Who Talks About What? Classifying Political Attention Across Levels

Yi Dou

Abstract

This paper studies how elected officials communicate across different levels of government under a federal system using social media text data. Drawing on posts from U.S. elected officials on X in 2024, I developed a pipeline that combined an unsupervised topic model with word embeddings to classify texts into federal, state, and local issue levels. Latent topics identified by an LDA model were mapped onto these three dimensions using embedding-based similarity measures with orthogonalization to address semantic overlap. Document-level labels were then constructed as weighted averages of topic-level scores and used to train supervised learning models, including a ridge multinomial logistic regression and a DistilBERT classifier. The models outperformed a naive baseline and produced predictions that align with institutional expectations when applied to an external data set of congressional tweets. Overall, the results suggest that embedding-based labeling offers a scalable and interpretable approach for studying multi-level political communication.

Introduction and Motivation

Coming from a unitary state, I have long been curious about how a federal system actually functions. In particular, I wonder whether officials under federalism serve primarily their states or the federal government itself. Do officials at different levels focus on different kinds of issues? Do these focus differs among parties or other characteristics? To find out the answers, I decided to build a classifier which classify the attention level of officials based on their social media texts.

Here's a overview of my work flow. First, I scraped data from X using the tweet id from the Digitally Accountable Public Representation (DAPR) data set. Then I applied a topic model to identify the latent topics. Each topic were then be categorized as corresponding to the federal, state, or local level according to the embedding of their word distribution. Next, I assigned each document as one of these category according to their topic distribution. These labeled data formed the basis of a supervised learning model designed to predict the governmental level of any given post. I then applied the model to another data set to see how it worked.

This project's methodology is inspired by some existing works. Barberá et al (2019) used a Latent Dirichlet Allocation (LDA) model to identify issue-specific topics in congressional speeches and social media posts, and then use these topics as substantive units to study agenda-setting dynamics between legislators and the mass public. By contrast, Motolinia (2021) also used topic models for subsequent anlysis, but classified the latent topics into a few meta-topic categories instead of labeling each of them. On the other hand, Rheault and Cochrane (2020) introduced word embeddings augmented with political metadata to capture semantic relationships in parliamentary texts, and scale ideological positions of parties and legislators. Together,

these studies illustrate two complementary approaches: topic models summarize political discourse into interpretable issue categories, while word embeddings preserve contextual semantics and enable flexible measurement of latent ideological dimensions. I applied a similar pipeline to conduct the research.

Data Collection

I used the Digitally Accountable Public Representation (DAPR) Database, which tracks and analyzes the online communication of federal, state, and local elected officials in the U.S. focusing on X/Twitter and Facebook. Specifically, I limited my research scope to the data from X in 2024. There are two main reasons for this decision. Firstly, the DAPR database does not include any data on Facebook posts beyond December 2021, which is a little outdated in the context of social media data. Secondly, in the X data beginning in 2024, DAPR collects data from a random sample of 1,000 elected officials each month, which makes the data more appropriate to aggregate and compare with each other (Tai et al, 2024).

The primary goal of this data collection process was to construct a data set of political social media posts which could later be used for text analysis and supervised learning. The initial data from DAPR consisted of TF-IDF representations stored across multiple files. However, these TF-IDF values were computed independently for each week, so they were not directly comparable with each other. More importantly, the original textual content was not included, which prevented the construction of a unified document–term matrix or the recalculation of global TF-IDF scores.

To address this limitation, I designed a data collection pipeline to reconstruct the original tweet texts associated with each observation. The key insight was that, although the raw text was missing, the data set included tweet identifiers (IDs) corresponding to each post. These tweet IDs could potentially be used to retrieve the original content from the X platform.

Because access to the official Twitter API is restricted and rate-limited, I adopted a browser-based web automation approach rather than an API-based method. Specifically, I used Python together with the *Playwright* library to automate a real Chrome browser session. By launching the browser with an existing authenticated user profile, the scraper was able to access tweets that were publicly visible to a logged-in user, while avoiding API authentication requirements. This approach also allowed the scraper to handle JavaScript-rendered content, which is essential for loading tweet text on X.

