Enhancing Partisanship Prediction in Congressional Speeches

Stanford CS224N Custom Project

Amelia Leon
Department of Computer Science
Stanford University
ameliait@stanford.edu

JB Jong Beom Lim
Department of Computer Science and Political Science
Stanford University
jlim216@stanford.edu

Sherry Yang
Department of Bioengineering
Stanford University
nyang19@stanford.edu

Abstract

This paper examines methods for improving partisanship prediction in congressional speeches by experimenting with transfer learning, feature engineering, and multiple transformer-based models. In doing so, we introduce novel datasets on abortion, firearms, and the Iraq War to understand the impact of domain-specific knowledge and dataset characteristics on model performance. Our results show that larger, more divisive datasets, such as those on abortion, noticeably enhance accuracy compared to smaller datasets like those on firearms. Additionally, incorporating temporal and geographic features further enhanced the model's accuracy. We compare baseline LSTM models with transformer-based models, including mDeBERTa-v3, DeBERTa-v3, and RoBERTa, to determine the most effective approach for political speech classification. Our findings underscore the relevance of dataset attributes and feature engineering in refining predictive models for political partisanship, providing valuable insights for future research in political text analysis.

1 Key Information to include

- Mentor: Shikhar Murty
- External Collaborators (if you have any): N/A
- Sharing project: N/A
- Team contributions: Everyone contributed to writing the report. **Amelia.** Implemented all stage 2 models (M1-M20), completed relevant analysis, and covered all of feature engineering. She also fine-tuned GPT-3.5 Turbo, completed relevant. **JB.** Implemented the mDeBERTa-v3 model in Stage 1 (4 models), completed relevant analysis for transfer learning. Created/labeled the novel datasets (abortion, firearms, Iraq War) and conceptualized the project. **Sherry.** Implemented LSTM model, conducted data-filtering for the aforementioned datasets. She also constructed GPU setup resources for testing llama and bert models, conceptualized and contributed to the ethics statement.

2 Introduction

In recent years, the United States has experienced a noticeable increase in political polarization, where ideological divides between parties have deepened on important issues including foreign policy, healthcare, and domestic regulations. A 2024 report revealed that Congressional polarization score has increased with politicians employing divisive rhetoric.Vanderbilt University (2024) This growing divide in Congress not only impacts the legislative process by hindering bipartisan cooperation but also fosters an "us versus them" mindset in the broader political discourse, often leading to a "steep rise in political violence."McCoy and Press (2022)

Within this context, accurately identifying party affiliation in congressional speeches can reveal the extent of these divides and provide insight into legislative behavior and partisan-based arguments. Prior research on political text classification has explored partisanship detection, leveraging advanced natural language processing models to identify partisan bias in news articles and congressional speeches (refer to Section 3.1). These studies have demonstrated the potential of language models in detecting linguistic nuances arising from political rhetoric.

However, the wide range of societal topics and limited availability of speeches on specific issues pose limitations to researchers, highlighting the need for transferable methods across different fields. Domain transfer learning has shown promise in addressing such challenges by leveraging knowledge learned from one domain to improve performance in another (refer to Section 3.2).

Building upon these insights, our project investigates whether knowledge of party preferences from distant yet rhetorically polarized domains can be effectively transferred to a researcher's specific area of interest. To do so, we first introduce novel datasets on congressional speeches related to the Iraq War, abortion, and gun control issues. We then explore the effectiveness of distant domain transfer learning in enhancing the party classification accuracy of baseline long short-term memory (LSTM) and more sophisticated transformer-based models. Furthermore, additional context features such as states or speaker names are incorporated to improve the models' performances. Our findings reveal that the characteristics of the transfer domain dataset, the inclusion of temporal and geographic features, and the type of model used notably impact classification accuracy.

3 Related Work

3.1 Political bias detection

With the growing popularity of natural language processing (NLP) models, there is increasing concern about the propagation of political bias. Although NLP has been successful in various text processing tasks, such as sentiment analysis, it also raises concerns about enhancing political bias.Feng et al. (2023) Bias can originate from multiple sources. Many language models, including GPT and BERT, use word embeddings from sources like Word2Vec to develop a general understanding of human language and word representations. Hinton et al. (2012); Devlin et al. (2018); Vaswani et al. (2017) These embeddings often serve as a baseline structure to improve downstream tasks. However, Bolukbasi et al. (2016) identified that Google News, which was used to train word embeddings, has demonstrated bias.Bolukbasi et al. (2016) In the context of political speech, where word choices can significantly influence political ideology, such contextualized word embeddings can greatly affect model performance.Radford et al. (2021) This issue is indirectly reflected in the tendency of large-scale language models to generate politically biased content.Haberlin and Others (2021)

