Text Classification Accuracy of Politically-Biased Language Models

Luke Zanuck
Williams College
lhz1@williams.edu

Rick Yanashita
Williams College
ry5@williams.edu

Abstract

Large language models (LLMs) have been shown to carry the political biases of their training data. We investigate how political bias influences an LLM's performance on the text classification task, focusing on political ideology detection (PID). Using a transformed version of the Ideological Books Corpus (IBC), we evaluate models in two phases: zero-shot evaluation and fine-tuned evaluation. In the zero-shot phase, we employ HuggingFace pipelines and the OpenAI API to test various zero-shot models on PID with both three-label [Liberal, Conservative, Neutral] and two-label [Liberal, Conservative] setups. We observe a tendency for the models to overpredict "Neutral" foremostly, with the label aligning with their inferred political bias (bias label) second. In the two-label setup, the bias label is overpredicted. In the second phase, we fine-tune BERT-base for the PID task. The fine-tuned model avoids the Neutral overprediction observed in zero-shot evaluation. Our findings suggest that political biases in LLMs influence performance on the PID task with respect to the bias label both in zero-shot and brief fine-tuning scenarios, with the possibility of being mitigated through additional thorough fine-tuning and larger models.[1]

1 Introduction

As LLMs have become increasingly complex and widely adopted for everyday use, reliance on their responses has grown significantly. While these models offer great utility, they are susceptible to biases, which oftentimes diminish their aptitude without the user realizing. In some cases, this poses serious risk (Blodgett et al., 2020). One area where such biases are particularly concerning is the realm of politics and political science, a domain characterized by heightened tensions and rampant misinformation. Due to the prevalence and accessibility of polarized source data online, specifically social media and news, LLMs tend to exhibit political leanings embedded in their training data (Feng et al., 2023). These leanings can potentially lead to biases in the model's outputs on downstream tasks. Ideally, LLMs would be trained on unbiased data and be politically neutral. However, this is not realistic nor feasible. While it is beyond this study's capability to "fix" these biases, identifying and understanding their impact is a crucial step forward.

In this paper, we investigate whether the political leanings of LLMs influence their performance on text classification tasks, specifically focusing on political ideology detection (PID). Text classification, and PID in particular, was chosen because it provides insight into the model's "stance" on politically charged statements, offering a meaningful foundation for evaluating the effects of political bias. To address this problem, we first review prior research on political bias in LLMs and their implications in the Related Work section. Next, we describe the dataset we used for the PID task in the Dataset section. The Methods section outlines our experimental setup, including the zero-shot and fine-tuning phases, as well as evaluation metrics. We then present and analyze our findings in the Results section. Finally, in the Ethics & Limitations section, we reflect on the ethical considerations and impacts of the results and discuss the limitations of the study.

2 Related work

The relationship between political bias and LLMs has been a growing area of research. Several studies have explored and analyzed the political leanings of LLMs, injection of political bias, and subsequent effects on performance on downstream tasks.

The work that serves as primary inspiration for this study is Feng et al. (2023)'s paper. They construct a systematic framework for evaluating the political leanings of LLMs, using a traditional po-

[1]Our codebase can be found at https://github.com/luhaza/pid-of-nlp-models.

litical compass test. The models were prompted with questions about political statements and opinions, their answers were mapped to [strongly disagree, disagree, agree, and strongly agree], and were ultimately plotted on the social-economic ideology graph. Furthermore, the study found that pre-training the models on different partisan data significantly shifted political leaning. Lastly, they evaluated performance on hate speech and misinformation detection.

The IBC has also been used to improve PID accuracy. For example, recursive neural networks, artificial neural networks created by applying the same weights recursively, have been utilized by Iyyer et al. (2014). Their approach improves on the state-of-art PID models at the time. Their work on ideology detection is foundational to our study, as it provides the dataset that we use in our experiments. Additionally, they use the Convote dataset, consisting of data of the U.S Congressional floor debate transcripts from 2005.

Work has been done regarding the analysis of the political bias of a specific model. GPT-3 by OpenAI (Brown et al., 2020) is an autoregressive language model with 175 billion parameters. Although deprecated now with the release of GPT-4o family, the analysis on the political bias of GPT-3 remains relevant. The model is found to replicate the ideology present in an input text, but also tends to bias results toward left-leaning-sentiment more frequently. Sentiment analysis and RNNs are used in this GPT-3 analysis, which was successful at classifying political speech on a left-right scale. To do so, GPT-3 was instructed to write essays on 50 contemporary political issues (Gover, 2023).

One study evaluates the political bias of LLMs concerning political issues in the European Union, focusing on perspectives from German voters. The models were assessed based on their agreement with statements from the Wahl-O-Mat, a tool designed to help voters understand how political parties align with their views. The findings indicate that larger models, such as Llama-3-70B, tend to align more closely with left-leaning political parties, while smaller models maintain a neutral stance (Rettenberger et al., 2024). This paper highlights the role of model size in influencing political bias, which is an interesting factor that has influenced our project in different ways.

