RooseBERT: A New Deal For Political Language Modelling

Anonymous ACL submission

Abstract

The increasing amount of political debates and politics-related discussions calls for the definition of novel computational methods to automatically analyse such content with the final goal of lightening up political deliberation to citizens. However, the specificity of the political language and the argumentative form of these debates (employing hidden communication strategies and leveraging implicit arguments) make this task very challenging, even for current general-purpose pre-trained Language Models. To address this issue, we introduce a novel pre-trained Language Model for political discourse language called RooseBERT. Pre-training a language model on a specialised domain presents different technical and linguistic challenges, requiring extensive computational resources and large-scale data. RooseBERT has been trained on large political debate and speech corpora (8K debates, each composed of several sub-debates on different topics) in English. To evaluate its performances, we fine-tuned it on four downstream tasks related to political debate analysis, i.e., stance detection, sentiment analysis, argument component detection and classification, and argument relation prediction and classification. Our results demonstrate significant improvements over general-purpose Language Models on these four tasks, highlighting how domain-specific pre-training enhances performance in political debate analysis. We release RooseBERT for the research community.

1 Introduction

Political debates, politics-related (online) discussions and opinions represent key elements of public discourse and deliberation, and they are fundamental for citizens to take informed decisions. The ability to automatically analyse such a kind of content is therefore central in Artificial Intelligence, where different aspects of political discourse have been investigated, e.g., performance evaluation (Liu et al., 2024), argumentative structures (Haddadan et al., 2019), sentiment analysis (Abercrombie and Batista-Navarro, 2020), and logical fallacies (Goffredo et al., 2022). These approaches rely on pre-trained Language Models (PLMs) originally developed for general-purpose corpora, e.g., BERT (Devlin et al., 2019). Although these models achieve good results in some of these tasks, they underperform when the goal is to capture the unique features of political content (e.g., domain-specific terminology, implicit argumentation, strategic communication patterns). In the literature, researchers have shown that domain-specific variations in vocabulary and semantics can significantly affect model performance (Nooralahzadeh et al., 2018), avoiding vocabulary mismatches and encoding semantic shifts (Lee et al., 2020).

To address this issue, in this paper, we introduce RooseBERT, a domain-specific language model pre-trained on English political debates. Built on the BERT architecture, widely adopted for domain adaptation tasks (Beltagy et al., 2019; Huang et al., 2019; Lee et al., 2020; Yang et al., 2020; Kawintiranon and Singh, 2022; Yang et al., 2023; Wada et al., 2024; Zeinali et al., 2024), RooseBERT is tailored to capture the distinctive features of political debates. Our aim is to improve the performance of state-of-the-art models on tasks such as sentiment analysis, stance detection, and argument mining which are especially relevant to political analysis. Encoder-based architectures like BERT are especially well-suited for such tasks due to their strength in encoding semantic relationships and contextual information and have proven strong performance on a range of downstream tasks relevant to political discourse analysis (Abercrombie and Batista-Navarro, 2020; Liu et al., 2021; Goffredo et al., 2023a).

The main contributions in this paper are:

- We train and release RooseBERT, and its train-

ing dataset composed of 8K English political debates ($\sim$ 5GB).

- We extensively evaluate the model across four downstream tasks in political discourse analysis: Stance Detection, Sentiment Analysis, Argument Component Detection and Classification, and Argument Relation Prediction and Classification. We compare RooseBERT's performance with both general-purpose models and other domain-specific PLMs, demonstrating its effectiveness.
- We show that RooseBERT generalises well across different types of political debates from various geopolitical contexts, including televised U.S. presidential debates and U.K. parliamentary discussions.

By providing a ready-to-use, domain-specific language model, we aim to boost research on this sensible topic, while alleviating the computational costs of training such models from scratch.

2 Related Work

Word embeddings are the foundation of most NLP systems, encoding the semantic and syntactic properties that downstream tasks rely on. BERT (Devlin et al., 2019) leverages a transformer architecture to produce highly context-sensitive embeddings. It was trained using two different objective learning functions, Masked Language Modelling (MLM) and Next Sentence Prediction (NSP), with the ultimate goal of learning bidirectional contextual representations and capturing sentence-level relationships. BERT was trained on a large general domain corpora (BooksCorpus and English Wikipedia) and it struggles to capture semantics related to specialised domains. To address this issue, other models have been proposed by pre-training BERT on specific domains, like ClinicalBERT for medical applications (Huang et al., 2019), SciBERT for scientific texts (Beltagy et al., 2019), BioBERT for the biomedical domain (Lee et al., 2020), FinBERT for finance (Yang et al., 2020), BERTweet for Twitter data (Nguyen et al., 2020) and Mol-BERT (Li and Jiang, 2021) for molecular representation.

In the political domain, well-constructed embeddings are particularly important for capturing domain-specific terminology, and different PLMs have been proposed for use in this context. Yang et al. (2023) introduced ConfliBERT, a domain-specific PLM for conflict and political violence. ConfliBERT has been trained using 12 datasets and has been evaluated on 18 tasks on which the authors show that the improvement of ConfliBERT is statistically significant in all tasks but three. Kawintiranon and Singh (2022) presented PoliBERTweet, a PLM trained from BERTweet (Nguyen et al., 2020) on 83M US 2020 election-related English tweets. Their model outperforms TweetEval (Barbieri et al., 2020) and BERTweet (Nguyen et al., 2020) on all tasks related to the political domain than general purpose models. While ConfliBERT and PoliBERTweet are valuable political-domain models, they are not tailored to the dialogic and argumentative nature of political debates. ConfliBERT focuses on political violence in parliamentary records, while PoliBERTweet targets electoral content on Twitter. Both models lack the turn-taking, rebuttals, and rhetorical complexity central to debates.