3.2 Domain transfer learning

Recent studies have increasingly highlighted the complexities of domain transfer learning. A research paper from Google Brain has demonstrated that domain transfer learning proves to be highly effective when the pre-training domain closely matches the target dataset but leads to negative transfer and poor performance when there is mismatch between the domains. Moreover, the team has also found that more pre-training data can actually make performance worse, if this additional data does not contribute any fine-tuning value. Hu et al. (2021)

In the specific domain of political party classification, a similar study investigated the generalizability of a transfer learning model for partisan detection in general media sources. The team concluded that models trained on Congress speech perform poorly when transferring from one text domain to another. Hassan et al. (2021) The level of predictability also varies notably across different subjects: topics such as tax policy demonstrate greater consistency across various domains in comparison to issues like abortion. However, another paper from Cambridge University Press published a differing opinion, where language models fine tuned on a small curated dataset performs better than a general cross-domain language model. Kershaw et al. (2021) Overall, these findings underscore the need for careful consideration of domain specificity and dataset composition in transfer learning to enhance model performance.

4 Approach

The diagram for the experiment architecture can be found in the Appendix 3.

4.1 Transfer learning

We adopted a transfer learning approach to evaluate the effectiveness of cross-domain knowledge transfer on political speech classification. The Iraq War dataset served as our primary subject, while the abortion and firearms datasets, chosen for their clear partisan divides, were used for transfer learning (refer to Section 5.1). All datasets were divided into training, validation, and test sets with a 70/10/20 split. For the non-treatment group, all models were fine-tuned directly on the Iraq War dataset and tested on the same dataset to establish a baseline. For the treatment group, the models underwent a fine-tuning phase on the abortion and firearms datasets for domain transfer, followed by a final fine-tuning phase on the Iraq dataset before testing. All models were ultimately tested on the Iraq dataset to evaluate their performance. We also explored merging the abortion and firearms datasets before the initial training phase to enhance the model's exposure to diverse political discourse.

Finally, using the baseline LSTM and mDeBERTa-v3 models, we determined that the abortion dataset provided superior transfer learning results compared to the firearms dataset. This finding guided our decision to prioritize the abortion dataset for subsequent fine-tuning, optimizing computational efficiency and leveraging the more pronounced partisan distinctions within the abortion dataset.

4.2 Feature engineering

Given that political polarization is known to have been increasing in the last several years and that political viewpoints are often separated by state lines, we experiment with using both temporal and geographic features. We leverage two different feature extractors that give the model more real-world context about the speech. We add temporal features by extracting the year of the speech and prepending the string "year {year}" to the speech text. We add geographical features by extracting the full state names and prepending the string "state state". We also experiment with adding the two features together by adding "year {year} state {state}" before the speech text.

4.3 Baseline

For the baseline model, we implemented a bidirectional LSTM architecture to compare with transformer-based models. We tokenized and padded the text data, using an embedding layer with vectors from a pre-trained Word2Vec dictionary. The model, comprising two bidirectional LSTM layers and a dense output layer with a sigmoid activation function, was optimized with binary cross-entropy loss. We then applied the same transfer learning experiments from Section 4.1 to assess the baseline model's performance across different datasets.

4.4 Models

For our main models we experiment with a variety of BERT-based models including mDeBERTa-v3 (multilingual), DeBERTa-v3, and RoBERTa He et al. (2020); Liu et al. (2019); He et al. (2021) for political text classification. DeBERTa-v3 and mDeBERTa-v3 use a transformer architecture with multi-head self-attention mechanisms. The architecture included a pre-trained mDeBERTa-v3 model with a linear projection layer, subsequently fine-tuned on targeted datasets. We employed a custom dataset class and data module for handling congressional speeches. Additionally, we relied on LoRA (Low-Rank Adaptation) to reduce the number of trainable parameters, significantly improving the model's efficiency.Hu et al. (2021) This method allowed us to fine-tune mDeBERTa with fewer resources, making it practical for transfer learning experiments using multiple domain datasets. Given this advantage, we applied the model to compare the value of transfer learning across different datasets, including those on abortion, firearms, and merged datasets.

