DOGE's Downsizing, Can AI Read the Reddit Room?

Kevin Linares
University of Maryland
and
Felix Baez-Santiago
and
Aria Lu
and
Gloria Zhou
University of Michigan

April 8, 2025

Abstract

We investigate public sentiment on Reddit regarding the Department of Government Efficiency's federal workforce reduction by classifying 400 labeled comments (28% favor, 18% neutral, 54% oppose) using supervised and unsupervised Large Language Models. Supervised models showed moderate success, particularly with "oppose" comments, but struggled with "favor" and "neutral" stances. Similarly, LLMs best identified "oppose" sentiment but exhibited low precision for "favor" and "neutral." These findings highlight the challenges of accurately gauging nuanced public opinion on government policy changes using automated methods on social media data.

Keywords: Reddit, Federal Government, DOGE

1 Introduction

The newly formed Department of Government Efficiency (DOGE) has reduced the federal workforce by almost 280,000 employees, a central pledge of the current administration. This action has generated significant apprehension among federal workers in regards to mental health and job security. To understand the impact of DOGE's actions on federal worker perceptions of job security, we labeled 400 Reddit comments on topics related to the current reduction in *federal* workforce by DOGE as whether the author favored, opposed, or had a neutral stance. We use these labels to build supervised learning models to predict stance. Additionally, we employ unsupervised large language models (LLM) to detect the stance of these Reddit comments from text, to further explore appropriate models for this current topic.

2 Methods

Data. We collected Reddit comments from subreddits related to the topic of interest in early March of 2024. This resulted in 12,553 comments which we preprocessed by removing web-URLs, replace special characters (i.e., replaced "@" with "at"), replaced numeric values with their spelling, and confirmed the absence of duplicate comments. We then randomly selected 400 comments (without replacement) and assigned them to four graduate students for coding, categorizing each comment as favoring, opposing, or neutral towards DOGE's approach to federal workforce reductions. Table 1 presents the breakdown in percent of our the comments we labelled and later use to train and evaluate our machine learning models. A review of the comments revealed more negative opinionated statements and reactions.

Table 1: Distribution of labeled Reddit comment data.

outcome	Percent
favor	28
neutral	18
oppose	54

We built two supervised machine learning models–K-nearest neighbor and random forest–to classify Reddit comments concerning stance on the current reduction in federal workforce into favor, oppose, or neutral. For predictors we used the Reddit score, up-votes, and down-votes each comment received at the time of the collection. Additionally, we used the keyword that was used to scrape these data along with the subreddit where the comment was posted. **DISCUSS ADDITIONAL FEATURE ENGINERRING SUCH AS TEXT EDITING THAT WENT INTO THE MODEL.** For the KNN model, we set the number of neighbors hyperparameter to three, and for the random forest, mtry was kept at a constant two as we did not have many features in these models. We took an 80/20 split for our training-testing sets that we used to train each model and evaluate.

In addition to our supervised machine learning models, we also applied to our comments two LLMs, gemma 3.12 and llama 3.2. 3B. Gemma, was developed by Google and is based on Gemini, is able to generate text output using instruction-tuned prompts from the user, such as providing answers on a user specified task reviewing text. We deployed gemma on a higher performance computing cluster (Great Lakes HCP) and processed text from Reddit comments on a GPU. We were also interested in deploying other lightweight LLMs locally on a machine with a dedicated-GPU and found llama, developed by Meta AI, to

be an excellent candidate. Llama excels at text classification which is how we used it to classify the stance for each comment text given a user specified prompt. We choose both of these LLMs as candidates to test against our supervised approaches because of their ability to apply locally and reputable performance on zero-shot classification tasks. Both of these models received the same prompt and tasked which were developed by the research team to reflect the utility of classifying stance within our topic of interest.