This motivates the introduction of RooseBERT, which to the best of our knowledge, is the first PLM specifically trained on full debate transcripts (presidential and parliamentary), capturing cross-speaker interactions, persuasive framing, and argument structures. Its debate-specific corpus and cross-national coverage enable stronger performance on tasks like argument mining and sentiment analysis, where the debate context is crucial.

3 Methodology

In this section, we first describe the datasets used for pre-training RooseBERT, then we present our pre-training strategy, followed by the description of the downstream tasks and the metrics applied to assess the model's performance.

3.1 Corpora

To pre-train RooseBERT on English political debates, we assembled a comprehensive corpus of debate transcripts and related material of approximately 5GB. Our complete dataset combines existing curated collections and also transcripts scraped from official debate websites. This approach ensures broad coverage of formats (televised exchanges, parliamentary debates, primary debates) on a comprehensive time period (1946–2024), capturing the evolution of rhetorical strategies and issue framing. All the datasets originate from authoritative speakers and official political settings (i.e., presidential debates and parliamentary sessions), where political decisions are discussed, contested, and communicated to the public. Each dataset was

pre-processed to retain only the relevant debate content, with extraneous elements such as hyperlinks, markup tags, and unrelated text removed to reduce noise. The datasets are:

US Presidential Debates *The American Presidency Project* [1] is an authoritative on-line source for presidential public documents. The website includes a list of presidential candidate debates from 1960-2024, including primary and general elections political debates.

United Nations General Debate Corpus (UNGDC) is a collection of the UN General Assembly debates (1946- 2023) with over 10k speeches from 202 countries (Baturo et al., 2017).

The House of Commons Parliamentary Debates contains data on all parliamentary debates held in the House of Commons between 1979 and 2019. In total, the data includes over 2.5 million speeches drawn from over 50 thousand parliamentary debates (Blumenau, 2021).

The Australian Parliamentary Debates corpus (Katz and Alexander, 2023) contains data on the proceedings from each sitting day in the Australian Parliament by the House of Representatives from March 1998 to September 2022.

The EUSpeech (Schumacher et al., 2016) is a dataset of 18,403 speeches from EU leaders (i.e., heads of government in 10 member states, EU commissioners, party leaders in the European Parliament, and European Central Bank and International Monetary Fund leaders) from 2007 to 2015.

The ParlEE Plenary Speeches dataset (Sylvester et al., 2022) is a collection of speeches from eight legislative European chambers (Austria, Belgium, France, Germany, Hungary, Ireland, Portugal and the United Kingdom) covering 2009-2019. Only Irish legislative chamber speeches were kept.

The ParlScot corpus (Braby and Stewart, 2021) contains full-text vectors for 1.8 million spoken contributions for the Scottish Parliament until 2021 (speeches in both parliament and in committees).

The UN Security Council Debates corpus (Schoenfeld et al., 2019) is a dataset of UN Security Council debates between January 1992 and December 2023. The corpus contains 106,302 speeches extracted from 6,233 meeting protocols.

[1] www.presidency.ucsb.edu

3.2 Pre-training Approaches

To effectively pre-train a language model, different approaches can be used (Chalkidis et al., 2020; Yang et al., 2023). One approach is to continue the pre-training (CONT) of the original model (BERT) by initialising the model with the original weights and vocabulary and training for additional steps on the domain-specific corpus. The second approach (SCR) consists in training BERT from scratch, starting with a random initialisation of its weights and creating a custom vocabulary using the domain-specific corpus.

RooseBERT was trained with both approaches (CONT and SCR), using *cased* and *uncased* vocabularies. When training the SCR version of RooseBERT, a custom *WordPiece Tokenizer* trained on the same training corpus of English political debates was used. Different studies show that SCR models gain substantial improvement in performance over CONT models (Beltagy et al., 2019; Gu et al., 2022). This performance gain is largely attributed to the use of a tokenizer with a domain-specific vocabulary, which can encode entire specialised terms as single tokens. In contrast, tokenizers based on general-purpose vocabularies tend to break these terms into multiple sub-tokens, potentially weakening the model's understanding of domain-specific content. For example, *RooseBERT-scr-uncased*'s vocabulary shares only 56.32% of its tokens with *bert-base-uncased*'s vocabulary. Words such as *deterrent*, *endorse*, and *bureaucrat* are represented as single tokens in RooseBERT's custom vocabulary, while they are absent from BERT's vocabulary and thus would be split into sub-tokens (see Table 1).

Word	BERT vocabulary	RooseBERT vocabulary
annuity	['an', '##nu', '##ity']	annuity
bureaucrat	['bureau', '##crat']	bureaucrat
deterrent	['de', '##ter', '##rent']	deterrent
undermining	['under', '##mini', '##ng']	undermining
endorse	['end', '##ors', '##e']	endorse
statutorily	['s', '##tat', '##uto', '##rily']	statutorily
consequential	['con', '##se', '##quential']	consequential
importer	['import', '##er']	importer

Table 1: Example of key political terms and their encoding in BERT and RooseBERT vocabularies.

3.3 Evaluation Tasks

To evaluate the performance of RooseBERT, we selected four downstream tasks: Stance Detection, Sentiment Analysis, Argument Component Detection and Classification, and Argument Relation Prediction and Classification. These tasks are central to NLP research on political debates, and publicly available annotated datasets enable systematic model evaluation. However, due to the limited availability of datasets focused specifically on political debates, we also included a dataset (for stance detection) that is not political-oriented. Accuracy was used as the evaluation metric for Sentiment Analysis, while Average Macro F1 was employed for the remaining tasks, in line with the metrics adopted by the original dataset authors.

