Comparing Prominent Generative Language Models for Classifying Political Alignment Of Limited Context Bigrams

Sankalp Singh

Abstract

Generative Language Models (GLMs) have transformed artificial intelligence by enabling human-like text generation across diverse applications. This study delves into GLM-generated content, focusing on the ability of GLMs to classify politically charged bigrams from congressional speeches with minimal context by creating a Python[1] script for each GLM to prompt the models en masse. The investigation studies three major GLMs: Google's Bard, OpenAI's GPT-3.5 Turbo, and OpenAI's GPT-4. Using prompts encompassing target bigrams, congress details, and polarity values, the study assesses the models' proficiency in aligning bigrams with left-leaning or right-leaning ideologies. The dataset originates from Stanford University, comprising of parsed political bigrams from congressional speeches and corresponding political polarity values for each bigram. Despite expected deviations from the exact Stanford benchmark polarity values, the GLMs show varying degrees of accuracy in political classification, with GPT-4 exhibiting the highest proficiency. The findings underline GLMs' capacity to consider context and infer political associations based on their training data. They also emphasize the complexities of language, ideology, and context. This research contributes to understanding GLMs' strengths, limitations, and implications in political discourse analysis.

Introduction

Generative language models represent a groundbreaking advancement in artificial intelligence (AI), revolutionizing the way machines comprehend and produce human-like text. Their significance spans across a spectrum of applications, encompassing creative content generation, language translation, text summarization, and even human-like conversation. By harnessing the power of deep learning techniques, GLMs have unlocked unprecedented capabilities, contributing to the evolution of various industries and sectors.

As versatile and impressive as GLMs are, they are not without their shortcomings. One of the glaring concerns associated with these models is their susceptibility to generating misinformation. This vulnerability arises from the nature of their training data, which, despite being extensive, can inadvertently encompass erroneous, biased, or unverified information. Consequently, GLMs may inadvertently propagate false narratives, as they may lack the discernment to differentiate between accurate and inaccurate content.

[1] https://github.com/cubicmight/GLMPoliticalTest

An illustrative study by Bolukbasi et al. (2016) brought to light the issue of bias in GLMs. In their work, they exposed how word embeddings trained on datasets of news articles could inadvertently perpetuate biased or prejudiced language, thereby fostering negativity toward certain societal groups. The implications of this discovery are profound, as it sheds light on the unintended consequences of deploying AI models in contexts that demand neutrality and accuracy.

Against this backdrop, this research elucidates the models' potential capacity to replicate realistic nuanced political discourse. We make this comparison because certain standout terms and bigrams in prompts to language models can greatly affect the information they output. The investigation focuses on evaluating the proficiency of three prominent GLMs in accurately classifying politically charged bigrams with limited contextual information. The contextual cues provided to the models include the specific congress during which the bigram was articulated, alongside the corresponding start and end dates of that legislative session. These four contextual elements, along with directions, collectively serve as the input prompt for the GLMs to generate a polarity value, indicating the model's assessment of the bigram’s alignment with left-leaning or right-leaning political ideologies.

To benchmark the generated polarity values against a credible and well-established source, the research employs the Congressional Record for the 43rd-114th Congresses: Parsed Speeches and Phrase Counts dataset provided by Stanford University. Even though the primary nature of this task involves regression, wherein the model generates a continuous numerical value on a scale, we are opting to treat it as a classification task. In doing so, we aim to ascertain whether the model arrives at the same conclusion as the original polarity values in terms of the political standing conveyed. Here, political standing is specifically defined by the ideological association attributed to a given bigram, signified by the polarity value's sign. In this context, negative values correspond to left-leaning affiliations, whereas positive values indicate a right-leaning inclination. We did this because of two key factors: First, the realization that the polarity values provided by Stanford might not comprehensively capture the entire spectrum of global political intricacies. Second, the constrained information environment within our study prevents the model from replicating the exact numerical values of the original polarity assessments.

Data

The data utilized for this research is obtained from Stanford University's dataset named "Congressional Record for the 43rd-114th Congresses: Parsed Speeches and Phrase Counts dataset." This extensive dataset is organized to encompass the proceedings of diverse congressional sessions, offering polarity values linked to specific bigrams. The core emphasis of this study centers on Congresses 95 through 114.

