Modeling Ideological Agenda Setting and Framing in Polarized Online Groups with Graph Neural Networks and Structured Sparsity

Valentin Hofmann[*‡], Janet B. 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

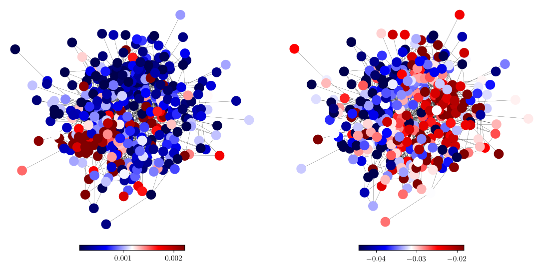
Abstract

The increasing polarization of online political discourse calls for computational tools that are able to automatically detect and monitor ideological divides in social media. Here, we introduce a minimally supervised method that directly leverages the network structure of online discussion forums, specifically Reddit, to detect polarized concepts. We model polarization along the dimensions of agenda setting and framing, drawing upon insights from moral psychology. The architecture we propose combines graph neural networks with structured sparsity learning and results in representations for concepts and subreddits that capture phenomena such as ideological radicalization and subreddit hijacking. We also create a new dataset of political discourse spanning 12 years and covering more than 600 online groups with different ideologies.

1 Introduction

The ideological polarization of online political discourse on platforms such as Twitter (Yardi and Boyd, 2010; Conover et al., 2011; Himelboim et al., 2013), Facebook (Bakshy et al., 2015), and Reddit (An et al., 2019; Marchal, 2020) has received increasing attention in the computational social sciences over the last years, particularly after the beginning of the Covid-19 pandemic (Green et al., 2020; Jing and Ahn, 2021). In NLP, a growing body of work on polarization has discovered typical mechanisms by which polarization manifests itself linguistically (An et al., 2018; Demszky et al., 2019; Shen and Rosé, 2019; Roy and Goldwasser, 2020; Tyagi et al., 2020; Vorakitphan et al., 2020). However, these studies rely on explicit information about the political orientation of text (e.g., manual labels), a requirement seldom met in dynamically evolving social media.

In this paper, we propose a minimally supervised method that directly leverages the network structure of online discussion forums, specifically Reddit, to detect polarized concepts. Building on prior work on political communication (Tsur et al., 2015; Field et al., 2018; Mendelsohn et al., 2021), we model the polarization of concepts along the dimensions of **agenda setting** (*what* concepts are discussed?) and **framing** (*how* are the concepts discussed?). For framing, we take into account insights about the psychological foundations of ideology (Haidt and Graham, 2007) and construct moral embedding subspaces that capture nuanced biases in the way concepts are discussed.



(a) Polarization of *fascist* (b) Polarization of *fact*

Figure 1: Examples of concepts polarized along the dimensions of agenda setting (a) and framing (b) in Reddit in 2019. Each circle is a subreddit. The values for agenda setting (a) are relative concept frequencies. The values for framing (b) are contextualized BERT embeddings projected into the moral authority/respect subspace. The relative frequency of *fascist* is higher in communist subreddits. The polarization of *fact* can be interpreted in the light of post-factual politics. We can diagnose such patterns using our method SLAP4SLIP in a minimally supervised way.

To automatically detect polarized concepts, we introduce a novel framework called SLAP4SLIP (Sparse language properties for social link prediction) that aims at finding linguistic features maximally informative about the edge topology of a social network. The model we propose for SLAP4SLIP combines graph neural networks with structured sparsity learning and identifies in a selfsupervised way (i) which concepts are the most polarizing ones for a social network, (ii) whether the polarization is due to differences in agenda setting or framing (or both), and (iii) which moral foundations are involved (when framing is relevant for a concept). For 2019, e.g., we find that *fascist* and *fact* are among the most polarized concepts with respect to agenda setting and framing, respectively (Figure 1). Our model also learns embeddings for individual subreddits that represent group-level ideology. We show that these embeddings capture ideological dynamics such as right-wing radicalization and subreddit hijacking.

Contributions. We introduce a novel framework for finding linguistic features maximally informative about the edge topology of a social network called SLAP4SLIP (Sparse language properties for social link prediction) and show that it can be used to detect polarized concepts in online discussion forums. We model polarization on the levels of agenda setting and framing, drawing upon insights from moral psychology. The architecture we propose combines graph neural networks with structured sparsity and learns rich representations for concepts and subreddits. We also create a new dataset of political discourse spanning 12 years and covering more than 600 online groups with different ideologies, making it a valuable resource for studies in the computational social sciences.[1]

2 Related Work

Our study is closely related to previous NLP studies on **polarization** (An et al., 2018; Demszky et al., 2019; Shen and Rosé, 2019; Roy and Goldwasser, 2020; Tyagi et al., 2020; Vorakitphan et al., 2020), but we try to avoid the need for explicit information about ideologies by leveraging the network structure of online discussion forums. There is also a large body of work on polarization in the computational social sciences (Adamic and Glance, 2005; Yardi and Boyd, 2010; Conover et al., 2011; Calais et al., 2013; Himelboim et al., 2013; Weber et al., 2013; Bakshy et al., 2015; Garcia et al., 2015; Garimella and Weber, 2017; Morales et al., 2019) and sociophysics (Baumann et al., 2020a,b; Prasetya and Murata, 2020). One insight of this line of work relevant for our study is that the structure of various types of online social networks reflects political polarization, which has been explained as a result of homophily (McPherson et al., 2001), i.e., nodes close to each other in the social network are likely to share similar views. More broadly, our study is related to NLP work on **ideological and political language** in general (Lin et al., 2008; Monroe et al., 2008; Gerrish and Blei, 2011; Sagi et al., 2013; Sim et al., 2013; Iyyer et al., 2014; Mejova et al., 2014; Volkova et al., 2014; Preotiuc-Pietro et al., 2017; Kulkarni et al., 2018)

Previous NLP work has shown that **agenda setting and framing** are two key mechanisms by which attention can be drawn to certain topics (agenda setting) or certain aspects of a topic (framing) during political communication (Card et al., 2015; Tsur et al., 2015; Card et al., 2016; Field et al., 2018; Demszky et al., 2019; Mendelsohn et al., 2021). For ideological framing in particular, the five moral foundations harm/care, fairness/reciprocity, ingroup/loyalty, authority/respect, and purity/sanctity from **moral foundations theory** (Haidt and Graham, 2007; Graham et al., 2009) have been shown to provide a suitable theoretical basis for analyzing what aspects of an issue tend to be highlighted by different ideologies (Garten et al., 2016; Fulgoni et al., 2016; Johnson and Goldwasser, 2018; Mokhberian et al., 2020). We follow this approach but as opposed to previous studies operate with contextualized embeddings that we project into moral embedding subspaces.

Methodologically, we draw on advances in **deep learning on graphs**, specifically graph convolutional networks and graph auto-encoders (Kipf and Welling, 2016, 2017). In NLP, such graph-based architectures are increasingly used to include information from social networks for downstream tasks (Yang and Eisenstein, 2017; del Tredici et al., 2019; Mishra et al., 2019; Hofmann et al., 2020). Our work differs from these studies in that we combine deep learning on graphs with **structured sparsity**, a form of regularization similar to ℓ_1 regularization (Tibshirani, 1996) that sets entire groups of parameters to zero (Liu et al., 2015; Alvarez and Salzmann, 2016; Lebedev and Lempitsky, 2016; Wen et al., 2016; Yoon and Hwang, 2017; Wen et al., 2018). Structured sparsity has been used in NLP before (Eisenstein et al., 2011; Martins et al., 2011; Murray and Chiang, 2015; Dodge et al., 2019), but not in connection with deep learning on graphs.