Sentiment Analysis Sentiment analysis is the task which aims to determine whether the polarity of a text is positive or negative. *ParlVote* (Abercrombie and Batista-Navarro, 2020) is a large annotated corpus of parliamentary debates annotated for the sentiment analysis task. This corpus consists of 34,010 UK parliamentary debates (17,721 annotated as positive, 15,740 as negative) labelled at the speech level.

Stance Detection Stance detection is the task of automatically identifying positions on a specific topic in text (Mohammad et al., 2017). The VAST dataset (Allaway and McKeown, 2020) is an annotated dataset collected from The New York Times Room for Debate section, part of the Argument Reasoning Comprehension (ARC) Corpus (Habernal et al., 2017). The VAST dataset comprises 23,525 sentences across 5,634 topics, each annotated as *pro* (supporting the topic), *con* (opposing the topic), or *neutral*.

Argument Component Detection and Classification In *Argument Mining* (AM), the detection and classification of argument components are challenging tasks that involve extracting argumentative units from natural language text by detecting their boundaries and classifying them as either *premises* or *claims* (Lawrence and Reed, 2019). In the literature, the most comprehensive resource for political debates annotated with both *claims* and *premises* is the *ElecDeb60to20* dataset (Goffredo et al., 2023b), a collection of 44 televised U.S. presidential debates from 1960 to 2020. The dataset comprises 38,667 argument components (25,078 claims and 13,589 premises).

Argument Relation Prediction and Classification In AM, argumentative components, i.e., *premises* and *claims*, can either *support* or *attack* other argument components, forming complex relational structures that are often represented as argumentation graphs (Lawrence and Reed, 2019). Identifying these relations is essential for understanding the argumentative structure of a debate. *ElecDeb60to20* provides annotations for argument relations. More precisely, 21,689 relations are annotated as support and 3,835 as attack.

4 Experimental Setup

In this section, we detail the design choice and technical elements necessary to train RooseBERT.

4.1 Pre-training Setup

RooseBERT was trained using both CONT and SCR approaches, in both *cased* and *uncased* versions. The architecture is the same as BERT-Base with 12 layers, 768 hidden units and 12 attention heads. See Appendix A for the *large* model training.

Initially, the models were trained to optimise both the Masked Language Modeling (MLM), with a masking probability of 15%, and the Next Sentence Prediction (NSP) objectives. However, as noted in several studies (Liu et al., 2019; Joshi et al., 2020; Yang et al., 2023), incorporating NSP alongside MLM did not lead to significant performance gains compared to training with MLM alone and it only slowed down the training. Therefore, we chose to discard the NSP objective in our training setup. Each pre-training dataset was split into 95% training and 5% testing, with the test set also reserved for perplexity evaluation. These splits were merged to form the final dataset, ensuring balanced representation across sources.

We experimented with a range of hyperparameters to identify the configuration that most effectively minimised the perplexity and the loss on the testing dataset. Specifically, we tested learning rates of $\{1, 3, 5\} \times e^{-4}$ and $\{1, 3, 5\} \times e^{-5}$. The learning rate was linearly increased (warmed up) over the first 10,000 steps, followed by a linear decay for the remainder of training. Once the learning rate was fixed, we experimented with varying number of training steps $\in$ {50K , 100K , 150K , 200K , 250K , 300K}. Following Devlin et al. (2019), each model was trained for 80% of the steps with a maximum sequence length of 128 and with a maximum sequence length of 512 for the remaining 20%. For

	BERT-base	ConfliBERT-cont	ConfliBERT-scr	PoliBERTweet	RooseBERT-cont	RooseBERT-scr
cased	25.281	4.095	4.115	156.390	**2.487**	2.710
uncased	9.295	4.662	4.095	N/A	**2.566**	2.747

Table 2: Perplexity evaluated on an held out corpus representing the overall distribution of political debates.

each training we used eight A100 GPUs and the Transformer Library of HuggingFace (Wolf et al., 2019) with Python 3.10. Each training run was optimised using DeepSpeed (Rasley et al., 2020) ZeRO-2 methods, gradient accumulation, and FP16 precision to reduce training time.

4.2 Fine-tuning Setup

Each PLM was fine-tuned on the four downstream tasks described in Section 3. Following the approach of Devlin et al. (2019), we experimented with various hyper-parameters: learning rates of $\{2, 3, 5\} \times e^{-5}$; batch sizes of 8, 16, and 32 and number of epochs of 2, 3, or 4. For each model, we used its maximum available sequence length. For all tasks, we used the original train, development and testing splits provided with each dataset.

The performances on the downstream tasks of RooseBERT were compared to the performances of BERT-base (used as a baseline) and of similar political PLMs, ConfliBERT and PoliBERTweet, each fine-tuned on the downstream tasks with the same hyper-parameters as RooseBERT.

For each run, we used one A100 GPU and the HuggingFace Framework with Python 3.10.

5 Evaluation

In this section, we detail the results of the pre-training and fine-tuning processes.