The core Python libraries used in the data collection process included:

a. *playwright* (for browser automation and page interaction),

b. *pandas* (for reading and processing CSV input files),

c. *asyncio* (to manage asynchronous page loading),

d. *csv* and *os* (for file input/output and directory traversal).

The scraping script iterated over all tweet IDs extracted from the original CSV files, visited the corresponding tweet URLs, and attempted to extract the tweet text. To improve robustness, the script incorporated several safeguards, including request timeouts, short delays between page visits to reduce the risk of rate limiting, and a checkpoint mechanism that recorded successfully processed tweet IDs. This checkpoint system enabled me to resume from the last completed position in case of interruption, ensuring that progress was not lost during long-running executions.

It is important to note that the majority of the Python code implementing this scraping pipeline was generated with the assistance of ChatGPT 5. My own contribution focused on adapting the generated code to my specific data structure, adjusting parameters such as timeout length and request delays, specifying file paths, and validating intermediate outputs. I also executed, monitored, and debugged the data collection process iteratively to ensure that the script ran successfully over a large number of observations.

In total, after running for more than 60 hours, the script attempted to retrieve over 52,358 tweets in 2024. Due to platform rate limiting, deleted posts, private accounts, and network timeouts, not all tweets could be recovered. Approximately 61% of the attempted tweets(31,870) were successfully retrieved with valid text content. Despite this level of missingness, the resulting data set contains a substantial number of high-quality textual observations and provides a consistent raw-text foundation for subsequent analysis[1].

Apart from the data obtained from X, I also used the tweets_congress data for application. I downloaded this data set from our course page. I'm not sure how to cite it, so I mentioned here as a footnote.

Method

Preprocessing

With these raw tweets collected, I began with the preprocess of data. Firstly, because tweets are extremely short and often fragmented, I first aggregated the data by author and by week. All tweets posted by the same official within the same week were merged into a single document. This gives me more stable linguistic units and reduces noise from casual one-off tweets. Once the aggregation was complete, I merged additional metadata about each official such as their party affiliation.

After preparing the text, I constructed a corpus with these aggregated tweets and preprocessed it through typical steps(tokenized the documents, removed punctuation, numbers, stopwords, and very short tokens, and converted everything to lowercase...). Then I built a document feature matrix (DFM) and trimmed extremely rare terms to reduce sparsity. Because the topic model implementation I used requires the document-term matrix format(DTM), I converted the DFM to that structure.

---

[1] Apart from the data obtained from X, I also used the tweets_congress data for application. I downloaded this data set from our course page. I'm not sure how to cite it, so I mentioned here as a footnote.

Building the Topic Model

With the document-term matrix ready, I fitted an LDA topic model. As a generative probabilistic model that represents documents as mixtures of latent topics, the goal of an LDA is to infer the latent topic structure—that is, the topic proportions for each document, the topic assignment for each word, and the word distributions that define each topic (Blei, Ng, and Jordan 2003).

Formally, the two key distributions in LDA are:

1. Document–topic distribution

$$\theta_d \sim Dirichlet(\alpha)$$

where θ_d is the topic proportion vector for document d, and α is a hyperparameter controlling how concentrated or diffuse topic mixtures are across documents.

2. Topic–word distribution

$$\phi_k \sim Dirichlet(\beta)$$

where ϕ_k is the word distribution for topic k, and β is a hyperparameter controlling how concentrated topics are over the vocabulary.

In an unsupervised machine learning modeling process, there is also a hyperparameter which must be manually set: the number of topics in the corpus (K). I used the same techniques as Barberá et al (2019) to do the cross validation and chose the K. It calculates two metrics of model performance and visualizes them as the ratio with regard to the worst value. The two metrics in Figure 1, log likelihood and perplexity are both optimal when the number of topics is around 30. Therefore, I chose K=30, which means the LDA model would contain 30 latent topics, and built the LDA model.

