On-Line Appendix for "Using Social Media Data to Reveal Patterns of Policy Engagement in State Legislatures"

Appendix A Training of the CNN topic classifier

We trained the model architecture described in the paper (Figure 1) with the four datasets that we describe again in Table A1. The first one is composed of publicly available data. The second one comes from the replication material of a published study, and we created two final dataset for the purpose of this study. In the first dataset (A) we combined all available CAP-labeled datasets for the United States available in the CAP website (789,004 observations in total). The second dataset (B) is comprised of 45,394 tweets from Senators who served during the 113th Congress and that were labeled by Russell (2018). The third set (C) consists of 18,088 tweets sent by media accounts and followers of our state legislators that we coded according to the CAP classification. The fourth dataset (D) consists of 3,368 tweets sent by the state legislators that we also coded.[18]

We trained the same CNN model nine times using the following data combinations, with the goal of taking advantage of transfer learning (Terechshenko et al. 2020) and training more accurate models than simply training the model with the tweets from state legislators that we had coded (so only set D): (1) only set A, (2) only set D, (3) set A and set D, (4) set D and a small sample of set A (1,300 observations), (5) set D and a smaller sample of set A (650 observations), (6) set D and set B, (7) set D and a small sample of set B (1,300 tweets), (8) set D and a smaller sample of set B (650 tweets), (9) set D and set C.

To assess the performance of these nine versions of the model we split the data used in each case into a train and test set. Moreover, we split set D (the tweets sent by state legislators that we coded) into a train, test, and validation set. This validation set is particularly useful for two reasons. First, although the test sets were not used for training the models, they were somewhat involved in the training process, as we decided the number of training iterations based on how well the CNNs predicted the coded documents both in the train and test tests.[19] Assessing accuracy based on a totally untouched validation set hence gives us a better indication of how the model will perform in predicting the topics of the unlabeled tweets. Furthermore, at the end of the day we wanted to specifically know how each CNN performed at predicting the topics in tweets sent by state legislators, rather than the documents in the test sets (which could be a combination of different types of documents: tweets, titles of congressional bills, newspaper headlines, etc.).

In Table A2 we report the accuracy of the nine versions of the model we trained (based on 3-fold cross-validation), based on held-out test sets, and on the validation set composed only of tweets sent by state legislators. We assess the test accuracy when predicting all tweets in the test split (*All*), and also when only predicting the tweets coded as being about one of

[18] See Footnote 12 for information on inter rater reliability for the tweets we coded in this study.

[19] We settled for fifty iterations, as at that point the accuracy based on the train set kept improving while test accuracy started declining.

Table A1: Public datasets coded using the CAP 21-issue classification, used for training and testing a classifier predicting *Policy Issues* in tweets from state legislators.

Set	Dataset	Time	N
A	Congressional Quarterly Almanac	1948-2015	14,444
	New York Times Front Page	1996-2006	31,034
	New York Times Index	1946-2014	54,578
	Congressional Bills	1947-2016	463,762
	Congressional Hearings	1946-2015	97,593
	Public Law Titles	1948-2011	33,644
	Public Laws	1948-2017	20,928
	Executive Orders	1945-2017	4,294
	Presidential Veto Rhetoric	1985-2016	1,618
	State of the Union Speeches	1946-2018	22,289
	Democratic Party Platform	1948-2016	15,953
	Republican Party Platform	1948-2016	19,836
	Supreme Court Cases	1944-2009	9,031
B	Tweets sent by Senators 113th Congress	2013-2015	45,394
C	Tweets sent by media accounts	2018	8,802
	Tweets sent by followers of state legislators	2018	9,286
D	Tweets sent by state legislators	2018	3,368
	Total	1944-2018	855,854

Table A2: Out of sample accuracy of the nine versions of the CNN model we trained predicting the political topics of the Comparative Agendas Project.

