

# GPolS: A Contextual Graph-Based Language Model for Analyzing Parliamentary Debates and Political Cohesion

Ramit Sawhney

Netaji Subhas Institute of Technology  
ramits.co@nsit.net.in

Arnav Wadhwa

MIDAS, IIIT Delhi  
arnavw96@gmail.com

Shivam Agarwal

Manipal Institute of Technology  
shivamag99@gmail.com

Rajiv Ratn Shah

MIDAS, IIIT Delhi  
rajivrtn@iiitd.ac.in

## Abstract

Parliamentary debates present a valuable language resource for analyzing comprehensive options in electing representatives under a functional, free society. However, the esoteric nature of political speech coupled with non-linguistic aspects such as political cohesion between party members presents a complex and underexplored task of contextual parliamentary debate analysis. We introduce GPolS, a neural model for political speech stance analysis jointly exploiting both semantic language representations and relations between debate transcripts, motions, and political party members. Through experiments on real-world English data, we provide a use case of GPolS as a tool for political speech analysis and polarity prediction.

## 1 Introduction

Politics is broadly defined as the set of activities associated with the governance of a country or a region. It involves various aspects that influence critical decisions having national importance. One such aspect is the conduct of parliamentary debates between political parties having ruling and opposition power. These debates discuss matters affecting the future development of a nation, such as economic and societal growth, policy reforms, and budget revisions. Records of such debates act as a valuable language resource as they provide a wealth of information regarding viewpoints of political representatives over critical societal factors (Abercrombie and Batista-Navarro, 2020b), and also for assessing political candidates and basing voting decisions (Utych, 2019).

Analyzing sentiment in such political discourse is propelled by the developments in fields like behavioral economics that bring the psychological aspects of parliamentary decision-making to the forefront (Rheault, 2016). However, under Parliament’s *Rules of Behavior*,<sup>1</sup> the language used in political debates is complex, laden with domain-specific procedural jargon used in the political realm, and obscure (Abercrombie and Batista-Navarro, 2018a). This esoteric and tedious nature of political debates makes their analysis complex, forming a barrier to ordinary citizen’s insights into political stances and wide-ranging consequences they entail (Edelman, 1985).

The good news is that natural language processing (NLP) shows promise for analyzing voluminous political debates and breaking the understanding barrier towards political ideology to help make informed voting decisions (Davoodi et al., 2020; Edelman et al., 2018). However, conventional language models (Hasan and Ng, 2013) may not generalize well on understanding the obscure linguistic styles of political debates. This complexity arises due to the lack of political context and procedural parliamentary jargon in generic corpora over which traditional text representation models are trained (Pennington et al., 2014; Bojanowski et al., 2017). For instance, in the political context, *Red State* is a state that primarily votes for Republicans, whereas a *Blue State* votes primarily for Democrats. Leveraging the success of unsupervised pre-training in NLP (Devlin et al., 2019; Liu et al., 2019), fine-tuning pre-trained models over the voluminous debate transcripts can lead to drastic advances in analyzing political debates.

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: <http://creativecommons.org/licenses/by/4.0/>.

<sup>1</sup><https://www.parliament.uk/documents/rules-of-behaviour.pdf>

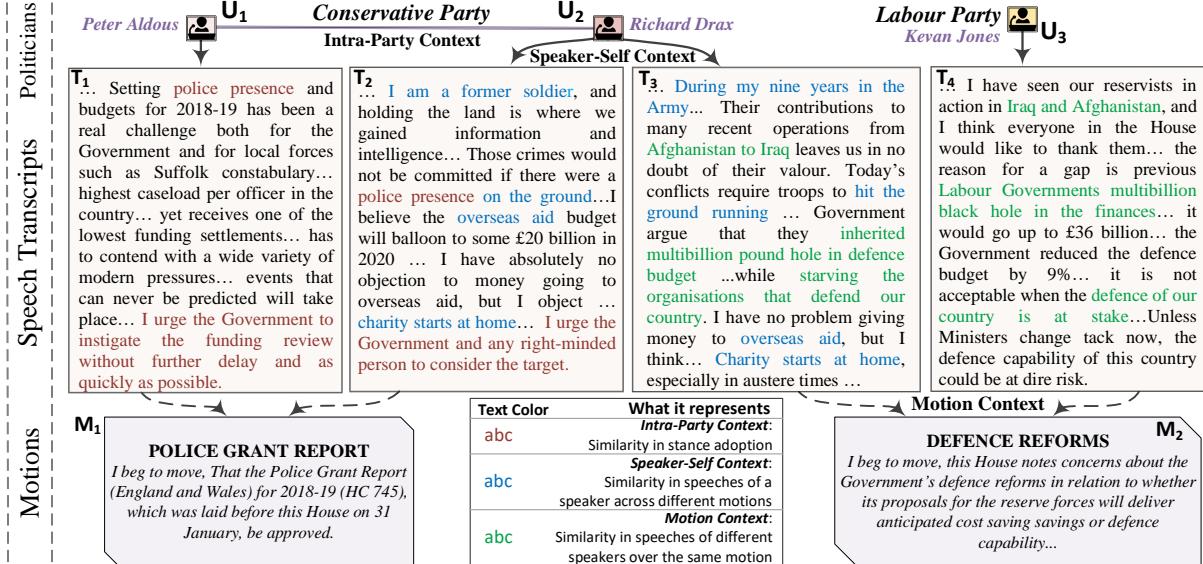


Figure 1: Four debate transcripts from the UK Parliament’s House of Commons, over two different motions. The colors indicate similarity between transcripts based across three different contexts.

The challenging aspect is that analyzing political debates involves multiple contextual elements beyond language, such as political party affiliations and topics of the debates. Consider Figure 1, where we present four speech transcripts over two different motions, from the UK House of Commons. The first two transcripts,  $T_1$  and  $T_2$  over the "Police Grant Report" motion, are from members  $U_1$  and  $U_2$  of the Conservative Party, who express similar viewpoints and support the motion under debate. Such a *intra-party context* and similarity is often indicative and an example of political cohesion within parties (Hug, 2009; Lai et al., 2019). Next, among transcripts  $T_2$  and  $T_3$ , we observe a remarkable similarity between two different transcripts from the same speaker  $U_2$ , debating over two different motions. Such a *speaker-self context* is characteristic of the linguistic styles of individual speakers that reflect across their speeches (WIEBE, 1994b; Cottam et al., 2015). Lastly, transcripts  $T_3$  and  $T_4$  highlight *motion context*, the similarity of speeches based on the motion under debate. The psychological impact of the topic of motion in debate and the stances adopted by peer-speakers, jointly tend to influence the stance a speaker likely has towards the motion. The *underlying connections* between the elements of a debate- such as participants, motions, and speeches; play an essential role in the outcome of the debate. Identifying such similarities in the parliamentary ecosystem unfolds the possibility to learn latent patterns among speakers, the speeches they deliver, their political affiliations, and how they target various motions.

