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 9, 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 (LLM). 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 Method

We collected a total of 12,553 Reddit comments in early March 2024 from subreddits relevant to the topic of interest. These comments were retrieved using keyword-based queries designed to capture posts of discussions that would reflect users' stances on the federal workforce reduction by DOGE. To prepare the data for analysis, a comprehensive preprocessing pipline were conducted in Pytho 3. This process involved removing HTML tags and web-URLs, stripping special characters and tagging artifacts such as "@", and replacing emojis with textual descriptions. Also, it entailed eliminating punctuations and converting numerical values into their spelled-out forms. The text was standardized to lowercase, lemmatized using the SpaCy library, and further refined by removing stopwords per the Natural Language Toolkit (NLTK) lists. Spelling corrections were performed to improve

textual coherence, while Duplicate comments were identified and removed to maintain data integrity.

From this cleaned dataset, 400 uniqe comments without replacement were randomly sampled for manual annotation. Four graudate students independently coded and labeled each comment with one of three possible stances:favor, opopose, or neutral toward DOGE's approach to federal workforce reductions. As shown in Table 1, there were 28% favor, 18% neutral, and 54% oppose. These labeled comments were subsequently used for training and evaluating both supervised and unsupervised classification models.

Table 1: Distribution of labeled Reddit comment data.

outcome	Percent
favor	28
neutral	18
oppose	54

The feature set used in model training and evluation consisted of two primary categories: metadata features and textual features. Metadata features included the Reddit comment's score, the subreddit in which it appeared in, and the search term that retrieved it. As the variables for score and upvotes were perfectly collinear in this dataset, only the score was retained. To ensure all values were non-negative, the score was adjusted by subtracting the dataset's minimum value from each score. The categorical variables, subreddit and search term, were numerically encoded using scikit-learn's label encoder.

Textual features were derived from each comment and represented using two distinct approaches: unigram term frequency and term freuqency-inverse document frequency (TF-

IDF). The unigram used a bag-of-words model to represent each word as a feature based on its frequency in comment. The TF-IDF transformation built upon these raw counts by accounting for term importance across the corpus. For both unigram and TF-IDF methods, the total number of text-derived features was 2,102. Combined with the three metadata features, each dataset contained a total of 2,105 features. Two datasets were constructed for classification: one combining the metadata with unigram features, and another with metadat and TF-IDF features. Each dataset was divded into 80% training data and 20% test data. All supervised learning models were implemented using Python 3, with scikit-learn and XGBoost as the primary machine learning (ML) techniques. Classification was conducted using a diverse set of algorithms, including Extra Trees Classifier, Random Forest Classifier, Logistic Regression, Ridge Classifier, K-Nearest Neightbors (KNN), Nearest Centroid, Multi-layer Perceptron Classifier, Bernoulli Naive Bayes, Multino- mial Naive Bayes, Linear Support Vector Classifier, Decision Tree Classifier, Extra Tree Classifier, and the XGBoost Classifier.

All models were trained using default hyperparameter confiurations. The number of neighbors for KNN model was set to three, and the mtry parameter for Random Forrest was fixed at two, reflecting the relatively small number of metadata features in those models. All experiments were conducted on a personal computer without the use of a GPU

In addition to traditional supervised models, we employed two instruction-tuned LLMs to perform zero- shot classification: Gemma 3.12B, developed by Google and based on the Gemini framework, and LLaMA 3.2B, developed by Meta AI. Gemma was deployed using the Great Lakes High-Performance Computing (HPC) cluster, while LLaMA was deployed locally on a GPU-enabled personal workstation using the rollama r pacakge. These models were selected for their strong zero-shot classification capabilities and ability to be run locally

or on cluster environments without reliance on commercial APIs.

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."

