TWEET LIKE A POLITICIAN: USING TWEETS TO PREDICT IDENTITY CHARACTERISTICS OF STATE-LEVEL POLITICAL FIGURES IN THE UNITED STATES

WHITE PAPER

Alexander Adams
McCourt School of Public Policy
Georgetown University
Washington, DC
aja149@georgetown.edu

May 9, 2022

ABSTRACT

An emergent phenomenon in American politics is the trend of nationalization of politics, the eliding of state or local variations in favor of unified national party lines. However, it is unclear how far this trend has progressed; it is also unclear if state-level officials, who are directly responsible for and responsive to a state, rather than the whole country, might be more immune to this polarization. I fit a BERTopic model to a corpus of tweets (n = 48,249) from governors, lieutenant governors, secretaries of state, attorneys general, and state treasurers in order to identify the prominent themes in their tweets. I then trained a linear support vector classifier to predict officials' state of residence, political office, and political party. The classifiers were most successful in predicting state of origin and political party, suggesting that these characteristics are more dominant in politicians' communications than the goings-on of their offices. I also trained a Wordfish model on the corpus, and found that it was more successful in estimating partisan ideal points for governors than for treasurers. Ultimately, I conclude that the these tweets show that the nationalization and increasing polarization of American politics is becoming visible at the state level, and that state identity and partisan identity are clearer signals in Twitter communications than political office.

Keywords Twitter · Politics · States · Linear Support Vector Classifier · Topic Modeling · Ideal Points

1 Introduction

The late congressman and Speaker of the House Thomas "Tip" O'Neill once famously quipped that "all politics is local". Every person, everywhere, has a particular accent and pattern of speech informed by the cultural context in which they grew up and currently live. For politicians, achieving victory in an election often depends in part on their ability to correctly reproduce the cadences of their electorate. However, politicians also face cross-pressures in communication. It is not merely enough to talk like a local; one must also speak like an officeholder, as well as a member of a particular political party. Furthermore, one must *tweet* like someone who belongs to those groups (Papacharissi, 2012).

Politicians at all levels of government must manage this balancing act. The increasing polarization and nationalization of American politics has meant that increasingly, some of the more local- or office-specific conventions may be eschewed in favor of partisan rhetoric, particularly in service of national interests (Conover et al., 2021), (Sievert & McKee, 2019), (Abramowitz & Webster, 2016). Among the consequences of this increased polarization is the progression of partisan sorting. In previous decades, there were factions of liberal Republicans and conservative Democrats, and the two parties routinely demonstrated ideological overlap. Increasingly, liberals associate themselves with the Democratic party, and conservatives with the Republican party. This results in fewer "moderate" or "centrist" candidates running and winning in many districts or states, particularly where partisanship is especially extreme (Thomsen, 2014).

However, state-level politics is often somewhat removed from the dynamics at play in the federal government. As such, if the goal is to determine if politicians from different states or in different offices communicate differently, it may make more sense to examine communications from state governments, as opposed to members of Congress. State-level politicians and their Twitter use have already been covered in various ways in the literature (Cook, 2017), (Kim et al., n.d.), (Casas et al., 2020); however, the bulk (if not the totality) of this literature focuses on state legislatures, rather than state executive branches. This is understandable, since there are more state legislator than governors or attorneys general, though it means that the Twitter presences of politicians in state executive branches constitute a gap in current knowledge.

The aim of this project is to determine the effectiveness of tweets as indicators of a given politician's state, office, and political party. If state-level politicians' communication styles are more informed by their states or office than their political party, then a classifier trained to predict each of those labels should be most effective at predicting a state.

2 Data Sources

For this analysis, I used a dataset of 49,717 tweets, scraped from 369 twitter accounts belonging to 226 state-level elected officials. 1,468 of these tweets were parsed by Twitter as being in a language other than English; these were excluded, leaving 48,249 tweets in the corpus. Tweets were scraped using the Python 'twint' library, and were scraped from politicians' official twitter accounts, campaign twitter accounts, and in four cases, their personal twitter accounts.[1] The offices included in this analysis were Governors, Lieutenant Governors, Secretaries of State, Attorneys General, and State Treasurers.[2] The earliest tweets included in the dataset are by Alabama State Treasurer Young Boozer (R), and are from 2010; over 99% of all tweets included in the dataset are from 2018 or later. Tweets were scraped on or prior to March 30, 2022, and the officials included in this dataset were all in office as of that date. None of the tweets or label metadata exhibit any amount of missingness; every tweet can be matched to a twitter account, which can in turn be matched to an individual. For each task, the tweets were tokenized (i.e. split into individual word elements), and extraneous text such as numbers or URLs were discarded.

The dataset used for this project also includes the name of each official, their state of residence, the office they hold, the date they took that office, and their political party. Politicians from the District of Columbia, Puerto Rico, American Samoa, Guam, the Northern Mariana Islands, or the U.S. Virgin Islands are not included in this dataset. Tables of officials' names, parties, offices, and relevant career dates were scraped from Wikipedia.

3 Methodology

I conducted five tasks for this analysis. The first task is a topic model of the whole corpus of tweets, to see if the content in the tweets mirrored national issues or contained more local themes. Next, I performed a multiclass classification task to predict the state associated with each elected official. Then, I performed another multiclass classification task to predict the office of each elected official. Following that, I performed a binary classification task to predict the political party of an elected official. Finally, I generated ideological ideal points based on tweets, broken down by office.

3.1 Topic Modeling: BERTopic

I began this project by fitting a topic model to the entire corpus of english-language tweets. I chose to use BERTopic(Grootendorst, 2022) with TF-IDF (term frequency-inverse document frequency) vector embeddings for this model. The BERTopic package implements a multi-step algorithmic process to generate topic-based clusters from documents. First, I used a regex pattern to drop all URLs from the tweets, and to exclude all tweets where the content consisted solely of a URL. I then generated TF-IDF embeddings using the TF-IDF vectorizer from the *scikit-learn* (Pedregosa et al., 2011) package, specifying five as the minimum number of documents (meaning a token had to occur in at least five tweets in order to be kept in the feature list). I did not specify a maximum number or proportion of documents, and I only used unigrams for this task.

This produces a large sparse matrix of features, where each feature is a unique token in the text. From there, BERTopic uses the UMAP algorithm to perform dimensionality reduction (i.e. to reduce the number of features in the matrix). Next, it uses HDBSCAN, a clustering algorithm, to generate clusters of documents based on the reduced-dimension

[1] The four officials with personal accounts I was able to find were Colorado Governor Jared Polis (D), Kentucky Attorney General Daniel Cameron (R), Texas Governor Greg Abbott (R), and Utah Governor Spencer Cox (R).