5.1 Pre-Training

Perplexity (PPL) (Jelinek et al., 2005) was used as a metric to determine the best performing pre-trained model given the hyper-parameters previously described. Perplexity is defined as the exponentiated average negative log-likelihood of a sequence $X = (x_0, x_1, x_2..x_t)$ as follows: $PPL(X) = exp\{-\frac{1}{t}\sum_i^t \; log \; p_\theta(x_i|x_{<i})\}$. Perplexity was used to evaluate the model's ability to predict masked tokens on an unseen test dataset representative of the overall corpus distribution. A lower perplexity indicates that the model assigns higher probabilities to the correct next words, i.e., it is more confident in its predictions, less *confused* by the text.

Table 2 reports the perplexity scores achieved by each model. For conciseness, we only report the best-performing configurations for RooseBERT. Specifically, for both RooseBERT-cont and RooseBERT-scr, a learning rate of $1e^{-4}$ consistently yielded the lowest perplexity on an held out portion of the testing dataset (see Section 3). Final evaluation loss also confirmed this trend. For the cased model, the lowest loss was 0.899 with $1e^{-4}$, followed by 0.901 with $3e^{-4}$, while all other rates (i.e., $5e^{-4}$, $\{1, 3, 5\} \times e^{-5}$) resulted in losses exceeding 0.925. For the uncased model, the lowest loss was 0.924 with $1e^{-4}$, increasing to 0.933 with $3e^{-4}$, and surpassing 0.951 for the remaining learning rates.

As shown in Table 2, BERT exhibits a high perplexity on a held-out portion of the dataset, underscoring the challenge that general-purpose models face in capturing the distribution of domain-specific data without further adaptation. In contrast, ConfliBERT and PoliBERTweet achieve lower perplexity scores, with ConfliBERT yielding results most closely aligned with RooseBERT. Among all models, RooseBERT demonstrates the lowest perplexity, indicating a more effective adaptation to the political debate domain. Notably, the CONT version of RooseBERT outperforms the SCR variant, suggesting that starting from a general-domain foundation may lead to better modelling of in-domain language patterns. Table 3 shows the hyper-parameters for the best CONT and SCR RooseBERT models.

Hyper-parameters	RooseBERT-cont	RooseBERT-scr
Warmup Steps	10,000	10,000
Total Training Steps	150K steps	250K steps
Total Batch Size	2048	2048
Learning Rate	$1e^{-4}$	$1e^{-4}$
Weight Decay	0.01	0.01
Adam β_1, β_2, ϵ	0.9, 0.98, $1e^{-6}$	0.9, 0.98, $1e^{-6}$
Vocabulary Type	BERT	Custom
Vocabulary Size	28,997 (*cased*) 30,523 (*uncased*)	28,996 (*cased*) 30,522 (*uncased*)

Table 3: Hyper-parameters used to pre-train RooseBERT with the *base* architecture setting.

In terms of training time, training the CONT models required 18 hours for 150k steps. Training the

SCR model as considerably more costly, with a total training time of 31 hours and 250k steps.

5.2 Fine-Tuning

BERT, RooseBERT, ConfliBERT and PoliBERTweet available versions (CONT, SCR, *cased* and *uncased*) were fine-tuned on four downstream tasks using the hyper-parameters declared in Section 4. In particular, ConfliBERT is available in both *cased* and *uncased* version, with CONT and SCR configurations. PoliBERTweet is only available in *cased* version with the CONT configuration. The results of the fine-tuning are reported in Table 4. Each reported score represents the average performance over 5 independent runs, each initialised with different seeds (42, 713, 168, 14 and 113).

As can be seen, RooseBERT, especially in its SCR version, outperforms its competitors across all tasks. More specifically, in sentiment analysis, RooseBERT-scr-uncased achieves a 5% improvement over the best-performing competitor. This gain is statistically significant, as demonstrated by paired t-tests conducted against BERT-base-uncased ($p<0.0001$) and ConfliBERT-cont-cased ($p<0.0001$). In the Stance Detection task, RooseBERT-cont-uncased achieves performance comparable to state-of-the-art models, which suggests that, although the VAST dataset is not specifically centered on political debates but on debates in general, RooseBERT can be successfully applied in this context.

In the Argument Component Detection task, RooseBERT-scr-uncased achieves a 2% improvement over best-performing baselines (BERT-base-uncased, $p<0.001$) and competitors (ConfliBERT-cont-uncased, $p<0.001$). In the Argument Relation Prediction task, RooseBERT-scr-uncased improves over of 1% over best performing baselines (BERT-base-uncased, $p<0.01$) and competitors (PoliBERTweet, $p<0.05$). These results confirm a positive trend induced by the adoption of RooseBERT even for complex natural language understanding tasks such as argument mining.

Overall, the encouraging results obtained on the different tasks emphasise the importance of domain-specific pre-training, where task-relevant embeddings contribute to a more refined understanding of argumentative structure.

5.3 Ablation Study

To better understand which portions of the dataset contribute most to our results, we conducted an ablation study. However, since the dataset consists of eight subsets, performing a leave-one-out ablation would have been computationally expensive. Instead, we clustered the dataset subsets based on the similarity of their content. This was achieved using cosine similarity computed over the average of the embeddings of the sentences of each dataset. To obtain the sentence embeddings we used `all-mpnet-base-v2`[2] Sentence Transformer. Figure 1 shows the results of the clustering, performed using Agglomerative Clustering with a distance threshold of 0.2.