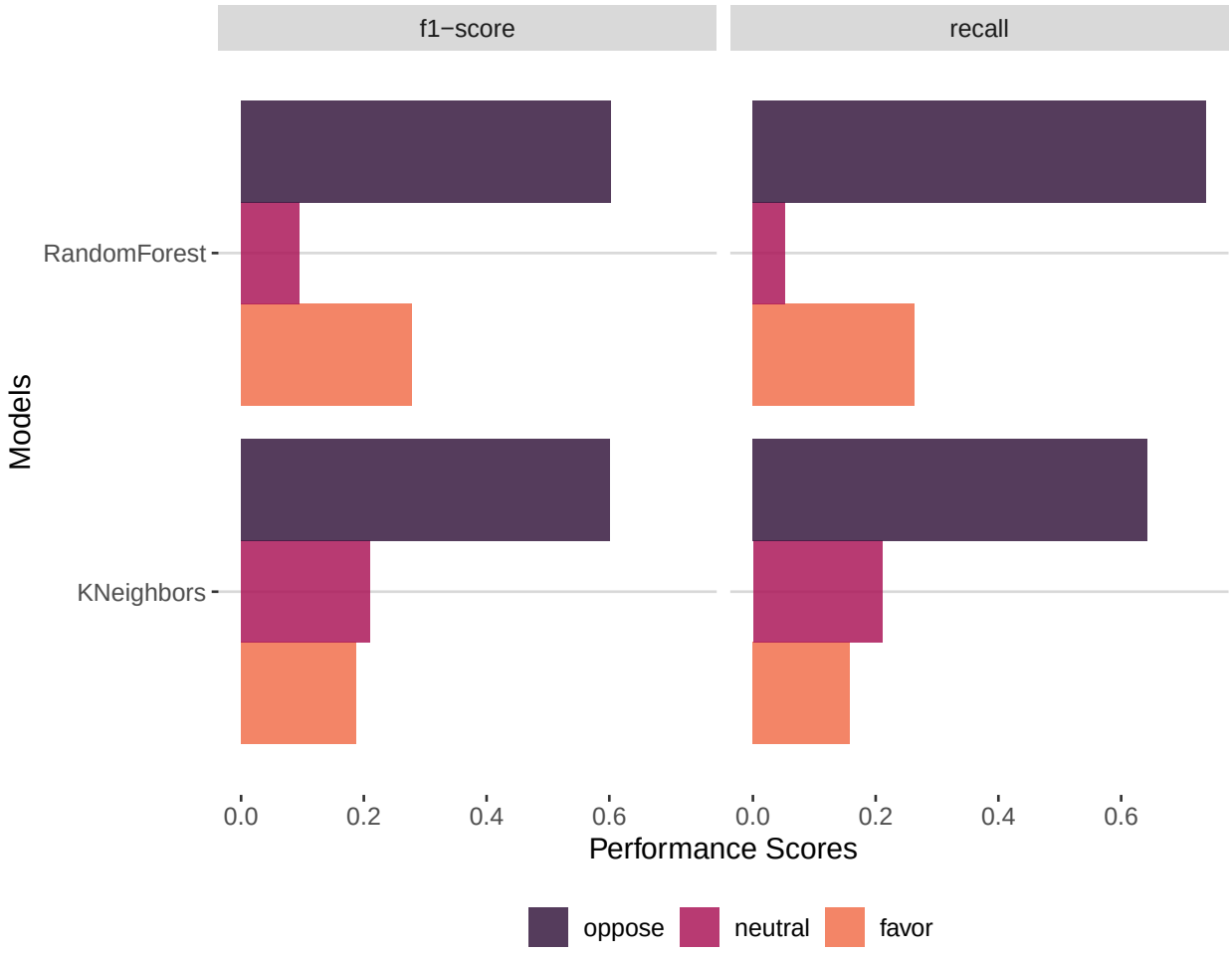
The model performance was evaluated by the F1-score and recall. The F1-score, a harmonic mean of precision and recall, serves as a balanced indicator of a model's ability to correctly identify relevant instances while minimizing both false positives and false negatives. Recall was also independently assessed to evaluate the proportion of true postive cases accurately identified by each model. While a high recall may increase the risk of false positives, its combination with the F1-score allows a more robust evaluation of model performance, especially for addressing the challenges inherent in imbalanced classification tasks

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.