (ICWSM)* 7, 2013.
- Guo, W. and Caliskan, A. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)* 4, 2021.
- Haidt, J. and Graham, J. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, 20(1):98–116, 2007.
- Haidt, J. and Joseph, C. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66, 2004.
- He, Z., Mokhberian, N., Câmara, A., Abeliuk, A., and Lerman, K. Detecting polarized topics in COVID-19 news using partisanship-aware contextualized topic embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021.
- Heckman, J. J. and Snyder, J. M. Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators. *The RAND Journal of Economics*, 28(0):142–189, 1997.
- Hewitt, J. and Manning, C. D. A structural probe for finding syntax in word representations. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HTL) 2019*, 2019.
- Heywood, A. *Political ideologies: An introduction*. Macmillan, London, UK, 2017.
- Himelboim, I., McCreery, S., and Smith, M. Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on Twitter. *Journal of Computer-Mediated Communication*, 18(2):40–60, 2013.
- Hofmann, V., Dong, X., Pierrehumbert, J. B., and Schütze, H. Modeling ideological salience and framing in polarized online groups with graph neural networks and structured sparsity. In *Findings of the Association for Computational Linguistics: NAACL 2022*, 2022a.
- Hofmann, V., Schütze, H., and Pierrehumbert, J. B. The Reddit politosphere: A large-scale text and network resource of online political discourse. In *International AAAI Conference on Web and Social Media (ICWSM) 16*, 2022b.
- Hopp, F. R., Fisher, J. T., Cornell, D., Huskey, R., and Weber, R. The extended Moral Foundations Dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior Research Methods*, 53(1):232–246, 2021.
- Iyyer, M., Enns, P., Boyd-Graber, J., and Resnik, P. Political ideology detection using recursive neural networks. In *Annual Meeting of the Association for Computational Linguistics (ACL)* 52, 2014.

- Jiang, M. and Fellbaum, C. Interdependencies of gender and race in contextualized word embeddings. In *Workshop on Gender Bias in Natural Language Processing (GeBNLP) 2*, 2020.
- Kingma, D. P. and Ba, J. L. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR) 3*, 2015.
- Kipf, T. N. and Welling, M. Variational graph auto-encoders. In *NIPS Bayesian Deep Learning Workshop*, 2016.
- Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR) 5*, 2017.
- Knoche, M., Popović, R., Lemmerich, F., and Strohmaier, M. Identifying biases in politically biased Wikis through word embeddings. In *ACM Conference on Hypertext and Social Media (HT) 30*, 2019.
- Kulkarni, V., Ye, J., Skiena, S., and Wang, W. Y. Multi-view models for political ideology detection of news articles. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2018*, 2018.
- Kumar, S., Hamilton, W., Leskovec, J., and Jurafsky, D. Community interaction and conflict on the web. In *The Web Conference (WWW) 27*, 2018.
- Lebedev, V. and Lempitsky, V. Fast ConvNets using group-wise brain damage. In *Conference on Computer Vision and Pattern Recognition (CVPR) 29*, 2016.
- Liu, B., Wang, M., Foroosh, H., Tappen, M., and Pensky, M. Sparse convolutional neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR) 28*, 2015.
- Massachs, J., Monti, C., Morales, G. D. F., and Bonchi, F. Roots of Trumpism: Homophily and social feedback in Donald Trump support on Reddit. In *ACM Conference on Web Science (WebSci) 12*, 2020.
- McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.
- McPherson, M., Smith-Lovin, L., and Cook, J. M. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.
- Mejova, Y., Zhang, A. X., Diakopoulos, N., and Castillo, C. Controversy and sentiment in online news. In *arXiv 1409.8152*. 2014.
- Mendelsohn, J., Tsvetkov, Y., and Jurafsky, D. A framework for the computational linguistic analysis of dehumanization. *Frontiers in Artificial Intelligence*, 3:55, 2020.
- Mendelsohn, J., Budak, C., and Jurgens, D. Modeling framing in immigration discourse on social media. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HTL) 2021*, 2021.
- Mokhberian, N., Abeliuk, A., Cummings, P., and Lerman, K. Moral framing and ideological bias of news. In *International Conference on Social Informatics (SocInfo) 12*, 2020.
- Nelson, T. E., Oxley, Z. M., and Clawson, R. A. Toward a psychology of framing effects. *Political Behavior*, 19(3):221–246, 1997.
- Nguyen, D., Rosseel, L., and Grieve, J. On learning and representing social meaning in NLP: A sociolinguistic perspective. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HTL) 2021*, 2021.
- Nguyen, K. A., Schulte im Walde, S., and Vu, N. T. Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction. In *Annual Meeting of the Association for Computational Linguistics (ACL) 54*, 2016.
- Nguyen, K. A., Schulte im Walde, S., and Vu, N. T. Distinguishing antonyms and synonyms in a pattern-based neural network. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL) 15*, 2017.
- Olson, R. S. and Neal, Z. P. Navigating the massive world of Reddit: Using backbone networks to map user interests in social media. *PeerJ Computer Science*, 2015.
- Parikh, N. and Boyd, S. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2013.
- Pretiuc-Pietro, D., Liu, Y., Hopkins, D. J., and Ungar, L. H. Beyond binary labels: Political ideology prediction of Twitter users. In *Annual Meeting of the Association for Computational Linguistics (ACL) 55*, 2017.
- Psylla, I., Sapiezynski, P., Mones, E., and Lehmann, S. The role of gender in social network organization. *PLoS ONE*, 12(12):e0189873, 2017.
- Roy, S. and Goldwasser, D. Weakly supervised learning of nuanced frames for analyzing polarization in news media. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2020*, 2020.
- Rozado, D. and al Gharbi, M. Using word embeddings to probe sentiment associations of politically loaded terms in news and opinion articles from news media outlets. *Journal of Computational Social Science*, 56(3):256, 2021.