Building on the interdependent nature of participating elements in parliamentary debates, we propose **GPolS**: Graph Political Sentiment analyzer: a neural framework for speech-level stance analysis of members of the parliament (MPs). First, we fine-tune BERT (Devlin et al., 2019), on a large corpus of debate speeches of the UK Parliament’s House of Commons. Through fine-tuning BERT, we extract semantically meaningful representations of political speeches and motions. We empirically validate the presence of similarities between transcripts across debates, motions, and speakers (**Sec. 3.2**). At the heart of GPolS, we model these similarities through relations between elements in a parliamentary setting as a graph. GPolS, through the use of a Graph Attention Network (GAT) (**Sec. 3.3**) aggregates features across the contextual relations between motions, transcripts, and speakers to hierarchically learn similarities across transcripts through semantic (*token-level attention*) and graph based (*graph attention*) representations. Through experiments (**Sec. 4**) on more than 33,000 transcripts from the UK House of Commons, we demonstrate GPolS’s ability for stance analysis in parliamentary debates (**Sec. 5**). Lastly, we visualize GPolS’s graph attention mechanism (**Sec. 5.3**) and token-level attention (**Sec. 5.4**) thus providing a use case for GPolS as a tool for parliamentary debate analysis.

## 2 Related Work

**Politics and Linguistics** Analyzing political data acts as a knowledge source that provides insights into cohesion within political parties, stances of MPs towards critical motions for both the general public, and across domains including humanities, and computational linguistics (Vilares and He, 2017; Sim et al., 2013; Slembrouck, 1992). A developing body of research at the intersection of Politics and Linguistics spans agreement detection (Menini and Tonelli, 2016; M. and M., 2018; Duthie and Budzynska, 2018), emotion analysis (Rheault, 2016; Dzieciatko, 2019), topic-opinion analysis (Nguyen et al., 2015; Abercrombie and Batista-Navarro, 2018b) and, debate stance classification (Proksch et al., 2019). Existing work focuses on these tasks through legislative speeches from the US Congress (Chen et al., 2017), the UK Parliament (Bhavan et al., 2019), and the EU Parliament (Glavaš et al., 2017; Frid-Nielsen, 2018) and through social media such as Twitter (Trilling, 2014; Boutyline and Willer, 2017). Recently, some tools for extracting and annotating political data have also been developed (Haddadan et al., 2019).

**Political Stance and Sentiment Analysis** NLP has seen a growth in analyzing and mining opinions from political discourse (Cabrio and Villata, 2018; Rheault, 2016; Fišer et al., 2020). Word embeddings have shown remarkable progress in analyzing political debates and text (Onyimadu et al., 2013; Abercrombie and Batista-Navarro, 2018a; Rheault and Cochrane, 2020). (Rudkowsky et al., 2018) conducted a textual analysis of Austrian parliamentary speeches, demonstrating the effectiveness of using word embeddings instead of conventional approaches such as Bag-of-Words. More recent approaches (Abercrombie et al., 2019; Abercrombie and Batista-Navarro, 2020a) show the ability of pre-trained transformers such as BERT in capturing domain-specific jargon better for feature extraction from debates. A promising new direction at the intersection of Politics and NLP is the inclusion of context such as political party affiliations and engagement in social circles. (Boutyline and Willer, 2017; Lai et al., 2019) study the linguistic patterns of politically engaged users on Twitter by inferring a user’s political inclination through the politicians and policy nonprofits they follow. They show that more conservative users exhibit a higher level of homophily and often stick to their political stances even when challenged. Recent works (Bhavan et al., 2020; Bhavan et al., 2019) have shown the presence of herd mentality in political stances through graph embeddings by identifying the linguistic similarity between members of the same political party over a set of 1,251 debates. (Davoodi et al., 2020) study the interactions between the content of a proposed bill and the legislative context in which it is presented. (Al Khatib et al., 2020) models debater characteristics to predict persuasiveness. GPolS builds on and differs from existing work across three pivots: 1) we analyze a substantially larger (34,461) set of parliamentary debates for a more meaningful analysis, 2) we analyze and empirically demonstrate political cohesion, and correlations between MP speeches for political stance analysis, and 3) Joint language and contextual modeling of parliamentary debates through token-level and graph attention.

## 3 Methodology

We first define the problem of analyzing parliamentary debates, and then present GPolS. Figure 2 presents an overview of GPolS. We fine-tune and use BERT to encode parliamentary debates and motions. We then motivate GPolS by analyzing the similarity across debate transcripts based on three types of context and model them through a heterogeneous graph. Lastly, we detail the Graph Attention Network (GAT) for propagating contextual information across debate transcripts, speakers and motions for classification.

**Problem Definition** We denote a debate transcript as  $t_i \in T = \{t_1, t_2, \dots, t_N\}$  corresponding to the speech made by a MP (speaker)  $s_j \in S = \{s_1, s_2, \dots, s_Z\}$  on one specific motion  $m_k \in M = \{m_1, m_2, \dots, m_Q\}$ . Each speaker  $s_i$  is affiliated to only one political party  $p_i \in P = \{p_1, p_2, \dots, p_R\}$ . On a debate motion  $m$ , given a transcript  $t$  spoken by a MP  $s$  with a political party affiliation  $p$ , the task is to classify the stance  $Y \in \{‘Aye’, ‘No’\}$  of the MP  $s$  on the motion  $m$  based on the transcript  $t$ . ‘Aye’ and ‘No’ represent a positive and negative stance on the motion, respectively.