3 SLAP4SLIP Framework

The key idea of this paper is to directly leverage the social network structure for determining polarized

[1] We will make all our code and data publicly available.

concepts. We introduce a novel framework called SLAP4SLIP (Sparse language properties for social link prediction) whose goal it is to model the structure of social networks in a data-driven way that obviates the need for extensive human annotation. SLAP4SLIP is a general framework to detect the most salient types of linguistic variablity in social networks and is in principle applicable in any scenario involving social networks with textual data attached to each node. In this paper, we show that for polarized online discussion forums, SLAP4SLIP can be used to find polarized concepts.

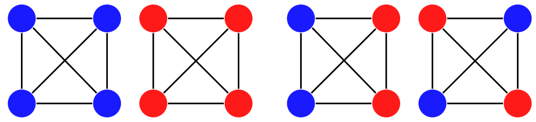
Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a network consisting of a set of nodes representing social entities, $\mathcal{V}$, and a set of edges, $\mathcal{E}$, representing connections between the social entities. We denote with $\mathbf{A} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$ the adjacency matrix of $\mathcal{G}$. Let further $\mathcal{C}$ be a set of word n-grams denoting concepts. Here, we confine ourselves to subreddits for $\mathcal{V}$ and unigrams and bigrams for $\mathcal{C}$, but SLAP4SLIP is applicable in other scenarios (e.g., for networks of people or concepts extracted from text in a more complex manner). We define a function $\psi_l : \mathcal{V} \times \mathcal{C} \to \mathbb{R}$ that assigns to each node $v_i \in \mathcal{V}$ and concept $c_j \in \mathcal{C}$ the value of a linguistic property l observed for c_j in v_i. ψ_l can be represented as a matrix in $\mathbb{R}^{|\mathcal{V}|\times|\mathcal{C}|}$,

$$\mathbf{\Psi}_l = \begin{bmatrix} \psi_l(v_1, c_1) & \dots & \psi_l(v_1, c_{|\mathcal{C}|}) \\ \vdots & \ddots & \vdots \\ \psi_l(v_{|\mathcal{V}|}, c_1) & \dots & \psi_l(v_{|\mathcal{V}|}, c_{|\mathcal{C}|}) \end{bmatrix},$$

where each column is a graph signal (Dong et al., 2020) over $\mathcal{G}$ determined by c_j and ψ_l. E.g., if we chose l to be the frequency count, ψ_l would indicate how often each concept occurred in the text attached to each node of the network.

The goal of SLAP4SLIP is to find the subset of concepts $\mathcal{C}^* \subseteq \mathcal{C}$ that best meets the following two desiderata: (i) given a linguistic property l, the signals imposed on $\mathcal{G}$ by ψ_l and the concepts in $\mathcal{C}^*$ should allow for optimal predictionsabout the structure of $\mathcal{G}$, specifically $\mathcal{E}$; (ii) the number of concepts in $\mathcal{C}^*$ should be minimal. In practice, we treat this as a constrained optimization problem (Bertsekas, 1982), i.e., we use (i) as the objective and impose (ii) as a hard constraint on $|\mathcal{C}^*|$.

As an example, consider the network in Figure 2. The network consists of eight nodes that fall into two fully connected components with no edges between the components. $\mathcal{C}$ consists of the two concepts c_1 and c_2. Taking the frequency count as linguistic property l and displaying it with the color of nodes, ψ_l results in the two signals shown in Figure 2. We can see that the signal of concept c_1 alone allows for a perfect prediction of the network structure according to the decision rule



(a) ψ_l for c_1 (b) ψ_l for c_2

Figure 2: Example for the prediction of graph structure from a linguistic property. The figures show ψ_l for concepts c_1 and c_2 on a toy graph, with l chosen to be the frequency count represented by node color (identical colors mean identical frequencies). The edges can be fully predicted from ψ_l for c_1 but not for c_2.

$$\mathbf{A}_{ij} = \begin{cases} 1 & \text{if } \psi_l(v_i, c_1) = \psi_l(v_j, c_1) \\ 0 & \text{otherwise.} \end{cases}$$

Since c_2 cannot achieve a perfect prediction, $\mathcal{C}^* = \{c_1\}$ is the optimal solution. Notice the variance of $\psi_l(v_i, c_j)$ is identical for both concepts and does not represent a good distinguishing factor.

Three important points need to be mentioned. First, in real-world social networks it will rarely be the case that one concept alone is enough for an optimal solution, or that the optimal solution allows for a perfect prediction. Second, the optimal solution is not necessarily unique: there might be another concept, c_3, with a similar or identical frequency count distribution as c_1 such that $\mathcal{C}^* = \{c_3\}$ would also be an optimal solution. Third, the choice of l is crucial for SLAP4SLIP. In this paper, we choose linguistic features$l$ that capture concept-level ideological agenda setting and framing (Section 5).

4 Reddit Politosphere Dataset

Reddit is an online discussion forum where people can create communities, so-called subreddits, devoted to certain interests or topics. Many of these subreddits are political discussion groups (e.g., `politics`), sometimes with explicit ideological orientation (e.g., `democrats`). Most previous work analyzing political discourse on Reddit has relied on a small number of hand-picked subreddits or external lists of political subreddits (An et al., 2019; Grover and Mark, 2019; Guimaraes et al., 2019), which provides an only incomplete picture of the Reddit political landscape.

To remedy this, we construct the Reddit Politosphere Dataset, a collection of comments from over 600 political subreddits spanning 12 years. Taking all comments from the Pushshift Reddit Dataset (Baumgartner et al., 2020) between 2008 and 2019, we first train year-specific Naive Bayes classifiers (Manning et al., 2008) to detect political comments. For each year, we take all comments from eight subreddits representing different points on the ideological spectrum (`Anarchism`, `Anarcho_Capitalism`, `Conservative`, `Libertarian`, `Republican`, `democrats`, `progressive`, `socialism`) as positive (political) examples and an equally-sized sample of comments from the default subreddits (a set of subreddits users used to be subscribed to automatically) as negative (non-political) examples.[2] We split the resulting comments into 80% train, 10% dev, and 10% test comments and train year-specific classifiers on train (with absolute discounting for smoothing) and evaluate them on test.[3] The performance of the classifiers on test as measured by accuracy is high for all years and lies between 81.3% (2008) and 84.1% (2012). We then predict for all comments from the Pushshift Reddit Dataset between 2008 and 2019 whether they are political or not. Based on these predictions, we classify a subreddit for a certain year as a political subreddit if the ratio of political and non-political comments is larger than two, i.e., there are at least twice as many political as non-political comments, and if there are at least 1000 comments and 100 users in the subreddit that year. We then manually go over the detected subreddits and remove subreddits that are not concerned with real-world politics (e.g., political simulation and gaming subreddits), resulting in a final list of 606 subreddits. To externally check the coverage of our dataset, we compare against a list of known ideological subreddits and find that the ones meeting our inclusion criteria are all among the found subreddits.[4]

We then construct year-specific social networks in the following way. For each year, we compute for every pair of subreddits the number of users that posted at least 10 comments in both subreddits, defining a weighted network over the subreddits. We use statistical backboning methods, specifically the noise-corrected filter (Coscia and Neffke, 2017), to find the edges that are significant above a significance threshold and discard all edges below the threshold, thus generating unweighted networks. We tune the threshold by examining the ratio of kept edges versus kept nodes (Serrano et al., 2009) and use the Kneedle algorithm (Satopää et al., 2011) to detect the knee point.

