Dean Casey
Meghane Saidenberg-Coriston

Do LLMs trained on different bodies of text generate headlines for news stories with different political biases?

With this project, we explored whether Large Language Models (LLMs) trained on different corpora of data generate politically biased headlines. In the context of an increasingly politically polarized US, the potential influence of social media companies (e.g. Facebook) over their users' political behavior through consumption of their algorithmic products is essential to consider within the context of data ethics. The manner in which one synthesizes information can rely heavily upon how one is primed (i.e. *anchoring bias* - filtering information through an initial framework one has mentally set up). As LLMs are increasingly freely available for online personal use, it is important to identify the risk of potential biases which can be imbued into their output. CNN and Fox News have a left and right-leaning bias respectively, influencing our decision to isolate 'Politics' articles from these sources as our corpora of data to investigate how data selection in training LLMs impacts their output.

We web-scraped headlines and body text for articles from 2022 for CNN and Fox News. For CNN, we used the Requests and BeautifulSoup modules to write a Python script to run locally. To generate Fox News data, we used a closed-source existing API from Apify by user hanatsai.

The LLM training process involved fine-tuning an existing open-source LLM (Mistrial 7B Instruct model and the Unsloth module) on the web-scraped CNN and Fox data (both n=7000). The 'bodies' refer to the article stories, and 'titles' their corresponding headlines. We trained each LLM such that it learned the writing style of the articles and corresponding titles, which we then fine-tuned, saved to Hugging Face, and used to generate new headlines on our test dataset. We imported the trained models back from Hugging Face, set up a pipeline to prompt the models, and retrieved their responses for each test article. These responses were then consolidated and grouped based on the model and test set they were generated from to be analyzed.

The first part of the analysis utilized the GPT-4 API to evaluate political bias sentiment. We utilized the OpenAI modules in a notebook such that GPT-4 could systematically 'opine' the generated headlines' political bias sentiment. For each model and test dataset, we counted the number of generated titles that were classified by GPT4 as "conservative", "liberal", and

“neutral”, and compared their distributions. This was bootstrapped with 20 samples and tested for significant distribution differences.

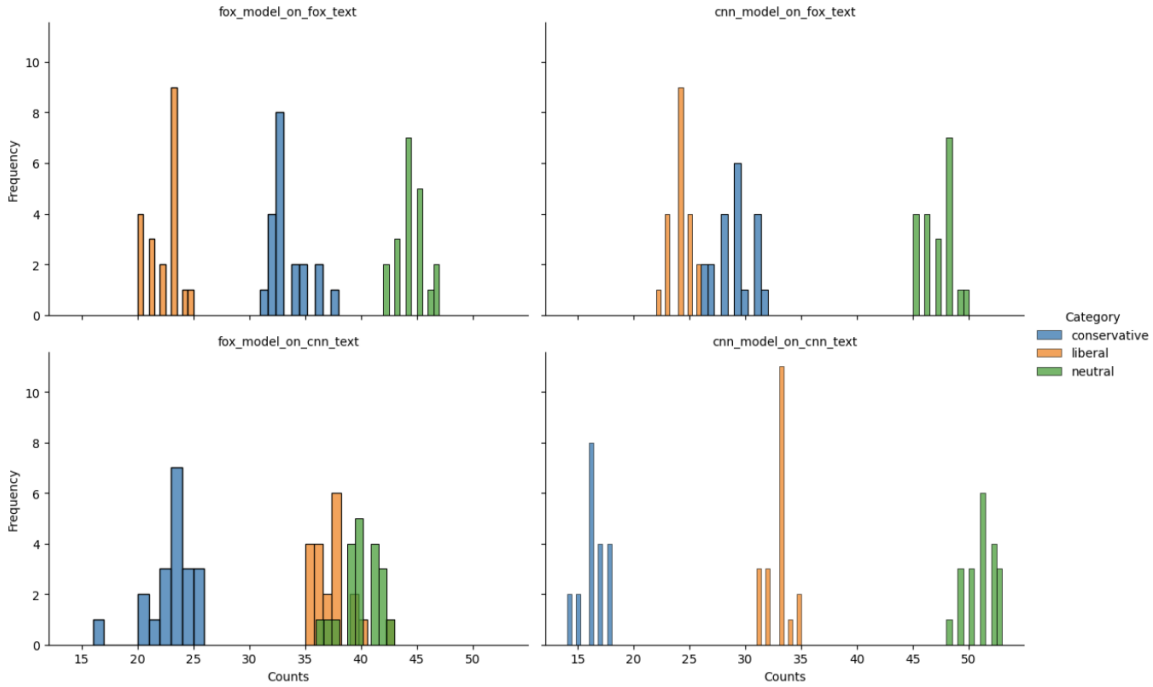

Figure 1: Histogram of distribution of political sentiment labels on each model/test set