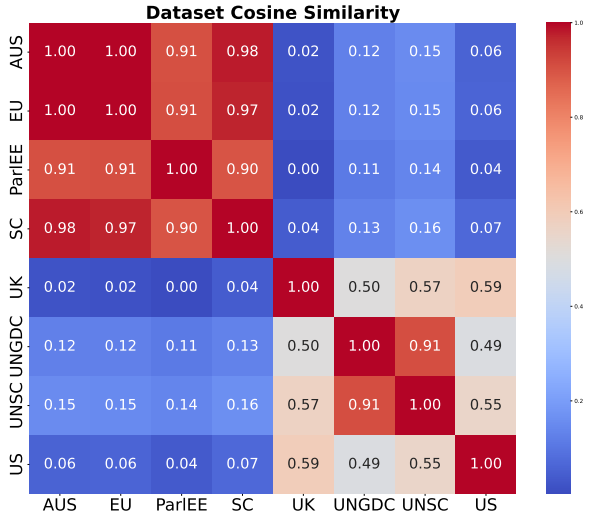

Figure 1: Cosine similarity of the pre-training datasets.

As shown in Figure 1, four distinct clusters emerge. One cluster contains only U.S. televised debates (*cluster#0*), which is expected since they are the only ones not held in a parliamentary setting. Another cluster groups the General Debate of the United Nations (UNGDC) with the United Nations Security Council debates (*cluster#1*), reflecting their shared institutional origin and likely overlap in terminology and thematic content. The remaining two clusters consist, respectively, of UK Parliament debates on one side (*cluster#2*), and debates from the European, Australian, Scottish, and Irish legislative chambers (*cluster#3*).

Once the clusters were identified, we pre-trained four separate RooseBERT models, each using a leave-one-out strategy with respect to the clusters. For instance, the model labelled *Cluster#1* was trained on all debates except those in Cluster 1 (i.e., UNGDC and UN Security Council debates). This setup allows us to assess the individual contribu-

[2] huggingface.co/sentence-transformers/all-mpnet-base-v2

Models	Sentiment Analysis Accuracy	Stance Detection Macro F1 Score	Arg. Component Detection Macro F1 Score	Arg. Relation Prediction Macro F1 Score
BERT-base-cased	0.620 ± 0.007	0.408 ± 0.003	0.464 ± 0.003	0.638 ± 0.025
BERT-base-uncased	0.623 ± 0.008	0.411 ± 0.011	0.466 ± 0.003	0.662 ± 0.007
ConfliBERT-cont-cased	0.658 ± 0.004	0.419 ± 0.008	0.460 ± 0.002	0.659 ± 0.006
ConfliBERT-cont-uncased	0.645 ± 0.006	0.414 ± 0.005	0.464 ± 0.003	0.659 ± 0.007
ConfliBERT-scr-cased	0.644 ± 0.012	0.422 ± 0.009	0.459 ± 0.005	0.621 ± 0.051
ConfliBERT-scr-uncased	0.633 ± 0.050	0.415 ± 0.009	0.458 ± 0.002	0.659 ± 0.007
PoliBERTweet	0.596 ± 0.013	0.407 ± 0.067	0.443 ± 0,004	0.667 ± 0.004
RooseBERT-cont-cased	0.693 ± 0.007	0.419 ± 0.012	0.473 ± 0.001	0.663 ± 0.007
RooseBERT-cont-uncased	0.699 ± 0.007	**0.425 ± 0.016**	0.466 ± 0.006	0.611 ± 0.109
RooseBERT-scr-cased	0.696 ± 0.004	0.339 ± 0.021	0.477 ± 0.005	0.669 ± 0.011
RooseBERT-scr-uncased	**0.702 ± 0.004**	0.424 ± 0.012	**0.480 ± 0.001**	**0.674 ± 0.003**

Table 4: Average and standard deviation results of the downstream tasks.

tion of each cluster to the overall performance by observing the impact of its exclusion.

To train the models, we selected the SCR configuration and the *uncased* version of the model. The SCR configuration was preferred over CONT because we aimed to evaluate the debates in isolation, without any influence from prior training on the general corpus. The choice of the *uncased* version was motivated by empirical observations on the previous results (see Table 4) showing that it generally outperforms the *cased* version across most settings. The chosen hyper-parameters are the same as RooseBERT-scr-uncased. Table 5 shows the perplexity results for each model on a held out portion of the dataset composed by all debates.

Cluster#0	Cluster#1	Cluster#2	Cluster#3	All datasets
2.748	2.839	2.750	3.270	**2.747**

Table 5: Perplexity evaluation of the different clusters.

As shown in Table 5, the exclusion of US Debates appears to have little impact on perplexity, likely because these debates differ in nature from the rest of the dataset, which is predominantly composed of parliamentary debates. On the other side, the exclusion of the European and Australian Parliamentary Debates has the biggest but moderate impact on the perplexity. Yet, all cluster perform worse or as the original model, confirming that most subsets contribute positively to the model's overall knowledge.

Table 6 shows the results of the fine-tuning of these models on the four different downstream tasks. Results are averaged on a total of 5 runs with different seeds. The performance of the models trained with different dataset configurations remains largely consistent across all downstream tasks, showing minimal to no degradation compared to the original RooseBERT model. This consistency highlights RooseBERT's strong generalisation capabilities. Notably, we hypothesised a performance drop in the argument component identification and detection, argument relation detection and classification and the sentiment analysis tasks for *Cluster#0* and *Cluster#2*, as these tasks are based on US and UK political debates that were withheld during the training of these models. However, the results show no significant decrease, suggesting that the structural patterns required for these tasks can be effectively learned from other parliamentary debates.

The reader might argue that the results of RooseBERT on downstream tasks are biased by data in the test datasets that is similar to, or identical to, that in the extensive corpora used for pre-training the PLM. However, as shown in Table 6, the results of the ablation study show the opposite, demonstrating that the model can generalise the discourse structure of political debates even when a key domain is withheld during pre-training.