We hypothesized that while mDeBERTa might initially underperform on English-centric tasks due to its generalized design, it could benefit significantly from transfer learning by acquiring more relevant linguistic context from targeted datasets. This method enables us to highlight the potential benefits of transfer learning in enhancing mDeBERTa's accuracy on a given political speech topic. To test this hypothesis, we first evaluate mDeBERTa's performance on the task without transfer learning and then compare it to the results obtained after applying transfer learning knowledge.

Following our experimentation with mDeBERTa, we apply the most effective transfer learning methods to the English-specific DeBERTa-v3 and RoBERTa models. Although the two models are inherently optimized for English-centric tasks, we explore the potential for further improvement by studying the impact of transfer learning and feature engineering on their performance. By comparing the results obtained from these approaches, we intend to understand the effectiveness of transfer learning and feature engineering in enhancing model accuracy. Our experiment design can be found in Section 5.3.

5 Experiments

5.1 Data

The dataset comes from the Stanford Congressional Record dataset, established by Gentzkow et al. in their 2018 publication.Gentzkow et al. (2018) It covers all congressional speeches from the 43rd to the 114th Congresses. Our analysis targeted a subset from the 100th to the 111th Congresses (1987-2010). For data processing, we joined each speech with corresponding speaker information to capture relevant party affiliations, dates of speeches, and other contextual details such as home state and chamber.

5.1.1 Preliminary Experiments

Our dataset originally focused on speeches related to foreign policy issues and international relations, filtering for texts containing the key words such as "Iraq", "Japan", "Canada," and "North Korea." However, preliminary analysis using models such as GPT-3.5 Turbo revealed poor performance in detecting partisanship within this raw dataset. As shown in Table 1, while the model performed well in classifying speeches as being about allies or adversaries, it struggled significantly with partisan classification. This highlighted the necessity of refining our dataset to better capture the nuances of partisan rhetoric, thereby improving the model's ability to distinguish between Democrat and Republican speeches.

Task	Accuracy	F1 Score	Precision	Recall
Ally/Adversary	0.8250	0.8158	0.8611	0.7750
Partisanship	0.5950	0.6161	0.5855	0.65

Table 1: Classification results from raw dataset (GPT-3.5 Turbo)

5.1.2 Constructing a novel dataset

Transfer Domain. We first consulted with political science professors and postdoctoral fellows to identify topics on which Democrats and Republicans clearly disagree. This led to five candidate datasets on abortion, defense spending, healthcare, Afghanistan, and gun control (i.e., firearms). We extracted all speeches containing these keywords and then randomly sampled 100 congressional speeches from each dataset for manual screening. Ultimately, we focused on speeches about firearms and abortion due to the distinct and polarized opinions held by the two political parties. The Democratic Party generally supports abortion rights and stricter firearm regulations, while the Republican Party opposes abortion and supports individual firearm ownership. These divisive arguments provided a clear basis for our analysis, unlike broader topics like defense spending, which covered a wider range of issues and were less suitable for our study. Examples of the speeches can be found in A.2.

Target Domain. For our target dataset (i.e. topic to apply transfer learning on), we manually scanned the original international relations dataset for topics on which Democrats and Republicans clearly disagree and ultimately identified the "Iraq War" as a contentious issue. The Republican Party predominantly supported the invasion, emphasizing national security concerns and the threat posed by weapons of mass destruction (WMDs). They also advocated for regime change in Iraq to neutralize these threats. Conversely, the Democratic Party grew increasingly opposed to the war over time, challenging its justification and demanding the withdrawal of U.S. troops. We focused on the speeches starting from 2003 when the war broke out. This led to the following breakdown in Table 3. Examples of the speeches can be found in Table 2.

5.2 Evaluation method

We evaluated our model using several metrics: accuracy, macro-precision, macro-recall, and macro-F1 score. First, given that our datasets were balanced, accuracy served as a reliable measure of overall performance. Additionally, since our task contained two classes representing different political parties, it was not appropriate to designate one class as positive or negative without considering the potential biases this may introduce. Therefore, macro-averaging, which calculates metrics for each class independently and averages them, ensured an unbiased evaluation of the model's performance. The macro-F1 score was computed as follows:

$$\text{Macro-F1} = \frac{2 \times \text{Macro-Precision} \times \text{Macro-Recall}}{\text{Macro-Precision} + \text{Macro-Recall}}$$