Together, these studies demonstrate the importance of examining political biases in LLMs, and provide valuable background work for our study.

3 Dataset

The Ideological Books Corpus (Sim et al., 2013) contains sentences annotated with political affiliation. We received the data with sub-sentential annotations (Iyyer et al., 2014). This means that the dataset was expressed as a tree, so we needed to transform the data for our purposes (.csv file, with [sentence, label] structure). We found that the transformed version of the IBC contained duplicates, some even with conflicting labels. It is unclear if the duplicates came from the original dataset, the tree structure from Iyyer et al., or our transformation. Regardless, these examples were removed before training and evaluation.

Table 1 details how we broke down the dataset. The IBC began with 4,326 examples. Removing 62 duplicates left us with 4,264 valid examples, split into 1997 Liberal, 1675 Conservative, and 592 Neutral. We randomly partitioned the remaining examples into three subsets; training (n=3000), evaluation (n=750), and sample (n=150). 364 examples were not used. This was to prevent any confounding factors due to unequally weighted training or evaluation, as there was an uneven distribution of labels in the original dataset. The training subset was split 80/20 into training/test sets and was used for finetuning models. The sample and evaluation subsets were both used for evaluating the models, with the sample subset being used for debugging and preliminary results and the evaluation subset for final results.

Evaluation does not include any Neutral examples. This was due to 1) a small number of examples relative to the other labels and 2) an attempt to mitigate overprediction of neutral, as seen in early results from sample. The "label" for the subsets takes both numeric ID and label form, depending

Dataset	*n*	Split
IBC (Original)	4326	2025/600/1701
IBC (Clean)	4264	1997/592/1675
Training + Test	3000	1250/500/1250
Evaluation	750	375/0/375
Sample	150	50/50/50

Table 1: Breakdown of the IBC dataset. Splits are given as Liberal/Neutral/Conservative.

on the use case.[2]

4 Methods

We chose models from across the political compass spectrum presented by Feng et al. (2023). Then, we evaluated each model's performance on the PID task twice. Rather than evaluating accuracy or some other similar metric, performance was determined to be raw prediction counts. We first evaluated the models on the PID three-label task, meaning the options were [Liberal, Conservative, Neutral]. Preliminary results showed a tendency for models to overpredict Neutral, so we removed Neutral as an option and ran the task again.

This was done in two phases, as some models could handle zero-shot tasks, while others needed fine-tuning.

4.1 Zero-Shot Evaluation

We began by evaluating the performance of zero-shot models, including GPT-4, GPT-4o, and GPT-4o mini, DeBERTa (He et al., 2021), and BART-large (Lewis et al., 2020). Of the GPT models, only GPT-4 was a part of the Feng et al. (2023) study. Thus, we are heuristically applying the label associated with GPT-4, Liberal, to the other GPT models. Similarly, DeBERTa was not a part of the original study, but builds on RoBERTa (Zhuang et al., 2021) and BERT (Devlin et al., 2019) (both Conservative) and is designed for zero-shot classification. We say more about this in the Ethics & Limitations section.

Zero-shot is ideal for our purposes as we are interested in the impact of political bias and not objective performance. Though beyond the extent of this study, thorough task-specific fine-tuning has the potential to mitigate the bias by improving performance on the PID task.

We utilized the OpenAI API to interact with the GPT models. We prepared a query containing 1) the task and 2) the example. This was repeated for each example in the evaluation subset. For the other models, we used the HuggingFace transformers library for pipelining the zero-shot classification task.

4.2 Fine-tuned Model Evaluation

Not all models are able to competently handle zero-shot tasks. To this end, we also fine-tuned BERT-base for PID. BERT-base was finetuned on the training subset with 4 epochs and batch sizes of 16. Furthermore, we elected to just train the classification head and topmost pooling layers in order to reduce computational cost. And unlike the zero-shot models, fine-tuned BERT-base did not overpredict Neutral, so we did not make a second pass. The same evaluation subset was used.

[2]Liberal: 0, Neutral: 1, Conservative: 2.

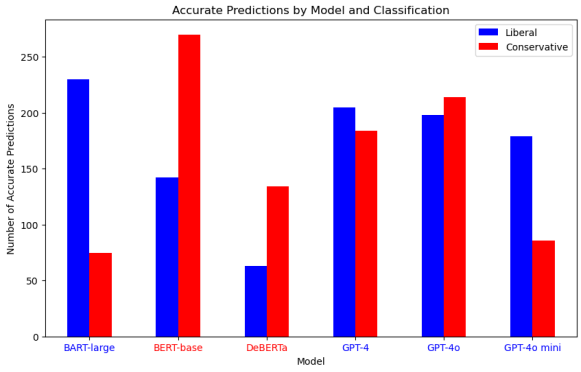

Figure 1: Accurate predictions by label, by model. Generally see higher accuracy of the bias label.