	Test Set		Validation Set			
	CNN		CNN		SVM	
Model version	All	Policy	All	Policy	All	Policy
(6) set D and B	0.78	0.79	**0.59**	**0.55**	0.38	0.40
(1) set A	0.73	0.73	0.27	0.53	0.23	0.47
(3) set D and A	0.73	0.73	0.36	0.52	0.44	0.45
(9) set D and C	0.77	0.49	0.66	0.43	0.61	0.31
(7) set D and small B	0.56	0.36	0.61	0.32	0.59	0.27
(4) set D and small A	0.55	0.32	0.60	0.28	0.58	0.27
(8) set D and smaller B	0.57	0.29	0.61	0.29	0.58	0.23
(5) set D and smaller A	0.56	0.28	0.60	0.27	0.59	0.22
(2) set D	0.60	0.26	0.59	0.22	0.57	0.19

the policy areas, so after excluding the non-policy tweets (*Policy*). The tweets not related to any policy area represented a large part of the tweets we coded from state legislators (set D) and we wanted to make sure that our model did well at both distinguishing overall policy relevance and at distinguishing between policy areas.

The model trained with the coded tweets by state legislators plus the coded tweets sent by Senators of the 113th Congress returned the best results. The test accuracy in both cases (all tweets, and just tweets we determine to be about policy areas) is close to 80%, and more importantly, the validation accuracy based on the untouched labeled tweets sent by state legislators is around 60% (very high given that the model is predicting 21 topic classes). In Table A2 we also compare the performance of our CNN models to a baseline n-gram based model, a Support Vector Machine (SVM).[20] We note that the CNN model clearly outperforms the SVM model in our validation set. To use an SVM with accuracy over all tweets (including non-policy tweets) as high as we achieve with our CNN classifier, we would have to choose an SVM trained on one of the sets listed in rows 4 through 9 of Table A2. However, none of those classifiers achieve an accuracy on policy tweets of over 0.31 (compared to an accuracy of 0.55 on policy tweets for our CNN). Thus the CNN model is much preferred to SVM here, and allows us to much more accurately assess which policy different tweets are about. Hence, we used the model reported in the first row of Table A2 to generate topic predictions for the rest of the tweets sent by state legislators and for the analysis and results presented in the paper.

[20] We chose an SVM as the baseline model because it outperformed other n-gram based models when we run some initial explorations and because previous research has shown SVM to perform the best out of the commonly used ngram/bag-of-words models, see Collingwood and Wilkerson (2012) and Hemphill and Schöpke-Gonzalez (2020).

Appendix B Additional figures

Figure B1: Construct Validity: Percentage of daily tweets predicted to be about Immigration.