Example phrases (Identified in manual screening)	**Partisanship**
This war was a **very bad mistake**. . . there [were] no nuclear or other WMDs at the sites identified by the CIA. . . Now we are **bogged down in a quagmire** with no end in sight.	Democrat
We want to leave Iraq, but we must leave Iraq **based on conditions where Iraq can sustain itself, defend itself, and govern itself.**	Republican
U.S. policy toward Iraq should be **focused on bringing home U.S. troops** as soon as possible while minimizing chaos in Iraq and maximizing Middle Eastern stability.	Democrat
Who can forget the **cheering of Iraqi citizens** in the streets as Baghdad was liberated and the **statue of Saddam Hussein toppled** to the ground?	Republican

Table 2: Filtered Dataset from Speeches on Iraq War

Split	Domain	Democrat	Republican	Total
Train	Firearms	1204	917	2121
	Abortion	3328	4237	7565
	Iraq	3715	3670	7385
Val	Firearms	274	257	531
	Abortion	833	1034	1867
	Iraq	531	408	939
Test	Iraq	1061	1021	2082
Total	-	10946	11544	22490

Table 3: Counts of party affiliations in the train/val/test sets

Finally, we conducted a qualitative assessment by manually reviewing misclassified examples to understand common error patterns and potential areas for improvement.

5.3 Experimental Design

To answer our research questions, we conduct experiments in stages. In stage 1, we validate whether transfer learning improves performance on the Iraq data by leveraging the abortion and firearms datasets. This stage allows us to test our hypothesis on the effectiveness of transfer learning. Next, in stage 2, we run extensive experiments with transfer learning and feature engineering, applying the learned values from stage 1 to standardize and optimize our models further.

5.3.1 Stage 1: Multilingual DeBERTa-v3

Our experiments utilized a batch size of 16, with the model trained over 10 epochs for both domain transfer and final fine-tuning phases. We employed the AdamW optimizer with a learning rate of 3e-5 and the Cosine Annealing LR scheduler. Training and validation datasets were pre-processed through tokenization and padding to a maximum length of 512 tokens. Experiments were conducted on a single GPU, and model checkpoints were saved based on validation loss to ensure the best-performing model was selected for evaluation.

5.3.2 Stage 2: Multilingual DeBERTa-v3, DeBERTa-v3 & RoBERTa

Each of our experiments (M5-M16) first trained over ten epochs on the domain transfer dataset (if it was used) and then fine-tuned for five epochs on the Iraq dataset. We also ran each of the models with data including features such as year and state. Experiments were conducted using an A100 GPU through Colab, and model checkpoints were saved based on validation loss to ensure the best-performing model was selected for evaluation.

Model	Transfer Domain	Accuracy	Macro F1	Macro Precision	Macro Recall
LSTM		0.6278	0.5992	0.6081	0.6015
LSTM	Firearm	0.5692	0.6519	0.7004	0.6873
LSTM	Abortion	0.6955	0.7210	0.7437	0.7321
mDeBERTa-v3		0.7918	0.8739	0.8741	0.8736
mDeBERTa-v3	Firearm	0.7822	0.8638	0.8622	0.8674
mDeBERTa-v3	Abortion	**0.8192**	**0.8947**	**0.8967**	**0.8942**
mDeBERTa-v3	Merged	0.7973	0.8779	0.8761	0.8815

Table 4: Stage 1 Results

5.4 Results

5.4.1 Stage 1 Results: Transfer Learning and Dataset Attributes

Despite variance across different models, the results revealed that transfer learning mostly enhances classification accuracy when applied to the correct transfer domain dataset. The baseline LSTM models showed substantial improvements with transfer learning, particularly for the abortion dataset. mDeBERTa-v3, while generally achieving lower accuracy than DeBERTa-v3, still demonstrated improvements with transfer learning. The highest performance for mDeBERTa-v3 was observed on the abortion dataset (0.8192) compared to its initial performance (0.7918). This suggested that mDeBERTa-v3, with its multilingual capabilities, benefitted from additional training on English-specific political datasets and captures language nuances better in divisive contexts.

The size of the transfer domain dataset also significantly impacted model performance. Our analysis showed that models using domain transfer from abortion speeches, which had a larger dataset (i.e., 9432 data points), were more accurate at predicting partisanship than those using domain transfer from gun control speeches with a smaller dataset (i.e., 2652 data points). When applied to the Iraq War dataset, the baseline LSTM model trained on the abortion dataset achieved an accuracy of 0.6955, compared to 0.5692 for the firearm dataset. The mDeBERTa-v3 model likewise achieved its highest accuracy on the abortion dataset (0.8192) and its lowest performance on the firearm dataset (0.7822).