The calculation of partisanship bigrams in the Stanford database is achieved by computing the expected posterior probabilities for a given bigram being associated with either the Republican or Democratic party. This involves incorporating parameters like probabilities of party affiliation (qR and qD) and a polarity function (ρt(xit)) that captures session-specific context. The partisanship of a particular bigram in a session is quantified as an average of these values across active speakers in that session. This measurement offers both the direction and magnitude of partisanship, with positive values representing Republican-leaning bigrams, negative values indicating Democratic-leaning bigrams, and the absolute value signifying the extent of partisanship. The model evaluates the political leaning of individual bigrams within specific legislative sessions.

To ensure the data's quality and suitability for analysis in this study, a series of preprocessing steps were applied. The initial phase involved addressing issues with misspelled words within bigrams that were present in Stanford's dataset. These misspellings likely occurred during the creation of the Stanford dataset when stemming and lemmatization techniques were applied, as most of the misspelled bigrams were bigrams with words that were simply roots and were not real words. This step was executed using the Python programming language in combination with the Natural Language Toolkit (NLTK) library. Additionally, the NLTK corpus was used to identify and exclude non-English bigrams and bigrams that contained non-English words. Consequently, the dataset was refined to encompass exclusively English-language bigrams.

Moreover, we implemented a systematic removal of bigrams that comprised words with minimal occurrence. This elimination process was driven by the criterion that all words forming a bigram must exhibit a minimum frequency of five instances across the entire corpus. This measure mitigated inaccuracies arising from infrequent, obscure, or misspelled words. Furthermore, to eliminate instances of tautological phrasing, bigrams comprised of identical words were expunged from the dataset.

By enforcing this criterion, we aimed to enhance the reliability of the bigrams used in the analysis. We aimed to ensure that the models' interpretation of these bigrams remains robust and accurate. Removing bigrams with unfavorable characteristics was intended to prevent the models from encountering mistakes or misclassifications caused by factors such as misspellings.

Following the preparatory stages, we compiled a refined list of bigrams that exhibited the highest degrees of polarity across all congresses. From the expansive dataset spanning congresses 95 through 114, we chose the forty bigrams with the largest magnitudes in polarity scores.

In total, there were fifteen left-leaning terms and twenty-five right-leaning terms. In terms of their polarity value ranges, left-leaning bigrams ranged from a minimum of eleven to a maximum of negative one hundred and four, while right-leaning bigrams spanned from a minimum of seven to a maximum of eighty-five. Finally, we associated each bigram with its corresponding congress and assigned start and end dates based on the congress number to finish preprocessing the data.

Methodology and Results

Methodology

A standardized prompt (Figure 1) was then formulated for the GLMs, which included the target bigram, the relevant congress, and its associated start and end dates. The prompt also provided a clear framework for interpreting polarity values, where positive, negative, and neutral numbers represented right-leaning Republican, left-leaning Democratic, and neutral bigrams, respectively. The prompt also directed the model to limit the polarity values to a range of -110 to 110, aligning with the original polarity values of the forty within that specific range. To facilitate comprehension, example outputs were included, transforming the task into a few-shot learning scenario. In few-shot learning, the model is presented with a limited number of examples or data points, as was the case here with the provided instances of target bigrams, congressional contexts, start and end dates, and their associated polarity values.

In this context, few-shot learning allowed the model to acquire the ability to generalize from these limited examples and apply its learned knowledge to make accurate predictions for similar cases that it had not explicitly encountered during training. This capability is particularly valuable when dealing with nuanced and context-dependent tasks like our classification of political bigrams, where each bigram may have unique characteristics that aren't easily captured with the limited amount of data points we provide the model to conclude from.

Model testing involved the selection of three prominent GLMs: Google's Bard, OpenAI's GPT-3.5 Turbo, and OpenAI's GPT-4. A consistent prompt was presented to each model to ensure uniform input conditions. To account for potential response variance, each bigram was subjected to three trials per model. This approach aimed to capture a comprehensive range of potential model-generated outcomes and establish an average result for subsequent analysis.

The analysis phase entailed the averaging of the three trial results for each bigram and model. These averaged predicted polarity values were then graphed against the corresponding polarity values from the Stanford dataset. This graphical representation provided a visual means of comparing the predictions of the selected GLMs with established reference values.