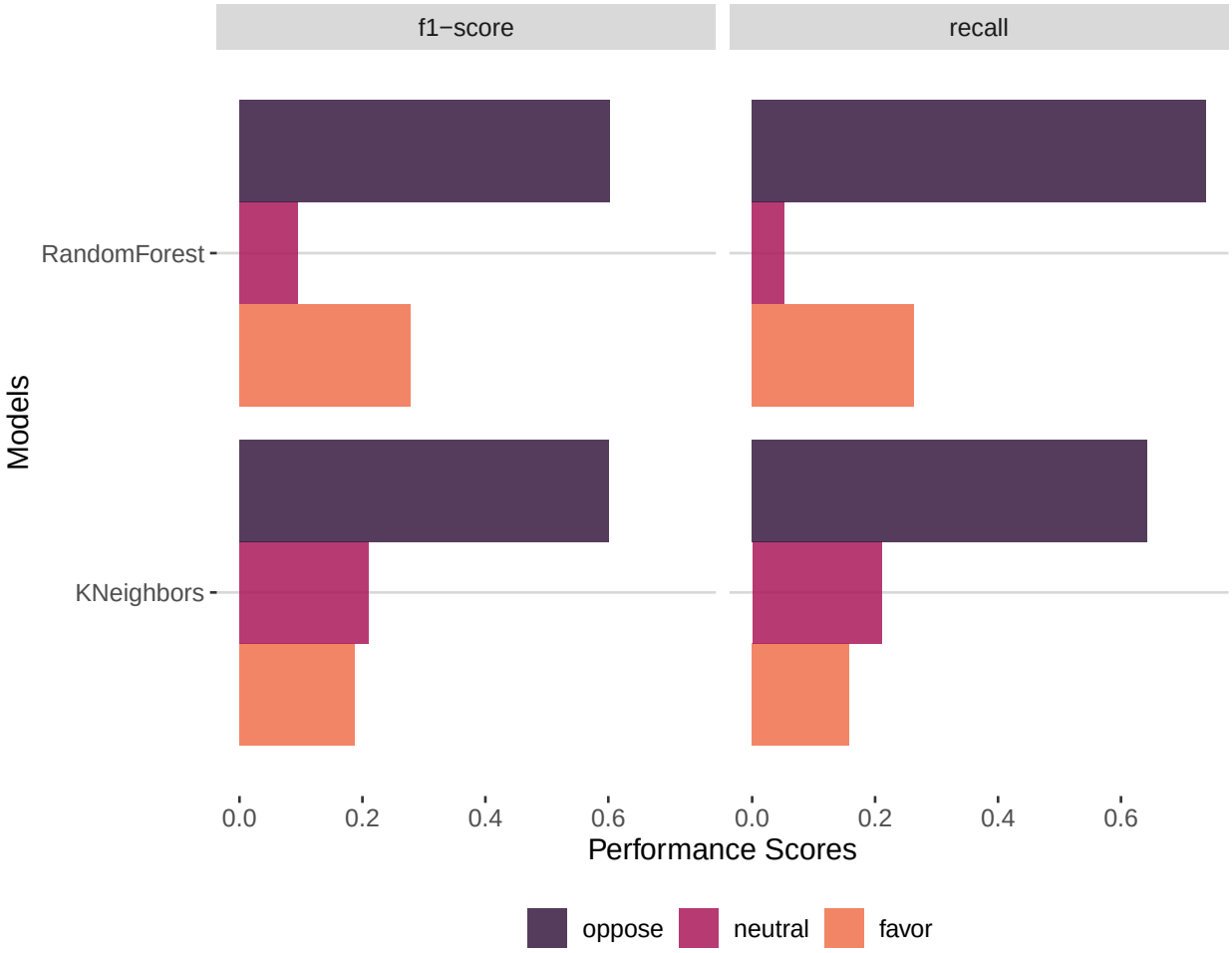
Prompt:	"Is this comment in 'favor', 'neutral', or 'oppose' the reduction in federal workforce? Provide one word answer only!
Task:	"You have assumed the role of a stakeholder that is presented with a reddit comment from a likely federal worker related to the current policies on reducing the federal workforce. Please determine the author's stance on this topic, and only provide the answer."