These results suggest that the more comprehensive abortion dataset provided better learning opportunities and led to models with enhanced accuracy. In contrast, the smaller firearm dataset possibly introduced more noise and less transferrable knowledge, undermining model performance. This aligns with existing literature highlighting the importance of dataset relevance in effective transfer learning. Although abortion and Iraq War topics seem distant in topical terms, the abundance of data and the divisive political rhetoric across party lines in the abortion dataset made it valuable for transfer learning in classifying partisanship.

5.4.2 Stage 2 Results: Feature Engineering and Optimizing Performance

Having learned that transfer learning works best with the abortion dataset, we applied this insight to optimize our models further in stage 2. DeBERTa-v3, already optimized for English, showed less improvements due to its robustness in handling English text without extensive transfer learning. Still, it achieved its peak accuracy of 0.9361 (M11) when utilizing the abortion dataset alongside state features. For RoBERTa, there was also a subtle increase in classification accuracy with domain transfer learning, even without additional feature engineering, as seen in models M13 and M17. These points collectively underline the effectiveness of transfer learning in improving model accuracy by adapting the models to specific, divisive domains. At the same time, the different performance levels across models reinforces the importance of selecting appropriate models based on their initial training and linguistic capabilities for corresponding tasks.

Appropriate feature engineering enhanced model performance for all transformer models. Overall, the highest-performing model was DeBERTa-v3 with state features and transfer knowledge from the abortion dataset, demonstrating that the combination of the correct model architecture, effective transfer learning, and well-chosen features enhances classification accuracy. This highlights the critical role of feature engineering in refining predictive models for political partisanship and reaffirms the importance of context in political speech classification tasks.

	Model	Transfer Domain	Features	Accuracy	Macro F1	Macro Precision	Macro Recall
B1	LSTM			0.6278	0.5992	0.6081	0.6015
B2	LSTM	Abortion		0.6955	0.6784	0.5941	0.6026
M1	mDeBERTa-v3			0.8713	0.8711	0.8760	0.8724
M2	mDeBERTa-v3		year	0.8602	0.8601	0.8607	0.8599
M3	mDeBERTa-v3		state	0.9155	0.9155	0.9163	0.9160
M4	mDeBERTa-v3		year+state	0.9198	0.9198	0.9198	0.9199
M5	DeBERTa-v3			0.8890	0.8888	0.8909	0.8884
M6	DeBERTa-v3		year	0.8934	0.8931	0.8950	0.8928
M7	DeBERTa-v3		state	0.5120	0.3514	0.5555	0.5026
M8	DeBERTa-v3		year+state	0.9270	0.9268	0.9289	0.9264
M9	DeBERTa-v3	abortion		0.8818	0.8813	0.8857	0.8809
M10	DeBERTa-v3	abortion	year	0.8828	0.8827	0.8832	0.8825
M11	DeBERTa-v3	abortion	state	**0.9361**	**0.9361**	**0.9361**	**0.9362**
M12	DeBERTa-v3	abortion	year+state	0.9111	0.9111	0.9111	0.9111
M13	RoBERTa			0.8650	0.8649	0.8676	0.8659
M14	RoBERTa		year	0.8554	0.8553	0.8590	0.8564
M15	RoBERTa		state	0.8718	0.8716	0.8753	0.8727
M16	RoBERTa		year+state	0.9087	0.9087	0.9088	0.9086
M17	RoBERTa	abortion		0.8742	0.8739	0.8757	0.8735
M18	RoBERTa	abortion	year	0.8722	0.8717	0.8757	0.8713
M19	RoBERTa	abortion	state	0.8814	0.8812	0.8823	0.8809
M20	RoBERTa	abortion	year+state	0.8602	0.8602	0.8602	0.8604

Table 5: Stage 2 Results

6 Analysis

We conducted a qualitative analysis on the outputs of M11, the best model, by looking at specific misclassified example, as well as conducting a comprehensive n-grams analysis on them. We found that the misclassified examples had little to no polarizing speech, and were generally neutral transcripts. For instance, these examples contained phrases such as "on the other side of the aisle" and "it is fraud to say its bipartisan." While evocative, these phrases do not contain context clues about the speakers' party affiliation.