Given the below phrase, the congress the phrase was said in, and the start and end date of that congress, and the fact that a positive number means that the phrase is a right-leaning Republican phrase, a negative number means that the phrase is a left-leaning Democratic phrase, and 0 means that it is a neutral phrase, output the polarity value for the given phrase. The polarity value should be on a scale of -110 to 110.
Example:
Phrase: red tape
Start Date: January 4, 1977
End Date: January 3, 1979
Congress: 95
Output: 30
Example:
Phrase: interest rate
Start Date: January 3, 2013
End Date: January 3, 2015
Congress: 113
Output: -100
Example:
Phrase: repeal afford
Start Date: January 3, 2013
End Date: January 3, 2015
Congress: 113
Output: -39
Phrase:
Start Date:
End Date:
Congress:

Figure 1: The prompt used for each of the GLMs

Results

The methodology employed in this research yields valuable insights into the performance and capabilities of the selected GLMs in accurately classifying politically charged bigrams with limited contextual information. The findings provide a comprehensive overview of the models' strengths, limitations, and potential real-world implications.

Firstly, it's noteworthy that the three GLMs chosen for this study – Google's Bard GLM, OpenAI's GPT-3.5 Turbo, and OpenAI's GPT-4 – demonstrate varying degrees of accuracy in predicting the political standing of the selected bigrams.

The analysis unveils a clear pattern in which Bard, demonstrating the lowest proficiency among the three, as depicted in Figure 2, incorrectly categorizes the political stance of 19 bigrams. GPT-3.5 Turbo displays a marginal performance improvement with 16 misclassifications, as depicted in Figure 3, whereas GPT-4 excels with the highest accuracy, misjudging only 8 bigrams, as depicted in Figure 4. This performance hierarchy indicates that more advanced iterations of GLMs are better equipped to infer the political leaning of the provided bigrams based on limited contextual cues

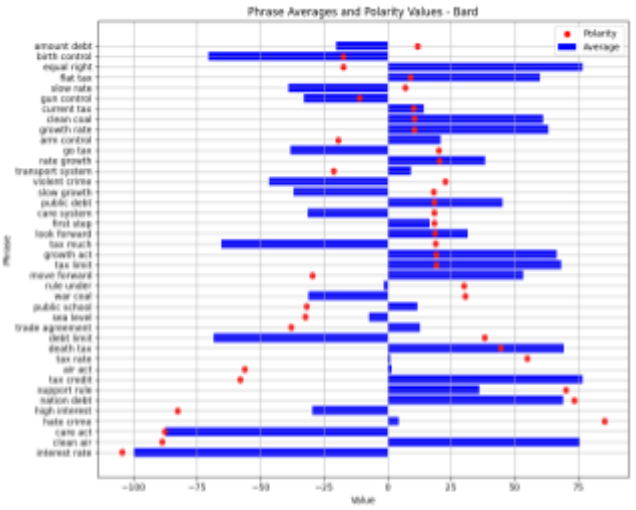

Figure 2: Bard Results. "Polarity" refers to the Stanford polarity value for each bigram. "Average" refers to the generated polarity value.

Additionally, it's noteworthy to examine the distribution of errors when considering left-leaning and right-leaning bigrams. Figure 2 shows that Bard incorrectly classifies 9 out of 15 left-leaning terms, yielding an error rate of 60%, and 10 out of 25 right-leaning terms, resulting in an error rate of 40%. As for GPT-3.5 Turbo, Figure 3 shows that the error rate for left-leaning terms stands at 6 out of 15, approximately 40%, while its error rate for right-leaning terms is 10 out of 25, approximately 40%. Finally, as seen in Figure 4, GPT-4's error rate for left-leaning terms is 4 out of 15, approximately 27%, while its error rate for right-leaning terms is 4 out of 25, approximately 16%.

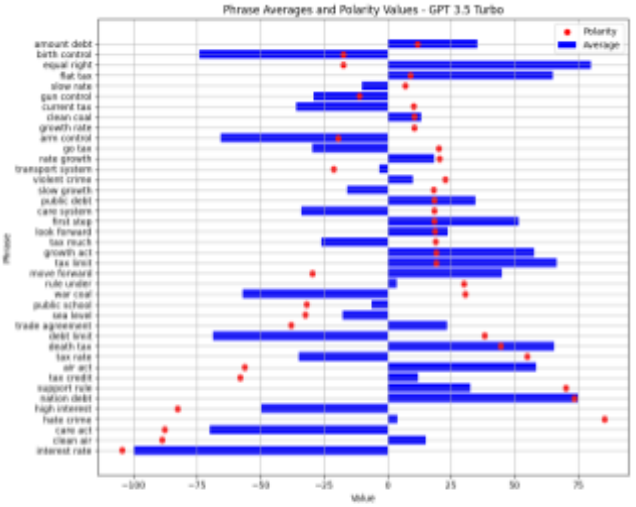

Figure 3: GPT-3.5 Turbo Results. "Polarity" refers to the Stanford polarity value for each bigram. "Average" refers to the generated polarity value.