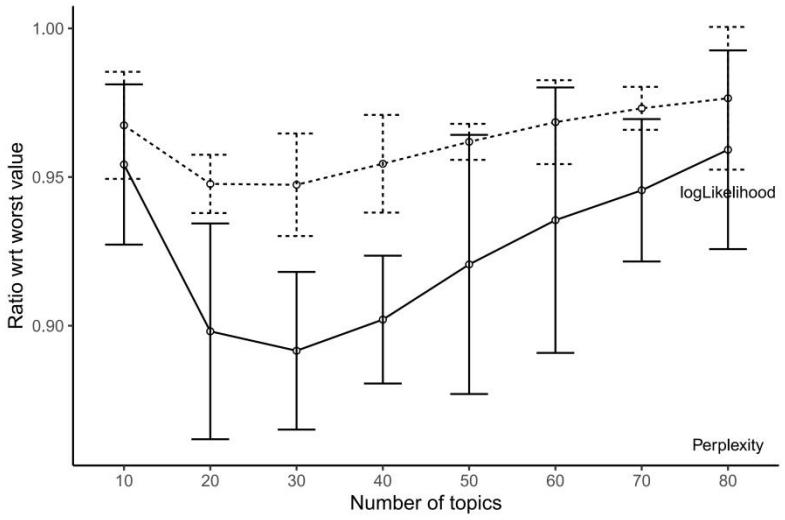

Figure 1

Labeling

At this stage, the model provides a purely statistical decomposition of the corpus, but it does not yet tell us anything about the federal, state, or local semantics we care about. To bridge that gap, I incorporated word embeddings. I used the GloVe 300-dimensional embeddings, which give vector representations of words based on their distributional semantics. To assign each topic a federal, state, or local score, I compared the topic's word distribution to the embeddings of the words "federal" ,"state" and "local". Concretely, for each topic, I weighted the cosine similarity between every word in the vocabulary and one of these three dimension words by the probability distribution of words in the topic (ϕ_k) . This produces a numerical measure of how closely a topic aligns with federal-level language, state-level language, or local-level language：

$$S_k^{(l)} = \sum_{\omega} \phi_k(\omega) \cdot \cos(e^{\omega}, \tilde{e}^{(l)})$$

where ω is each word in the topic k, and $\tilde{e}^{(l)}$ is the word embeddings of the dimension words, $l \in \{\text{federal,state,local}\}$.

However, I quickly discovered a problem: these three dimension words are highly correlated in embedding space. "Federal," "state," and "local" often appear in similar contexts, so their vectors overlap heavily. Because of this, as Figure 2 shows, the topic scores across the three dimensions were almost collinear. To solve this issue, I orthogonalized the three embedding vectors. In simpler terms, I forced each dimension to represent only the semantic information unique to itself, by removing the parts that could be explained by the other two vectors. As Figure 2 shows, after orthogonalization, the three dimensions became much more distinct, which allowed the topic scores to spread out in a meaningful way.