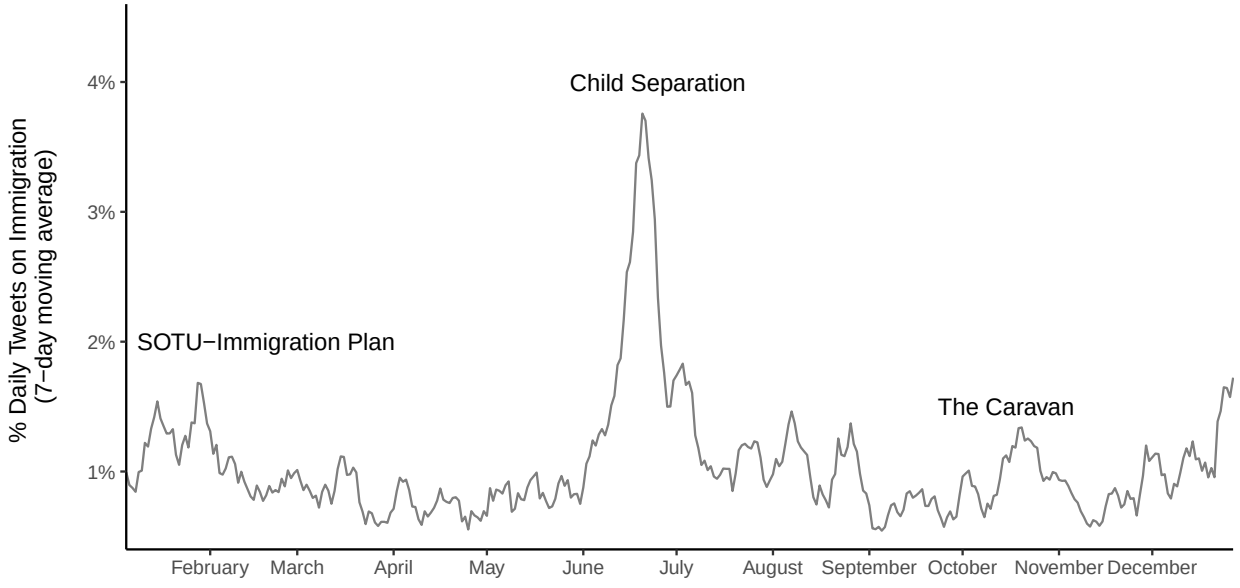

Figure B2: Logistic regression (left panel) and linear models (two right panels) predicting which legislators are on Twitter (binary outcome), how active they are on the platform (count variable) and how often they use it to discuss policy issues (proportion of tweets about one of the CAP policy areas). Replication of the pooled models in Figure 2 in which we replaced the *Legislative professionalization* score with the number of *Staff* members available for the entire legislature in each state (state-level covariate). Source of the new *Staff* variable: https://www.ncsl.org/research/about-state-legislatures/staff-change-chart-1979-1988-1996-2003-2009.aspx. Standard errors are clustered by state. Coefficient tables for these models are available in Table B1 (see Version 2 of the models).

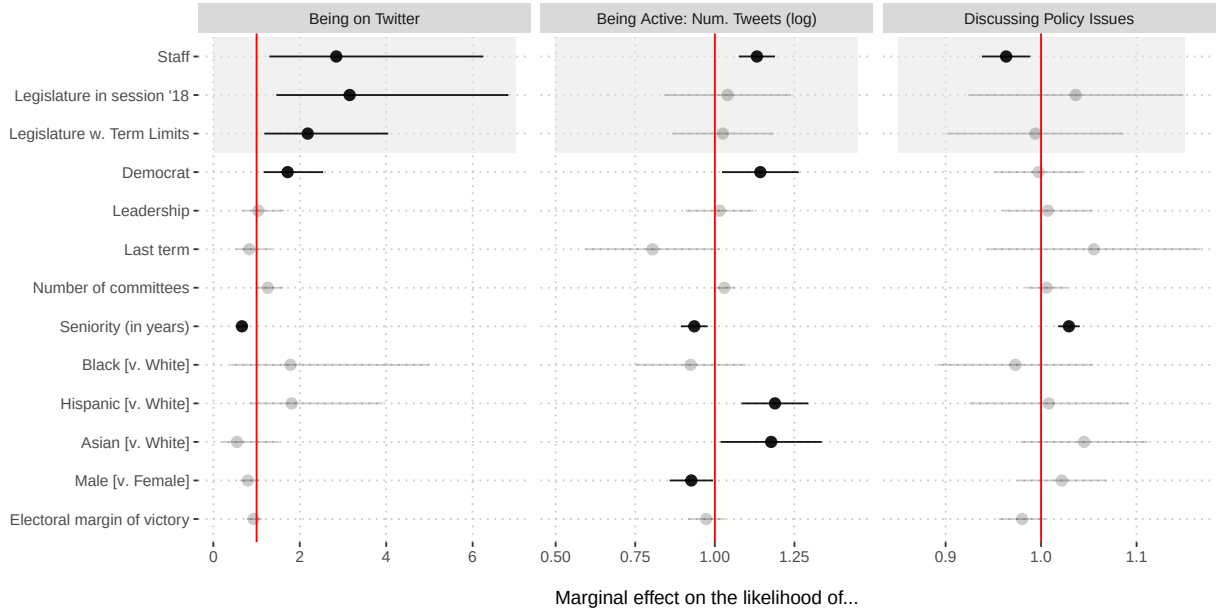

Table B1: Coefficient tables for the models in Figure 2 and Figure B2. In version 1 and 2 of the models we cluster standard errors by state. In version 3 we run multilevel models with state random intercepts.