5.4 Experiments with LLMs

Large instruction-tuned language models (LLMs) such as LLaMA-3 (Grattafiori et al., 2024), GPT4 (OpenAI et al., 2024), Mistral 7B (Jiang et al., 2023) or Gemma 3 (Team et al., 2025) have recently been explored for political discourse analysis through prompting strategies and few-shot learning (Törnberg (2023), Kuila and Sarkar (2024)). While full fine-tuning of these models is often computationally prohibitive and unsuitable for resource-constrained environments, zero-shot and few-shot prompting make them a more practical choice for tasks such as sentiment analysis and argument component detection and classification. However, even

Models	Sentiment Analysis Accuracy	Stance Detection Macro F1 Score	Arg. Component Detection Macro F1 Score	Arg. Relation Prediction Macro F1 Score
Cluster#0	0.706 ± 0.011	0.415 ± 0.008	0.475 ± 0.003	0.667 ± 0.005
Cluster#1	0.707 ± 0.005	0.421 ± 0.009	0.477 ± 0.003	0.676 ± 0.005
Cluster#2	0.695 ± 0.007	0.343 ± 0.115	0.477 ± 0.002	0.669 ± 0.009
Cluster#3	0.706 ± 0.005	0.411 ± 0.005	0.477 ± 0.002	0.663 ± 0.004
RooseBERT-scr-uncased	*0.702 ± 0.004*	*0.424 ± 0.012*	*0.480 ± 0.001*	*0.674 ± 0.003*

Table 6: Average and standard deviation results for the downstream tasks for the different clusters.

Models	Sentiment Analysis Accuracy	Stance Detection Macro F1 Score	Arg. Component Detection Macro F1 Score	Arg. Relation Prediction Macro F1 Score
`gemma-3-4b-it` (zero-shot)	0.485	0.232	**0.219**	0.392
`Mistral-7B-Instruct-v0.3` (zero-shot)	**0.503**	0.302	0.200	**0.446**
`Llama-3.1-8B-Instruct` (zero-shot)	0.467	0.395	0.202	0.186
`gpt-4.1-nano` (zero-shot)	0.488	**0.401**	0.215	0.427
`gemma-3-4b-it` (few-shot)	0.484	0.229	0.212	0.354
`Mistral-7B-Instruct-v0.3` (few-shot)	0.485	0.302	0.199	0.364
`Llama-3.1-8B-Instruct` (few-shot)	0.486	0.340	0.197	0.311
`gpt-4.1-nano` (few-shot)	**0.498**	**0.404**	**0.213**	**0.389**
RooseBERT's best model	*0.702*	*0.425*	*0.480*	*0.674*

Table 7: Performance of LLMs on the downstream tasks in the zero-shot and 3-shot settings.

with prompting, LLMs struggle to capture domain-specific features of political debates, such as implicit argumentation and turn-taking structures.

Table 7 reports the results for *gemma-3-4b-it*, *Mistral-7B-Instruct-v0.3*, *Llama-3.1-8B-Instruct*, and *gpt-4.1-nano* in the downstream tasks RooseBERT was tested for. All models were tested with a temperature of 0.0 to ensure deterministic outputs. In the Argument Component Detection and Classification task, the models were asked to generate the target sentence with *<claim></claim>* and *<premise></premise>* tags while in the other tasks, the models were asked to predict the labels. As can be seen, RooseBERT outperforms LLMs on the four downstream tasks. More specifically, in the stance detection task, the performance of the LLMs is comparable to that of RooseBERT: the VAST dataset used for this task comprises a range of general debatable topics and is not focused on politics, therefore the general-purpose nature of LLMs may be more suitable. However, we observe a performance drop for LLMs compared to RooseBERT when the tasks become more specialized, such as sentiment analysis and argument mining, both in zero-short and few shot settings. Notable errors across tasks included: predicting tags or relations that were not part of the schema (e.g., generating *neutral* as a sentiment label), hallucinations (such as producing random letters or irrelevant sentences) and disregarding the structure or constraints specified in the prompt. We also notice that the models performed best with sentences comprised between 100 and 400 characters.

6 Conclusion

We introduced RooseBERT, a domain-specific pre-trained language model for political debates in English. RooseBERT is developed through two approaches: continued pre-training of BERT, and training from scratch with randomly initialised weights, in both *cased* and *uncased* versions, and across both base and large architectures. To build RooseBERT, we curated a comprehensive dataset including transcripts from US presidential debates, UK, Australian, European and Scottish Parliament, United Nations General Assembly and Security Council. Through an extensive evaluation on four downstream tasks (sentiment analysis, stance detection, argument component detection, and argument relation prediction), we showed that RooseBERT consistently outperforms both general-purpose baselines, other domain-adapted models and LLMs, demonstrating its effectiveness in analysing political discourse. To the best of our knowledge, RooseBERT is the only language model tailored specifically for political debates.

Future work will focus on expanding RooseBERT to other languages, enabling cross-lingual analysis of political discourse and comparative studies across different political argumentation styles.

Limitations

While RooseBERT demonstrates substantial improvements over general-purpose PLMs and related political-domain models, our work is subject to several limitations. Our pre-training corpus, though comprehensive, is restricted to English-language debates and speeches. This excludes a large proportion of political deliberation conducted in other languages, which limits RooseBERT's applicability in multilingual or cross-lingual settings. Furthermore, despite spanning multiple geopolitical contexts (e.g., US presidential debates, UK and Australian parliaments, UN and EU proceedings), the corpus primarily reflects formal institutional settings. Informal or grassroots political discourse (e.g., social media discussions, activist debates) remains underrepresented.