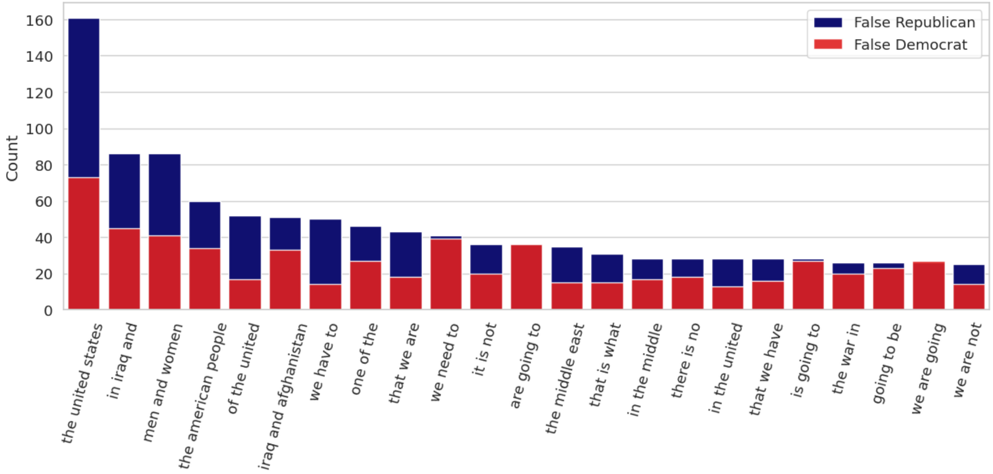

Figure 1: Analysis of top 3-grams in misclassified examples

There were 75 false Democrat, i.e. the model falsely predicted a Republican speaker as being Democrat, and 140 false Republican, i.e. the model falsely predicted a democratic speaker as being republican. Please see Fig. 2, for the confusion matrix. To better understand why examples may have been misclassified, we used an n-grams model for $n = 3, 5, 7$ on false Democrat and false Republican. We found that majority of the top-20 n-grams were largely shared across both the false positives and false negatives. In Fig. 1, we show the top-20 3-grams for False Republican and False Democrat. Notice that the almost all of the 3-grams are shared across both groups.

We also noticed that in the 7-grams analysis that False Republican examples in particular are generally more light-hearted and feature bipartisan language to refer to the other party. Some examples included, "on the other side of the aisle" and "light at the end of the tunnel". This implied that the model might have misclassified less adversarial and more collaborative language as coming from Democrats, possibly due to an inherent bias in the training data where such language is more frequently associated with Democratic speeches. Overall, these perspectives underscored the importance of considering tone and sentiment alongside specific phrases when training models for political text classification.

On the other hand, the analysis for False Democrat examples revealed recurring phrases related to specific incidents or locations, often referring to the enemy, which may have contributed to the misclassification. These included phrases such as "result of enemy action in al Anbar," "of enemy action in al Anbar province," and "as result of enemy action in al.” These phrases on military actions are less distinctly partisan and possibly led the model to incorrectly attribute them to Democratic speakers, reflecting a potential challenge in distinguishing partisanship based solely on context-specific language. This suggests that while the model can generally differentiate between party affiliations, it sometimes struggles with contextually ambiguous or less partisan language.

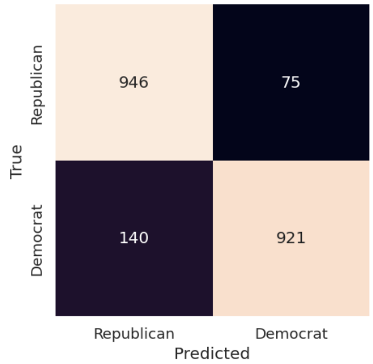

Figure 2: Confusion matrix for best model

7 Conclusion

Our findings uncover that the properties of the transfer domain dataset, the addition of temporal and geographic features, and the choice of model architecture collectively enhance classification accuracy. The combination of effective transfer learning, appropriate model choice, and the incorporation of correct features, as demonstrated by M11 in Table 5, contributed to its status as the best-performer. Models that leveraged knowledge from the larger, more divisive abortion dataset outperformed those utilizing the smaller firearm dataset, which illuminates the impact of dataset size and divisiveness. Such findings indicate that transfer learning from an unrelated domain can be beneficial if the dataset itself is well-defined and informative. Additionally, feature engineering further improved model performance, underscoring the importance of contextual information. The choice of model architecture (even across transformer-based models) also significantly influenced outcomes and demonstrated the necessity of careful model selection.

A limitation of this study involves the varying performance across different topics and models, which underlines that the benefits of transfer learning and feature engineering are not universal. Despite these limitations, our contribution remains valuable by highlighting the importance of dataset attributes, knowledge transfer, and contextual features in enhancing model performance. Future work should explore additional features and domains to refine these models further. Overall, our research provides a foundation for more effective political speech analysis and classification.