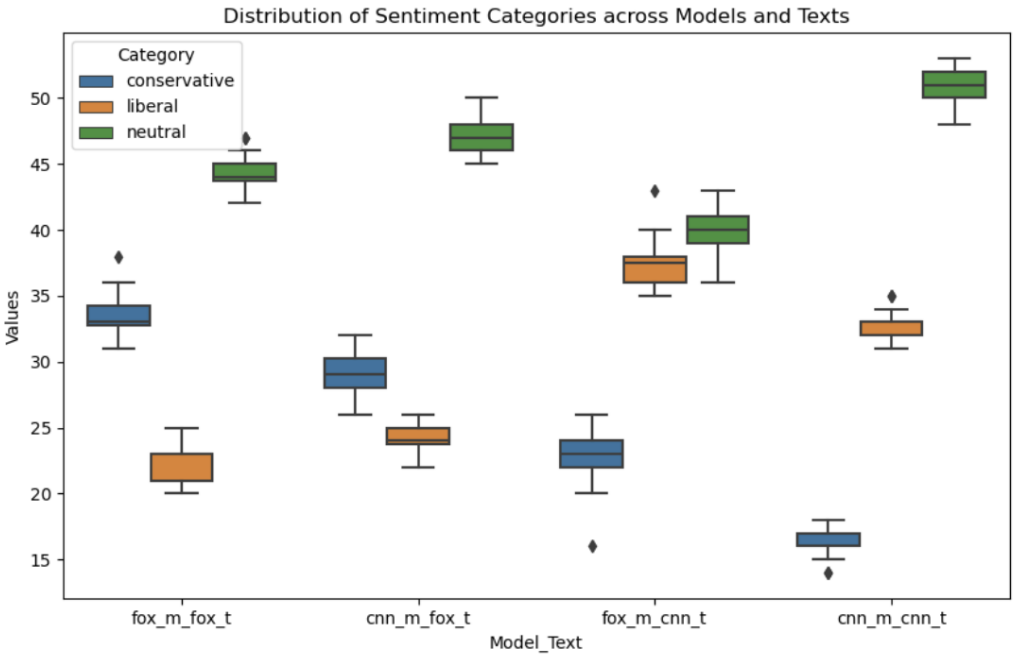

Figure 2: Boxplot of distribution of political sentiment labels on each model/test set

		Values		
		mean	std	median
Model_Text	Category			
cnn_model_on_cnn_text	conservative	16.30	1.218282	16.0
	liberal	32.80	1.105013	33.0
	neutral	50.90	1.447321	51.0
cnn_model_on_fox_text	conservative	28.90	1.713722	29.0
	liberal	24.10	1.020836	24.0
	neutral	47.00	1.450953	47.0
fox_model_on_cnn_text	conservative	22.65	2.254236	23.0
	liberal	37.35	2.007224	37.5
	neutral	40.00	1.747178	40.0
fox_model_on_fox_text	conservative	33.55	1.700619	33.0
	liberal	22.15	1.460894	23.0
	neutral	44.30	1.380313	44.0,

Figure 3: Summary statistics on distribution of political sentiment labels on each model/test set

	Text	Category	T-statistic	P-value
0	fox_text	conservative	8.613366	1.821160e-10
1	fox_text	liberal	-4.893141	1.852708e-05
2	fox_text	neutral	-6.029455	5.186994e-07
3	cnn_text	conservative	11.082692	1.817780e-13
4	cnn_text	liberal	8.880691	8.303584e-11
5	cnn_text	neutral	-21.485668	7.361741e-23

Figure 4: Statistical significance between Fox & CNN models for each test dataset (fox_text and cnn_text)

Much of the variation in political sentiment seems to come from the test dataset when comparing the outputs from the heterogeneous model/text combinations (e.g. mean ‘liberal’-labeled headlines for CNN-model-CNN-text=32.8; corresponding mean for CNN-model-Fox-text=24.1).

The interesting part of the analysis comes from comparing the opposing models that were prompted using the same test data. All categories across both test datasets were significantly different, indicating that both models carried significant weight/bias (e.g. mean ‘conservative’-labeled headlines for CNN-model-CNN-text=16.3; corresponding mean for Fox-model-CNN-text=22.65). We were surprised that the Fox model yielded a higher ‘liberal’-label count than the CNN model (37.35 vs 32.80) on CNN-text.

However, one cannot extrapolate news sources to ideology; ‘liberal’, ‘conservative’ and ‘neutral’ labels are the genesis for understanding one's political belief system rather than its end point. Attempts to boil down myriad beliefs to nebulous labels risk being reductive, as is the case with the limited scope of the GPT-generated labeling output.

The second part of the analysis focused on the generated headlines’ distinguishability by the training corpus provenance of their LLMs (i.e. can we predict which model the generated headline comes from). We used the sklearn module to vectorise the data for classification (Multinomial and Bernoulli Naive Bayes, Logistic Regression, SVM), and

tested significance for models' performance over 100 simulations (p=0.05) compared to a naive model (prob{'Fox', 'CNN'}=0.5) with an 80/20 split for the simulated train/test sets.