Year	$\lvert\mathcal{D}\rvert$	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	μ_d	μ_π	ρ
2008	736,175	9	9	2.00	2.31	.250
2009	1,021,653	14	17	2.43	2.55	.187
2010	1,961,843	21	28	2.67	2.54	.133
2011	3,875,384	56	143	5.11	2.70	.093
2012	5,691,828	86	271	6.30	2.78	.074
2013	6,306,458	108	324	6.00	3.08	.056
2014	6,664,567	132	335	5.08	3.86	.039
2015	9,230,022	168	493	5.87	3.87	.035
2016	34,801,075	255	1,318	10.34	3.14	.041
2017	38,278,685	295	1,572	10.66	3.14	.036
2018	40,222,627	316	1,604	10.15	3.17	.032
2019	46,590,000	412	2,536	12.31	3.20	.030

Table 1: Dataset statistics. $\lvert\mathcal{D}\rvert$: number of comments; $\lvert\mathcal{V}\rvert$: number of nodes (subreddits) in network; $\lvert\mathcal{E}\rvert$: number of edges; μ_d: average node degree; μ_π: average shortest path length; ρ: network density. In this study, we only use data starting from 2013.

Table 1 provides year-wise summary statistics. In this study, we only use data starting from 2013 (the first year in which the social network has more than 100 nodes). See Appendix A.1 for details about data preprocessing.

5 Model

We adopt the SLAP4SLIP framework (Section 3) to model ideological polarization along the dimensions of agenda setting and framing in the Reddit Politosphere Dataset (Section 4). Specifically, we propose a neural architecture that uses information about concept-level agenda setting and framing to predict links between subreddits while reducing the number of considered concepts as far as possible. This has the effects of (i) resulting in a compact set of concepts that is maximally informative about the social network structure and can be used for analytical purposes and (ii) providing neural representations of subreddits that combine linguistic and social information. At the same time, the performance on link prediction makes it straightforward to compare the quality of different models.

[2] We remove `r/news` and `r/worldnews` from the default subreddits since they also contain political content.

[3] We tune the discounting parameter on the dev comments of 2008 and use the best value for all other years.

[4] We use the partisan subreddits from https://web.archive.org/web/20190502124604/https://www.reddit.com/r/politics/wiki/relatedsubs.

5.1 Determining Concepts

To obtain the concepts $\mathcal{C}$, we compute for each year mutual information scores for unigrams and bigrams based on the political comments and an equally-sized sample of comments from the default subreddits.[5] We only consider unigrams and bigrams that appear more often within than outside of noun phrases as detected by a noun phrase chunker to remove unigrams and bigrams typical of discussions but not relevant to agenda setting and framing (e.g., *regarding*, *dont think*). We then take for $\mathcal{C}$ the top 1000 unigrams and bigrams according to mutual information. This and all other other steps are done separately for each year, i.e., we extract year-wise concepts and train year-wise models to detect the polarizing concepts.

5.2 Modeling Agenda Setting and Framing

The first part of the architecture models the function ψ_l, i.e., it extracts linguistic information related to concept-level agenda setting and framing from the subreddits and maps them to scalar representations. In the resulting matrix $\mathbf{\Psi}_l$, each column is a signal on the entire graph defined by one concept, and each row is a feature vector for one subreddit defined by all concepts in $\mathcal{C}$ (Section 3).

To model concept-level ideological agenda setting, we measure the relative frequency of concepts within subreddits,

$$a(v_i, c_j) = \frac{n(v_i, c_j)}{\sum_k n(v_i, c_k)},$$

where $n(v_i, c_j)$ is the frequency count of concept c_j in subreddit v_i. Variations in the relative frequency of a concept that are strongly correlated with the social network structure indicate that the concept is used with systematically higher frequency in certain regions of the social network, potentially caused by ideological agenda setting.

To model concept-level framing, we use a pretrained language model, specifically BERT (Devlin et al., 2019), and obtain contextualized embeddings for the concepts.[6] For each subreddit v_i and concept c_j, we then compute the average contextualized embedding, $\mathbf{e}(v_i, c_j)$. Contextualized embeddings capture fine semantic nuances of the sentence context (Field and Tsvetkov, 2019; Wiedemann et al., 2019), making them a good starting point for modeling ideological framing. To distill the ideologically relevant information, we further project the average contextualized embeddings into five ideological subspaces corresponding to the five moral foundations of moral foundations theory (Haidt and Graham, 2007; Graham et al., 2009). Specifically, for each moral foundation m_k, we use BERT to obtain contextualized embeddings for the ten highest-ranked words of both poles according to Frimer et al. (2017), and compute average contextualized embeddings for each word.[7] We then perform PCA on the average contextualized embeddings for each m_k and use the first principal component as the subspace representation, $\mathbf{e}(m_k)$. This allows us to project the subreddit-specific average contextualized concept embeddings $\mathbf{e}(v_i, c_j)$ into the five moral subspaces,

$$s_k(v_i, c_j) = \cos\left(\mathbf{e}(v_i, c_j), \mathbf{e}(m_k)\right).$$

$s_k(v_i, c_j)$ reflects how relevant the moral foundation m_k is for the ideological framing of concept c_j in subreddit v_i. Of course, the moral foundations are expected to be relevant for the framing of concepts to differing degrees. We therefore compute concept-specific weighted sums,

$$f(v_i, c_j) = \sum_k \beta_k^{(c_j)} s_k(v_i, c_j),$$

where $\sum_k \beta_k^{(c_j)} = 1$ and $\beta_k^{(c_j)} \geq 0$. $f(v_i, c_j)$ is an aggregate indicator of how important moral framing is for concept c_j in v_i. The parameters $\beta_k^{(c_j)}$ are optimized during training.

Since agenda setting and framing can be of different importance for different concepts, we combine $a(v_i, c_j)$ and $f(v_i, c_j)$ in a weighted sum,

$$u(v_i, c_j) = \gamma^{(c_j)} a(v_i, c_j) + (1 - \gamma^{(c_j)}) f(v_i, c_j),$$

where $0 \leq \gamma^{(c_j)} \leq 1$ is again a concept-specific parameter that is optimized during training. Two important points must be stressed. First, $\beta_k^{(c_j)}$ and $\gamma^{(c_j)}$are specific for concepts but identical for all subreddits: e.g., if a concept $c_j$has $\gamma^{(c_j)} = 1$, this means that only information from $a(v_i, c_j)$ is used for all subreddits. Second, values for $u(v_i, c_j)$ are only comparable across subreddits but not across

[5] We again remove `r/news` and `r/worldnews`.