We evaluate model performance using the F1-score which is a harmonic mean balance between precision and recall. This metric considers both the ability to correctly identify positive instances, known as recall, and the ability to avoid incorrectly labeling negative instances as positive (i.e., precision). A high F1-score indicates a model with well balanced recall and precision that both minimize false negativesand false positives. The second performance metric we used is recall, as this determines how many of the actual positive cases the model can correctly identify. Recall can be prone to incorrectly labeling negative cases as positive, suggesting more false positives, which is why we intereggoate on two performance metrics in this study.

3 Results

Our machine learning models reveal a mix of performance results in classifying the stance of Reddit comment into favor, neutral, oppose. Both models exhibit low recall and F1-scores for the "favor" stance (see Figure 1), yet the random forest model (recall .26: f1-score .28) slightly outperforms KNN (recall .16: f1-score .19). The "neutral" class presented a significant challenge for the random forest with very low scores (recall .05: f1-score .10), while KNN exhibited slightly better performance (recall .21: f1-score .21). Both models showed the strongest performance in identifying comments expressing opposition. The random forest correctly identified 74 percent of all the actual Reddit comments expressing opposition compared to KNN at 64 percent, meaning that KNN missed a larger proportion of true "oppose" comments. The f1-scores were similar for both models at 60 percent, this implies that KNN may have a slightly higher precision score given that it had a lower recall score than the random forest model. This implies that when KNN predicts comments for the "oppose" class, it is as likely as the random forest to be correct.

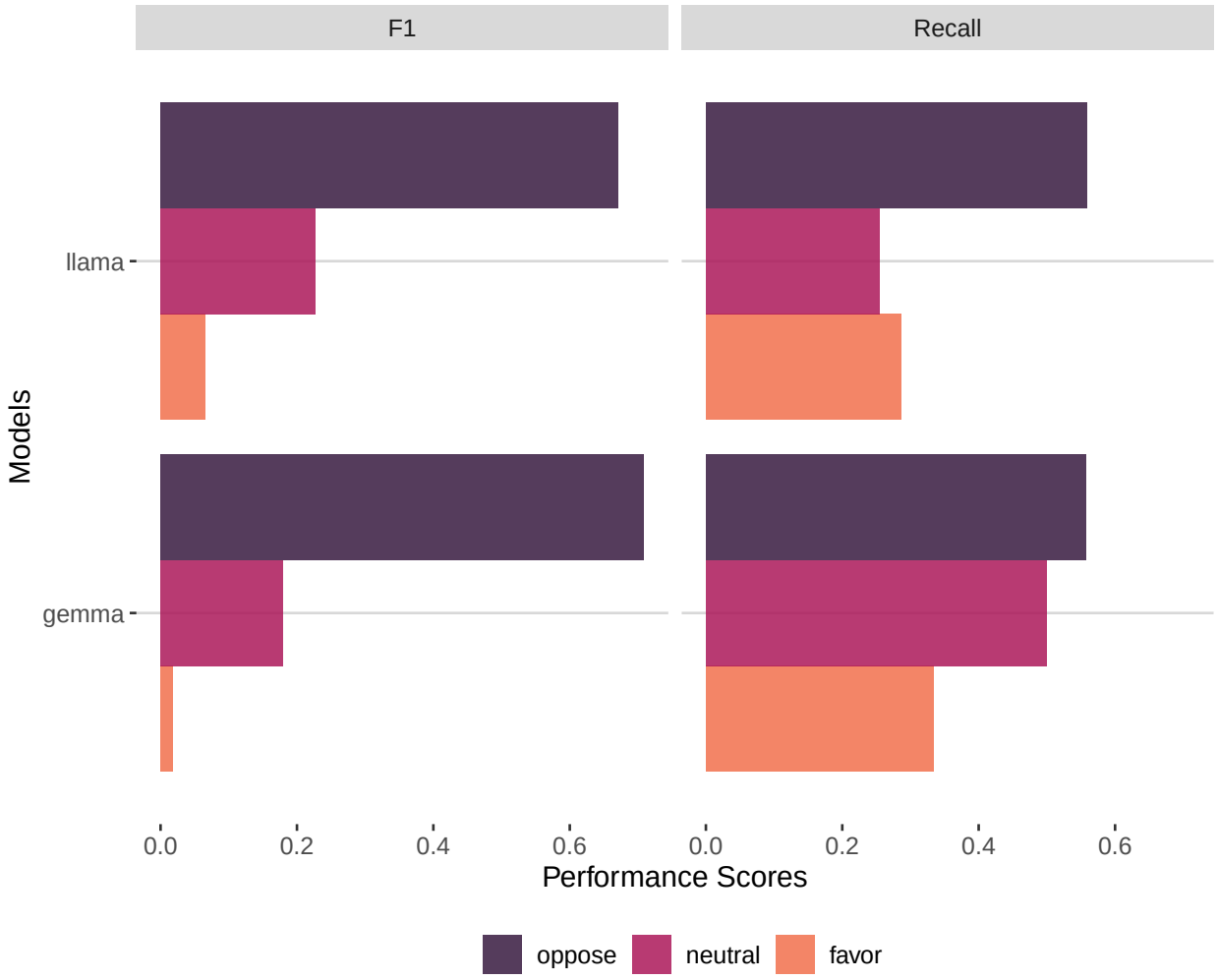
Figure 1. Recall and precision scores for random forest and KNN models.