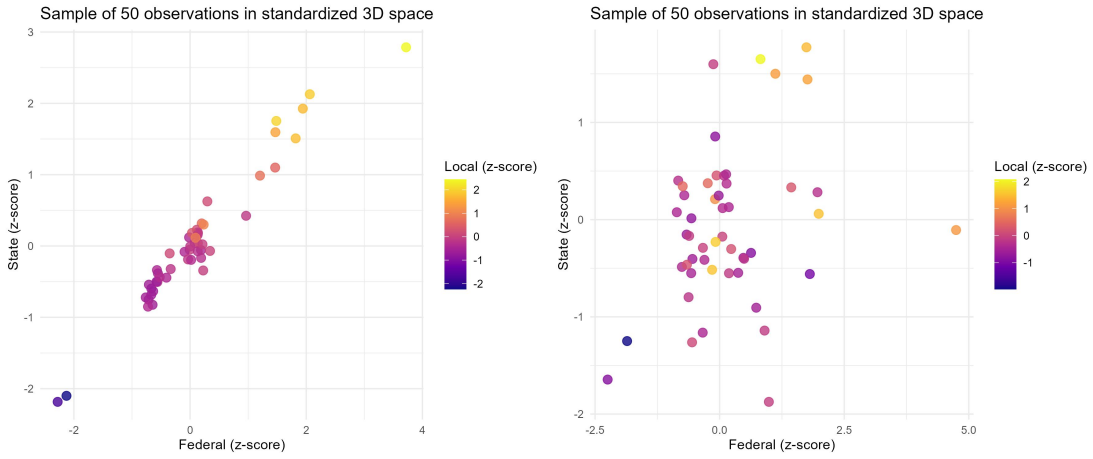

Figure 2

Once I had orthogonalized topic-level scores, I propagated them to the document level. I multiplied each document's topic distribution (θ_d) by the topic-level scores, so each document received its own federal, state, and local semantic scores:

$$S_d^{(\iota)} = \sum_k \theta_d(k) \cdot S_k^{(\iota)}$$

where k is each topic in the document d.

Finally, using the three document-level scores, I assigned each document a predicted label. The prediction is simply the dimension with the highest score among federal, state, and local. This gives us a direct, interpretable way to classify political language and prepared for next step's classification.

Here are two clarifications I must make about this labeling strategy. The first one is why I didn't label each topic like the Barberá paper and many of my classmates did. There are both theoretical and practical considerations. Theoretically, the topics provided by the LDA model are latent, and they are not necessarily the perfect matching for human-readable topics. For examples, topic 27 may have a 70% similarity with foreign policy, so foreign policy is the best label for it. However, if we label it as foreign policy, we will lose the 30% information of it which are not related to foreign policy. Practically, as a non-native English speaker, I am not confident about manually labeling for the topics. Actually I tried to use large language models to do it for me. However, the labeling results provided by the AI are heavily influenced by the prompt words, which deepened my doubts about this method.

The second clarification is why I didn't label some of the observations manually. The advantage of manual labeling is to provide a classification which is based on human understanding of the texts instead of statistical inference. However, this understanding is based on the semantic meanings of the texts, which could also be captured by word embeddings. In fact, if the following assumption stands, it would be rational to trust the label provided by the calculation based on word embeddings and not label by myself:

The GloVe 300-dimensional embeddings can capture the semantic meanings of the tweets of U.S elected officials better than me.

It's just an assumption and can not easily be proved. Based on my understand about word embeddings and myself, I chose to trust this assumption.

Supervised Learning

After labeling the data, I moved to the supervised learning models. The motivation to build the supervised learning models is to classify the observations based on the texts, so that when we apply the models to new data, we can infer their issue level more conveniently and don't need to do the time costing topic models again. Therefore, the dependent variable of the supervised learning models was the label provided by the previous steps, and the inputs were the texts.

I tried to build two different supervised learning models to do the classifications: a ridge multinomial logistic regression model and a DistilBERT model. The initial inputs of these two models were both the aggregated raw texts, but there is a slight

difference: I calculated the TF-IDF values for the ridge model fitting, while fed the raw text directly to the DistilBERT model as it needed contexts. The performance and application of these models will be shown in the result section.

Results

LDA Models and Labeling

Based on the modeling and labeling strategy shown above, I classified all of the observations into the three categories: federal, state and local. Figure 3 shows the proportion of each category. Since I used z-scores to standardize the scores of the three categories, they are more comparable in scale, and thus have a relatively balanced distribution.

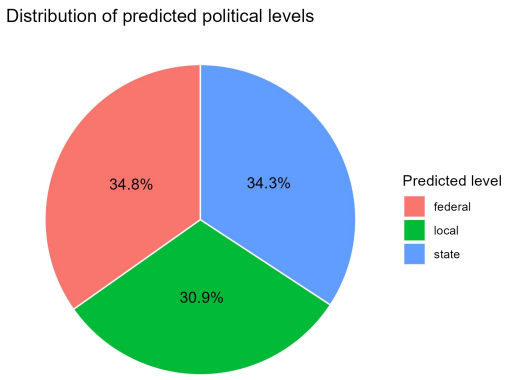

Figure 3

Figure 4 shows the average scores of these three categories grouped by official characteristics. In Figure 4a, the observations are grouped by the office level of officials. We can see in general, the higher is the office level, the more is the attention towards all of these three categories. Local officials have the lowest interest in these issues, state officials are around the baseline, and federal officials have the highest scores for federal and local issues. We can also see that officials of each office levels focus most on the issues of their own level, which validate the labeling strategy. In Figure 4b, the observations are grouped by party affiliations. We can see that Democrats and Republicans have little difference on these scores, while they both have much higher scores than officials from other parties or independent candidates. This implies the propaganda on social media of the major parties have great advantages towards people from marginal positions in the U.S. politics political system.