From these error rates, we can again see that GPT 4 performed marginally better than the other two GLMs in both right and left-leaning contexts. However, we must take into account the fact that the amount of left and right-leaning bigrams is not equal. Because there are more right-leaning terms, the models had more chances to correctly classify terms, which means that the results are somewhat skewed.

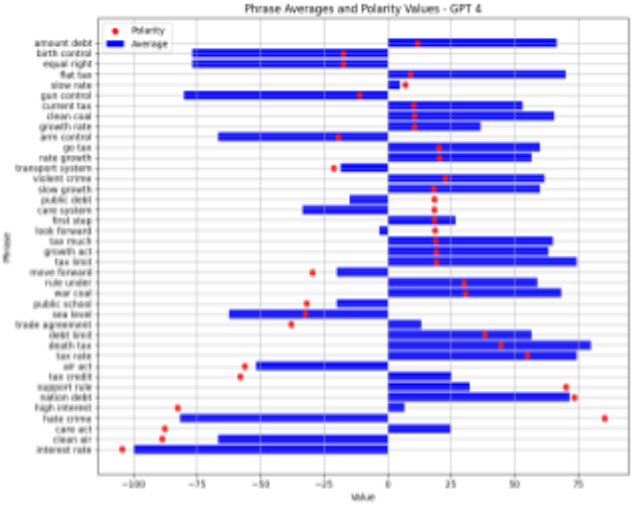

Figure 4: GPT-4 Results. "Polarity" refers to the Stanford polarity value for each bigram. "Average" refers to the generated polarity value.

However, by calculating error rates as percentages, we are standardizing the measure of errors across

different sample sizes. This normalization allows us to directly compare the models' accuracy in both left-leaning and right-leaning contexts. The key advantage of this approach is that it accounts for the fact that there are more right-leaning terms than left-leaning terms, ensuring that the models are not unfairly penalized for having to classify a larger number of right-leaning terms. Therefore, in low-context environments, it can be concluded that Bard demonstrates a relatively better performance in classifying right-leaning terms, although both of its error rates are relatively high. GPT-3.5 Turbo, on the other hand, exhibits roughly similar accuracy levels for both left-leaning and right-leaning terms, albeit with somewhat elevated error rates. Finally, GPT-4 excels in categorizing right-leaning terms with a notably lower error rate, with the left-leaning error rate not far behind.

The fact that these models exhibit any degree of accuracy in classifying the political standing of the bigrams is a testament to their ability to understand and interpret context (Zhang et al., 2023). The fundamental principle of context-based learning is evident here, where the models recognize patterns and associations between certain bigrams and political ideologies based on the training data. This phenomenon underscores the models' capacity to leverage extensive datasets to draw meaningful conclusions, even when the provided contextual cues are minimal during inference, as highlighted in Devlin et al.'s work on BERT (Devlin et al., 2018). BERT, a deep bidirectional transformer model for language understanding, demonstrated the significance of pre-training models on massive text corpora to develop a rich contextual understanding of language. Such contextual understanding enables models to capture nuanced relationships between words and bigrams, allowing them to make informed predictions in diverse contexts. The success of the models in predicting the general political standings of these bigrams suggests that they have effectively learned from the real-world usage of these bigrams in context, both in congressional discourse and online discussions.

These results highlight each of the models' capability to capture the nuanced ways in which these bigrams are employed in political communication across different contexts. It is worth noting that these bigrams in and of themselves are not politically charged. Rather, it is the contexts in which these bigrams are used which give them underlying connotations. By depriving the models of those politically charged contexts and simply providing them with a bigram, we constrain the models' understanding of the broader historical, socio-political, and cultural contexts that may influence the political connotations of a bigram and make them draw conclusions based on their training data. Due to the vast amount of training data, these models were able to use the context that these bigrams generally appeared in within their training data and draw conclusions based on those contexts, resulting in all three of them showing satisfactory levels of proficiency at this task.