- Sagi, E., Diermeier, D., and Kaufmann, S. Identifying issue frames in text. *PloS ONE*, 8(7):e69185, 2013.
- Satopää, V., Albrecht, J., Irwin, D., and Raghavan, B. Finding a “kneedle” in a haystack: Detecting knee points in system behavior. In *International Conference on Distributed Computing Systems (ICDCS) 31*, 2011.
- Shen, Q. and Rosé, C. The discourse of online content moderation: Investigating polarized user responses to changes in Reddit’s quarantine policy. In *Workshop on Abusive Language Online 3*, 2019.
- Shwartz, V., Santus, E., and Schlechtweg, D. Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL) 15*, 2017.
- Silverstein, M. Indexical order and the dialectics of sociolinguistic life. *Language & Communication*, 23(3-4): 193–229, 2003.
- Sylwester, K. and Purver, M. Twitter language use reflects psychological differences between democrats and republicans. *PloS ONE*, pp. e0137422, 2015.
- Tan, Y. C. and Celis, L. E. Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems (NeurIPS) 33*, 2019.
- Tripodi, R., Warglien, M., Sullam, S. L., and Paci, D. Tracing antisemitic language through diachronic embedding projections: France 1789-1914. In *International Workshop on Computational Approaches to Historical Language Change 1*, 2019.
- Tyagi, A., Field, A., Lathwal, P., Tsvetkov, Y., and Carley, K. M. A computational analysis of polarization on Indian and Pakistani social media. In *International Conference on Social Informatics (SocInfo) 12*, 2020.
- Vargas, F. and Cotterell, R. Exploring the linear subspace hypothesis in gender bias mitigation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2020*, 2020.
- Vorakitphan, V., Guerini, M., Cabrio, E., and Villata, S. Regexit or not regexit: Aspect-based sentiment analysis in polarized contexts. In *International Conference on Computational Linguistics (COLING) 28*, 2020.
- Vorontsov, E., Trabelsi, C., Kadoury, S., and Pal, C. On orthogonality and learning recurrent networks with long term dependencies. In *International Conference on Machine Learning (ICML) 34*, 2017.
- Waller, I. and Anderson, A. Quantifying social organization and political polarization in online platforms. *Nature*, 600 (7888):264–268, 2021.
- Walter, T., Kirschner, C., Eger, S., Glavaš, G., Lauscher, A., and Ponzetto, S. P. Diachronic analysis of German parliamentary proceedings: Ideological shifts through the lens of political biases. In *arXiv 2108.06295*. 2021.
- Warriner, A., Kuperman, V., and Brysbaert, M. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, 45(4):1191–1207, 2013.
- Weber, I., Garimella, K., and Batayneh, A. Secular vs. islamist polarization in Egypt on Twitter. In *International Conference on Advances in Social Networks Analysis and Mining (ASONAM) 2013*, 2013.
- Webson, A., Chen, Z., Eickhoff, C., and Pavlick, E. Do “undocumented workers” == “illegal aliens”? Differentiating denotation and connotation in vector spaces. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2020*, 2020.
- Wen, W., Wu, C., Wang, Y., Chen, Y., and Li, H. Learning structured sparsity in deep neural networks. In *Advances in Neural Information Processing Systems (NIPS) 30*, 2016.
- Wiedemann, G., Remus, S., Chawla, A., and Biemann, C. Does BERT make any sense? interpretable word sense disambiguation with contextualized embeddings. In *arXiv 1909.10430*. 2019.
- Xie, J. Y., Ferreira Pinto, R., Hirst, G., and Xu, Y. Text-based inference of moral sentiment change. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2019*, 2019.
- Yardi, S. and Boyd, D. Dynamic debates: An analysis of group polarization over time on Twitter. *Bulletin of Science, Technology & Society*, 30(5):316–327, 2010.
- Yoon, J. and Hwang, S. J. Combined group and exclusive sparsity for deep neural networks. In *International Conference on Machine Learning (ICML) 34*, 2017.
- Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, 68(1):49–67, 2006.
- Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., and Chang, K.-W. Gender bias in contextualized word embeddings. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HTL) 2019*, 2019.