[6] We extract the mean-pooled embedding if the concept is split into multiple WordPiece tokens. For energy considerations (Strubell et al., 2019), we sample a maximum of 100 occurrences per subreddit and concept.

[7] We sample 1000 occurrences per word.

Model	DEV 2013	2014	2015	2016	2017	2018	2019	$\mu \pm \sigma$	TEST 2013	2014	2015	2016	2017	2018	2019	$\mu \pm \sigma$
AF-SGAE	.857	.893	.911	.921	.923	.913	.921	.906±.022	.890	.895	.895	.923	.937	.908	.934	.912±.018
A-SGAE	.833	.868	.872	.864	.883	.865	.904	.870±.020	.886	.890	.853	.875	.894	.864	.925	.884±.022
F-SGAE	.832	.880	.863	.861	.884	.868	.894	.869±.019	.875	.893	.878	.885	.905	.875	.917	.890±.015
AF-SLAE	.712	.812	.772	.771	.778	.729	.748	.760±.031	.653	.810	.754	.781	.764	.729	.752	.749±.046

Table 2: Performance (AUC = area under the ROC curve) on edge prediction. Best score per column in gray, second-best in light-gray. AF-SGAE outperforms baselines that use only agenda setting information (A-SGAE), only framing information (F-SGAE), or lack graph convolutions (AF-SLAE).

Model	DEV 2013	2014	2015	2016	2017	2018	2019	$\mu \pm \sigma$	TEST 2013	2014	2015	2016	2017	2018	2019	$\mu \pm \sigma$
AF-SGAE	.846	.881	.909	.924	.926	.914	.929	.904±.028	.905	.902	.893	.928	.935	.918	.942	.918±.017
A-SGAE	.813	.862	.880	.871	.893	.875	.909	.872±.028	.905	.899	.837	.883	.898	.875	.933	.890±.028
F-SGAE	.844	.851	.850	.862	.896	.869	.905	.868±.022	.895	.905	.879	.890	.910	.886	.925	.899±.015
AF-SLAE	.744	.811	.803	.777	.782	.741	.762	.774±.025	.685	.804	.740	.787	.763	.736	.752	.753±.036

Table 3: Performance (AP = average precision) on edge prediction. Best score per column in gray, second-best in light-gray. AF-SGAE generally outperforms baselines that use only agenda setting information (A-SGAE), only framing information (F-SGAE), or lack graph convolutions (AF-SLAE).

concepts: since $\beta_k^{(c_j)}$ and $\gamma^{(c_j)}$ differ between concepts, differences in $u(v_i, c_j)$ are not meaningful for two different concepts.

To get the final concept representation that is passed to subsequent parts of the model, we set $\psi_l = u$, i.e., each entry in $\mathbf{\Psi}_l$ contains the value of $u(v_i, c_j)$ for subreddit v_i and concept c_j.

5.3 Graph Auto-encoder

We use a graph auto-encoder (Kipf and Welling, 2016) to predict the links in $\mathcal{G}$. The graph auto-encoder takes as input the matrix $\mathbf{\Psi}_l$ as well as the adjacency matrix $\mathbf{A}$ of $\mathcal{G}$.

The encoder consists of a two-layer graph convolutional network (Kipf and Welling, 2017). In each layer, the subreddit representations $\mathbf{H}^{(d)}$ are updated according to the propagation rule

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

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ is $\mathcal{G}$'s adjacency matrix with added self-loops, $\tilde{\mathbf{D}}$ is the degree matrix of $\tilde{\mathbf{A}}$, and $\mathbf{W}^{(d)}$ is the weight matrix of layer d. σ is the activation function, for which we use a rectified linear unit (Nair and Hinton, 2010) after the first and a linear activation (no non-linearity) after the second layer. We set $\mathbf{H}^{(0)} = \mathbf{\Psi}_l$. In our architecture, $\mathbf{Z} = \mathbf{H}^{(2)}$ is the output of the encoder.

Intuitively, a graph convolution takes embeddings of all neighbors of a subreddit and the embedding of the subreddit itself, transforms them, and accumulates them by a normalized sum. This form of neural message passing (Dai et al., 2016; Gilmer et al., 2017) between neighboring nodes has been shown to be mathematically equivalent to Laplacian smoothing (Li et al., 2018), which is an important property for our architecture: if a concept does not occur in a subreddit, the Laplacian smoothing property of the graph convolutions ensures that the subreddit can still receive a representation by means of message passing.

In the decoder, we compute the reconstructed adjacency matrix, $\hat{\mathbf{A}}$, according to

$$\hat{\mathbf{A}} = \sigma\left(\mathbf{Z}\mathbf{Z}^\top\right),$$

where we use the sigmoid for σ. $\hat{\mathbf{A}}$ is then used to compute a prediction loss, $\mathcal{L}^{(\text{pred})}$.

5.4 Structured Sparsity

Following the SLAP4SLIP framework, we want to reduce the number of concepts in $\mathcal{C}$. In the described architecture, this amounts to reducing the number of columns in $\mathbf{\Psi}_l$. We want to achieve this as part of training, using structured sparsity learning, specifically group lasso regularization (Yuan and Lin, 2006; Jenatton et al., 2011), to set entire columns of the weight matrix $\mathbf{W}^{(0)}$ to zero. Writing $\mathbf{W}^{(0)} = [\mathbf{w}_1^{(0)}, \dots, \mathbf{w}_{|\mathcal{C}|}^{(0)}]$ as a series of column vectors, we define the regularization penalty as

$$\mathcal{L}^{(\text{reg})} = \sum_{j=1}^{|\mathcal{C}|} \|\mathbf{w}_j^{(0)}\|_2.$$

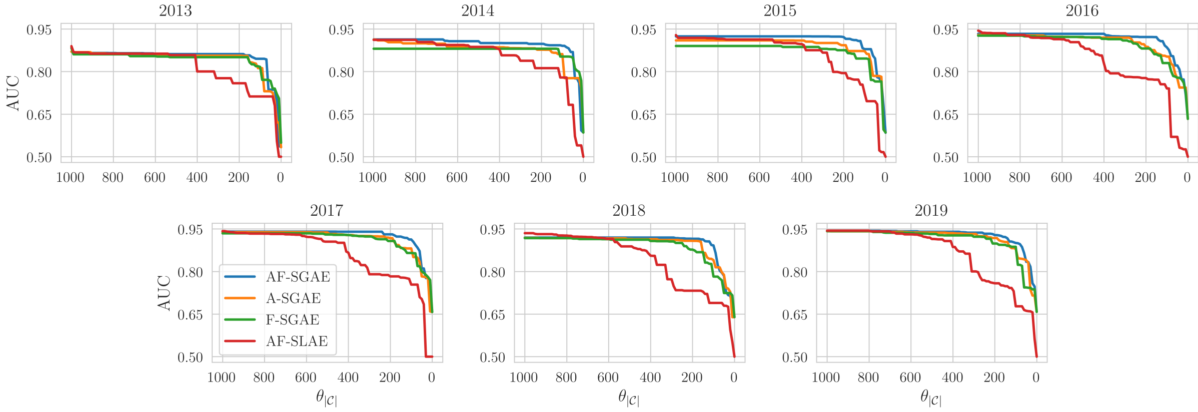


Figure 3: Impact of the sparsity threshold $\theta_{|\mathcal{C}|}$ on model performance. The plots show the performance as measured in AUC on the dev edges. In all years, AF-SGAE performs better than any other model in the sparse regime ($\theta_{|\mathcal{C}|} \leq 200$), showing that it better captures ideological polarization in online discussion forums.