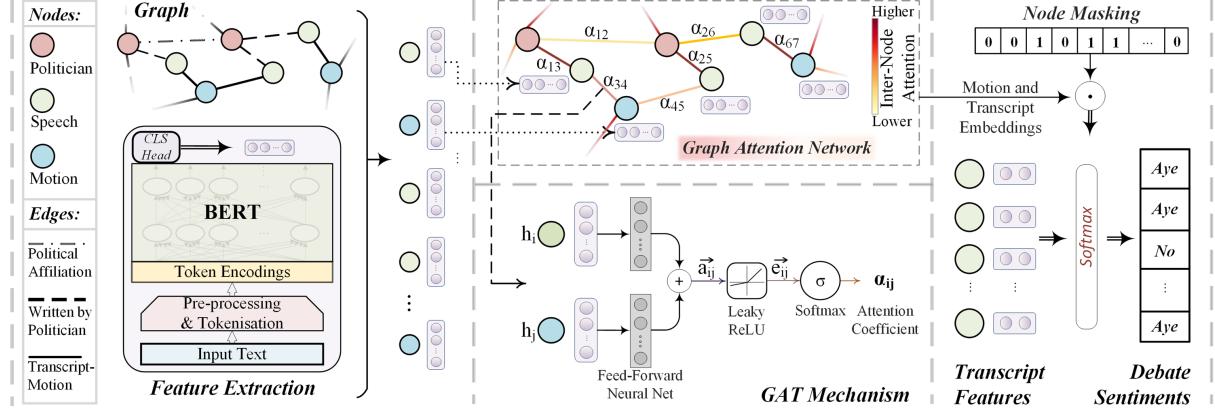


Figure 2: An overview of the proposed context graph, GPolS model and its components.

### 3.1 Encoding Parliamentary Debates and Motions through BERT

A debate comprises *Motions*, i.e., expressions over policy positions taken by the government, Members of Parliament (MP), etc. The motion is followed by complementary or counter-responses from other MPs, often termed as *Utterances*, which collectively comprise a *Speech*. We adopt a transformer language model: Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) to encode debate transcripts  $t$  and motion descriptions  $m$ . We fine-tune the BERT-Base-Cased (Wolf et al., 2019)<sup>2</sup> model architecture. Often, fine-tuning BERT on task-specific corpus yields performance gains on downstream NLP tasks (Alsentzer et al., 2019; Sun et al., 2019). The domain-specific nature of political speeches and complex political jargon motivate us to *fine-tune* the pre-trained BERT language model.

We fine-tune BERT on the ParlVote dataset (Abercrombie and Batista-Navarro, 2020a), a corpus of debates from the UK Parliament’s House of Commons comprising of 33,461 transcripts of debate speeches from May 7th, 1997 to November 5th, 2019.<sup>3</sup> Following (Abercrombie and Batista-Navarro, 2020a), we only consider the first 512 tokens of speeches in ParlVote (18,253 speeches have less than 512 tokens)<sup>4</sup>. All motions are less than 512 tokens, and describe the topic for a given debate. For each transcript  $t \in T$  and, each motion  $m \in M$ , we obtain feature vectors  $h_t = \text{BERT}(t)$  and  $h_m = \text{BERT}(m) \in \mathbb{R}^F$ . The feature vector is a one-dimensional vector of size  $F = 768$  obtained the output of the [CLS] token from the final BERT layer. The pre-processing, training, and hyperparameter tuning are detailed in Section 4.

### 3.2 GPolS: Graph Political Stance analyzer: Context Modeling and Graph Creation

Decision-makers are subtly influenced by the environment around them (Bode et al., 2014). We identify three major types of contexts in parliamentary debates— *intra-party context*, *speaker-self context* and *motion context*. The first, *Intra-Party Context*, captures the influence of the same political affiliations and fellow party members over the speeches of a speaker. During debates, some speakers tend to express their *raw* individual opinions, while some may exhibit *homophily*: the likeliness of associated individuals to adopt similar viewpoints (Boucek, 2002). *Speaker-Self Context* captures the unique linguistic style of a speaker based on the similarity between the speeches they deliver (Johnstone, 2009). Lastly, we also present a *Motion Context*, that captures the relationship between a speech and the motion under debate. Motivated by (Chen et al., 2019a) we use heterogeneous graphs to model such contextual information.

**Hypothesising the presence of contextual information in parliamentary debates** We perform an experiment to examine the similarity between debate transcripts that are related based on their motion, speaker, and speaker affiliation. For this experiment, we compute the cosine similarity between the fine-tuned BERT embeddings of each of the debate transcripts in ParlVote. We then compare the distributions

<sup>2</sup><https://huggingface.co/bert-base-cased>

<sup>3</sup>Available in XML format at <https://www.theyworkforyou.com/> under an Open Parliament License.

<sup>4</sup>We also experiment with LongFormer (Beltagy et al., 2020) to reduce the trimming of tokens in long speeches, but empirically observe no statistically significant difference in performance between BERT and LongFormer over the ParlVote corpus.

Nodes
Speaker (1, 346)
Motion (1, 995)
Transcript (33, 461)
Linkages (Contexts)
Speaker-Speaker (335, 621)
Motion-Transcript (33, 461)
Speaker-Transcript (33, 461)

Table 1: Graph statistics

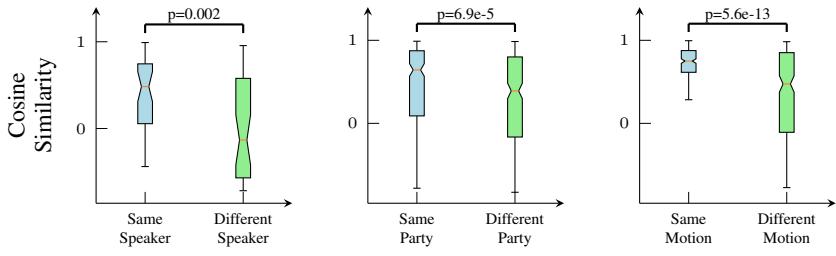


Figure 3: Box plots show the distribution of pairwise cosine similarity between transcripts (based on speakers, speaker affiliation, and motion) with their corresponding confidence intervals (notch).

of cosine similarities between transcript embeddings (BERT) from the same speaker, and those from different speakers. Similarly, we compare cosine similarity distributions across transcripts spoken by members of the same party, and across different parties, as well as across the same motion, and different motions. We present the distribution in Figure 3, and observe that transcripts related to each other (*based on speaker, transcript and motion*) are significantly ( $p < 0.005$ ) more similar than that from different origins under the Wilcoxon signed-rank test (Woolson, 2007). Based on these empirical similarities between debates, we model these contexts in the form of a heterogeneous graph, which we describe next.