Table 6: Training statistics. μ_m , σ_m : mean and standard deviation of MAUC performance on dev for all hyperparameter search trials; r : learning rate; λ_o : orthogonality constant (only \mathcal{X}_*); λ_s : sparsity constant (only \mathcal{X}_*); τ : runtime in seconds for full hyperparameter search.

Space	Year	μ_m	σ_m	r	λ_o	λ_s	τ
\mathcal{X}_*	2013	.619	.021	1e-04	3e-03	1e-02	39,633
	2014	.599	.045	3e-04	1e-03	1e-02	40,489
	2015	.625	.016	3e-04	1e-03	1e-02	40,757
	2016	.664	.016	3e-04	1e-03	1e-02	43,315
	2017	.662	.025	1e-04	1e-03	3e-02	44,341
	2018	.636	.015	1e-04	1e-03	1e-02	44,665
	2019	.705	.019	1e-04	1e-03	3e-02	47,159
\mathcal{X}	2013	.623	.024	1e-04	—	—	3,297
	2014	.651	.032	1e-03	—	—	3,333
	2015	.648	.024	3e-04	—	—	3,335
	2016	.665	.030	3e-04	—	—	3,543
	2017	.674	.021	1e-04	—	—	3,553
	2018	.655	.016	1e-04	—	—	3,536
	2019	.721	.019	3e-04	—	—	3,717

A. Appendices

A.1. Hyperparameters and Training Details

Both hidden layers of the graph auto-encoder have 10 dimensions. The number of trainable parameters is 7,800 (\mathcal{X}) and 597,624 (\mathcal{X}_*), with the latter shrinking during training as a result of the sparsity penalty.

Table 6 provides training statistics such as mean and standard deviation of the MAUC performance on dev, best hyperparameter configurations, and runtimes. Experiments are performed on a GeForce GTX 1080 Ti GPU (11GB).

A.2. Nominal and Verbal Semantic Axes

Nominal and verbal semantic axes with highest and lowest s_a scores are provided in Tables 7 and 8. Similar to adjectives, while top axes tend to have abstract evaluative meanings, this is not the case for bottom axes.

Table 7: Top and bottom nominal semantic axes. For each year, the table shows the four nominal semantic axes with highest and lowest s_a scores.