This is a mixed ℓ_1/ℓ_2 regularization (the ℓ_1 norm of the column ℓ_2 norms) that leads to sparsity on the level of columns. When all entries in a column $\mathbf{w}_j^{(0)}$ are zero, this has the effect of removing concept c_j from $\mathcal{C}$. We compute the final loss as

$$\mathcal{L}^{(\text{total})} = \mathcal{L}^{(\text{pred})} + \lambda \mathcal{L}^{(\text{reg})},$$

where $\lambda > 0$ is a hyperparameter controlling the intensity of the ℓ_1/ℓ_2 regularization. Since $\mathcal{L}^{(\text{reg})}$ is non-differentiable, and subgradient methods (Wen et al., 2016) are not guaranteed to lead to sparse solutions (Bach et al., 2011), we use proximal gradient methods (Parikh and Boyd, 2013; Deleu and Bengio, 2021) for optimization.

6 Experiments

6.1 Experimental Setup

For each year, we split $\mathcal{E}$ into 60% training, 20% dev, and 20% test edges. For dev and test, we randomly sample non-edges $(v_i, v_j) \notin \mathcal{E}$ as negative examples such that edges and non-edges are balanced in both sets (50% positive, 50% negative). For training, we sample non-edges in every epoch (i.e., the set of sampled non-edges changes in every epoch). During the test phase, we rank all edges according to their predicted scores.

In this paper, we use sparsity as a hard constraint on the number of concepts with non-zero column weights in $\mathbf{W}^{(0)}$, i.e., we only consider models for which $|\mathcal{C}| \leq \theta_{|\mathcal{C}|}$, where $\theta_{|\mathcal{C}|}$ is the sparsity threshold. We initially set $\theta_{|\mathcal{C}|} = 150$ but later analyze its impact in greater detail.

Since we use both the Adam optimizer (Kingma and Ba, 2015) and proximal gradient descent (Section 5.4), we need to compute the weighted proximal operator of the ℓ_1/ℓ_2 norm, which cannot be evaluated in closed-form in general. We therefore use the approximation based on the Newton-Raphson algorithm recently proposed by Deleu and Bengio (2021). We evaluate the model using average precision (AP) and area under the ROC curve (AUC). AP has been shown to emphasize the correctness of the top-ranked edges (Su et al., 2015) more than AUC. See Appendix A.2 for details about hyperparameters.

We refer to our model as **AF-SGAE** (Agenda Setting/Framing Sparse Graph Auto-encoder).

6.2 Baselines

We compare the main model (AF-SGAE) against three baselines: a model where we only use information from agenda setting, i.e., $\psi_l = a$ (A-SGAE), a model where we only use information from framing, i.e., $\psi_l = f$ (F-SGAE), and a model in which we use information from both agenda setting and framing but replace the graph convolutions with linear layers (AF-SLAE).

6.3 Overall Performance

AF-SGAE clearly outperforms the baseline models, on some years even substantially (Tables 2 and 3). This shows that jointly modeling agenda setting and framing captures ideological polarization on online discussion forums better than only modeling one of the two. Among A-SGAE and F-SGAE, there is no clear winner even though F-SGAE performs slightly better and even beats AF-SGAE in one year. AF-SLAE performs substantially worse than all other models, which indicates that the Laplacian

Year	$\gamma^{(c_j)} = 1$	$\gamma^{(c_j)} = 0\ (m_k)$	$0 < \gamma^{(c_j)} < 1\ (m_k)$
2019	*lefties*	*mainstream* (p/s)	*white* (p/s)
	fascist	*fact* (a/r)	*lies* (h/c)
	donald	*illegal* (a/r)	*women* (h/c)
2018	*free market*	*migration* (i/l)	*firearms* (h/c)
	voter fraud	*sjw* (p/s)	*scotus* (a/r)
	marxist	*peace* (h/c)	*far right* (h/c)

Table 4: Example concepts with different $\gamma^{(c_j)}$ values for the two most recent years in the dataset. For concepts with $\gamma^{(c_j)} > 0$, we also give the foundation with maximum $\beta_k^{(c_j)}$. h/c: harm/care; i/l: ingroup/loyalty; a/r: authority/respect; p/s: purity/sanctity.

smoothing in the form of graph convolutions is a crucial component of the model.

6.4 Quantitative Analysis

How does the sparsity threshold $\theta_{|\mathcal{C}|}$ impact model performance? This question is of theoretical interest since it indicates how many concepts are required to capture the central ideological divides on the social network.

We vary the threshold $0 \leq \theta_{|\mathcal{C}|} \leq 1000$ and measure the performance (AUC) of the four models on the dev set (Figure 3) . First, we find that for the models using graph convolutions (AF-SGAE, A-SGAE, and F-SGAE), reducing $|\mathcal{C}|$ to approximately 200 concepts does not hurt performance. In other words, having more than 200 concepts does not result in better performance. For the model without the graph convolutions (AF-SLAE), on the other hand, performance starts to drop already around 400 concepts. This makes intuitive sense: given that the graph convolutions act as a form of Laplacian smoothing, AF-SLAE needs more concepts to have a reliable feature vector for each subreddit. Second, we observe that the advantage of AF-SGAE does not only lie in its higher performance in the sparse regime but also in its ability to reduce $|\mathcal{C}|$ much further than any of the other models given a performance threshold. E.g., in 2016, the performance of AF-SGAE starts to drop more than 100 concepts later compared to A-SGAE and F-SGAE as we reduce $|\mathcal{C}|$. This again demonstrates that a joint model of agenda setting and framing results in richer information for each concept, making it possible to reduce the number of concepts further than for the other models.

6.5 Qualitative Analysis

The key goal of the SLAP4SLIP framework is to identify the concepts that are most polarized along the dimensions of agenda setting and framing. Here, we analyze which concepts are selected by AF-SGAE, i.e., we take a look at $\mathcal{C}^*$.

We first examine the weight distribution of $\gamma^{(c_j)}$ for all $c_j \in \mathcal{C}^*$. Recall that the magnitude of $\gamma^{(c_j)}$ indicates the importance of agenda setting (versus framing). We notice that $\gamma^{(c_j)}$ tends to be either 0 or 1, i.e., the model makes a clear decision for most concepts whether to use information about agenda setting or framing. Furthermore, for the majority of concepts (approximately 80%) $\gamma^{(c_j)} = 1$, i.e., the model uses more information about agenda setting than framing. Inspecting the concepts with $\gamma^{(c_j)} = 1$ for each year (Table 4), we find that many of them are clearly associated with political ideologies such as names of politicians (e.g., *donald*), designations of ideological orientation (e.g., *marxist*), or words and phrases typical for an ideology (e.g., *voter fraud*). We notice that these patterns are very similar to observations made in previous studies on polarization based on data with ideological labels (Adamic and Glance, 2005; Mejova et al., 2014; Jing and Ahn, 2021). Words with $\gamma^{(c_j)} \neq 1$ (most of which have in fact $\gamma^{c_j} = 0$) tend to be more general concepts such as *peace* and *mainstream* that are not directly connected to a certain ideology. However, taking into account the moral subspace in which the framing of a certain concept is polarized, we find interesting patterns. One of the most polarized concepts 2019, e.g, is *fact* (polarized in the authority/respect subspace), which can be interpreted in the light of post-factual politics (MacMullen, 2020).