	Being on Twitter (Binary) (Logistic regression)			Being Active: Num Tweets (Logged count) (Linear model)			Discussing Policy Issues (Proportion) (Linear model)		
Variable	Version 1	Version 2	Version 3	Version 1	Version 2	Version 3	Version 1	Version 2	Version 3
(Intercept)	0.89 (0.883)	0.165 (0.491)	1.561 (0.422)*	4.353 (0.561)*	3.858 (0.35)*	4.558 (0.279)*	0.621 (0.044)*	0.628 (0.033)*	0.717 (0.018)*
Leg. Prof.	0.995 (0.361)*			0.432 (0.154)*			-0.007 (0.009)		
Staff		1.045 (0.401)*			0.51 (0.112)*			-0.023 (0.008)*	
Leg. in session '18	0.532 (0.728)	1.149 (0.394)*		-0.179 (0.552)	0.155 (0.394)		0.026 (0.048)	0.023 (0.036)	
Leg. w. Term Limits	0.347 (0.547)	0.781 (0.315)*		-0.203 (0.335)	0.096 (0.314)		0.004 (0.037)	-0.004 (0.03)	
Democrat	0.476 (0.221)*	0.542 (0.199)*	0.476 (0.183)*	0.484 (0.285)	0.551 (0.237)*	0.575 (0.165)*	-0.005 (0.014)	-0.002 (0.015)	-0.019 (0.011)
Leadership	-0.147 (0.248)	0.038 (0.233)	0.019 (0.195)	-0.026 (0.182)	0.059 (0.211)	0.105 (0.179)	0.008 (0.017)	0.004 (0.015)	0.002 (0.012)
Last term	-0.25 (0.205)	-0.184 (0.248)	-0.1 (0.325)	-0.77 (0.414)	-0.754 (0.414)	-0.646 (0.33)	0.041 (0.038)	0.035 (0.036)	0.051 (0.024)*
Num. committees	0.238 (0.143)	0.231 (0.129)	0.131 (0.111)	0.057 (0.095)	0.115 (0.069)	-0.043 (0.095)	0.006 (0.007)	0.004 (0.007)	0 (0.006)
Seniority (in years)	-0.36 (0.084)*	-0.414 (0.085)*	-0.461 (0.086)*	-0.228 (0.095)*	-0.249 (0.083)*	-0.275 (0.085)*	0.017 (0.004)*	0.018 (0.004)*	0.015 (0.006)*
Black [v. White]	0.798 (0.724)	0.578 (0.776)	0.342 (0.756)	-0.225 (0.353)	-0.294 (0.342)	-0.315 (0.45)	-0.019 (0.026)	-0.017 (0.026)	-0.006 (0.031)
Hispanic [v. White]	0.799 (0.415)	0.593 (0.392)	0.363 (0.404)	0.79 (0.247)*	0.728 (0.206)*	0.656 (0.249)*	0 (0.027)	0.005 (0.027)	0.007 (0.016)
Asian [v. White]	-0.421 (0.364)	-0.602 (0.522)	-0.717 (0.701)	0.652 (0.312)*	0.682 (0.313)*	0.717 (0.56)	0.021 (0.019)	0.028 (0.021)	0.025 (0.036)
Other [v. White]	-1.215 (1.005)	-1.474 (1.273)	-1.368 (1.164)	-0.007 (1.444)	-0.01 (1.428)	-0.214 (1.591)	-0.029 (0.162)	-0.042 (0.162)	-0.091 (0.1)
Male [v. Female]	-0.257 (0.142)	-0.222 (0.143)	-0.161 (0.19)	-0.297 (0.145)*	-0.285 (0.134)*	-0.311 (0.161)	0.012 (0.015)	0.014 (0.015)	0.013 (0.011)
El. margin of victory	0.022 (0.076)	-0.07 (0.094)	0.032 (0.09)	-0.036 (0.101)	-0.105 (0.112)	-0.102 (0.083)	-0.014 (0.009)	-0.012 (0.008)	-0.005 (0.006)
State intercepts									
AZ			1.561			4.558			0.717
CA			2.979			4.879			0.659
FL			2.583			3.844			0.577
IL			1.036			3.349			0.669
MA			1.699			4.335			0.661
MT			-0.462			2.859			0.721
ND			-1.364			3.036			0.668
NJ			1.663			3.53			0.685
NV			1.275			4.117			0.601
NY			2.049			4.905			0.634
OH			1.484			3.839			0.655
TX			3.207			4.847			0.58
UT			0.767			3.847			0.646
VA			1.498			4.857			0.669
WY			-0.794			3.119			0.63
N	1267	1267	1267	998	998	998	829	829	829
Log Likelihood	-577.25	-552.96	-542.03			-2224.24			418.06
AIC	1184.51	1135.92	1110.05			4462.99			-886.48
R-squared				0.1	0.11		0.03	0.05	
Adjusted R-Squared				0.08	0.1		0.02	0.04	

Figure B3: Percentage of State Legislators with a Twitter Account, by State and Party.