Year	Top	Bottom
2013	<i>assets/liabilities</i>	<i>dislike/liking</i>
	<i>objective/subjective</i>	<i>dwarf/giant</i>
	<i>divestment/investment</i>	<i>nay/yea</i>
	<i>immorality/morality</i>	<i>husband/wife</i>
2014	<i>patriot/traitor</i>	<i>inside/outside</i>
	<i>citizen/foreigner</i>	<i>permanent/temporary</i>
	<i>impossibility/possibility</i>	<i>higher/lower</i>
	<i>ability/inability</i>	<i>external/internal</i>
2015	<i>assets/liabilities</i>	<i>major/minor</i>
	<i>objective/subjective</i>	<i>king/queen</i>
	<i>belief/skepticism</i>	<i>defeat/victory</i>
	<i>demand/supply</i>	<i>decision/knockout</i>
2016	<i>impossibility/possibility</i>	<i>inside/outside</i>
	<i>credit/debit</i>	<i>closing/opening</i>
	<i>guilt/innocence</i>	<i>comedy/drama</i>
	<i>freeman/slave</i>	<i>east/west</i>
2017	<i>evolution/revolution</i>	<i>husband/wife</i>
	<i>analysis/synthesis</i>	<i>afternoon/morning</i>
	<i>barbarism/culture</i>	<i>closing/opening</i>
	<i>objective/subjective</i>	<i>tomorrow/yesterday</i>
2018	<i>anarchy/government</i>	<i>afternoon/morning</i>
	<i>barbarism/culture</i>	<i>known/unknown</i>
	<i>deflation/inflation</i>	<i>higher/lower</i>
	<i>immorality/morality</i>	<i>tomorrow/yesterday</i>
2019	<i>immorality/morality</i>	<i>everybody/nobody</i>
	<i>barbarism/culture</i>	<i>afternoon/morning</i>
	<i>deflation/inflation</i>	<i>standing/working</i>
	<i>client/server</i>	<i>anything/nothing</i>

Table 8: Top and bottom verbal semantic axes. For each year, the table shows the four verbal semantic axes with highest and lowest s_a scores.

Year	Top	Bottom
2013	<i>criminalize/decriminalize</i>	<i>acknowledge/deny</i>
	<i>ameliorate/exacerbate</i>	<i>hate/love</i>
	<i>plummet/skyrocket</i>	<i>shout/whisper</i>
	<i>complicate/simplify</i>	<i>bless/damn</i>
2014	<i>plummet/skyrocket</i>	<i>prefix/suffix</i>
	<i>deport/repatriate</i>	<i>acknowledge/deny</i>
	<i>emigrate/immigrate</i>	<i>avoid/confront</i>
	<i>disqualify/qualify</i>	<i>couple/decouple</i>
2015	<i>agree/disagree</i>	<i>forget/remember</i>
	<i>decrypt/encrypt</i>	<i>cease/continue</i>
	<i>conform/deviate</i>	<i>get/miss</i>
	<i>plummet/skyrocket</i>	<i>move/stay</i>
2016	<i>criminalize/decriminalize</i>	<i>guess/know</i>
	<i>entice/frighten</i>	<i>lock/unlock</i>
	<i>plummet/skyrocket</i>	<i>enter/exit</i>
	<i>hasten/postpone</i>	<i>head/tail</i>
2017	<i>centralize/decentralize</i>	<i>lock/unlock</i>
	<i>generalize/specialize</i>	<i>irritate/please</i>
	<i>dehumanize/humanize</i>	<i>relax/stress</i>
	<i>overpay/underpay</i>	<i>hate/love</i>
2018	<i>deport/repatriate</i>	<i>relax/stress</i>
	<i>criminalize/decriminalize</i>	<i>guess/know</i>
	<i>centralize/decentralize</i>	<i>forget/remember</i>
	<i>detain/liberate</i>	<i>keep/let</i>
2019	<i>entice/frighten</i>	<i>guess/know</i>
	<i>generalize/specialize</i>	<i>relax/stress</i>
	<i>centralize/decentralize</i>	<i>agree/disagree</i>
	<i>elevate/relegate</i>	<i>forget/remember</i>