We further analyze which moral subspaces are most important for the polarized framing of concepts in general by examining the learned values of $\beta_k^{(c_j)}$ (Section 5.2). We first notice that most concepts have one moral foundation for which $\beta_k^{(c_j)}$ is much larger compared to the other moral foundations (Figure 5). The three moral foundations that most frequently have the highest $\beta_k^{(c_j)}$ value are ingroup/loyalty (30%), purity/sanctity (27%), and authority/respect (21%), followed by harm/care (18%) and fairness/reciprocity (3%). Interestingly, ingroup/loyalty, purity/sanctity, and authority/respect are the three moral foundations on which democrats and republicans exhibit the greatest differences (Haidt and Graham, 2007; Graham et al., 2009), indicating that this ideological divide is also a central axis for the polarized framing of concepts on Reddit.

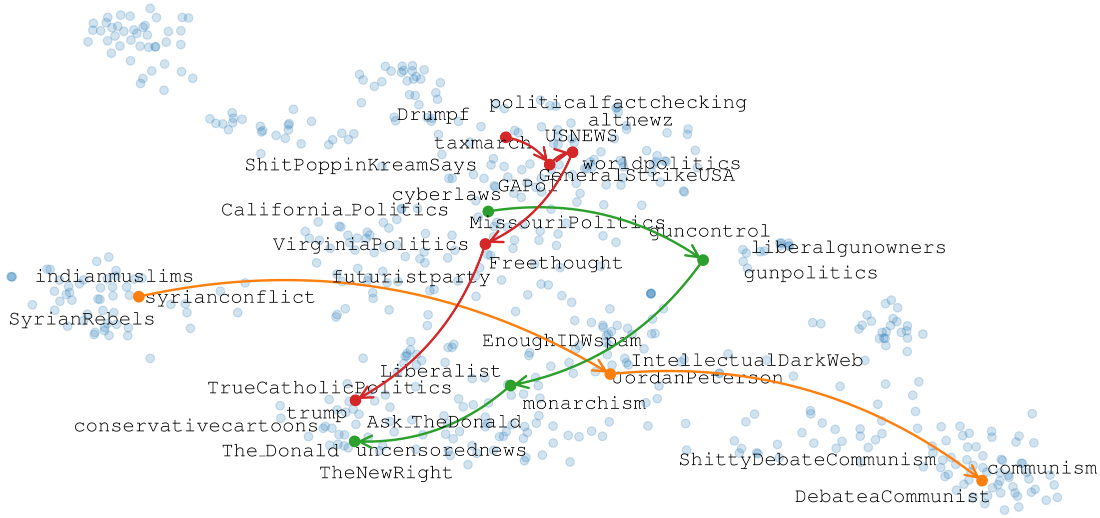


Figure 4: Temporal ideological dynamics in the Reddit politosphere. The figure shows three subreddits that experienced a pronounced shift in their ideology. Orange: `Sino`, a subreddit originally devoted to geopolitical discussion about China that was hijacked by communist users; green and red: `FreeSpeech` and `POLITIC`, two originally moderate subreddits that moved to a more right-wing position in ideology space.

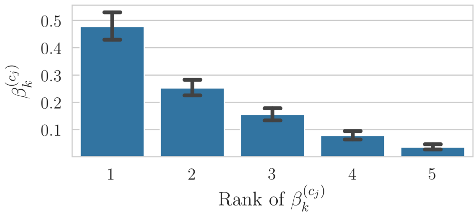


Figure 5: $\beta_k^{(c_j)}$ as a function of rank. The figure shows that for individual concepts, the top-ranked moral foundation typcically has a much larger value of $\beta_k^{(c_j)}$ than moral foundations on lower ranks.

6.6 Ideological Dynamics

The embeddings $\mathbf{Z}$ learned by our model are subreddit representations that combine linguistic information with information from the social network. Here, we analyze what types of temporal ideological dynamics are captured by $\mathbf{Z}$.

We map the embeddings $\mathbf{Z}$ for all years into a common embedding space using orthogonal Procrustes (Schönemann, 1966; Hamilton et al., 2016) and measure for each subreddit the cosine similarities between its embedding in the first year and its embedding in all subsequent years. If the resulting time series of cosine similarities is continuously decreasing, this indicates a change in ideology away from the subreddit's original position. To detect such shifts automatically, we compute for each subreddit Pearson's r between the time series of years and the time series of cosine similarities. Examining the subreddits with the lowest values of r, we observe that most of them experienced a pronounced shift in their ideological orientation over the years (Figure 4). We discern two main patterns: ideological radicalization, where a subreddit is starting at a relatively moderate position in ideology space and then moves into a more extreme (typically right-wing) position (e.g., `FreeSpeech`), and subreddit hijacking, where a subreddit is conquered by users of a certain ideology, resulting in a shift of its position. This is the case for `Sino`, a subreddit originally devoted to geopolitical discussion that was later hijacked by communist users that uncritically praise the Chinese government.

6.7 Limitations

One limitation of our method is that its success depends on how accurately polarization is reflected by the social network structure, which means that care must be taken when selecting the network (for explicit networks) or constructing the network (for implicit networks). For example, on Reddit, user overlap can also be due to conflict between subreddits (Datta et al., 2017; Kumar et al., 2018; Datta and Adar, 2019). While we do not find this to affect our results, it might be a limitation if the degree of homophily in the network is too low.

7 Conclusion

We introduce a novel framework for finding linguistic features maximally informative about the structure of a social network called SLAP4SLIP (Sparse language properties for social link prediction) and show that it can be used to detect polarized concepts in online discussion forums. We model polarization along the dimensions of agenda setting and framing. For framing, we project concept representations into embedding subspaces inspired by moral foundations theory. Our main architecture combines graph neural networks with structured sparsity and learns rich representations for concepts and subreddits. We also release a new dataset of political discourse covering 12 years and more than 600 online groups with different ideologies. We see our study as an exciting first step towards bringing together computational social science research on online polarization, NLP work on political language, and graph-based deep learning.

Acknowledgements

This work was funded by the European Research Council (ERC #740516). We are grateful to Jeremy Frimer for making the MFD 2.0 z-scores available to us. We also thank Xiaowen Dong for extremely helpful feedback.

References

Lada A. Adamic and Natalie Glance. 2005. The political blogosphere and the 2004 U.S. Election: Divided they blog. In *International Workshop on Link Discovery (LinkKDD) 3*.

Jose M. Alvarez and Mathieu Salzmann. 2016. Learning the number of neurons in deep networks. In *Advances in Neural Information Processing Systems (NIPS) 30*.

Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2018. SemAxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. In *Annual Meeting of the Association for Computational Linguistics (ACL) 56*.

Jisun An, Haewoon Kwak, Oliver Posegga, and Andreas Jungherr. 2019. Political discussions in homogeneous and cross-cutting communication spaces. In *International AAAI Conference on Weblogs and Social Media (ICWSM) 13*.

Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. 2011. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106.