**Graph Creation for Context Modeling** We represent the relations between transcripts, speakers and motions in the form of a graph  $G = (V, E)$ .  $V$  and  $E$  represent the nodes and edges in the graph. The nodes  $V$  are consisted of the set of debate transcripts  $T$ , speakers  $S$  and motions  $M$ . The edges  $E$  have three types: Speaker-Transcript edges  $E_{st}$  based on speaker-self context, Speaker-Speaker edges  $E_{ss}$  based on intra-party context, and Transcript-Motion edges  $E_{tm}$  based on motion context.  $G$  is a heterogeneous graph as it has different types of nodes and edges. We now describe these three relations one by one that capture different contexts based on the similarities observed through Figure 3.

- **Speaker-self context** captures the relationship between a speaker and the transcript of their speech in a given debate. Each speech is personalized and reflects the mentality of a speaker across different speeches, motions and times (Wiebe, 1994a; Layman et al., 2006). Contrastingly, self-speaker context can also capture the domain expertise of the speakers. Speaker-self context captures the similarity between speeches by the same speaker. It is formally represented as a Speaker-Transcript edge  $E_{st}$  between a transcript  $t$  and the MP  $s$  whose speech that transcript corresponds to.
- **Intra-party context** models the relationship between a MP and the party they belong to. We build on the hypothesis that speakers are influenced by other party members, and there exists a partisan mentality like political cohesion within parties. (Lai et al., 2019; Owens, 2003; Chartash et al., 2018). Formally, intra-party context is represented by a Speaker-Speaker edge  $E_{ss}$  between two MPs (speakers)  $s_i, s_j \in S$  if both  $s_i$  and  $s_j$  are affiliated to the same political party  $p \in P$ .
- **Motion level context** encodes the relation between a transcript and a motion. In debates, a speech is based on the current motion of discussion. Formally, motion context is represented as a Transcript-Motion edge  $E_{tm}$  between a transcript  $t$  and a motion  $m$  corresponding to that debate transcript.

We summarize the statistics of the created graph  $G$  in Table 1. We now describe the graph attention network for context propagation across nodes and the optimization for political stance analysis.

### 3.3 Graph Attention Network: Semi-Supervised Node Classification for Stance Prediction

We treat stance prediction ('Aye' or 'No') as a node classification problem. As labels are available for a subset of nodes in our graph (i.e., speech nodes (transcripts)), we frame the node classification problem in a semi-supervised learning setting, allowing the model to distribute gradient information from the supervised loss on the labeled transcript nodes. The inclusion of unlabeled motion and speaker

nodes allow us to not only capture the structural traits of relations between similar transcripts based on contextual relations but also the contextual text representation of each motion. As each context has a different degree of influence on a speaker’s speech, it is important that the graph encoding suitably weighs more relevant relations between transcripts, speakers and motions. To this end, we use GATs, that are graph neural networks with node level attention popularly used for node classification (Veličković et al., 2017).

We first describe a *single graph attention layer* (Veličković et al., 2017) which is used throughout the GAT component. The input to this layer is a set of node features  $h = \{x_1, x_2, \dots, x_{|V|}\}, x_i \in \mathbb{R}^F$  where,  $F$  is the number of features for each node. For each debate transcript  $t$ , we set the node feature as the embedding extracted  $h_t$  from the text transcript using BERT. Similarly, for each motion  $m$ , the node feature is set as  $h_m$ . As there is no text information available for speakers  $s \in S$ , we initialize their feature vectors to zero embeddings.<sup>5</sup> The node features are transformed to context dependent features,  $h' = [q_1, q_2, \dots, q_{|V|}]$ ;  $q_i \in \mathbb{R}^{F'}$  based on the influence by its neighbors during training. Following (Veličković et al., 2017) we first apply a shared linear transform parameterized by  $W \in \mathbb{R}^{F' \times F}$  to all the nodes. Then, we apply a shared self-attention mechanism to each node  $i$  in its neighborhood  $\mathcal{N}_i$ . For each node  $j \in \mathcal{N}_i$ , we compute normalized attention coefficients  $\alpha_{ij}$  which shows the importance of context between nodes  $i$  and  $j$ . Formally,  $\alpha_{ij}$  is given as:

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(a_w^T [Wx_i \oplus Wx_j]))}{\sum_{k \in \mathcal{N}_i} \exp(\text{LeakyReLU}(a_w^T [Wx_i \oplus Wx_k]))} \quad (1)$$

where,  $.^T, \oplus$  represent transpose and concatenation, respectively. We use LeakyReLU as the activation function throughout our framework to mitigate vanishing gradients (Maas et al., 2013). Following (Veličković et al., 2017)  $a_w \in \mathbb{R}^{2F'}$  is the parameter matrix of a single layer layer feed forward neural network. The attention coefficients are used to weigh and aggregate contextual information from neighboring nodes. Following (Veličković et al., 2017) we use multi-head attention to stabilize training (Vaswani et al., 2017). Formally,  $U$  independent executors apply the attention mechanism. Their output features are concatenated to yield:

$$q_i = \bigoplus_{k=1}^U \text{LeakyReLU} \left( \sum_{j \in \mathcal{N}_i} \alpha_{ij}^k W^k x_j \right) \quad (2)$$

where  $\alpha_{ij}^k$  and  $W^k$  denote the attention coefficients and linear transform weight matrix computed by the  $k^{th}$  attention head. We apply a 2-layer GAT model, the first layer consists of  $U = 8$  attention heads calculating  $F' = 8$  features per node. The second layer is used for classification with one attention head and  $F' = 2$  ("Aye" or "no") followed by a softmax activation (Nwankpa et al., 2018).

$$y_i = \text{Softmax} \left( \sum_{j \in \mathcal{N}_i} \beta_{ij} W_2 \bigoplus_{k=1}^8 \text{LeakyReLU} \left( \sum_{j \in \mathcal{N}_i} \alpha_{ij}^k W^k x_j \right) \right) \quad (3)$$

where,  $\beta_{ij}$  and  $W_2 \in \mathbb{R}^{2 \times 64}$  denote the attention coefficients and linear transform weight matrix computed by the second attention layer. The entire framework is trained in an end-to-end fashion to minimise the cross entropy loss of labeled nodes using the Adam optimizer (Kingma and Ba, 2014), as:

$$L_{cse} = - \sum_{i=1}^{|V|} Y_i \ln(y_i) + (1 - Y_i) \ln(1 - y_i) \quad (4)$$

where,  $Y_i$  is the true stance and,  $y_i$  is the estimated probability of "Aye" or "No". During training, we only have labeled transcript nodes for those transcripts in the graph that are part of the train set. Speaker and motion nodes are masked so that losses are propagated only for the labeled nodes.