5 Results

We classified the results from each model into six groups, based on the distribution of the frequency counts (Table 2). Only one model (GPT-4o) does not predict the bias label more often than the opposite label. Four out of the six (GPT-4o mini, BERT-base, DeBERTa, and BART-large) overpredict the bias label significantly at a rate greater than or equal to 2x. GPT-4 sees a lesser ratio compared to those four, but the disparity does increase in the two-label task.

We also looked at the number of accurate predictions by class, by model (Figure 2). These results reflect those of Table 2, generally models accurately predict the bias label more often than the opposite label.

The open-source models are more distinctly biased than the closed-source GPT models. This could be due to a number of reasons that are open to further investigation, but one hypothesis is that the weights have changed since the Feng et al. (2023) paper's release. Again, more of this is discussed in the Ethics & Limitations section. We are also curious as to why the results of GPT-4o and GPT-4o mini are vastly different. According to the OpenAI website, GPT-4o mini is intended to be a high-accuracy, low-cost option with less parameters. Feng et al. (2023) note that differently-sized models within the same family could have non-negligible differences in political leanings, but

		3 Labels			2 Labels		
Model	**Bias**	**Liberal**	**Neutral**	**Conservative**	**Liberal**	**Other**	**Conservative**
GPT 4	Liberal	259	263	228	404	47	299
GPT 4o	Liberal	242	241	267	360	1	389
GPT 4o mini	Liberal	248	402	100	540	1	209
BERT-base	Conservative	216	60	474	–	–	–
DeBERTa	Conservative	143	302	305	237	0	513
BART-large	Liberal	461	125	164	556	0	194

Table 2: Results from the models across 3 Labels and 2 Labels. Purple denotes instances where the bias label count is $\geq$ 2x the opposite label count, pink denotes instances where the bias label count is $\geq$ the opposite label count, red denotes instances where the bias label count is < the opposite label count, blue denotes instances where Neutral is not the leading classification, and white is none of the above.

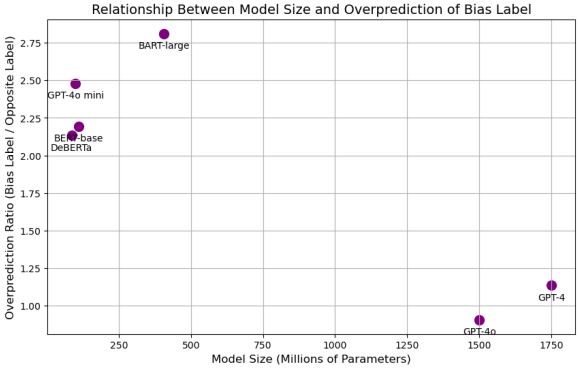

Figure 2: Relationship between model size and overprediction ratio. OpenAI does not disclose model size, so the GPT numbers are estimates gathered from online.

their examples were a lot closer than this difference would suggest.

To this end, we plotted the relationship between overprediction ratio and model size (Figure 2), and found that most of the smaller models exhibited a higher overprediction ratio. Although the evaluation of more models is necessary to generalize this result, this shows that there is a significant amount of variability among models due to other factors outside of training data, including size and complexity.

6 Ethics & Limitations

Based on our results, an important ethical consideration is the risk of these language models reinforcing harmful biases and stereotypes due to overprediction. If the models consistently overpredict on their political-leaning, it might cause neutral content to be incorrectly classified, distorting the perception of the text. For users of LLMs not politically informed, it may mean that their perceptions of certain political groups may be affected. Furthermore, the overprediction bias could disproportionately affect certain groups by mislabeling their stance or intention.

There are also ethical considerations about our study as a whole. There must exist a balance between pushing to further the horizons of LLM technology and ensuring that these models follow ethical principles while remaining as neutral as possible. While ethical guidelines are sometimes seen as constraints, they must remain and serve as foundational pillars that guide the development of LLMs in ways that will help, not hurt, all members of society. That being said, there are elements of this study that could be more fair. For example, we could evaluate the training data for language models more carefully, and ensure that our results cannot be misinterpreted. Though the the models overpredict the bias label in the PID task by a decently large margin, they were restricted to a relatively generalized 3-label set [“Conservative”, “Neutral”, “Liberal”]. The political ideology spectrum is wide and complex, so this generalization could suppress fine-grained differences and amplify certain political biases in unrealistic ways.