The LLMs, showed poor performance in classifying the "favor" stance, with very low f1-scores (llama .07: gemma .02) despite moderate recall (llama .29: gemma .33), indicating low precision (see Figure 2). For the "neutral" class, both models showed modest results, with llama achieving a slightly better F1-score (.23) compared to gemma (0.18), although gemma had a higher recall (.50 vs .25), again suggesting lower precision for gemma. Similar to the supervised models, both LLMs performed best in classifying the "oppose" stance, with relatively high and similar F1-scores (llama .67: gemma .70) and comparable recall (llama .56: gemma .56), suggesting a better balance between precision and recall for this category. Both models correctly identify 67 to 70 percent of all the actual Reddit comments expressing opposition. Overall, both LLMs struggled with the

"favor" and "neutral" classes but demonstrated a stronger ability to identify opposing comments.

Figure 2. Recall and precision scores for LLMs.



Gemma and llama demonstrated only fair agreement in stance classification (Cohen's Kappa = .24). The contingency table, Table 2 below, reveals highest agreement for "oppose" (.80), indicating some consistency in identifying a strong negative stance. However, agreement was substantially lower for "neutral" (.03) and "favor" (.01), highlighting divergent interpretations of more nuanced language. Off-diagonal values further illustrate these discrepancies, suggesting fundamental differences in how the models process ambiguous cues and establish classification boundaries, particularly for less extreme stances.

Table 3: Aggreement between LLMs, Cohens Kappa, 0.24

	favor	neutral	oppose
favor	3	1	10
neutral	0	10	49
oppose	0	5	317

4 Conclusion

We explored public sentiment on Reddit regarding the DOGE federal workforce reduction using supervised learning (KNN, Random Forest) and unsupervised LLMs (gemma 3.12b and Llama 3.2 3B) for stance classification. Our supervised learning models achieved moderate success in this classification task, with the strongest performance in identifying opposing viewpoints, which were also the most prevalent in our labeled data. Moreover, both models struggled with the "favor" and "neutral" stances, suggesting limitations in our initial features to capture these nuances. Our unsupervised learning models also showed the best ability to classify "oppose" comments, yet exhibited significant challenges with the "favor" and "neutral" categories, particularly demonstrating low precision.

This study had limitations such as the small size of our labelled dataset of 400 Reddit, imbalance in the stance categories, and few features in the supervised approach. The zero-shot capabilities of the LLMs are suitable for topic exploration but may require fine-tuning for optimal performance on this specific domain, particularly with nuance language used in "favor" comments. Future research should continue to understand federal worker perceptions and improve stance classification by expanding labeled dataset and incorporate

more contextual text rather than single comment text. Fine tuning LLMs on a larger corpus with relevant comments could render more accuracy in the classification task.