Eytan Bakshy, Solomon Messing, and Lada A. Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 384(6239):1130–1132.

Fabian Baumann, Philipp Lorenz-Spreen, Igor M. Sokolov, and Michele Starnini. 2020a. Emergence of polarized ideological opinions in multidimensional topic spaces. In *arXiv 2007.00601*.

Fabian Baumann, Philipp Lorenz-Spreen, Igor M. Sokolov, and Michele Starnini. 2020b. Modeling echo chambers and polarization dynamics in social networks. *Physical Review Letters*, 124(4):048301.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift Reddit dataset. In *arXiv 2001.08435*.

Dimitri P. Bertsekas. 1982. *Constrained optimization and Lagrange multiplier methods*. Academic Press, New York, NY.

Pedro H. Calais, Wagner Meira, Claire Cardie, and Robert Kleinberg. 2013. A measure of polarization on social media networks based on community boundaries. In *International AAAI Conference on Weblogs and Social Media (ICWSM) 7*.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Annual Meeting of the Association for Computational Linguistics (ACL) 53*.

Dallas Card, Justin H. Gross, Amber E. Boydstun, and Noah A. Smith. 2016. Analyzing framing through the casts of characters in the news. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2016*.

Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Goncalves, Alessandro Flammini, and Filippo Menczer. 2011. Political polarization on Twitter. In *International AAAI Conference on Weblogs and Social Media (ICWSM) 5*.

Michele Coscia and Frank M. Neffke. 2017. Network backboning with noisy data. In *IEEE International Conference on Data Engineering (ICDE) 33*.

Hanjun Dai, Bo Dai, and Le Song. 2016. Discriminative embeddings of latent variable models for structured data. In *International Conference on Machine Learning (ICML) 33*.

Srayan Datta and Eytan Adar. 2019. Extracting intercommunity conflicts in reddit. In *International AAAI Conference on Weblogs and Social Media (ICWSM) 13*.

Srayan Datta, Chanda Phelan, and Eytan Adar. 2017. Identifying misaligned inter-group links and communities. *Proceedings of the ACM on Human-Computer Interaction*, 1:1–23.

Marco del Tredici, Diego Marcheggiani, Sabine Im Schulte Walde, and Raquel Fernández. 2019. You shall know a user by the company it keeps: Dynamic representations for social media users in nlp. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2019*.

Tristan Deleu and Yoshua Bengio. 2021. Structured sparsity inducing adaptive optimizers for deep learning. In *arXiv 2102.03869*.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Matthew Gentzkow, Jesse Shapiro, and Dan Jurafsky. 2019. 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*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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*.

Jesse Dodge, Roy Schwartz, Hao Peng, and Noah A. Smith. 2019. RNN architecture learning with sparse regularization. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2019*.

Xiaowen Dong, Dorina Thanou, Laura Toni, Michael Bronstein, and Pascal Frossard. 2020. Graph signal processing for machine learning: A review and new perspectives. *IEEE Signal Processing Magazine*, 37(6):117–127.

Jacob Eisenstein, Noah A. Smith, and Eric Xing. 2011. Discovering sociolinguistic associations with structured sparsity. In *Annual Meeting of the Association for Computational Linguistics (ACL) 49*.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and agenda-setting in Russian news: A computational analysis of intricate political strategies. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2018*.

Anjalie Field and Yulia Tsvetkov. 2019. Entity-centric contextual affective analysis. In *Annual Meeting of the Association for Computational Linguistics (ACL) 57*.

Jeremy Frimer, Jonathan Haidt, Jesse Graham, Morteza Dehghani, and Reihane Boghrati. 2017. Moral foundations dictionaries for linguistic analyses, 2.0.

Dean Fulgoni, Jordan Carpenter, Lyle H. Ungar, and Daniel Preotiuc-Pietro. 2016. An empirical exploration of moral foundations theory in partisan news sources. In *International Conference on Language Resources and Evaluation (LREC) 10*.

David Garcia, Adiya Abisheva, Simon Schweighofer, Uwe Serdült, and Frank Schweitzer. 2015. Ideological and temporal components of network polarization in online political participatory media. *Policy and Internet*, 7(1):46–79.

Kiran Garimella and Ingmar Weber. 2017. A long-term analysis of polarization on Twitter. In *International AAAI Conference on Weblogs and Social Media (ICWSM) 11*.

Justin Garten, Reihane Boghrati, Joe Hoover, Kate Johnson, and Morteza Dehghani. 2016. Morality between the lines: Detecting moral sentiment in text. In *IJCAI Workshop on Computational Modeling of Attitudes (WCMA) 2016*.

Sean Gerrish and David Blei. 2011. Predicting legislative roll calls from text. In *International Conference on Machine Learning (ICML) 28*.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. 2017. Neural message passing for quantum chemistry. In *International Conference on Machine Learning (ICML) 34*.

Jesse Graham, Jonathan Haidt, and Brian A. Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5):1029–1046.

Jon Green, Jared Edgerton, Daniel Naftel, Kelsey Shoub, and Skyler Cranmer. 2020. Elusive consensus: Polarization in elite communication on the COVID-19 pandemic. *Science Advances*, 6:eabc2717.

Ted Grover and Gloria Mark. 2019. Detecting potential warning behaviors of ideological radicalization in an alt-right subreddit. In *International AAAI Conference on Weblogs and Social Media (ICWSM) 13*.

Anna Guimaraes, Oana Balalau, Erisa Terolli, and Gerhard Weikum. 2019. Analyzing the traits and anomalies of political discussions on Reddit. In *International AAAI Conference on Weblogs and Social Media (ICWSM) 13*.

Jonathan Haidt and Jesse Graham. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, 20(1):98–116.

William Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Annual Meeting of the Association for Computational Linguistics (ACL) 54*.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In *Annual Meeting of the Association for Computational Linguistics (ACL) 49*.

Itai Himelboim, Stephen McCreery, and Marc Smith. 2013. 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.

Valentin Hofmann, Janet B. Pierrehumbert, and Hinrich Schütze. 2020. Dynamic contextualized word embeddings. In *arXiv 2010.12684*.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Annual Meeting of the Association for Computational Linguistics (ACL) 52*.

Rodolphe Jenatton, Jean-Yves Audibert, and Francis Bach. 2011. Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research*, 12:2777–2824.

Elise Jing and Yong-Yeol Ahn. 2021. Characterizing partisan political narratives about Covid-19 on Twitter. In *arXiv 2103.06960*.

Kristen Johnson and Dan Goldwasser. 2018. Classification of moral foundations in microblog political discourse. In *Annual Meeting of the Association for Computational Linguistics (ACL) 56*.

Diederik P. Kingma and Jimmy L. Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR) 3*.

Thomas N. Kipf and Max Welling. 2016. Variational graph auto-encoders. In *NIPS Bayesian Deep Learning Workshop*.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR) 5*.

Vivek Kulkarni, Junting Ye, Steven Skiena, and William Y. Wang. 2018. Multi-view models for political ideology detection of news articles. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2018*.

Srijan Kumar, William Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community interaction and conflict on the web. In *The Web Conference (WWW) 27*.

Vadim Lebedev and Victor Lempitsky. 2016. Fast ConvNets using group-wise brain damage. In *Conference on Computer Vision and Pattern Recognition (CVPR) 29*.