[2] While all 50 states have a governor, the other four offices each only exist in 43-47 states. In some cases, a state may have a differently-titled office which fulfills the same responsibilities (i.e. Pennsylvania has a Secretary of the Commonwealth instead of a Secretary of State); these were treated equivalently to their peer offices.

feature matrix. It then finds the most common or representative words in that cluster, and uses those as topic labels. The total number of topics can also be reduced further, for increased interpretability.

3.2 Classification: Linear Support Vector Machine

For analyses 2-4 for this project, I used a Linear Support Vector Machine (SVM), also from the *scikit-learn*(Pedregosa et al., 2011) package. To preprocess the tweet data for the SVM, I used a TF-IDF Vectorizer from *scikit-learn*, with the following parameters specified: a token must occur in a minimum of 5 documents (where each tweet is a document) and a maximum of 80% of the total number of documents, and the ngram range should include unigrams and bigrams.[3] This vectorizer first tokenizes input text, splitting each document into constituent words. It then removes stopwords (words which do not contribute meaningful information in a document, such as "the", "an", or "so"). Finally, it calculates the TF-IDF scores for each token for each document, resulting in a large sparse matrix where all non-zero values are ratios of term frequency in that document to term frequency across the corpus.

I chose to use an SVM because, unlike many other types of classifiers such as decision trees, random forests, or artificial neural networks, SVMs are robust to overfitting (Awad & Khanna, 2015). They also routinely demonstrate a high degree of effectiveness at classification. Additionally, unlike tree-based classifiers or nearest-neighbor classifiers, SVMs can achieve this without the need for hyperparameter tuning, thus saving time during the investigative process. Additionally, using a support vector machine for all three classification tasks removes the potential for bias which could arise from the use of different algorithms, and allows for a greater validity of cross-comparison.

3.3 Ideal Points: Wordfish

The final analysis I performed for this project uses wordfish. Wordfish is a scaling algorithm which generates a Poisson distribution from word frequencies in texts. It then uses that Poisson distribution to calculate unidimensional scores, which can be interpreted as ideal points for political ideology (Proksch & Slapin, 2008). Like other algorithms, Wordfish takes in a document-term matrix; however, this time it is merely a standard DTM, and not a TF-IDF matrix. It also requires two documents to be specified in order to clarify the direction of the unidimensional Poisson distribution.[4] Unlike tasks 1-4, which I executed in Python, task 5 was executed in R, due to the lack of ideal point generation libraries for Python.

Due to the size of the resultant matrices, it was not possible to run a wordfish model on the entire corpus. Instead, I aggregated tweets by elected official, and then fit the wordfish model to subsets of the data, filtered by elected office (i.e. just governors or just lieutenant governors). In this way, I was able to assess not only if tweets indicate political party (see Results: Question 4), but if tweets can indicate partisanship relative to other peer officials.

4 Results

4.1 Question 1 (Topic Modeling): What Do State-Level Elected Officials Tweet About?

The United States' federalist system of government means that while every state is affected by national issues, each state also has its own set of internal political issues with which it must contend. To that end, an initial object of my analysis was to determine what topics or themes were present in the corpus. These topics would indicate what kinds of things state-level politicians communicate about on social media.

When the BERTopic model with TF-IDF embeddings was fit to the corpus of tweets, it originally produced 374 topics. To facilitate interpretation, these were then further simplified down to 10 topics. Additional testing using the "auto" argument[5] for the number of clusters simplified the clusters down to approximately 180. Figures 1 and 2 below show these clusters represented in two dimensions.

[3] These parameters are adjustable through DVC.

[4] This informs how users should interpret the output. For this project, points further to the left are more liberal, while those further to the right are more conservative. See section 4.5 for more information.

[5] Using "auto" for topic reduction with BERTopic causes it to simplify only topics with a greater than 90% overlap. Simplifying further down to 10 topics requires BERTopic to lower that threshold.

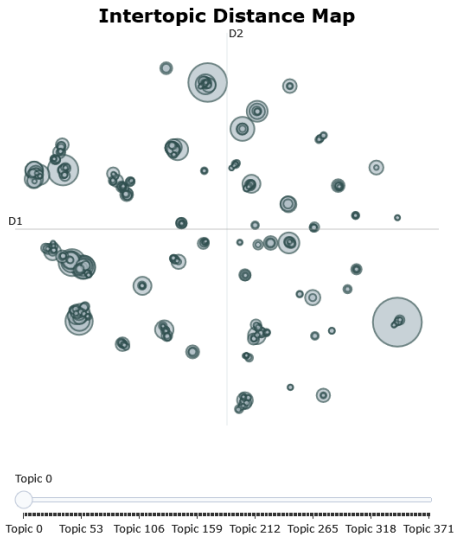


Figure 1: All 374 topics

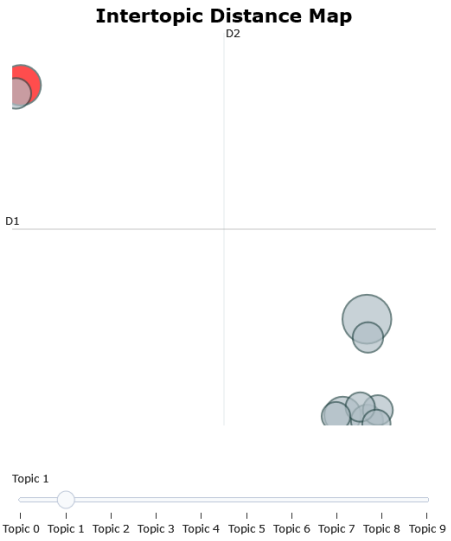


Figure 2: 10 most prominent topics

It is clear from the figures above (especially figure 2) that there is significant thematic distance between some of these clusters. Viewing the top words associated with each offers further clarity:

Table 1: Top words and frequencies for 10 most prominent topics

Topic	Count	Top Words
1	1645	0_vote_your_election_ballot
2	1126	1_christmas_happy_merry_and
3	937	2_thank_you_for_your
4	825	3_read_more_statement_the
5	556	4_happy_birthday_day_patricks
6	550	5_covid19_get_vaccine_and
7	541	6_gas_energy_tax_the
8	521	7_tune_live_watch_covid19
9	477	8_veterans_who_the_served
10	474	9_care_health_to_access

What stands out the most here is how anodyne many of these topics are. Topic 2 is focused on Christmas, and topic 8 is likely derived from the messages which accompany livestreams (i.e. "Tune in now to watch Governor XYZ discuss..."). COVID-19 also appears as a topic, specifically with words related to getting vaccinated. The main political issues present here besides COVID-19 are topic 1 (elections and voting), topic 7 (energy and gas taxes), and topic 10 (access to health care).