---

<sup>5</sup>It is possible to initialize user features using external information about the politician, however, this is beyond the scope of this work, and forms our future direction. We follow related literature on user profiling (Chen et al., 2019b; Mishra et al., 2019) in heterogeneous graph attention network settings and set user features as zero embeddings.

## 4 Experimental setup

### 4.1 Dataset and Preprocessing

We evaluate GPolS on the ParlVote dataset consisting of 33,461 debate transcripts. On an average, a speech in ParlVote has  $760.2 \pm 901.3$  tokens (*max* 20,730). Following (Abercrombie and Batista-Navarro, 2020a), we remove non-speech elements, tokenize motions and transcripts and, preserve the texts' original casing. The speaker names and party affiliations are obtained using *TheyWorkForYou*.<sup>6</sup> The dataset is fairly balanced with 53.57/46.43% *Aye/No* labels. The transcripts are labeled based on a speaker's vote to their speech, with votes for 'Aye' and 'no' representing positive and negative sentiment. Following (Abercrombie and Batista-Navarro, 2020a), the dataset is divided into 5 subsets for experimenting with various corpus and debate transcript sizes as: (1) *Large* is the complete pre-processed subset. (2) *Medium-any* is a random sample with half (18,253) of the total instances. (3) *Medium* ( $\leq 512$ ) consists of all speeches and motions in *medium-any* with 512 tokens or fewer. (4) *Small-any* is a random sample of the same size as the corpus used by (Abercrombie and Batista-Navarro, 2018a)—1,251 examples. (5) *Small* ( $\leq 512$ ) contains speeches+motions in *small-any* with 512 tokens or fewer.

### 4.2 Training setup

All experiments were performed on an NVIDIA Tesla P100 GPU. For fine-tuning BERT, we explore the following hyperparameters: learning rates  $\in \{2 \cdot 10^{-5}, 5 \cdot 10^{-5}, 2 \cdot 10^{-4}\}$ , batch sizes  $\in \{8, 16, 32\}$ , epochs  $\in \{3, 4, 5, 6, 7, 8, 9, 10\}$ . The optimal hyperparameters for BERT were selected based on validation accuracy, as: learning rate =  $5e-5$ , batch size of 8, with the AdamW (Loshchilov and Hutter, 2019) optimizer, for 7 epochs on the ParlVote dataset. We use the default dropout (Srivastava et al., 2014) rate (0.1) on self attention layers but do not use additional dropout at the top linear layer.

**Training GPolS:** We use grid search for hyperparameter selection for all models over all variants of the dataset individually, and select optimal values based on validation accuracy. All intermediate GAT layers are used with 8 attention heads and an output space of 8 with a dropout=0.6. The Adam optimiser is set with default values  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ ,  $\epsilon = 1e-8$ , weight-decay =  $5e-4$  and an initial learning rate of 0.001. We use a exponential learning rate scheduler with a decay rate of 0.67 (Li and Arora, 2019) and early stopping with a patience of 10 epochs. Following (Abercrombie and Batista-Navarro, 2020a), we evaluated all models using the same randomly selected 80/10/10 training-validation-testing split of the data for each subsection of the corpus, and report mean results over 10 different runs.

### 4.3 Baselines

Following Abercrombie (2020a), we compare GPolS with baselines on classification accuracy.<sup>7</sup>

**Majority class:** The majority class in the training set as the predictions for test set. This baseline does not use any textual or contextual features.

**Support Vector Machines (SVM):** A bag-of-words (BoW) model that uses unigram features as input with term frequency-inverse document frequency feature selection (TF-IDF). We use SVM with a linear kernel, L2 regularisation and optimise squared hinge loss (Gentile and Warmuth, 1999).

**Multi-layer Perceptron (MLP):** A BoW model that utilises only unigram textual features from transcripts as input with TF-IDF selection. We use a MLP with 1 hidden layer containing 100 units followed by ReLU activation. We use L-BFGS optimisation (Liu and Nocedal, 1989) for training 200 epochs

**BERT-MLP:** BERT (Devlin et al., 2019) embeddings are used on the ParlVote dataset followed by a MLP with the same settings as described above. It is a text only model with no additional context.

**Deepwalk:** Concatenates speaker-speaker graph (speaker-self context) embeddings with a set of language features followed by a MLP (same as above). TF-IDF based BoW features (upto trigrams) were concatenated with subjectivity scores for each speech computed using the Harvard General Inquirer lex-

<sup>6</sup>Available at <https://github.com/mysociety/>

<sup>7</sup>We replicated all baselines based on the configurations given in (Abercrombie and Batista-Navarro, 2020a) for the SVM, MLP and BERT-MLP, (Bhavan et al., 2020) for the Deepwalk baseline.

Model	Small - 1,251 transcripts		Medium - 18,253 transcripts		Large - 33,461 transcripts
	Any speech	#Tokens $\leq 512$	Any speech	#Tokens $\leq 512$	All speeches
Majority class	$0.53 \pm 1e-3$	$0.53 \pm 3e-3$	$0.50 \pm 2e-3$	$0.52 \pm 5e-3$	$0.50 \pm 4e-3$
SVM	$0.51 \pm 4e-3$	$0.57 \pm 6e-4$	$0.68 \pm 1e-3$	$0.63 \pm 3e-3$	$0.66 \pm 2e-3$
MLP	$0.50 \pm 2e-3$	$0.56 \pm 1e-3$	$0.63 \pm 4e-3$	$0.63 \pm 8e-4$	$0.65 \pm 3e-3$
BERT + MLP	$0.64 \pm 1e-3$	$0.53 \pm 2e-3$	$0.61 \pm 9e-4$	$0.61 \pm 4e-3$	$0.67 \pm 7e-3$
Deepwalk	$0.73 \pm 4e-3$	$0.73 \pm 3e-3$	$0.72 \pm 9e-4$	$0.72 \pm 1e-3$	$0.72 \pm 8e-4$
GPolS	<b><math>0.80 \pm 5e-4</math></b>	<b><math>0.80 \pm 4e-4</math></b>	<b><math>0.77 \pm 6e-4</math></b>	<b><math>0.77 \pm 5e-4</math></b>	<b><math>0.76 \pm 3e-4</math></b>

Table 2: Classification accuracy averaged over 10 different runs. **Bold** denotes the best results.