8 Ethics Statement

Social media has become a mainstream platform for spreading political ideology. An NLP model designed to identify partisan-specific language, trained on a broad range of domain knowledge, can be used for both beneficial purposes and weaponization. Commercially, such a model can be exploited to detect users' political leanings and feed content that aligns with their views. Additionally, it can be used to learn and replicate the language of a political party, fabricating political ideas to manipulate people. This poses a threat by potentially escalating polarized political opinions and increasing societal tension. One way to mitigate such problems is to establish model accountability. By incorporating prompts and interactive features, the use of NLP models can be monitored to prevent misuse for inappropriate purposes. This might require implementing a governance model, similar to an honor-code violation detection system, to ensure responsible usage. Kim et al. (2020); Li et al. (2021)

As a related issue, classification models are not infallible and will inevitably make errors. This is particularly true for topics with less polarized opinions, as shown in our previous analysis. If the model is open-sourced and available to the public, any false positive or false negative results could introduce bias to individuals. On a more serious note, inaccuracies in the model could spread misinformation. Additionally, since the model is designed to detect either Democratic or Republican affiliation, it is compelled to produce a partisan result. This could inadvertently contribute to increased political polarization. A mitigation strategy is to modify the model to also identify opinions shared by both parties and provide information about their common grounds and the contexts for each party's positions. This approach could promote peaceful political discourse rather than contentious dynamics. Ravenscroft et al. (2020); D'Angelo et al. (2020)

References

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *arXiv preprint arXiv:1607.06520*.

Giovanni D'Angelo, Federico Martini, Sorzano Carlos Oscar Cristina, Pablo Blas Vicente, Elias George, Oscar Manuel Rueda, Joan Segura, and et al. 2020. A machine learning approach for identifying gene-gene interactions in high-dimensional genomic data. *BMC Bioinformatics*, 21(1):361.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv preprint arXiv:1810.04805.

Shangbin Feng, Chan Young Park, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. Montreal AI Ethics Institute.

Matthew Gentzkow, Jesse M. Shapiro, and Matt Taddy. 2018. Congressional record for the 43rd-114th congresses: Parsed speeches and phrase counts. `https://data.stanford.edu/congressional_record`.

M. M. Haberlin and A. N. Others. 2021. Title of the article. *Journal Name*, Volume Number(Issue Number):Page Numbers.

Sayed Hassan, Ninareh Mehrabi, Tal Yarkoni, Bo Pang, Michael V Arnold, Karen Levy, and Oliver Hammond. 2021. The congressional classification challenge: Domain specificity and partisan intensity. *The Becker Friedman Institute for Economics Working Paper Series*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654.

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. ArXiv preprint arXiv:1301.3781.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Daniel Kershaw, Jonathan Zirn, Ryan Shorey, and William Muirhead. 2021. Topic classification for political texts with pretrained language models. *Political Analysis*.

Buomsoo Kim, Jinsoo Park, and Jihae Suh. 2020. Transparency and accountability in ai decision support: Explaining and visualizing convolutional neural networks for text information. *Decision Support Systems*, 134:113302.

Zizhao Li, Miaojing Shi, Jian Cheng, Zhiyong Cheng, and Jiaxiang Wu. 2021. Addressing label noise in zero-shot learning via semantic adversarial adaptation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pages 1161–1170. ACM.

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.

Jennifer McCoy and Benjamin Press. 2022. What happens when democracies become perniciously polarized?

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2110.01804*.

James Ravenscroft, Aurélie Mahalatchimy, Sharon Gilad, Gill Green, Enrico Rossi, Christopher Birchall, Till Metzler, and James Fatchett. 2020. Algorithmic governance and the politics of operationalising ai for public value: A case study of the uk government's data science campus. *Big Data Society*, 7(2):20539517231179199.

Vanderbilt University. 2024. Latest vanderbilt unity index shows the u.s. continuing its trend toward increased political polarization.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. ArXiv preprint arXiv:1706.03762.

