

# Public Opinion Classification on Government Policy Using Social Media: An Exploration of ChatGPT's Capabilities and Limitations

Tammy Babad-Falk, Soon Ae Chun

City University of New York-College of Staten Island, NY 10314, USA  
tammybabad516@gmail.com, Soon.Chun@csi.cuny.edu

## Abstract

Gauging the public's sentiments and opinions toward policies is a critical task for policy makers. Social media posts offer a wealth of information for such a task but also pose unique challenges for achieving reasonable accuracy and transparency. While a well-trained machine learning model can offer accurate classifications of public perspectives, it cannot offer its reasoning for that classification or any further abstraction, which makes it difficult to use such solutions in practice. ChatGPT offers a possible solution; if the LLM has the requisite accuracy, the ability to explain the reasoning behind its decisions is built in. In this poster, we demonstrate ChatGPT's potential for such a task by documenting its accuracy and reasoning on opinion classification for posts regarding government policy in zero-shot and few-shot scenarios and compare the results to those of well-respected Natural Language Inference (NLI) models and ground truth human labels.

## Introduction<sup>1</sup>

Social media offers massive amounts of data to study public opinion; specifically, a variety of sentiment analysis called stance detection/opinion classification is used frequently (Brunham 2024; Pelicon et al. 2020; Abbonato et al 2024; Lan et al. 2024). In this paper, we explore opinion classification of posts on government policy, a challenging task due to the indirect referencing of policies, importance of context, and the complex nature of policy topics. Supervised and NLI classifiers are often recommended for these tasks for their high accuracies (Brunham 2024), but such solutions offer no easy form of explainability or method to transform results into human-interpretable outputs.

Since ChatGPT can provide some of the reasoning behind its decisions and be used further for summarization

and abstraction (Achiam et al. 2023), if the LLM could produce accurate enough results for the classification step, it could bridge this gap. In this paper, we use long-form Reddit posts about the Trump administration's federal funding policies and have two NLI models and ChatGPT determine if a given post is 'pro' or 'anti' the policy. The length of the Reddit posts, broad nature of the policy, and context-based classification scheme offer a challenging stance detection task to test ChatGPT's capabilities.

**Related Works:** A few papers have studied NLI-based stance detection (Brunham et al. 2024; Abbonato 2024). Brunham's 2024 study investigates stance detection tasks for political science with in-context models (like GPT-4) and NLI classifiers. Lan (2024) used GPTs to detect stances about political people and ideas.

## Research Questions and Methodology

We attempt to answer the research questions: Can ChatGPT be used to classify a social media post's political stance without any training (zero-shot ChatGPT)? How good is ChatGPT's accuracy on this task when compared to a dedicated NLI political stance detection model? Can ChatGPT with a few shots perform better than zero-shot ChatGPT or NLI classifiers? If ChatGPT achieves better or equivalent accuracy, can it provide reasonable explainability compared to the NLI classifiers? To answer these questions, we set up experiments to test ChatGPT's accuracy against human labels and NLI classifiers in the specific use case of long-form policy-based Reddit posts.

**Data:** The data for this project was gathered from Reddit using the Reddit API to access Reddit's landing page, using keywords like 'nsf', 'federal funding', 'funding cuts', and 'DOGE' etc. These results were then augmented

---

<sup>1</sup>Copyright © 2025 by the authors.  
This open access article is published under the Creative Commons Attribution-NonCommercial 4.0 International License.

with posts from specific republican sub-communities to balance the data, which, by nature of Reddit's demographic, leaned primarily left. Once the posts for the dataset were gathered, a human filtered out irrelevant posts and gave ground truth 'pro' or 'anti' labels to those remaining. Posts labeled 'pro' expressed opinions aligned with the Trump administration's federal funding policies, and posts labeled 'anti' expressed opinions against those policies. Cleaned and labeled, the dataset consisted of 64 rows and 2 columns, 'text' and 'label', where 51 posts were labeled 'anti' and 13 were labeled 'pro'.

**Experimental Setup:** To establish how well a pre-trained stance detection NLI model could perform on our task, we first used the 'deberta-v3-large-zeroshot-v2.0' model (Laurer et al. 2023). This model has shown accuracies near or better than supervised classifiers and ChatGPT (Burnham 2024) when tested on similar tasks, and is well respected in the political science field. We also tested the 'Political\_DEBATE\_large\_v1.0' model, a version of Laurer's NLI classifier that was fine-tuned on political data (Burnham et al. 2024). Both models were trained with premise-hypothesis pairs and labels. In our case, we provided each NLI with the hypothesis: 'This text is {label} the Trump administration funding cuts', the labels 'pro' and 'anti', and the text of a post (premise).

The third and fourth experimental setups tested ChatGPT's ability to label the same posts as 'pro' or 'anti' in zero-shot and few-shot scenarios with basic prompts. Lastly, we prompted ChatGPT with few-shot queries that also necessitated sharing the reasoning behind its classifications. The ChatGPT experiments were done using the OpenAI API and the basic chat completions object. The exact System and User messages used for each experiment are detailed in Figure 1, and all four experimental setups are summarized in Table 1, shown in the appendix.

## Results and Discussions

**Accuracy:** Accuracy was measured using the SciKit-Learn standard accuracy function as shown in Table 2 in the appendix. The standard NLI model provided a baseline accuracy of 88%, and the NLI classifier fine-tuned on political data achieved the highest accuracy, 92%. The standard NLI classifier, while lacking training on political data, is well respected and widely used for stance detection, making it a good starting point for comparison. The DEBATE model, being newer and more specifically tuned, is less popular but shows the absolute best a pre-trained model should be able to achieve on this task.