Training RooseBERT—particularly the from-scratch (SCR) variant—required significant computational resources (multi-GPU setups, tens of hours of training). This limits the accessibility of further experimentation (e.g., larger architectures, multilingual versions) for resource-constrained researchers.

Finally, while RooseBERT can support the study of political discourse, it may also be misused for manipulative purposes such as automated propaganda generation, speaker profiling, or biased framing detection. Responsible deployment requires careful consideration of these risks.

References

Gavin Abercrombie and Riza Batista-Navarro. 2020. Parlvote: A corpus for sentiment analysis of political debates. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 5073–5078. European Language Resources Association.

Emily Allaway and Kathleen R. McKeown. 2020. Zero-shot stance detection: A dataset and model using generalized topic representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8913–8931. Association for Computational Linguistics.

Francesco Barbieri, José Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1644–1650. Association for Computational Linguistics.

Alexander Baturo, Niheer Dasandi, and Slava J Mikhaylov. 2017. Understanding state preferences with text as data: Introducing the UN general debate corpus. *CoRR*, arXiv:1707.02774.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3613–3618. Association for Computational Linguistics.

Jack Blumenau. 2021. House of commons parliamentary debates, 1979-2019. https://reshare.ukdataservice.ac.uk/854292/. Accessed: 2024-12-01.

Daniel Braby and Fraser Stewart. 2021. parlScot: a dataset of 1.8 million spoken contributions from the Scottish Parliament 1999-2021. https://doi.org/10.7910/DVN/EQ9WBE. Accessed: 2024-12-01.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: the muppets straight out of law school. *CoRR*, abs/2010.02559.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Pierpaolo Goffredo, Elena Cabrio, Serena Villata, Shohreh Haddadan, and Jhonatan Torres Sanchez. 2023a. Disputool 2.0: A modular architecture for multi-layer argumentative analysis of political debates. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 16431–16433. AAAI Press.

Pierpaolo Goffredo, Mariana Espinoza, Serena Villata, and Elena Cabrio. 2023b. Argument-based detection and classification of fallacies in political debates. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 11101–11112. Association for Computational Linguistics.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious argument classification in political debates. In *Proceedings of the Thirty-First International Joint*

Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022, pages 4143–4149. ijcai.org.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Heal.*, 3(1):2:1–2:23.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2017. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. *arXiv preprint arXiv:1708.01425*.

Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019. Yes, we can! mining arguments in 50 years of US presidential campaign debates. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4684–4690. Association for Computational Linguistics.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *CoRR*, arXiv:1904.05342.

F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker. 2005. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Trans. Assoc. Comput. Linguistics*, 8:64–77.

Lindsay Katz and Rohan Alexander. 2023. A new, comprehensive database of all proceedings of the australian parliamentary debates (1998-2022). `https://doi.org/10.5281/zenodo.8121950`. Accessed: 2024-12-01.

Kornraphop Kawintiranon and Lisa Singh. 2022. Polibertweet: A pre-trained language model for analyzing political content on twitter. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 7360–7367. European Language Resources Association.

Alapan Kuila and Sudeshna Sarkar. 2024. Deciphering political entity sentiment in news with large language models: Zero-shot and few-shot strategies. *CoRR*, abs/2404.04361.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Comput. Linguistics*, 45(4):765–818.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240.

Juncai Li and Xiaofei Jiang. 2021. Mol-bert: An effective molecular representation with BERT for molecular property prediction. *Wirel. Commun. Mob. Comput.*, 2021:7181815:1–7181815:7.

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. *CoRR*, abs/1907.11692.

Zhengliang Liu, Yiwei Li, Oleksandra Zolotarevych, Rongwei Yang, and Tianming Liu. 2024. LLM-POTUS score: A framework of analyzing presidential debates with large language models. *CoRR*, abs/2409.08147.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. Crossner: Evaluating cross-domain named entity recognition. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13452–13460. AAAI Press.

Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–23.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 9–14. Association for Computational Linguistics.

Farhad Nooralahzadeh, Lilja Øvrelid, and Jan Tore Lønning. 2018. Evaluation of domain-specific word embeddings using knowledge resources. In *Proceedings of the Eleventh International Conference on*

Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 3505–3506. ACM.

Mirco Schoenfeld, Steffen Eckhard, Ronny Patz, Hilde van Meegdenburg, and Antonio Pires. 2019. The UN Security Council Debates. https://doi.org/10.7910/DVN/KGVSYH. Accessed: 2024-12-01.

Gijs Schumacher, Martijn Schoonvelde, Tanushree Dahiya, and Erik de Vries. 2016. EUSpeech. https://doi.org/10.7910/DVN/XPCVEI. Accessed: 2024-12-01.

Christine Sylvester, Zachary Greene, and Benedikt Ebing. 2022. ParlEE plenary speeches data set: Annotated full-text of 21.6 million sentence-level plenary speeches of eight EU states. https://doi.org/10.7910/DVN/ZY3RV7. Accessed: 2024-12-01.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *CoRR*, abs/2304.06588.

Shoya Wada, Toshihiro Takeda, Katsuki Okada, Shirou Manabe, Shozo Konishi, Jun Kamohara, and Yasushi Matsumura. 2024. Oversampling effect in pretraining for bidirectional encoder representations from transformers (BERT) to localize medical BERT and enhance biomedical BERT. *Artif. Intell. Medicine*, 153:102889.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Wooseong Yang, Sultan Alsarra, Luay Abdeljaber, Niamat Zawad, Zeinab Delaram, Javier Osorio, Latifur Khan, Patrick T. Brandt, and Vito D'Orazio. 2023. Conflibert-spanish: A pre-trained spanish language model for political conflict and violence. In *7th IEEE Congress on Information Science and Technology, CiSt 2023, Agadir - Essaouira, Morocco, December 16-22, 2023*, pages 287–292. IEEE.