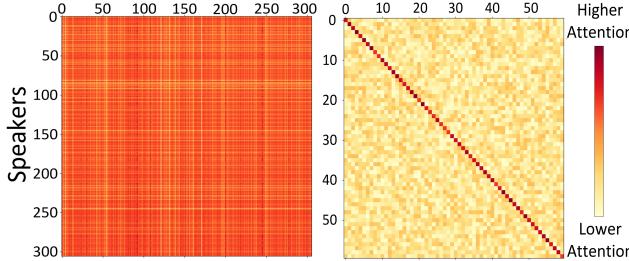


Figure 4: Attention coefficients amongst MPs of Conservative Party (left) and Liberal Democrats (right)

icon (Roberts, 1997). Deepwalk (Perozzi et al., 2014) was used to obtain speaker embeddings from a speaker graph based on party affiliation and concatenated with the text features from transcripts.<sup>8</sup>

## 5 Results and Analysis

### 5.1 Performance Comparison with Baselines

We compare the performance of GPolS with baseline methods in terms of classification accuracy over 10 different runs in Table 2. We note that BERT+MLP significantly ( $p < 0.05$ ) outperforms Majority class and BoW (TF-IDF) based approaches: SVM and MLP. We postulate this to fine-tuning BERT to obtain rich embeddings that better capture the context within each debate transcript. We observe that graph-based models (Deepwalk, GPolS) outperform text-only models (SVM, MLP, BERT-MLP), reiterating the presence of similarity between related transcripts. GPolS outperforms all baselines significantly ( $p < 0.05$ ) under the Wilcoxon signed-rank test, by a large margin greater than 6.5%. We attribute this improvement to two aspects: 1) Fine-tuning BERT for domain-specific embeddings, and 2) Graph attention mechanism in GPolS. First, BERT is able to better model political jargon in debate transcripts post fine-tuning. Second, GPolS enhances text features through a context propagation mechanism via graph attention by modeling speaker-self, intra-party, and motion level context. The additional context that GPolS adds by learning the latent patterns between related transcripts, sets GPolS apart from all the baselines. We further analyze these in the following subsections, first through an ablation study, and then by analyzing BERT’s token-level attention on debates, and GPolS’s graph attention mechanism.

### 5.2 Ablation Study: Quantifying the Impact of Context

We perform an ablation study over the different kinds of context that GPolS models, in Table 3. All performance differences are statistically significant ( $p < 0.05$ ) under Wilcoxon’s signed rank test. We remove the contexts one by one and find that the speaker-self context leads to an improvement owing to the similarity in transcripts by the same speaker. We also note that the accuracy drop on removing motion context is small but statistically significant ( $p < 0.05$ ). The small performance change also shows that while there is a content-based similarity between transcripts of the same motion, such context may not be as relevant from the perspective of stance analysis. We also note that speaker-self and intra-party context

<sup>8</sup>We also experiment with node2vec from (Bhavan et al., 2019), however, as reported in their following work (Bhavan et al., 2020), there is no significant difference between both, so we only report results based on the more recent method for brevity.

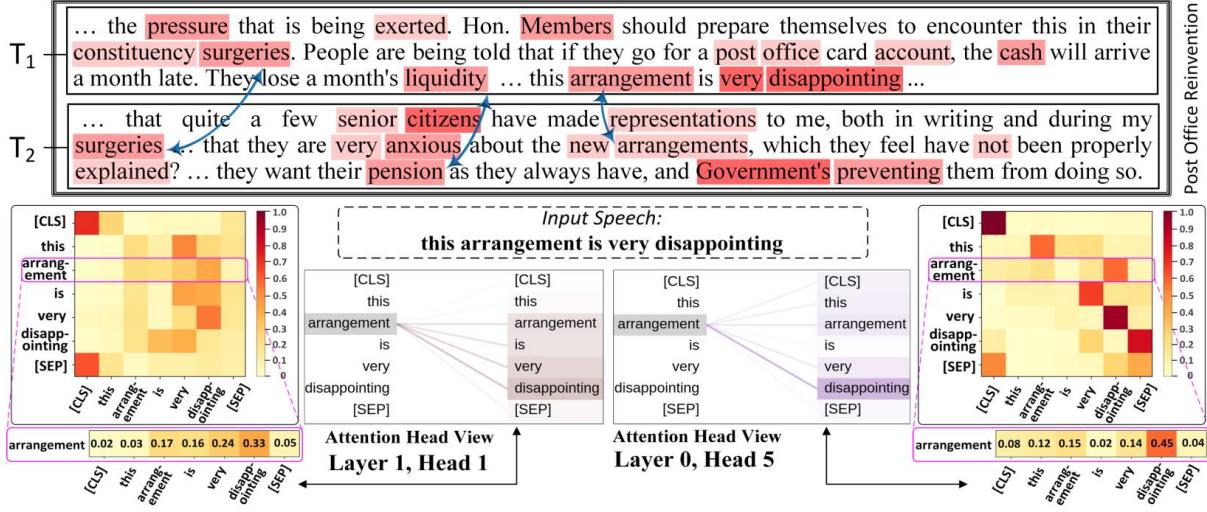


Figure 5: Attention visualisation using BERT. Intense red denotes higher attention. Arrows represent similar terms across transcripts. Heatmaps represent BERT heads that encode semantic links in speech.

are slightly more important as compared to motion-context. This is potentially due to existence of larger political cohesion and partisan mentality in political parties. We analyze this next.

### 5.3 Analyzing Political Cohesion and Partisan Identities: Visualizing GPolS’s Graph Attention

Multiple studies have highlighted the presence of political cohesion, and a partisan-like herd mentality in UK-based political parties (Huddy, 2003; Bowler et al., 1999). To analyze this political cohesion, and GPolS’s intra-party context, we first calculate the attention scores amongst each member in two well-known UK political parties, as shown in Figure. 4. A higher attention score between a speaker and their neighbors indicates a higher degree of peer influence on the speaker and a similarity in their political stances. We first analyze the Conservative party, one of the oldest and largest parties in the UK. We observe in Figure 4 (left) that high attention scores exist between speakers and their peers, indicating a high degree of political cohesion between MPs. High political cohesion in the conservative party is further consolidated by the age and large size of the party, as goals and decisions are much more common between members of such large and old parties (Hayton, 2012). On the other end of the spectrum, we analyze the Liberal Democrats, a relatively newer and smaller party as compared to the Conservative party, in Figure 4 (right). We note high attention scores along the diagonal, indicating a large self-dependency of speakers or a lower degree of political cohesion. Through these examples, we show that GPolS effectively learns and weights such latent patterns through context propagation over the speaker-speaker and speaker-transcript relations. The attention mechanism also provides insights into political cohesion across large volumes of transcripts to aid interpretability in fields such as Digital Humanities.