The BERTopic library includes a visualization tool to plot topic frequencies over time. Figure 3 below shows those trends:

Figure 3: Topics over time, January 2018-March 2022

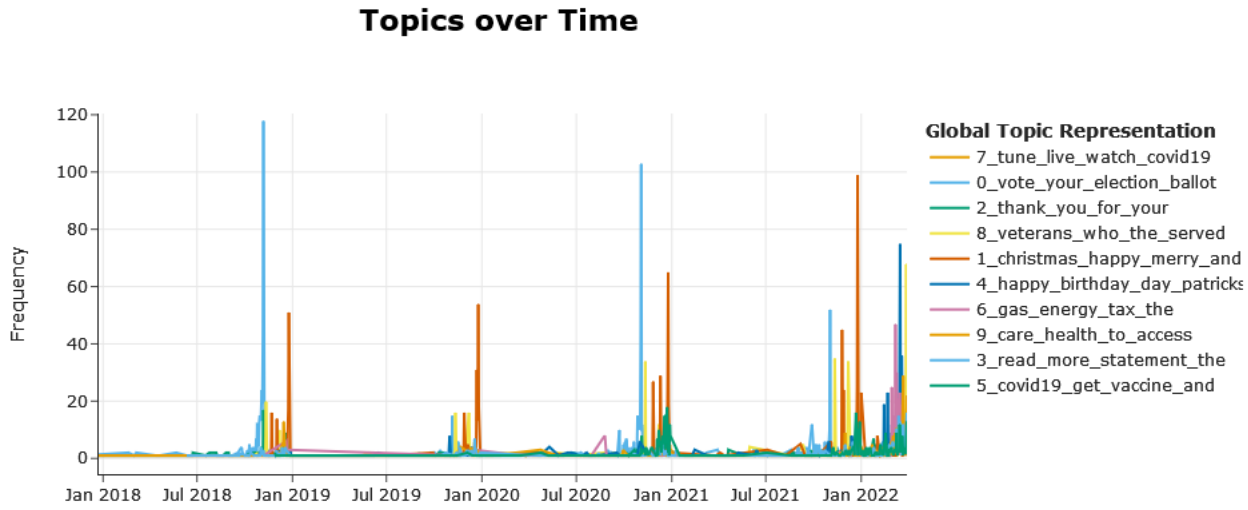


Figure 3: *Line graph showing the prevalence of each topic listed in table 1 at different points from January 2018 to March 2022.*

Figure 3 shows that all 10 topics spike during winter, at the end of each year. Topic 1, which encompasses elections and voting, consistently spikes in and around November. This is not surprising, since most U.S. elections are held in November. Topic 7, about veterans and military service, spikes at similar times. Again, this makes sense, because Veterans' Day in the United States is on November 11. The topic related to Christmas spikes slightly later, right before the end of the year, when Christmas occurs. Topic 6, with keywords "covid19" and "vaccine", mainly tracks the ebb and flow of the pandemic (which tends to line up with seasonal changes; note the relative spike in late 2021, during the Omicron wave).

These topics seem like they could have been generated from any politician's tweets, at any level of American government. The political issues present in these keywords (elections, energy and gas, COVID-19, and healthcare) are universal issues. It does not appear that any local topics were significant enough to break through the deluge of national issues, or that any region-specific trends are present at the highest, most abstracted levels of this corpus.[6] The model card for this model can be found in Appendix B.1.

4.2 Question 2 (Classification): Can tweets be used to predict the state an official leads?

This question is a multiclass classification problem with 50 target classes. To address it, I used the *scikit-learn* TF-IDF vectorizer to tokenize and vectorize the corpus of tweets. From there, I fit the linear support vector machine to the corpus. This classifier was extremely effective. The mean precision and recall scores were both approximately 0.98, and the precision and recall scores for all 50 states were greater than 0.9 (the lowest being the precision score for Colorado, which was 0.911). The full precision and recall score output, along with the absolute numbers of correctly- and incorrectly-classified tweets, can be found in the appendix. The column "errors" displays the labels generated as false negatives.[7]

This suggests that individual states are distinctive enough that in the majority of cases, given no identifying information besides tweets, a classifier can correctly assign a corresponding state. This would seem to support the hypothesis that politicians from different states tweet differently.

However, there is a confounding element to this. After local and state identity, another prominent form of geographic identification is regional identity, with many Americans identifying to varying degrees as southerners, midwesterners, new englanders, or with other regional groups. Given this, one would expect that if a tweet from, for example, an

[6] i.e. No topics which only apply to the South, Midwest, West Coast, etc.

[7] Ex. For the first row, Alaska, 225 tweets were classified correctly as originating from an Alaskan politician, and 1 was classified incorrectly. The incorrectly-classified tweet was mislabeled as originating from Rhode Island.

Alabama politician were misclassified, it would be more likely to be misclassified as coming from another southern state than from a non-southern state. This does not appear to be the case. An examination of the false negatives in the appendix shows that the only southern states which were false negatives for Alabama are Kentucky and North Carolina. If there is a pattern in which states get incorrectly predicted as labels, it is not immediately visible. This also suggests that regional identity may be too abstracted of a linguistic characteristic to be observable through social media communications, though more inquiry is necessary to establish the validity of this assertion. These trends are further obscured by the fact that California and Colorado show up as false negatives for a large number of states. The model card for this model can be found in Appendix B.2.

4.3 Question 3 (Classification): Can tweets predict the office a politician holds?

It seems reasonably clear that politicians tweet like their states, but how much do they tweet like their offices? I again preprocessed the tweet corpus using the TF-IDF vectorizer, and fit the linear support vector classifier to the corpus. This time, the labels were politicians' offices (Governor, Lieutenant Governor, Secretary of State, Attorney General, or State Treasurer).

Figure 4 below shows the confusion matrix for this classifier. Overall, the classifier performs relatively well, though it performs worse than the state classifier in section 4.2. The largest error category are tweets by lieutenant governors being predicted as being by governors. This may be due to class imbalance (there are almost twice as many tweets from governors in the corpus as tweets from lieutenant governors). Across all non-governor offices, governor is the most likely error category, and treasurer the least common.