Qimai Li, Zhichao Han, and Xiao-Ming Wu. 2018. Deeper insights into graph convolutional networks for semi-supervised learning. In *Conference on Artificial Intelligence (AAAI) 32*.

Wei-Hao Lin, Eric Xing, and Alexander Hauptmann. 2008. A joint topic and perspective model for ideological discourse. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD) 2008*.

Baoyuan Liu, Min Wang, Hassan Foroosh, Marshall Tappen, and Marianna Pensky. 2015. Sparse convolutional neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR) 28*.

Ian MacMullen. 2020. Survey article: What is "post–factual" politics? *Journal of Political Philosophy*, 28(1):97–116.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge University Press, Cambridge, UK.

Nahema Marchal. 2020. The polarizing potential of intergroup affect in online political discussions: Evidence from Reddit r/politics. In *SSRN: 3671497*.

André F. Martins, Noah A. Smith, Pedro M. Aguiar, and Mário A. Figueiredo. 2011. Structured sparsity in structured prediction. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2011*.

Miller McPherson, Lynn Smith-Lovin, and James M. Cook. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444.

Yelena Mejova, Amy X. Zhang, Nicholas Diakopoulos, and Carlos Castillo. 2014. Controversy and sentiment in online news. In *arXiv 1409.8152*.

Julia Mendelsohn, Ceren Budak, and David Jurgens. 2021. 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*.

Pushkar Mishra, Marco del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Abusive language detection with graph convolutional networks. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HTL) 2019*.

Negar Mokhberian, Andrés Abeliuk, Patrick Cummings, and Kristina Lerman. 2020. Moral framing and ideological bias of news. In *International Conference on Social Informatics (SocInfo) 12*.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Alfredo J. Morales, Xiaowen Dong, Yaneer Bar-Yam, and Alex Pentland. 2019. Segregation and polarization in urban areas. *Royal Society Open Science*, 6(10):190573.

Kenton Murray and David Chiang. 2015. Auto-sizing neural networks: With applications to n-gram language models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2015*.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted Boltzmann machines. In *International Conference on Machine Learning (ICML) 27*.

Neal Parikh and Stephen Boyd. 2013. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):123–231.

Hafizh A. Prasetya and Tsuyoshi Murata. 2020. A model of opinion and propagation structure polarization in social media. *Computational Social Networks*, 7(1):1201.

Daniel Preotiuc-Pietro, Ye Liu, Daniel J. Hopkins, and Lyle H. Ungar. 2017. Beyond binary labels: Political ideology prediction of Twitter users. In *Annual Meeting of the Association for Computational Linguistics (ACL) 55*.

Shamik Roy and Dan Goldwasser. 2020. Weakly supervised learning of nuanced frames for analyzing polarization in news media. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2020*.

Eyal Sagi, Daniel Diermeier, and Stefan Kaufmann. 2013. Identifying issue frames in text. *PloS ONE*, 8(7):e69185.

Ville Satopää, Jeannie Albrecht, David Irwin, and Barath Raghavan. 2011. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *International Conference on Distributed Computing Systems (ICDCS) 31*.

Peter H. Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 36(1).

M. Ángeles Serrano, Marián Boguñá, and Alessandro Vespignani. 2009. Extracting the multiscale backbone of complex weighted networks. *PNAS*, 106(16):6483–6488.

Qinlan Shen and Carolyn Rosé. 2019. The discourse of online content moderation: Investigating polarized user responses to changes in Reddit's quarantine policy. In *Workshop on Abusive Language Online 3*.

Yanchuan Sim, Brice D. Acree, Justin H. Gross, and Noah A. Smith. 2013. Measuring ideological proportions in political speeches. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2013*.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Annual Meeting of the Association for Computational Linguistics (ACL) 57*.

Wanhua Su, Yan Yuan, and Mu Zhu. 2015. A relationship between the average precision and the area under the ROC curve. In *International Conference on the Theory of Information Retrieval (ICTIR) 2015*.

Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–288.

Oren Tsur, Dan Calacci, and David Lazer. 2015. A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Annual Meeting of the Association for Computational Linguistics (ACL) 53*.

Aman Tyagi, Anjalie Field, Priyank Lathwal, Yulia Tsvetkov, and Kathleen M. Carley. 2020. A computational analysis of polarization on Indian and Pakistani social media. In *International Conference on Social Informatics (SocInfo) 12*.

Svitlana Volkova, Glen Coppersmith, and Benjamin van Durme. 2014. Inferring user political preferences from streaming communications. In *Annual Meeting of the Association for Computational Linguistics (ACL) 52*.

Vorakit Vorakitphan, Marco Guerini, Elena Cabrio, and Serena Villata. 2020. Regrexit or not regrexit: Aspect-based sentiment analysis in polarized contexts. In *International Conference on Computational Linguistics (COLING) 28*.

Ingmar Weber, Kiran Garimella, and Alaa Batayneh. 2013. Secular vs. islamist polarization in Egypt on Twitter. In *International Conference on Advances in Social Networks Analysis and Mining (ASONAM) 2013*.

Wei Wen, Yiran Chen, Hai Li, Yuxiong He, Samyam Rajbhandari, Minjia Zhang, Wenhan Wang, Fang Liu, and Bin Hu. 2018. Learning intrinsic sparse structures within long short-term memory. In *International Conference on Learning Representations (ICLR) 6*.

Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. 2016. Learning structured sparsity in deep neural networks. In *Advances in Neural Information Processing Systems (NIPS) 30*.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does BERT make any sense? interpretable word sense disambiguation with contextualized embeddings. In *arXiv 1909.10430*.

Yi Yang and Jacob Eisenstein. 2017. Overcoming language variation in sentiment analysis with social attention. *Transactions of the Association for Computational Linguistics*, 5:295–307.

Sarita Yardi and Danah Boyd. 2010. Dynamic debates: An analysis of group polarization over time on Twitter. *Bulletin of Science, Technology & Society*, 30(5):316–327.

Jaehong Yoon and Sung J. Hwang. 2017. Combined group and exclusive sparsity for deep neural networks. In *International Conference on Machine Learning (ICML) 34*.

Ming Yuan and Yi Lin. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, 68(1):49–67.

A Appendices

A.1 Data Preprocessing

We remove comments by deleted users as well as known bots and spammers. We further remove URLs from the comments. We follow Han and Baldwin (2011) in reducing repetitions of more than three letters to three letters.

A.2 Hyperparameters

The specific BERT variant we use to extract the contextualized concept embeddings is $\text{BERT}_{\text{BASE}}$ (uncased) (Devlin et al., 2019).

For all tested models, the input layer has 1000 dimensions (which are sparsified during training), the first hidden layer 100 dimensions, and the second hidden layer 10 dimensions. We perform grid search for the number of epochs $n_e \in \{1, \ldots, 1000\}$, the learning rate $r_l \in \{1 \times 10^{-4}, 3 \times 10^{-4}, 1 \times 10^{-3}, 3 \times 10^{-3}\}$ and the regularization constant $\lambda \in \{1 \times 10^{-4}, 3 \times 10^{-4}, 1 \times 10^{-3}, 3 \times 10^{-3}\}$.

Models are trained with binary cross-entropy as the loss function. Experiments are performed on a GeForce GTX 1080 Ti GPU (11GB).