The Zero-shot ChatGPT experiment provided a below-baseline accuracy of 83%. This was mainly because even though the model was explicitly told to respond with only the words 'pro' or 'anti', on certain, less clear-cut posts, ChatGPT refused to give a concrete classification and instead returned nonanswers like 'the post does not

explicitly state a pro or anti stance...', which brought down the model's overall accuracy. Secondly, the model severely under-classified 'pro' examples; a shortfall shared by all four experiments to differing degrees.

The few-shot ChatGPT scenario performed better, with an accuracy of 91%. It surpassed the basic NLI classifier primarily in true positives, performing with 60% more accuracy on the more complicated 'pro' posts. While it made different mistakes than the DEBATE NLI classifier, it only made one more mistake overall, and 75% of all its errors were on relatively difficult examples. Due to the small size of the dataset, these accuracies are relatively close together but might diverge more given more data of a better distribution. Even with a small dataset of relatively poor distribution, the few-shot ChatGPT model performed better than a pre-trained model designed for similar tasks and only slightly worse than a powerful model designed for this domain-specific use case. We can, therefore, conclude that a few-shot ChatGPT can achieve the requisite accuracy to be a valid building block for future systems relying on traceable, clear political opinion classification.

**Explainability:** While it is possible to explain NLI classifiers through such involved methods as gradients, extracted rationale, and counterfactuals (Lyu 2024), explanations of this variety are difficult to implement and understand without domain-specific knowledge. In contrast, by adding a few lines to our original few-shot prompt, ChatGPT provided a detailed rationale for its chosen label that used universally understandable, nontechnical language and indicated the key points that informed its decision. Below is one such explanation:

*The post describes an event organized in response to perceived threats to research funding and academic freedoms, indicating concern and opposition to such funding cuts. The use of phrases like "under attack" and the call to "stand up against these threats" clearly demonstrates that the author...is against funding cuts*

## Conclusion and Future Studies

In this paper, we created 4 experimental setups to test how well ChatGPT performs on an opinion classification task for policy-related posts as compared to human ground truths and established NLI models. The ChatGPT few-shot experiment provided the second-highest accuracy, displaying only one more error than the dedicated political stance detection model. This provides a promising implication that few-shot ChatGPT can be used without resorting to domain-specific classifiers. For instance, we can summarize only those social media posts that have opinions fundamentally 'anti' a policy. We plan to experiment more with few-shot ChatGPT by classifying different user opinions and domain-specific perspectives while providing more explainable reasoning capability.

## References

- Abbonato, D. 2024. Public sentiments on the fourth industrial revolution: An unsolicited public opinion poll from Twitter. *arXiv preprint arXiv:2411.14230*.
- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Burnham, Michael. 2024. Stance Detection: A Practical Guide to Classifying Political Beliefs in Text. *Political Science Research and Methods*: 1-18. <https://doi.org/10.1017/psrm.2024.35>.
- Burnham, M., Kahn, K., Wang, R. Y., & Peng, R. X. 2024. Political debate: Efficient zero-shot and few-shot classifiers for political text. *arXiv preprint arXiv:2409.02078*.
- Pelicon, Andraž, Marko Pranjic, Dragana Miljkovic, Blaž Škrlić, and Senja Pollak. 2020. Zero-Shot Learning for Cross-Lingual News Sentiment Classification *Applied Sciences* 10, no. 17: 5993. <https://doi.org/10.3390/app10175993>
- Lan, X., Gao, C., Jin, D., & Li, Y. 2024. Stance Detection with Collaborative Role-Infused LLM-Based Agents. *Proceedings of the International AAAI Conference on Web and Social Media*: 18(1), 891-903. <https://doi.org/10.1609/icwsm.v18i1.31360>
- Laurer, M., van Atteveldt, W., Casas, A., & Welbers, K. 2023. Building efficient universal classifiers with natural language inference. *arXiv preprint arXiv:2312.17543*.
- Lyu, Q., Apidianaki, M., & Callison-Burch, C. 2024. Towards faithful model explanation in nlp: A survey. *Computational Linguistics*, 50(2), 657-723.

## Appendix

Setup 1	Setup 2	Setup 3	Setup 4
Trained NLI classification model	Trained NLI political classification model	ChatGPT in ZeroShot Prompting	ChatGPT in FewShot Prompting (with and without explanation)

Table 1 Experimental Setup

NLI classifier	NLI Political Classifier	Zero-shot ChatGPT	Few-shot ChatGPT
88%	92%	83%	91%

Table 2 Accuracy of Experiments

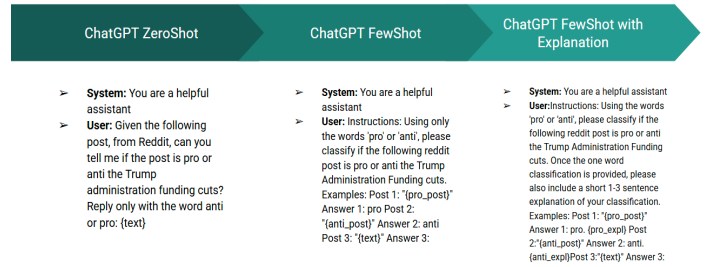


Figure 1. System and User Instructions for ChatGPT Experiments