Figure 4: Office Classifier Confusion Matrix

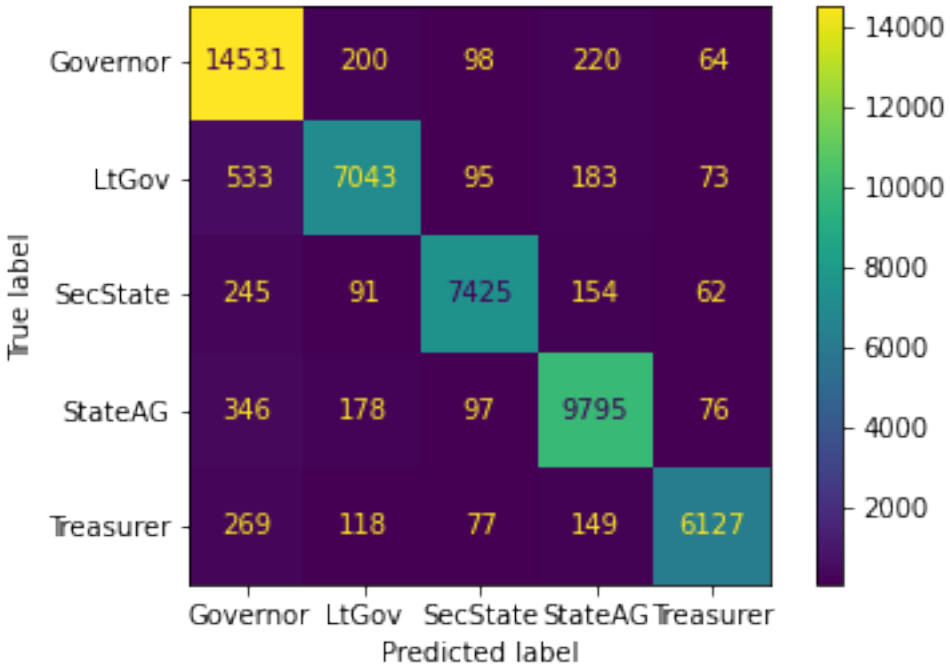


Figure 4: *Confusion matrix for a classifier which predicts political office based on tweets.*

Table 2 shows the precision and recall statistics for each office. The mean recall score for this classifier is slightly higher than the mean precision score, but both are 0.05 less than the corresponding scores for the state classifiers. This occurs in spite of the significant reduction in classes. While there are approximately the same number of tweets from both lieutenant governors and secretaries of state, the classifier was notably more accurate when classifying tweets by secretaries of state, and exhibited correspondingly higher precision and recall. As such, it is possible that lieutenant governors have less distinctive tweets than other state-level officials. On some level this makes sense: secretaries of state are often in charge of election administration and other administrative duties, attorneys general focus on law, and

state treasurers are in charge of their states' finances. Lieutenant governors, in contrast, have less clearly defined roles, or at least responsibilities which are less constrained to one particular domain of governance. Additionally, the past two years have seen secretaries of state and state attorneys general take on more prominent roles as Republicans in those roles have promoted conspiracy theories and specious legal claims about the 2020 election; in contrast, lieutenant governors have not achieved that same salience. The model card for this model can be found in Appendix B.2.

Table 2: Evaluation Metrics for Office Classifier

Office	Correct	Incorrect	Support	Precision	Recall
Governor	14531	582	15113	0.961490	0.912522
LtGov	7043	884	7927	0.888482	0.923067
SecState	7425	552	7977	0.930801	0.952900
StateAG	9795	697	10492	0.933568	0.932768
Treasurer	6127	613	6740	0.909050	0.957045

4.4 Question 4 (Classification): Can tweets predict the political party of a state-level political official?

The previous two classifiers indicate that state identity is easier to perceive through tweets than political office. However, this project rests in part on the idea that state-level officials are less partisan than their peers in the federal government. For this to be true, a classifier should be less effective at predicting political party than at predicting other labels (state or political office in this case). This is a binary classification problem, where the two categories are Democrat and Republican.[8] Figure 5 below shows the confusion matrix for this classifier.

Figure 5: Political Party Classifier Confusion Matrix

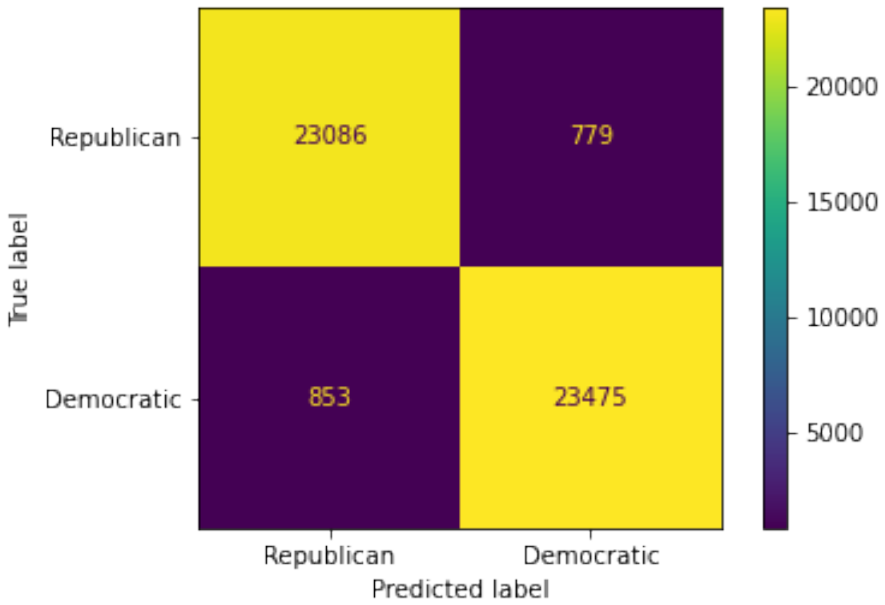


Figure 5: *Confusion matrix for a classifier which predicts political party based on tweets.*

Figure 5 shows that the SVM performed equivalently at classifying tweets from Democrats and Republicans; the numbers of correct and incorrect predictions for each party are almost identical. In this case, the classes are balanced,

[8]Michigan Treasurer Rachael Eubanks and Virginia Treasurer Manju Ganeriwala are Independents; they were excluded from this portion of the analysis. Minnesota Governor Tim Walz, Lieutenant Governor Peggy Flanagan, Secretary of State Steve Simon, and Attorney General Keith Ellison are all members of the Democratic Farmer-Labor Party, Minnesota's state affiliate of the Democratic party. For the purposes of this analysis, all four of these officials were recoded as belonging to the Democratic party.

and unlike with political offices (where some offices might be more distinctive than others), the two parties are comparatively distinct. The precision and recall for both parties are each approximately 0.98, a stronger performance than the political office classifier in section 4.3. This suggests that the hypothesis that partisanship among state-level officials is not strong enough to dominate other factors, such as state identity or political office, is incorrect. That this hypothesis turns out to be unsupported is not entirely unsurprising; the two major political parties diverge on nearly every issue, and even on details as fundamental as support for democratic governance. It is fully plausible that such differences would manifest themselves through politicians' tweets to an extent detectable by standard machine-learning algorithms. In particular, the algorithm's ability to effectively classify tweets according to partisan lean may be informed by the COVID-19 pandemic. COVID-19 was one of the top 10 topics in this corpus (see section 4.1), and Democrats and Republicans across the United States responded very differently to the spread of the virus. Leaders diverged on public heath mandates, mitigation efforts, relief, resource allocation, community guidance, and personal conduct and rhetoric. While there are many issues where the left and right are far apart, COVID-19 stands out as perhaps the most salient, and perhaps the most present. Few other political issues demonstrate such a clear separation between the parties; it is possible that if this same analysis were conducted on tweets from a different time period, the classifier would be less successful, owing in part to the lack of such a clarifying issue. The model card for this model can be found in Appendix B.2.