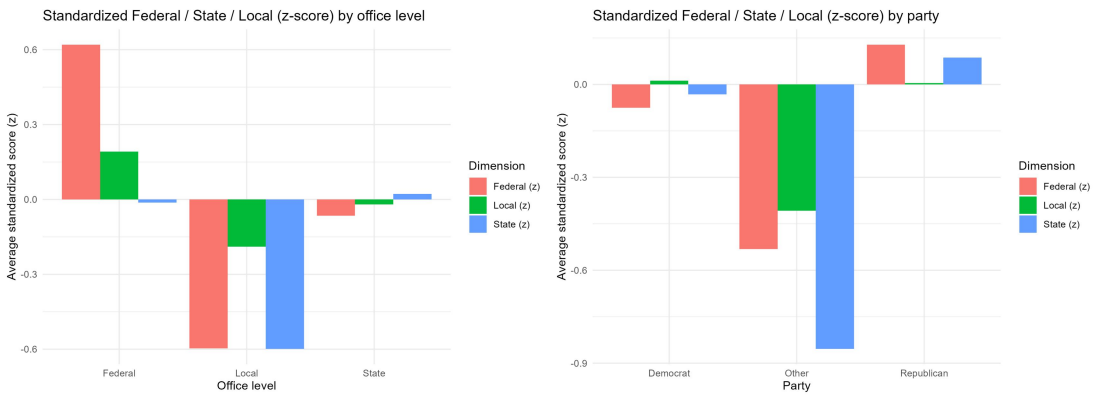

Figure 4

Supervised Learning

After training the ridge and DistillBERT models mentioned before, I evaluated them on a test set and exported the classification report to compare its performance with the baseline (guess all observations as federal, which has the highest proportion in the sample) and each other. As we can see in Figure 4, though the accuracy of these two models are only o.65 and 0.67, they almost double compared to the baseline. Considering that this is a three-class instead of a two-class classification, I think the performance is acceptable. As for the comparison between the two models, the total accuracy of the DistillBERT model is a little less than the ridge model, but it's more balanced across the three levels, and precision and recall. Considering the difference of the computation cost of these two models, the ridge model is more economic.