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications. *CoRR*, arXiv:2006.08097.

Nahid Zeinali, Alaa Albashayreh, Weiguo Fan, and Stephanie Gilbertson White. 2024. Symptom-bert: Enhancing cancer symptom detection in ehr clinical notes. *Journal of Pain and Symptom Management*, 68(2):190–198.e1.

A Experiments with Large Architectures

In addition to the *base* architecture, RooseBERT was also trained in CONT and SCR approaches, each in both *cased* and *uncased* variants using the *large* (i.e., *BERT-large*) configuration, as larger models have been shown to capture more complex linguistic patterns. The *large* model follows BERT-Large, with 24 layers, 1024 hidden units, and 16 attention heads.

To the best of our knowledge, the current state of the art does not include pre-trained language models for political debates which utilise a *large* architecture. Therefore, the only possible comparison for RooseBERT-large is its baseline, BERT-large.

Pre-Training. Table 8 display the perplexity results of the *large* architectures in both cased and uncased version on an held-out portion of the dataset.

	BERT-large	**RooseBERT-cont large**	**RooseBERT-scr large**
cased	27.505	**2.203**	2.531
uncased	7.832	**2.280**	2.558

Table 8: Perplexity of large architectures evaluated on an held out corpus representing the overall distribution of political debates.

The BERT-large model struggles without a specific pre-training on the political terminology with perplexity values that are comparable to the perplexity of the *base* architecture. On the other side, RooseBERT performs significantly better than BERT. The perplexity for the CONT configurstion is again lower than its counterpart SCR, in both versions.

Models	Sentiment Analysis Accuracy	Stance Detection Macro F1 Score	Arg. Component Detection Macro F1 Score	Arg. Relation Prediction Macro F1 Score
BERT-large-cased	0.528 ± 0.000	0.316 ± 0.096	0.469 ± 0.005	0.576 ± 0.198
BERT-large-uncased	0.544 ± 0.036	0.352 ± 0.032	0.466 ± 0.002	0.410 ± 0.283
RooseBERT-large-cont-cased	0.719 ± 0.007	0.367 ± 0.117	0.475 ± 0.002	0.499 ± 0.264
RooseBERT-large-cont-uncased	0.689 ± 0.090	0.300 ± 0.128	0.410 ± 0.135	0.598 ± 0.210
RooseBERT-large-scr-cased	0.709 ± 0.006	$\mathbf{0.419 \pm 0.011}$	0.475 ± 0.002	0.687 ± 0.008
RooseBERT-large-scr-uncased	$\mathbf{0.715 \pm 0.005}$	0.412 ± 0.008	$\mathbf{0.479 \pm 0.002}$	$\mathbf{0.695 \pm 0.006}$

Table 9: Average and standard deviation results of the downstream tasks for the *large* architectures.

Training the *large* architecture required more computational power with respect to the *base* architecture, with a total training time of 35 hours for both the CONT and SCR versions. As a consequence, due to time constraints, it was only possible to train the *large* versions of RooseBERT for 125K steps. The hyper-parameters used to train the *large* version of RooseBERT (see Table 10) are the same used to train the *base* version with the exception of the number of steps.

Hyper-parameters	RooseBERT-cont large	RooseBERT-scr large
Learning Rate Decay	Linear	Linear
Warmup Steps	10,000	10,000
Total Training Steps	125K steps	125K steps
Total Batch Size	2048	2048
Learning Rate	$1e^{-4}$	$1e^{-4}$
Weight Decay	0.01	0.01
Adam β_1, β_2, ϵ	0.9, 0.98, $1e^{-6}$	0.9, 0.98, $1e^{-6}$
Vocabulary Type	BERT	Custom
Vocabulary Size	28,997 (*cased*) 30,523 (*uncased*)	28,996 (*cased*) 30,522 (*uncased*)

Table 10: Hyper-parameters used to pre-train RooseBERT with the *large* architecture setting.

Fine-Tuning. Due to constraints arising from the computational power required to train the *large* architectures, for the fine-tuning of RooseBERT-Large, we fixed the max sequence length to 256 tokens while leaving other hyper-parameters unaltered. Table 9 reports the results of the fine-tuning of BERT-large and RooseBERT-large on the four downstream tasks and hyper-parameters discussed in Section 4. Each result is averaged over five runs with different seeds.

In this case, we again observe the superiority of the SCR configuration over the CONT one, with RooseBERT-large-scr-uncased outperforming all other models on three out of four tasks. Notably, in the *large* architecture, the BERT-Large models appeared more prone to overfitting under certain random seeds, especially in tasks such as sentiment analysis and stance detection, leading to inconsistent or degraded performance.

Performance gains from scaling up to the large architecture were modest when compared to the base model. An exception is relation classification, where RooseBERT-large achieved a Macro F1 score of 0.695, improving over the base model's score of 0.674. This findings suggests that, in some cases, a well-optimised base model is sufficient to capture most of the relevant patterns in the data, and that increasing model size does not always guarantee substantial improvements, particularly when data complexity or volume does not fully leverage the larger parameter space. Furthermore, these results underscore the difficulty of tasks that rely on datasets of political debates, highlighting the need for PLMs specifically trained on such data, like RooseBERT.