4.5 Question 5 (Ideal Points): Can tweets predict *how* partisan an elected official is?

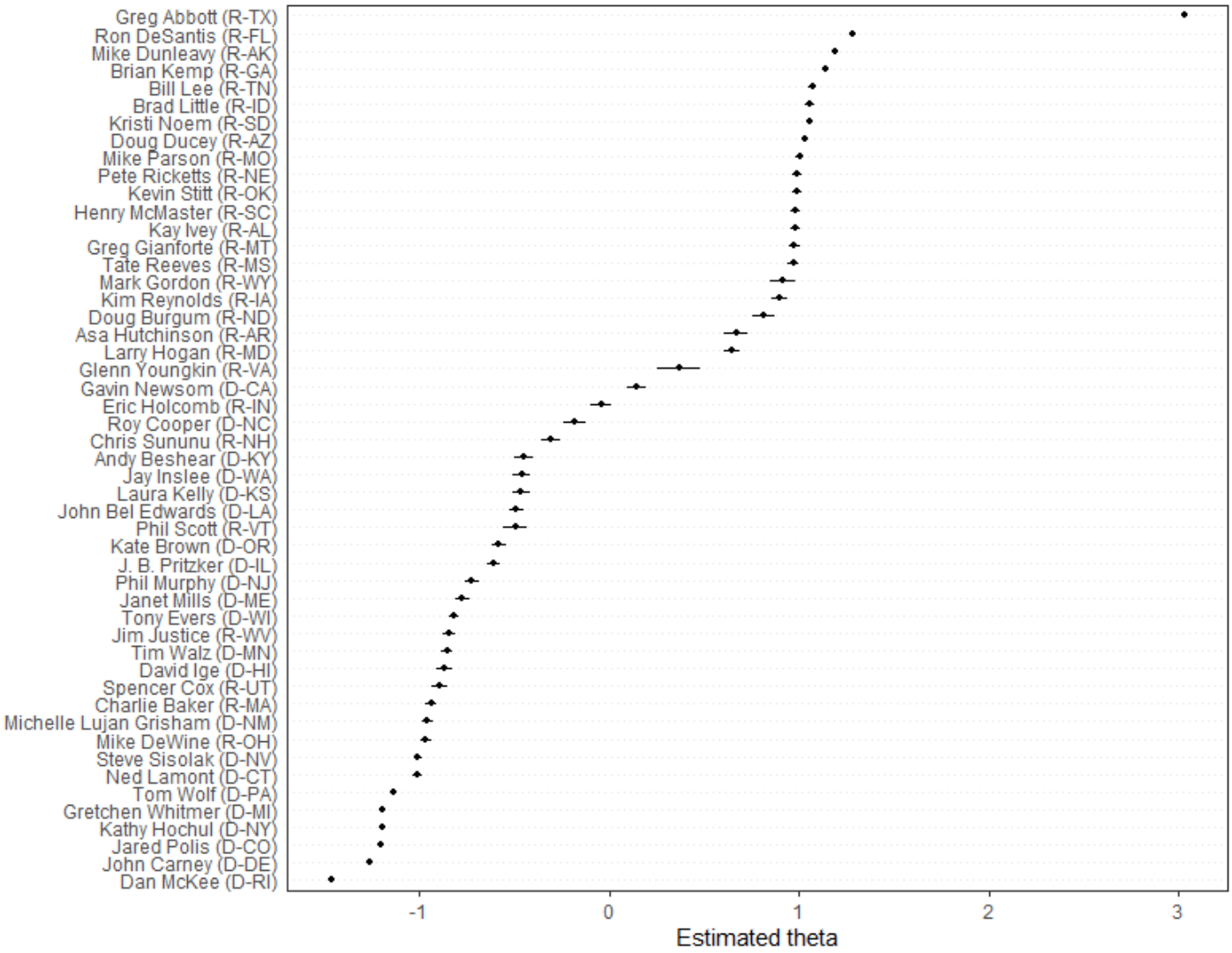
Section 4.4 establishes that tweets can reliably be labeled as originating from Democrats or Republicans with a high degree of both precision and recall. Granting this, it is then worth asking if tweets can describe *how* Democratic or Republican the tweeter is. This value is known in political science as an "ideal point". The median voter theorem is predicated in part on ideal points: the theory proposes that, given a choice between a field of candidates, a voter will select the candidate whose ideology is closest to their ideal point, regardless of the direction of the difference. As such, the optimal strategy for a candidate for office is to tailor their positions to be closer than their opponents' to the ideal point of the median voter (i.e. the voter whose vote secures them a majority).

I initially planned to fit the *wordfish* model on the entire corpus; however, the resultant tokenized matrix was too large for R to hold in memory. To address this, I split up the dataset by political office. The algorithm failed to converge for lieutenant governors, secretaries of state, and attorneys general, meaning that ideal points were unable to be generated for those offices. The algorithm was able to successfully calculate ideal points for governors and state treasurers.

To generate ideal points, Wordfish requires two indices α and β which correspond to documents in the corpus, such that $\theta(\alpha) < \theta(\beta)$. To estimate ideal points for governors, I specified Oregon Governor Kate Brown (D) as α and Alabama Governor Kay Ivey (R) as β. Values further to the left indicate a more liberal ideal point, while values further to the right indicate a more conservative ideal point. The clearest outlier on this graph is Texas Governor Greg Abbott (R), whose estimated ideal point is over 3 (meaning very conservative). In contrast, the most liberal governor based on this corpus is Rhode Island Governor Dan McKee (D), whose ideal point is close to -1.5. It is notable how much this graph conforms to popular perceptions: the governors with the most liberal ideal points are Democrats, and those with the most conservative ideal points are Republicans. Governors known for being politically moderate, like North Carolina Governor Roy Cooper (D), New Hampshire Governor Chris Sununu (R), and Kentucky Governor Andy Beshear (D), appear in the middle of the graph. Some governors appear to be out of place; California Governor Gavin Newsom has an ideal point close to zero and almost identical to that of Virginia Governor Glenn Youngkin (R),[9] when in reality he is among the more liberal governors in the U.S. Certain Republicans who are perceived as moderate or even liberal, like Vermont Governor Phil Scott (R) and Massachusetts Governor Charlie Baker (R), have ideal points to the left of several Democrats. Notably, the furthest left Republican governor on this plot is Ohio Governor Mike DeWine, who is not especially liberal, or even particularly moderate. Additionally, several of the Democratic governors with the furthest left ideal points, like New York Governor Kathy Hochul and Colorado Governor Jared Polis, are considered moderates or even centrists relative to the rest of their party. These values can be seen in Figure 6 below:

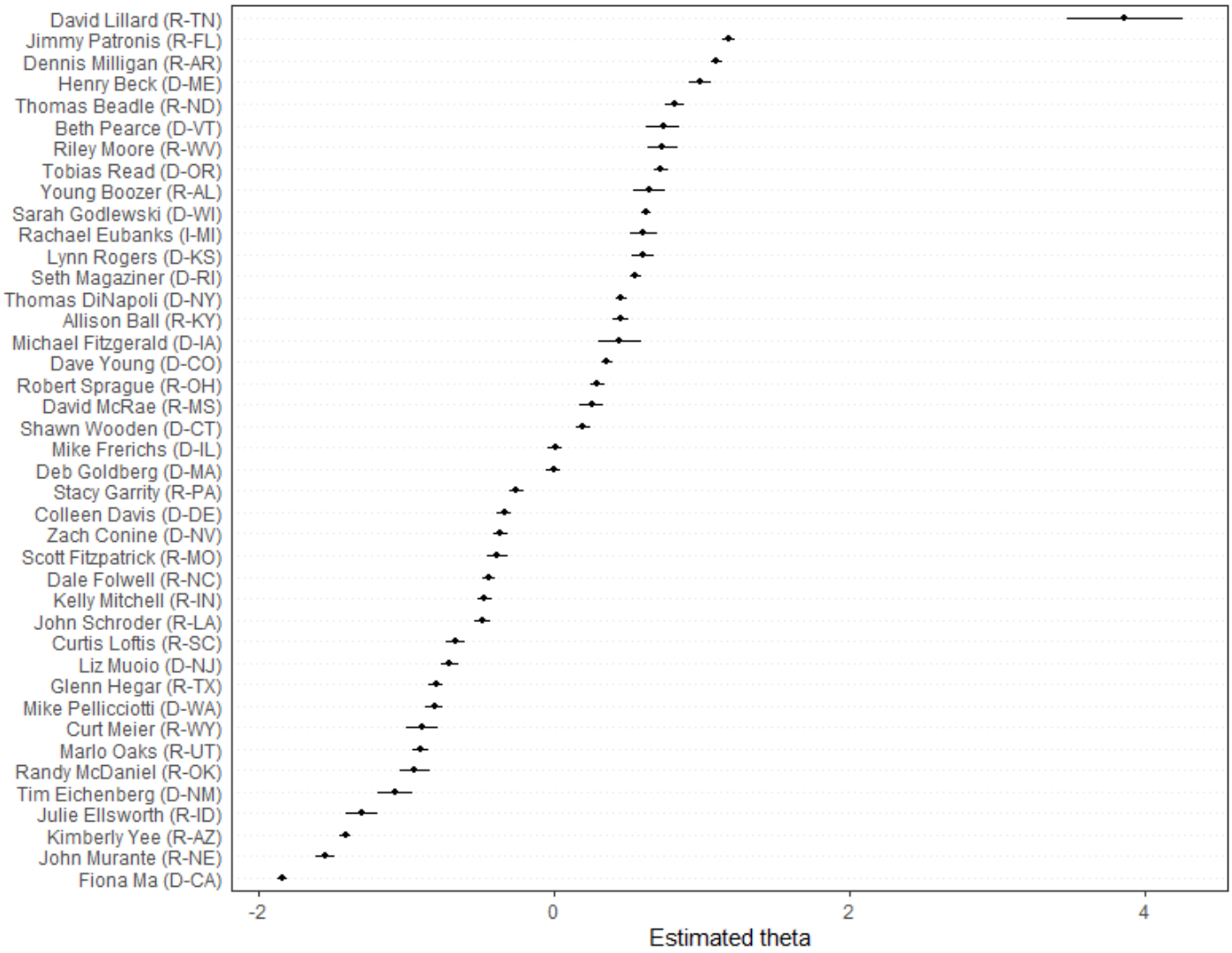
[9] Youngkin's relative lack of ideological polarization could be due to the relatively small number of tweets he has posted since taking office, rather than his actual ideological leanings.

Figure 6: Left-Right Ideal Points, Governors



The values shown in Figure 6 should be contrasted with those shown in Figure 7, which shows the estimated ideal points for state treasurers. To generate these ideal points, California State Treasurer Fiona Ma was used as α, and Alabama State Treasurer Young Boozer was used as β. While the governors are approximately correctly ordered from left to right, the state treasurers are not ordered in any way which corresponds neatly to partisanship. Tennessee State Treasurer David Lillard (R) has an estimated ideal point of almost 4, significantly more conservative than even Texas Governor Greg Abbott. Maine State Treasurer Henry Beck (D) is the fourth-most conservative, as per these ideal points, and Vermont State Treasurer Beth Pearce (D) is the sixth-most conservative. In comparison, Nebraska State Treasurer John Murante (R) and Arizona State Treasurer Kimberly Yee (R) are the second- and third-most liberal state treasurers, respectively. If these ideal points are so thoroughly divorced from ground-truth markers of partisanship (i.e. party identification), then perhaps partisanship is less present in tweets from state treasurers than in tweets by governors. A possible explanation for this could be that state treasurer is not always an elected position, whereas all governors are popularly elected. As such, governors have a greater incentive to engage in partisan politics, since the amplification and promotion of partisan stances is integral to their continued employment. The task of state treasurer, meanwhile, is a largely nonpartisan one; all states need a well-composed budget and strong financial planning, regardless of their partisan lean. Thus, while some state treasurers may be political appointees, selected to their roles by their governors, and while they may participate in certain partisan projects as political figures in a state government, they are largely not engaged in the day-to-day work of politicking. They might not be as strong partisans as governors because they simply do not need to be, and that may be reflected in their tweets, and the associated incoherence produced when one attempts to divine a coherent ideological continuum from said tweets. This model can be found in Appendix B.3.

Figure 7: Left-Right Ideal Points, Treasurers



5 Conclusions

Politicians' tweets are rich sources of information. They communicate essential details about current affairs, announce major political events, and help clarify ideological positions. These tweets also indicate with relative clarity the state represented by an official, that person's political office, and their political party. Political tweets can even be used to position politicians on a left-right spectrum, with varying degrees of success. It appears that state-level politics are colored by partisan polarization, and that partisan identification is more easily deduced from a tweet than that person's job.

One limitation of this project is that while I gathered tweets from multiple accounts per official to the greatest extent possible (meaning official accounts, campaign accounts, and in some cases personal accounts), I did not have time to analyze any differences in these accounts or consider the effects of differential communication modes (formal and official versus campaign trail messaging) in any of my analyses. While I expect that politicians tweet differently on different accounts, I did not gather evidence during this project which supports or undermines that hypothesis. I was also unable to find twitter handles for some politicians, particularly in lower-visibility executive cabinet positions in rural states in the mountain west. There is no real way to get around this other than choosing to exclude or include certain other tweets from peer officials in other states, but future analyses could choose to process this data differently. Also, twint, while fast (and not requiring a Twitter Developer account or an API key), sometimes does not scrape all

possible tweets, even accounting for deleted tweets or tweets under different usernames. Thus it is possible that there are tweets I could have included in the corpus which I was unable to retrieve due to the tools available to me. I ran multiple twint scrapes; the data used in this project are the unique tweets from all the scrapes I conducted.