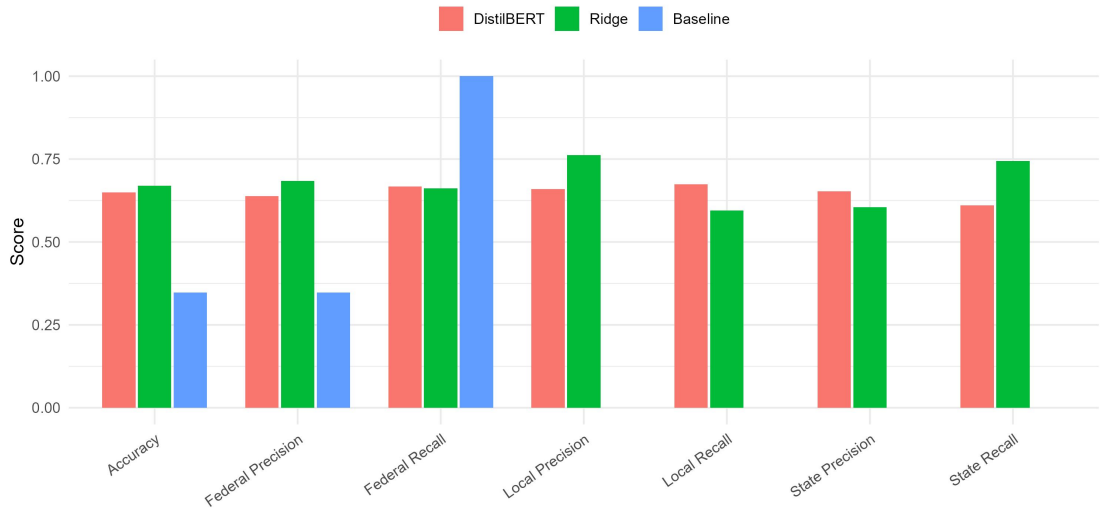

Figure 5

Therefore, I applied the ridge model to a new data set (the tweets_congress data set). I processed the new text data using the same pipeline as the training data, predicted the level for each tweet, and added the predicted labels back into the data set. Figure 6 shows the distribution of predicted topic levels across different subgroups, including party affiliation and chamber. It is important to note that all observations in this data

set are drawn from members of the congress. Given this institutional context, a very small share of content addressing explicitly local-level issues is expected. The consistently low proportion of the local category (approximately 2% across all groups) therefore aligns well with the data-generating process and strengthens, rather than undermines, the credibility of the model's predictions.