### 5.4 Token-level Attention Visualisation using BERT

As a concluding use-case, we analyze token-level attention over political debate transcripts. Our goal with this attention analysis is not to study causation or an explanation of GPolS’s predictions, but rather analyzing the impact of fine-tuning BERT on ParlVote. In Figure 5, we present two snippets of real-world debates over the motion: ‘Post Office Reinvention Program’. In  $T_1$ , BERT allots higher attention to terms relevant to motion, such as *post office account*, *cash*, *liquidity*, etc. Similarly in  $T_2$ , BERT allots more attention to the context of *senior citizens* being *very anxious* about the *new arrangements*. BERT also allots high attention to words indicative of stance adoption over motions, such as the *arrangement* being *very disappointing* in  $T_1$ . We present two such heads and analyse their corresponding self-attention heat-maps over an input speech snippet from  $T_1$ . Higher attention scores link the ‘*arrangement*’ as being ‘*very disappointing*’. Notably, the term *surgeries* in  $T_1$  and  $T_2$ , get assigned a high attention weight. In the political domain, surgeries are held by MPs where they meet people and discuss matters of concern.

## 6 Conclusion

We propose GPolS, that enhances linguistic analysis of debates with context for political stance detection. Fine-tuning BERT, we encode debates and motions from the UK Parliament’s House of Commons to capture semantics in political jargon. We show GPolS practical applicability and interpretability in parliamentary debate settings. Through this work, we hope to complement digital humanities, media, social scientists, political linguistics and any members of the public who wish to scrutinize the activities of their elected representatives. GPolS makes a step towards facilitating the understanding of MP’s opinions through analyzing latent correlations and sentiments in voluminous parliamentary debates.

## Acknowledgement

Rajiv Ratn Shah is partly supported by the Infosys Center for AI and the Center of Design and New Media at IIIT Delhi.

## References

- Gavin Abercrombie and Riza Batista-Navarro. 2018a. ‘aye’ or ‘no’? speech-level sentiment analysis of hansard UK parliamentary debate transcripts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).
- Gavin Abercrombie and Riza Theresa Batista-Navarro. 2018b. Identifying opinion-topics and polarity of parliamentary debate motions. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 280–285, Brussels, Belgium, October. Association for Computational Linguistics.
- Gavin Abercrombie and Riza Batista-Navarro. 2020a. ParlVote: A corpus for sentiment analysis of political debates. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5073–5078, Marseille, France, May. European Language Resources Association.
- Gavin Abercrombie and Riza Batista-Navarro. 2020b. Sentiment and position-taking analysis of parliamentary debates: a systematic literature review. *Journal of Computational Social Science*, 3(1):245–270, Apr.
- Gavin Abercrombie, Federico Nanni, Riza Batista-Navarro, and Simone Paolo Ponzetto. 2019. Policy preference detection in parliamentary debate motions. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 249–259, Hong Kong, China, November. Association for Computational Linguistics.
- Khalid Al Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein. 2020. Exploiting personal characteristics of debaters for predicting persuasiveness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7067–7072, Online, July. Association for Computational Linguistics.
- Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly available clinical bert embeddings.
- Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.
- Anjali Bhavan, Rohan Mishra, Pradyumna Prakhar Sinha, Ramit Sawhney, and Rajiv Ratn Shah. 2019. Investigating political herd mentality: A community sentiment based approach. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 281–287, Florence, Italy, July. Association for Computational Linguistics.
- Anjali Bhavan, Mohit Sharma, Ramit Sawhney, and Rajiv Ratn Shah. 2020. Analysis of parliamentary debate transcripts using community-based graphical approaches (student abstract). In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13753–13754. AAAI Press.
- Stefan Bode, Carsten Murawski, Chun Siong Soon, Philipp Bode, Jutta Stahl, and Philip L. Smith. 2014. Demystifying “free will”: The role of contextual information and evidence accumulation for predictive brain activity. *Neuroscience & Biobehavioral Reviews*, 47:636–645, November.

- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Francoise Boucek. 2002. The structure and dynamics of intra-party politics in europe. *Perspectives on European Politics and Society*, 3(3):453–493.
- Andrei Boutyline and Robb Willer. 2017. The social structure of political echo chambers: Variation in ideological homophily in online networks. *Political Psychology*, 38(3):551–569.
- Shaun Bowler, David M Farrell, Richard S Katz, et al. 1999. *Party discipline and parliamentary government*. The Ohio State University Press.
- Elena Cabrio and Serena Villata. 2018. Five years of argument mining: a data-driven analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5427–5433. International Joint Conferences on Artificial Intelligence Organization, 7.
- David Chartash, Nicholas J Caruana, Markus Dickinson, and Laura B Stephenson. 2018. When the team’s jersey is what matters. *Party Politics*, page 135406881879519, August.
- Wei Chen, Xiao Zhang, Tengjiao Wang, Bishan Yang, and Yi Li. 2017. Opinion-aware knowledge graph for political ideology detection. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3647–3653.
- Weijian Chen, Yulong Gu, Zhaochun Ren, Xiangnan He, Hongtao Xie, Tong Guo, Dawei Yin, and Yongdong Zhang. 2019a. Semi-supervised user profiling with heterogeneous graph attention networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 2116–2122. International Joint Conferences on Artificial Intelligence Organization, 7.
- Weijian Chen, Yulong Gu, Zhaochun Ren, Xiangnan He, Hongtao Xie, Tong Guo, Dawei Yin, and Yongdong Zhang. 2019b. Semi-supervised user profiling with heterogeneous graph attention networks. In *IJCAI*, volume 19, pages 2116–2122.
- Martha L Cottam, Elena Mastors, Thomas Preston, and Beth Dietz. 2015. *Introduction to political psychology*. Routledge.
- Maryam Davoodi, Eric Waltenburg, and Dan Goldwasser. 2020. Understanding the language of political agreement and disagreement in legislative texts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5358–5368, Online, July. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.
- Rory Duthie and Katarzyna Budzynska. 2018. A deep modular rnn approach for ethos mining. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4041–4047. International Joint Conferences on Artificial Intelligence Organization, 7.
- Mariusz Dzieciatko. 2019. Application of text analytics to analyze emotions in the speeches. In Ewa Pietka, Paweł Badura, Jacek Kawa, and Wojciech Wieclawek, editors, *Information Technology in Biomedicine*, pages 525–536, Cham. Springer International Publishing.
- Murray Edelman. 1985. Political language and political reality. *PS*, 18(1):10–19.
- Vladimir Eidelman, Anastassia Kornilova, and Daniel Argyle. 2018. How predictable is your state? leveraging lexical and contextual information for predicting legislative floor action at the state level. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 145–160, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.
- Darja Fišer, Maria Eskevich, and Franciska de Jong, editors. 2020. *Proceedings of the Second ParlaCLARIN Workshop*, Marseille, France, May. European Language Resources Association.
- Snorre Sylvester Frid-Nielsen. 2018. Human rights or security? positions on asylum in european parliament speeches. *European Union Politics*, 19(2):344–362.
- Claudio Gentile and Manfred K. K Warmuth. 1999. Linear hinge loss and average margin. In M. J. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems 11*, pages 225–231. MIT Press.

- Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2017. Unsupervised cross-lingual scaling of political texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 688–693, Valencia, Spain, April. Association for Computational Linguistics.
- Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019. Disputool – a tool for the argumentative analysis of political debates. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6524–6526. International Joint Conferences on Artificial Intelligence Organization, 7.
- Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1348–1356, Nagoya, Japan, October. Asian Federation of Natural Language Processing.
- Richard Hayton. 2012. *Reconstructing Conservatism?: The Conservative party in opposition, 1997–2010*.
- Leonie Huddy. 2003. Group identity and political cohesion.
- Simon Hug. 2009. Selection effects in roll call votes. *British Journal of Political Science*, 40(1):225–235, October.
- Barbara Johnstone. 2009. Stance, style, and the linguistic individual. *Stance: sociolinguistic perspectives*, 29:52.
- Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.
- Mirko Lai, Marcella Tambuscio, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. 2019. Stance polarity in political debates: A diachronic perspective of network homophily and conversations on twitter. *Data & Knowledge Engineering*, 124:101738.
- Geoffrey C. Layman, Thomas M. Carsey, and Juliana Menasce Horowitz. 2006. PARTY POLARIZATION IN AMERICAN POLITICS: Characteristics, causes, and consequences. *Annual Review of Political Science*, 9(1):83–110, June.
- Zhiyuan Li and Sanjeev Arora. 2019. An exponential learning rate schedule for deep learning.
- Dong C. Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1-3):503–528, August.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.
- Ahmadalinezhad M. and Makrehchi M. 2018. Detecting agreement and disagreement in political debates. In: Thomson R., Dancy C., Hyder A., Bisgin H. (eds) *Social, Cultural, and Behavioral Modeling. SBP-BRIMS*, 2018.
- Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3.
- Stefano Menini and Sara Tonelli. 2016. Agreement and disagreement: Comparison of points of view in the political domain. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2461–2470, Osaka, Japan, December. The COLING 2016 Organizing Committee.
- Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Abusive language detection with graph convolutional networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2145–2150.
- Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, and Kristina Miler. 2015. Tea party in the house: A hierarchical ideal point topic model and its application to republican legislators in the 112th congress. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1438–1448, Beijing, China, July. Association for Computational Linguistics.
- Chigozie Nwankpa, Winifred Ijomah, Anthony Gachagan, and Stephen Marshall. 2018. Activation functions: Comparison of trends in practice and research for deep learning.

- Obinna Onyimadu, Keiichi Nakata, Tony Wilson, David Macken, and Kecheng Liu. 2013. Towards sentiment analysis on parliamentary debates in hansard. In *Revised Selected Papers of the Third Joint International Conference on Semantic Technology - Volume 8388*, JIST 2013, page 48–50, Berlin, Heidelberg. Springer-Verlag.
- John E. Owens. 2003. Part 1: Cohesion. *The Journal of Legislative Studies*, 9(4):12–40, December.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*.
- Sven-Oliver Proksch, Will Lowe, Jens Wäckerle, and Stuart Soroka. 2019. Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches. *Legislative Studies Quarterly*, 44(1):97–131.
- Ludovic Rheault and Christopher Cochrane. 2020. Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis*, 28(1):112–133.
- Ludovic Rheault. 2016. Expressions of anxiety in political texts. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 92–101, Austin, Texas, November. Association for Computational Linguistics.
- Carl W Roberts. 1997. *Text analysis for the social sciences : methods for drawing statistical inferences from texts and transcripts*. Mahwah NJ. : Erlbaum. Includes bibliographical references and indexes.
- Elena Rudkowsky, Martin Haselmayer, Matthias Wastian, Marcelo Jenny, Štefan Emrich, and Michael Sedlmair. 2018. More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures*, 12(2-3):140–157.
- Yanchuan Sim, Brice DL Acree, Justin H Gross, and Noah A Smith. 2013. Measuring ideological proportions in political speeches. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 91–101.
- Stef Slembrouck. 1992. The parliamentary hansard ‘verbatim’ report: the written construction of spoken discourse. *Language and literature*, 1(2):101–119.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, January.
- Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? *China National Conference on Chinese Computational Linguistics*, pages 194–206.
- Damian Trilling. 2014. Two different debates? investigating the relationship between a political debate on TV and simultaneous comments on twitter. *Social Science Computer Review*, 33(3):259–276, July.
- Stephen M. Uttych. 2019. Speaking style and candidate evaluations. *Politics, Groups, and Identities*, 0(0):1–19.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.
- Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2017. Graph attention networks.
- David Vilares and Yulan He. 2017. Detecting perspectives in political debates. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 1573–1582.
- Janyce M. Wiebe. 1994a. Tracking point of view in narrative. *Comput. Linguist.*, 20(2):233–287, June.
- JM WIEBE. 1994b. Tracking point of view in narrative. *Computational linguistics-Association for Computational Linguistics (Print)*, 20(2):233–287.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrick Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.
- RF Woolson. 2007. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials*, pages 1–3.