Future inquiries into politicians' tweets could seek to compare tweets by state-level officials to those by federal government officials. They could also attempt to scrape tweets from accounts before and after an official is inaugurated, to track changes in social media rhetoric over time. Given more time, I would have liked to explore differences in topics present among tweets by Democrats versus Republicans, or across different political offices; however, such analysis was outside the scope of this particular project.

6 Acknowledgements

The author would like to thank Professor Thurston Sexton, Adjunct Professor at the McCourt School of Public Policy at Georgetown University. The author would also like to thank the program directors of the Master of Science in Data Science for Public Policy program at the McCourt School, as well as the faculty, staff, and administrators of the McCourt School.

The GitHub repository for this project can be found here: https://github.com/aadams149/ppol628-final-project

The DagsHub repository for this project can be found here: https://dagshub.com/aadams149/ppol628-final-project

References

Abramowitz, A. I., & Webster, S. (2016). The rise of negative partisanship and the nationalization of u.s. elections in the 21st century. *Electoral Studies*, *41*, 12–22. https://doi.org/https://doi.org/10.1016/j.electstud.2015.11.001

Awad, M., & Khanna, R. (2015). Support vector machines for classification. *Efficient learning machines: Theories, concepts, and applications for engineers and system designers* (pp. 39–66). Apress. https://doi.org/10.1007/978-1-4302-5990-9_3

Casas, A., Payson, J., Nagler, J., Bonneau, R., & Tucker, J. A. (2020). Using social media data to reveal patterns of policy engagement in state legislatures. *Available at SSRN 3698990*.

Conover, M., Ratkiewicz, J., Francisco, M., Goncalves, B., Menczer, F., & Flammini, A. (2021). Political polarization on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, *5*(1), 89–96. https://ojs.aaai.org/index.php/ICWSM/article/view/14126

Cook, J. M. (2017). Twitter adoption and activity in u.s. legislatures: A 50-state study. *American Behavioral Scientist*, *61*(7), 724–740. https://doi.org/10.1177/0002764217717564

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Kim, T., Nakka, N., Gopal, I., Desmarais, B. A., Mancinelli, A., Harden, J. J., Ko, H., & Boehmke, F. J. (n.d.). Attention to the covid-19 pandemic on twitter: Partisan differences among u.s. state legislators. *Legislative Studies Quarterly*, *n/a*(n/a). https://doi.org/https://doi.org/10.1111/lsq.12367

Papacharissi, Z. (2012). Without you, i'm nothing: Performances of the self on twitter. *International Journal of Communication (19328036)*, *6*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Proksch, S.-O., & Slapin, J. B. (2008). Wordfish: Scaling software for estimating political positions from texts. *Version*, *1*, 323–344.

Sievert, J., & McKee, S. C. (2019). Nationalization in us senate and gubernatorial elections. *American Politics Research*, *47*(5), 1055–1080.

Thomsen, D. M. (2014). Ideological moderates won't run: How party fit matters for partisan polarization in congress. *The Journal of Politics*, *76*(3), 786–797.

7 Appendix A: Classification Output for State Classifier

State	Correct	Incorrect	Support	Precision	Recall	Errors
AK	225	1	226	0.965665	0.995575	[RI]
AL	1050	12	1062	0.988701	0.988701	[CO, KY, MA, MN, NC, ND, OR, RI, VT]
AR	1147	15	1162	0.982021	0.987091	[AL, CA, DE, LA, ME, MI, MO, NC, NM, NY, OH, OK, OR, RI, SD]
AZ	873	15	888	0.995439	0.983108	[CA, CO, IL, MA, MI, OH, OR, RI, UT]
CA	1347	23	1370	0.944600	0.983212	[AK, CO, IL, MA, ME, MN, MS, NC, NJ, NM, NV, NY, OH, RI, VT, WV]
CO	1512	24	1536	0.910843	0.984375	[CA, FL, IL, IN, LA, MA, MO, NC, NJ, NV, PA, RI, VT, WA, WI]
CT	1372	32	1404	0.990614	0.977208	[AR, CA, CO, IA, IN, KY, MI, MN, MO, NC, NJ, NM, RI, UT, WA, WI]
DE	848	8	856	0.989498	0.990654	[CO, CT, LA, MD, ME, NV, OH, SC]
FL	1258	18	1276	0.989772	0.985893	[AL, AR, CA, CO, LA, RI, VT]
GA	859	13	872	0.993064	0.985092	[AL, CA, CO, IA, MA, ME, MI, MT, NM, RI, TX]
HI	353	2	355	0.994366	0.994366	[LA, NY]
IA	1021	27	1048	0.985521	0.974237	[AK, AL, CA, CO, DE, GA, IN, KY, MD, MO, MS, ND, RI, WA, WI, WY]
ID	520	7	527	0.998081	0.986717	[CO, IN, KY, LA, MO, NY, WY]
IL	1478	15	1493	0.984021	0.989953	[AL, CA, CO, FL, IA, IN, KY, MN, MO, OR, TX, UT]
IN	1713	38	1751	0.989601	0.978298	[AR, CA, CO, CT, IA, IL, KY, LA, MI, MO, ND, NE, NM, OH, RI, SC, TX, UT, VT]
KS	857	4	861	0.998834	0.995354	[MI, MO, TN, WA]
KY	1103	29	1132	0.981317	0.974382	[AR, CA, CO, IL, MA, MI, MT, NM, OR, PA, RI, UT, VA, WI, WV, WY]
LA	1111	22	1133	0.991964	0.980583	[AR, CO, CT, FL, GA, IA, IL, KY, ME, MS, NM, PA, RI, TX, VT, WA]
MA	1018	29	1047	0.977906	0.972302	[AL, CA, CO, MI, MO, NC, ND, NE, NJ, OH, OK, PA, RI]
MD	807	10	817	0.993842	0.987760	[AL, CO, FL, IA, MI, NM, OH, VA, WA]
ME	476	21	497	0.981443	0.957746	[CA, CO, MN, MO, NM, NV, NY, RI, UT, WI, WY]
MI	1358	31	1389	0.982634	0.977682	[AR, CA, CO, CT, GA, IN, KY, MN, MS, NC, ND, NM, NV, OH, RI, VT, WA, WI, WV]
MN	1023	23	1046	0.989362	0.978011	[CA, CO, FL, GA, IL, KY, MD, MI, NV, PA, TX, VT, WA, WV]
MO	1040	25	1065	0.980207	0.976526	[AR, CA, CO, IA, IL, IN, MD, ME, MN, NC, NY, OR, RI, WA, WI]
MS	627	9	636	0.990521	0.985849	[AR, ME, NE, NM, NV, RI, UT]
MT	501	5	506	0.990119	0.990119	[AR, MN, ND, OH, RI]
NC	1148	16	1164	0.988803	0.986254	[CO, IL, IN, KY, MA, MI, MO, NM, OH, WI]
ND	564	17	581	0.975779	0.970740	[CA, CO, IL, MA, NJ, RI, WA, WY]
NE	603	9	612	0.991776	0.985294	[AK, AL, CA, MO, NC, ND, TX]
NH	331	1	332	1.000000	0.996988	[OH]
NJ	1150	23	1173	0.989673	0.980392	[CA, CO, ID, IL, KS, MA, ME, NY, OH, PA, RI, TX, VT, WA]
NM	897	21	918	0.977124	0.977124	[AL, CA, CO, CT, FL, MA, MO, ND, NY, PA, WY]
NV	1228	25	1253	0.991122	0.980048	[AR, CA, CO, HI, IA, IN, MI, MN, MO, NJ, OK, OR, PA, RI, VA, WA, WI]
NY	1642	13	1655	0.990350	0.992145	[AK, AZ, CA, DE, MD, MI, NM, OH, OR, RI]
OH	1213	23	1236	0.982982	0.981392	[AR, CA, CO, CT, GA, IN, LA, MI, MS, NJ, NY, RI, UT, VT, WI, WY]