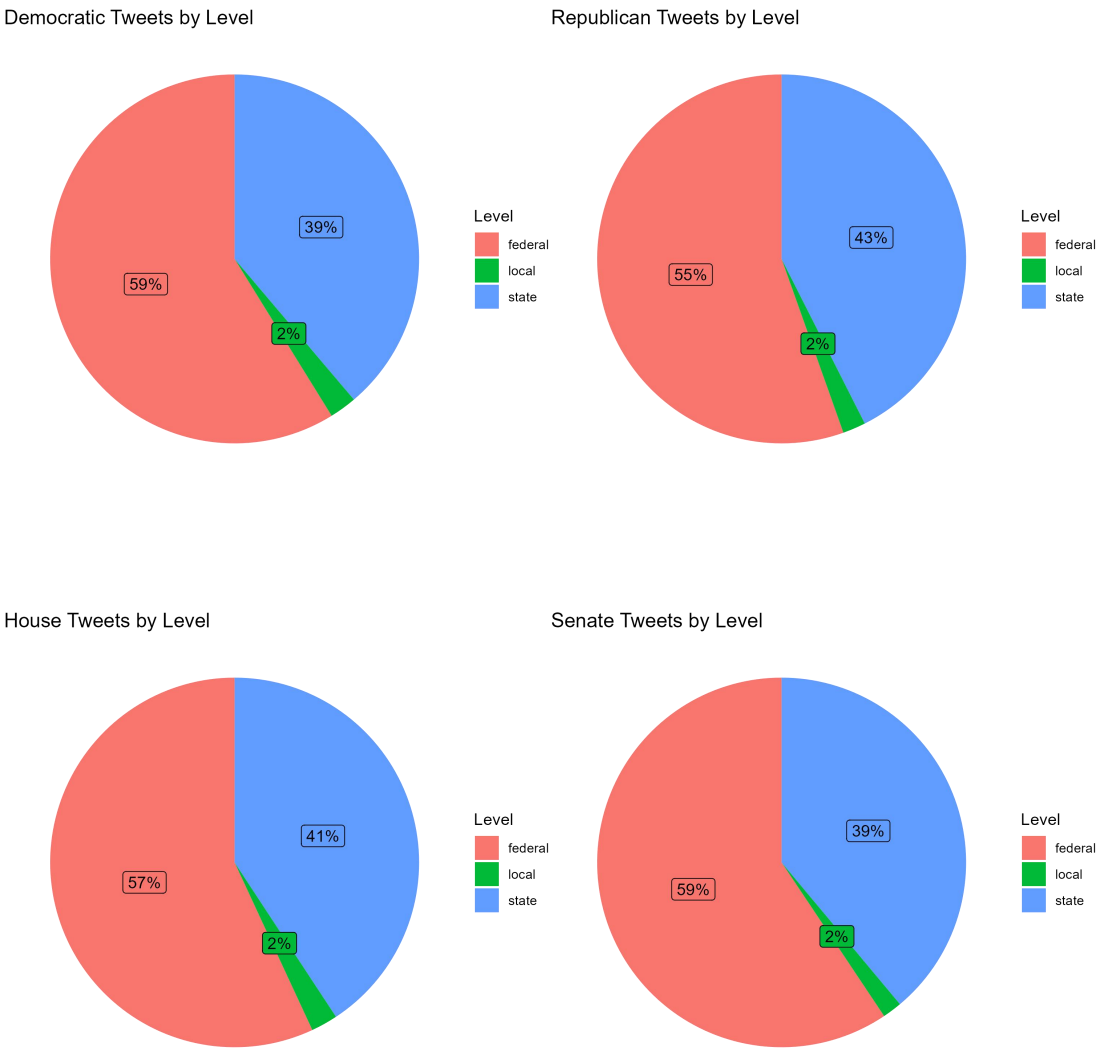

Figure 6

Some little differences in topic distribution can be observed across groups. Republicans exhibit a slightly higher share of federal level topics compared to Democrats, while Democrats are relatively more concentrated at the state level. Similarly, senators show a marginally higher proportion of federal -level issues, while House members display a greater emphasis on state level. However, these differences are small in magnitude and do not reveal clear or systematic patterns that would suggest fundamentally distinct topic-level strategies across parties or chambers. A more conservative interpretation is that these variations likely reflect model

uncertainty or sampling noise rather than substantively different underlying behaviors. Overall, the distributions remain highly similar across groups, indicating that topic-level emphasis is broadly consistent within this dataset.

Discussion

This project develops a multi-stage pipeline to classify the issue level of social media texts under a federal system, with a particular focus on distinguishing federal, state, and local level attention. Rather than relying on manual annotation or predefined labels, the core contribution of this study lies in its labeling strategy. By combining an unsupervised topic model with word embeddings, I constructed labels that were grounded in the latent semantic structure of the text while remaining interpretable in substantive political terms.

Specifically, latent topics inferred by the LDA model were mapped onto federal, state, and local dimensions through embedding-based cosine similarity. This approach avoided the need to assign discrete human-readable labels to topics, thereby preserving the probabilistic nature of topic membership. Each document was then labeled based on a weighted aggregation of topic-level semantic scores, allowing documents to reflect mixtures of issues rather than forcing sharp categorical boundaries at the topic level. The orthogonalization of embedding dimensions further improved this process by addressing semantic overlap among closely related concepts, resulting in more distinct and informative document-level scores.

There are several promising directions for future research. First, expanding the training data would likely improve model stability and classification performance. Second, the labeling procedure itself could be refined by incorporating richer sets of reference terms, domain-specific embeddings trained on political corpora, or alternative methods for disentangling correlated semantic dimensions. Finally, the current framework could be extended beyond the federal–state–local distinction to study other latent dimensions of political communication, such as policy domains or governance functions. Taken together, these extensions would further strengthen the generalizability and usefulness of the approach proposed in this study.

References

Barberá, P., Casas, A., Nagler, J., Egan, P. J., Bonneau, R., Jost, J. T., & Tucker, J. A. (2019). Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data. *American Political Science Review*, 113(4), 883 – 901. https://doi.org/10.1017/S0003055419000352

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 993 – 1022. https://www.jmlr.org/papers/v3/blei03a.html

Motolinia, L. (2021). Electoral accountability and particularistic legislation: Evidence from an electoral reform in Mexico. *American Political Science Review*, 115(1), 97 – 113. https://doi.org/10.1017/S0003055420000672

Rheault, Ludovic, and Christopher Cochrane. (2020).Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora.*Political Analysis*, 28(1), 112 – 133. https://doi.org/10.1017/pan.2019.26

Tai, Y. C., Nakka, N., Patni, K. N., Rajtmajer, S., Munger, K., Lin, Y.-R., & Desmarais, B. (2024). Digitally Accountable Public Representation (DAPR) data (Version 8) [Dataset]. *Harvard Dataverse*. https://doi.org/10.7910/DVN/A9EPYJ