A Appendix

A.1 Diagram architecture

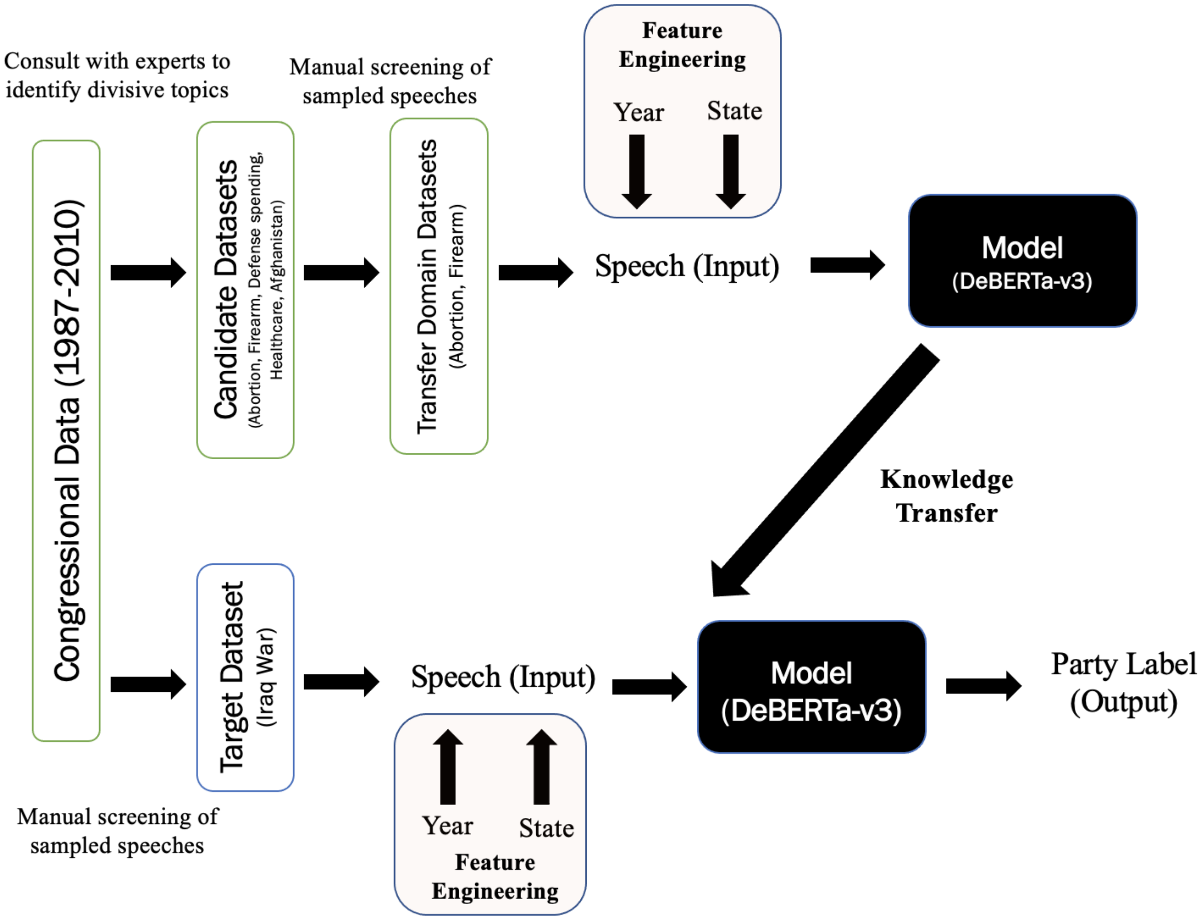

Figure 3: Experiment architecture (DeBERTa-v3 Model)

A.2 Examples from Domain Transfer Dataset

Table 6: Sample Speeches by Category and Partisanship

Example phrases (Identified in manual screening)	**Category**	**Partisanship**
You would never do this to the women of America because **they have control over their lives.**	Abortion	Democrat
Mr. Speaker, **healthcare reform should not be used** as an opportunity to use Federal funds to pay for **elective abortions.**	Abortion	Republican
I am **against abortion** and everybody cheers	Abortion	Democrat
Mr. President. on January 6. 1987. I introduced a bill to **eliminate the tax exempt and tax deductible status** of organizations that perform finance or **provide facilities for abortion.**	Abortion	Republican
I would urge President Obama to veto the Credit Cardholders Bill of Rights and send it back to Congress to **take the guns out.**	Firearm	Democrat
This historic Congress should take a stand in support of the rights of gun owners and all people who cherish freedom.	Firearm	Republican
I hope we **support closing the gun show loophole.** I also hope we support the assault weapons ban.	Firearm	Democrat
We have an individual right to protect ourselves. We have **an individual right to own a firearm.**	Firearm	Republican