State	Correct	Incorrect	Support	Precision	Recall	Errors
OK	853	22	875	0.994172	0.974857	[AK, CA, CO, CT, FL, GA, IA, IN, KY, MA, MO, NC, NE, NM, VT]
OR	890	22	912	0.987791	0.975877	[AR, AZ, CA, CO, IL, MA, MI, MT, NY, RI, TN, WA, WV]
PA	1311	19	1330	0.991679	0.985714	[AZ, CA, CO, IL, LA, MA, MI, NC, ND, NY, RI, VA, WA]
RI	1208	26	1234	0.937161	0.978930	[AR, CA, CO, CT, IA, IL, KY, MA, MI, MO, NJ, NM, NV, OH, WA, WY]
SC	699	8	707	0.997147	0.988685	[AR, IL, MN, NC, ND, NM, RI]
SD	526	5	531	0.996212	0.990584	[CA, CT, MO, OK, TX]
TN	712	10	722	0.995804	0.986150	[DE, HI, IN, MS, NJ, NV, RI, SD, VA]
TX	1670	20	1690	0.994048	0.988166	[AR, CA, CO, FL, IA, KY, MN, MO, NJ, RI]
UT	679	22	701	0.982634	0.968616	[AL, CA, CO, CT, FL, IN, KY, MA, MI, MO, NM, OK, PA, RI, TN, VT, WY]
VA	494	5	499	0.989980	0.989980	[IL, MA, RI, WI]
VT	604	11	615	0.971061	0.982114	[AK, CO, CT, DE, ME, ND, NE, RI, UT]
WA	1116	14	1130	0.984127	0.987611	[CO, DE, IA, MI, NM, NY, OH, VT]
WI	1236	15	1251	0.985646	0.988010	[CA, CO, CT, IL, KY, MI, ND, NV, NY, RI, UT]
WV	735	19	754	0.987903	0.974801	[AK, CA, CO, FL, KY, OR, RI, WY]
WY	407	12	419	0.976019	0.971360	[CO, IL, KY, MT, NC, NM, RI, WV]

8 Appendix B: Model Cards

8.1 Model 1: BERTopic

- Model Details: Topic model using BERT (Bidirectional Encoder Representations from Transformers). Tokenizes text and generates embeddings from tokens, then applies clustering algorithms to embeddings to identify clusters, which are then labeled as topics for human interpretability. See (Grootendorst, 2022).
- Intended Use: Generate topics from text, similar to existing libraries like gensim or spacy.
- Factors: N/A. BERTopic is capable of processing text in non-english languages, and does not require a minimum or maximum document size or corpus size.
- Metrics: Number of topics generated, number of documents assigned to each topic
- Evaluation Data: Tweets from state-level politicians in the United States. Tweets were preprocessed using TfidfVectorizer from scikit-learn. Tweets were also filtered to include only english-language tweets, and all tweets where the sole text was a URL were dropped.
- Training Data: Same as evaluation data. BERTopic was trained on tweets.
- Quantitative Analyses: Number of topics generated, number of documents assigned to each topic, consistency of topics across multiple iterations of BERTopic on training data
- Ethical Considerations: None known.
- Caveats and Recommendations: BERTopic can take a long time to run, so it may be worth considering other topic modelers for large corpuses or long documents.

8.2 Model 2: Linear Support Vector Classifier

- Model Details: Machine learning algorithm which calculates probabilities that a given instance belongs to each of a specified number of target classes. See (Awad & Khanna, 2015).
- Intended Use: Generate predicted class labels for instances in a dataset based on input data. Can also be used to generate expected values for regression tasks
- Factors: N/A. Research evaluating SVCs compared to other machine learning classifiers finds that SVCs are consistently among the top-performing algorithm.
- Metrics: Number of instances classified correctly, false positives, false negatives, recall, precision, F1 score
- Evaluation Data: Full corpus of tweets. Tweets were preprocessed using TfidfVectorizer from scikit-learn, and were also filtered to include only english-language tweets. Tweets consisting of solely a URL were dropped. Labels were either state (section 4.2), political office (section 4.3), or political party (section 4.4).

- Training Data: 50% of tweet corpus, randomly selected.
- Quantitative Analyses: See confusion matrices and precision and recall scores in sections 4.2, 4.3, and 4.4.
- Ethical Considerations: None known.
- Caveats and Recommendations: SVCs become significantly slower as corpus size and document size increase. However, as they are robust to overfitting (unlike many other algorithms), they are likely a powerful option for classification tasks for many smaller corpuses and documents.

8.3 Model 3: Wordfish

- Model Details: One-dimensional Poisson distribution, based on sparse feature matrix. Assigns ideal points along one dimension at the document level. See (Proksch & Slapin, 2008).
- Intended Use: Generate ideal points representing ideological positions on a left-right continuum based on text data. Conceptually similar to DW-NOMINATE, which finds ideal points based on binary yea/nay roll-call votes in legislatures.
- Factors: Must be able to identify at least two documents in the corpus where the expected ideal point of document α is to the left of β.
- Metrics: Ideal point for each document (symbol: theta), effects of each feature on left-right continuum
- Evaluation Data: Tweets from governors and state treasurers in the United States. Tweets were preprocessed to be tokenized, and URLs and non-word tokens were removed.
- Training Data: Same as evaluation data, no training-test split required since evaluation and training are same task.
- Quantitative Analyses: Do estimated ideal points approximate expectations based on domain knowledge, range of ideal points, distance between ideal points
- Ethical Considerations: None known.
- Caveats and Recommendations: Size. Wordfish struggles to hold sufficiently large datasets or feature matrices in memory, so corpuses and documents should be kept relatively small and short respectively. Also, some corpuses will not converge, and so ideal points cannot be estimated.