Furthermore, we acknowledge the existance of a few limitations of our experiments. First and foremost, as undergraduate students we are constrained by computational, monetary, and time costs. It takes a significant amount of time for our computers to run processes. To expand further, we would need more time with high-powered GPUs. This would allow us to train and evaluate models on larger datasets like NewB (Wei, 2023), and incorporate a greater quantity of larger LLMs such as Llama-3.2-1B. Additionally, the cost of using the OpenAI API is not insignificant, especially for GPT-4 (unfortunate, as it is the only model from the Feng paper that is not defunct). Larger queries

would only continue to impose a financial burden on us.

If we were to extend this work, we would likely exclude OpenAI models from our experiments not only are they expensive to use, but they are harder to work with and analyze due to being closed-source. To this end, it is uncertain if the GPT-4 of 2023 is the same as we are testing now, just before 2025. It is much more straightforward to draw inferences from the other open-source models in our study. We also spent considerable time attempting to instruction-tune Llama, which turned out to be improbable. In hindsight, we would instead fine-tune another model, like BART-base, the model closest to center.

Some other considerations are that the IBC is an older dataset created in the early 2010s. The politics of now are almost certainly different from then, specifically Post-Trump, as highlighted by Feng et al. (2023). This could make the IBC dataset outdated for modern analysis as the definitions of what is Conservative or Liberal may have evolved. Finally, most models have poor performance around or below the baseline (50%); it would be remiss to not mention the possibility of some of the discrepancies being due to the unfamiliarity with the task.

7 Conclusion

Our results reveal that language models both consistently overpredict and have a higher accuracy for their political leanings on the PID task. Additionally, the models exhibit relatively low accuracy in correctly identifying political ideologies, indicating that these tasks are challenging for even state-of-the-art language models.

We have learned that language models have political biases, and that these political biases influence their text-classification abilities. We have also learned that this is an ever-growing and developing subfield of computational social science, which only serves to continue to do so in our current political climate.

This project has shown us that people should generally avoid interacting with LLMs about politics. Individuals should make their own electoral decisions. It is promising, however, that there is work being done to decrease the political bias inherent in LLMs and inform the general population of this bias.

If we had six more months to continue our project, there are a few things we would like to try. Firstly, we would like to instruction-tune a quantized Llama model on the political ideology detection task. We attempted to instruction-tune the base Llama-3.2-1B model using the Hugging Face transformers package, and also tried using the Peft package to perform low-rank adaptation (LoRA) on the Llama-3.2-1B and the unsloth/Llama-3.2-1B-Instruct-bnb-4bit model. Unfortunately, we were unsuccessful, running into issues with runtime and padding (`https://github.com/meta-llama/llama3/issues/42`). Even with some help from our professor, Katie, we were unable to resolve the issues before the deadline of the project. Given the number of articles about issues with fine-tuning/instruction tuning Llama, it seems Meta is trying to limit access to its models. Given more time on this project, we would like to complete the instruction tuning of Llama on the IBC and evaluate its text classification accuracy.

Furthermore, we would like to evaluate more models in order to generalize our findings. Specifically, it would be interesting to fine-tune the BART-base model because of its neutral political ideology. As a "neutral" model, it would be an interesting baseline to compare the political-biases of more partisan models to.

We would also like to evaluate PID with more fine-grained ideologies (more labels), more inclusive than "Conservative", "Liberal", and "Neutral." This may yield interesting results. For example, labels could be in the realm of: "Traditional Conservative", "Fiscal Conservative", "Neoconservative", "Classical Liberal", "Modern Liberal", "Progressive", "Moderate", "Populist", or "Technocratic." It could be interesting to see if models have more specific political leanings, or if these extra labels introduce too much complexity to this NLP task.

Lastly, the full IBC was on the order of 4000 sentence-label pairs. This dataset was perfect for the scope of the project and the resources we were given. However, if we had more time and computational power, we would like to use the Newspaper Bias Dataset (NewB). It contains more than 200,000 sentences for political bias detection and is a corpus made of sentences from eleven news sources regarding Donald Trump (Wei, 2023). This dataset is revolutionary and modern in two ways. First, it has eleven labels for the eleven news sources it gets data from. Each news source's political views will be slightly different, providing

a more complex labeling structure. Second, it is made of sentences regarding Donald Trump, an extremely polarizing figure in modern politics. It makes the dataset very relevant to today's society, unlike the IBC.

References

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

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. Association for Computational Linguistics.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Lucas Gover. 2023. Political bias in large language models. *The Commons: Puget Sound Journal of Politics*, 4(1):2.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Luca Rettenberger, Markus Reischl, and Mark Schutera. 2024. Assessing political bias in large language models. *arXiv preprint arXiv:2405.13041*.

Yanchuan Sim, Brice D. L. 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, Seattle, Washington, USA. Association for Computational Linguistics.

Jerry Wei. 2023. Newb: 200,000+ sentences for political bias detection.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.