	Multinomial Naïve Bayes	**Naïve Bayes**	**Logistic Regression**	**Linear Support Vector Machine**
All test data (Fox model with CNN stories, Fox model with Fox stories, CNN model with Fox Stories, CNN model with CNN stories)	mean of simulations: 0.554 standard deviation: 0.04105151973712146 t-statistic: 1.315420241340535 p-value: 0.09494495251892054	mean of simulations: 0.549125 standard deviation: 0.03400158752978241 t-statistic: 1.4447854811770415 p-value: 0.07504941355509742	mean of simulations: 0.571125 standard deviation: 0.041503278135101525 t-statistic: 1.7137200528708552 p-value: 0.04407258968463945	mean of simulations: 0.57075 standard deviation: 0.03706481491995482 t-statistic: 1.9088183807956869 p-value: 0.028865870584248277
Heterogenous model + text Combination (CNN model with Fox stories and Fox stories with CNN stories)	mean of simulations: 0.5295000000000001 standard deviation: 0.07065530211648463 t-statistic: 0.4175199753780038 p-value: 0.3383752505190907	mean of simulations: 0.5259999999999999 standard deviation: 0.06935125059864428 t-statistic: 0.37490311675083443 p-value: 0.3540668877230435	mean of simulations: 0.5315000000000001 standard deviation: 0.06312765645975629 t-statistic: 0.49898890227425513 p-value: 0.3091705672755751	mean of simulations: 0.5262499999999999 standard deviation: 0.06320311573319656 t-statistic: 0.41532762579001203 p-value: 0.33917590031648026
Homogenous model + text combinations (CNN model with CNN stories and Fox model with Fox stories)	mean of simulations: 0.606 standard deviation: 0.06186495556672394 t-statistic: 1.713409458213779 p-value: 0.044101193919563575	mean of simulations: 0.5742499999999999 standard deviation: 0.06699304577206638 t-statistic: 1.1083239930995856 p-value: 0.13453319293883492	mean of simulations: 0.61675 standard deviation: 0.072217317238256 t-statistic: 1.6166482564676816 p-value: 0.053773171289072774	mean of simulations: 0.62125 standard deviation: 0.06594064901044873 t-statistic: 1.8387747439487152 p-value: 0.033723063262548036

Figure 5: Table of summary statistics for classification models and model/text combinations. Text in red signifies statistically significant results.

We observed how the models containing all the test data (n=320) appear to yield more statistically significant means (LR, SVM) with more concentrated distributions compared to the smaller heterogeneous and homogeneous isolated data (both n=160). The accuracy and significance of the homogenous data (MNB, SVM) versus the heterogeneous data could suggest a comparatively more cohesive news-source data generation and was subsequently more highly weighted in the 'all-test-data' result generation. However, the heterogeneous data is unlikely to just be noise, as accuracy improved with its inclusion. It is worth noting that we used a small training corpus of data with crude success criteria rather than a neutral 'control' test set. We wish to explore this further with more data to see if this (potential) slight divergence increases with a larger corpus of data.

The limitations of our project's results are numerous. The web scraping processes for Fox News and CNN diverged due to differing data management systems. Fox data availability was contingent on buying compute, further hampered by inability to limit the crawler date range. Furthermore, we only used one base LLM to train our models on ~7,000 data points (~45MB of data), which is very little compared to general LLM training corpus size.

Additionally, it's important to note that we only used a small subset of data: 'Politics' articles from 2022. One could argue that these would be the most politically biased articles, and wouldn't be representative of a diverse corpus of training data. Finally, due to compute/time constraints, we don't show any 'control'; we don't have an LLM trained on both datasets, and our testing data is inherently biased, with no objective 'neutral' news source to test on.

Our conclusion is that there are *possible* indicators of politically biased output based on LLM training corpus, but our results contain several caveats due to data availability and computational limitations. This is a promising start to what we hope will be a more all-encompassing project going forward.

Bibliography

Agudo, U., & Matute, H. (2021). The influence of algorithms on political and dating decisions. *PloS one*, *16*(4), e0249454. https://doi.org/10.1371/journal.pone.0249454

Pilat D., & Sekoul D. (2021). Anchoring Bias. The Decision Lab. Retrieved May 16, 2024, from https://thedecisionlab.com/biases/anchoring-bias

Teovanović P. Individual Differences in Anchoring Effect: Evidence for the Role of Insufficient Adjustment. Eur J Psychol. 2019 Feb 28;15(1):8-24. doi: 10.5964/ejop.v15i1.1691. PMID: 30915170; PMCID: PMC6396698.

Weatherly, J. N., Petros, T. V., Christopherson, K. M., & Haugen, E. N. (2007). Perceptions of political bias in the headlines of two major news organizations. *Harvard International Journal of Press/Politics*, *12*(2), 91–104. https://doi.org/10.1177/1081180x07299804

Nodematic Tutorials. “Mistral Fine Tuning for Dummies (with 16K, 32K, 128K+ Context).” *YouTube*, YouTube, 14 Mar. 2024, www.youtube.com/watch?v=rANv5BVcR5k&ab_channel=NodematicTutorials.

Joomi K. (n.d.). *Sentiment Analysis on Twitter Data*. Retrieved May 13, 2024, from https://joomik.github.io/sentiment/

CHATGPT Pricing, openai.com/chatgpt/pricing. Accessed 16 May 2024.