Model performance in natural language processing tasks has shown correlation with the number of model parameters. Generally, larger models with more parameters tend to exhibit improved

performance due to their enhanced capacity to capture complex patterns and nuances in language (Brown et al., 2020). In this study, the same correlation holds true. Bard, with the least satisfactory performance, also has the least parameters, at 137 billion. GPT 3.5-turbo has the second most, at 175 billion, and GPT 4 has by far the most, at 1.7 trillion. Bard and GPT 3.5-turbo most likely performed similarly because they had fewer contexts to rely on when classifying the terms, whereas GPT 4 had a much larger sample size to conclude from.

With regard to the actual polarity values generated by the models, it's important to note that the polarity values provided by Stanford serve as benchmarks and do not encapsulate the full complexity of global political nuances. These values are, in fact, the output of a model themselves, and as such, they reflect the predictions made by that specific model. This distinction is important because it underscores that we are comparing one model's output to another's in this study. Therefore, we are not directly measuring each of the GLM's ability to match the "real world" political landscape. Despite this, the Stanford dataset is the largest, most comprehensive, and realistic polarity value database and, thus, was the best dataset for this study.

As a result, the models' generated polarity values reflect their interpretations based on the given prompt, and they are shaped by the patterns and associations they have learned from their training data. The deviations from the Stanford values can stem from a variety of factors, including the models' inherent biases, the nuances of political discourse not fully captured in the prompt, and the complex interplay of political ideologies that cannot be entirely distilled into a single polarity value.

Despite the deviations from the benchmark polarity values, these values remain significant for several reasons. First, they serve as a foundation for evaluating the GLMs' ability to generalize political associations from limited context. The fact that the models can even approximate the political leaning of bigrams based on these prompts underscores their capacity to understand the patterns of language use across different contexts in their training data.

Also, while the polarity values may not capture the entirety of global political intricacies, they represent an attempt to quantify and categorize political associations within a constrained framework. These values, though simplified, offer a measure of political alignment that can be useful for comparative analysis and for understanding the broad ideological connotations of bigrams.

Furthermore, the divergence between the models' predictions and the benchmark values highlights the complexities of language, ideology, and context. It emphasizes that the accurate interpretation of political sentiment requires a rich understanding of historical, cultural, and socio-political factors that extend beyond the immediate textual cues. This recognition

underscores the need for models to be continually refined and developed to better grasp the intricate layers of meaning embedded in human communication.

One of the significant takeaways from this analysis is the real-world application potential of GLMs. Their ability to interpret and generate contextually appropriate content has wide-ranging implications. Beyond the scope of this research, these models could be harnessed for tasks like fact-checking, content generation, and political analysis. However, the inherent limitations of context-driven learning must also be acknowledged, highlighting the importance of ongoing research and fine-tuning to enhance the models' discernment capabilities.

Conclusion

In conclusion, this research delves into the intricate landscape of generative language models (GLMs) and their ability to interpret and classify politically charged language with limited contextual information. The study's methodology involved a rigorous analysis of three prominent GLMs—Google's Bard GLM, OpenAI's GPT-3.5 Turbo, and OpenAI's GPT-4—utilizing a standardized prompt framework to assess their proficiency in predicting the political standing of selected bigrams. The results underscore a hierarchy of accuracy among the models, with GPT-4 exhibiting the highest classification accuracy, followed by GPT-3.5 Turbo and Bard. This performance hierarchy aligns with the models' respective parameter sizes and highlights their capacity to draw on extensive training data to understand nuanced political nuances, even in scenarios with limited context.

The findings underscore the potential of GLMs in comprehending and generating contextually appropriate content. However, the research also highlights the models' susceptibility to biases and inaccuracies inherent in their training data, underscoring the need for ongoing refinement and vigilance to ensure responsible deployment. The study not only sheds light on the capabilities and limitations of GLMs but also emphasizes the intricate layers of meaning embedded in human language, political ideology, and context. As these models continue to shape the landscape of AI and information dissemination, further research is imperative to mitigate the risks associated with misinformation and to maximize the potential for responsible and effective use.

References

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. Congressional Record for the 43rd-114th Congresses: Parsed Speeches and Phrase Counts. Palo Alto, CA: Stanford Libraries [distributor], 2018-01-16. https://data.stanford.edu/congress_text

Zhang, Y., Wang, X., Li, Y., & Liu, Z. (2023). How do generative language models learn from the internet? A survey of methods and challenges. arXiv preprint arXiv:2303.18223.

Matthew Gentzkow & Jesse M. Shapiro & Matt Taddy, 2019. "Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech," Econometrica, Econometric Society, vol. 87(4), pages 1307-1340, July.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In Advances in neural information processing systems (pp. 4349-4357).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.