Two Computational Models for Analyzing Political Attention in Social Media

Libby Hemphill, Angela M. Schöpke
School of Information
University of Michigan
Ann Arbor, MI 48104
{libbyh, aschopke}@umich.edu

Abstract

Understanding how political attention is divided and over what subjects is crucial for research on areas such as agenda setting, framing, and political rhetoric. However, existing methods for measuring attention, such as manual labeling according to established codebooks, are expensive and restrictive. We describe two computational models that automatically distinguish topics in politicians' social media content. Our models - one supervised classifier and one unsupervised topic model - provide different benefits. The supervised classifier reduces the labor required to classify content according to pre-determined topic lists. However, tweets do more than communicate policy positions. Our unsupervised model uncovers both political topics and other Twitter uses (e.g., constituent service). Together, these models are effective, inexpensive computational tools for political communication and social media research. We demonstrate their utility and discuss the different analyses they afford by applying both models to the tweets posted by members of the 115th U.S. Congress.

Questions about what political topics are receiving attention and how that attention is distributed are central to issues such as agenda setting and framing. Knowing what politicians are talking about and how those topics differ among various populations (e.g., Democrats and Republicans) and over time could enable advances in political communication research and has potential to increase constituents' knowledge.

Our study is motivated by two primary challenges in studying political attention and congressional communication. First, existing methods for studying political attention, such as manual topic labeling, are expensive and restrictive. Labeling content by hand requires tremendous human effort and substantial domain knowledge. Using pre-defined codebooks assumes a particular set of topics and therefore cannot effectively classify content that falls outside those predefined areas. Quinn and colleagues (Quinn et al. 2010) provide a more detailed overview of the challenges associated with labeling according to known topics and by hand; they also provide a topic-modeling approach to classifying speech in the Congressional Record that is similar to ours.

Second, Congress increasingly uses social media as a mechanism for speaking about and engaging with constituencies (Straus et al. 2013), but our tools for studying their social media use have not kept pace. Existing studies of MCs' social media use rely on the human labeling and domain knowledge mentioned above (Russell 2017; 2018; Evans, Cordova, and Sipole 2014; Frechette and Ancu 2017) or focus on the frequency (LaMarre and Suzuki-Lambrecht 2013) or type of use (Golbeck et al. 2018; Hemphill, Otterbacher, and Shapiro 2013) rather than the content of messages. MCs' social media speech can be used for understanding polarization (Hemphill, Culotta, and Heston 2016; Hong and Kim 2016) and likely impacts political news coverage (Shapiro and Hemphill 2017; Moon and Hadley 2014), and mechanisms for estimating attention and style would provider richer views of activity and enable new analyses.

To address these methodological challenges, we developed two computational models for estimating political attention—one supervised model that leverages human labels to label texts at scale and a second unsupervised model that uses readily-available data and low computational overhead to automatically label social media posts according to their policy topic and communication style. By providing the model and its codebook, we make it possible for others to label political texts efficiently and automatically, reducing the total effort required to label data for analysis. This enables more efficient and nuanced studies of Congress's political attention and communication style and facilitates comparative studies of political attention across media (i.e., social media, congressional hearings, news media). We demonstrate the utility of these approaches by applying the models to the complete corpus of tweets posted by the 115th U.S. Congress and briefly discussing the insights gained. We also discuss the costs and benefits of supervised and unsupervised models and provide aspects of each to consider when choosing a tool for analysis. In summary, our contributions are

1. a supervised model for assigning tweets to policy topics
2. an unsupervised model for assigning tweets to policy topics and communication uses
3. trade offs to consider when choosing supervised and/or unsupervised approaches to topic labeling

Why Twitter Existing work in political attention relies largely on political speeches (Laver, Benoit, and Garry 2003;

Oliver and Rahn 2016; Yu, Kaufmann, and Diermeier 2008), the Congressional Record (Quinn et al. 2010), and party manifestos (Gabel and Huber 2000; Slapin and Proksch 2008). At the same time, politicians around the world increasingly use social media to communicate, and researchers are examining the impacts of that use on elections (Bossetta 2018; Karlsen 2011), the press (Murthy 2015; Shapiro and Hemphill 2017), and more. Given its prevalence among politicians and in the public conversation about politics, especially during the Trump presidency in the U.S., politicians' behavior on Twitter demands our attention. We also expect that Twitter's ready availability and frequent updating will enable us to study political attention more efficiently and at a larger scale than prior data.

In order to understand how U.S. Congressional tweets may help us to answer these social and political research questions, we must first understand what U.S. MCs are saying. Given the sheer volume of content that MCs put out on Twitter, manual tweet-by-tweet labeling is neither time efficient nor affordable in the long term. We therefore explore whether an unsupervised topic model could help us understand MC conversations on Twitter in a more cost and time efficient way than manual labeling.

Computational study of MCs use of Twitter will eventually allow us to (1) situate the MCs' political attention patterns on Twitter within the broader media context, (2) understand whether MCs' political attention patterns on Twitter reflect attention patterns in a broader media context, and (3) understand the unique value that the use of supervised classification and unsupervised topic modeling techniques can contribute to our overall understanding of MCs' political attention patterns. However, tools or models for performing these analysis are currently unavailable.

In the sections that follow, we discuss how this process helps us understand important ideas about how MCs communicate their public facing personae, MCs' agenda setting practices in public facing discourse, and how to ask more specific questions about MCs' public-facing communications across different media platforms. We propose that Twitter posts offer unique insights that can be studied computationally at relatively low cost and higher time efficiency compared to other datasets.

Data

In this section, we describe the data we used to train our supervised classifier and unsupervised topic model, and to compare our supervised model's performance against our unsupervised model's performance.

Tweets from the 115th Congress

Using the Twitter Search API, we collected all tweets posted by official MC accounts (voting members only) during the 115th Congress which ran January 3, 2017 to January 3, 2019. We identified MCs' Twitter user names by combining the lists of MC social media accounts from the United

	Senate		House		Total
	Dem	GOP	Dem	GOP	
Men	32	48	128	207	415
Women	16	5	64	24	109
Total	48	53	192	231	524

Table 1: Number of accounts by party, chamber, and gender. Two independent senators who caucus with Democrats have been grouped with them.

	Senate		House	
	Dem	GOP	Dem	GOP
Men	143,522	171,083	441,890	356,870
Women	74,905	19,867	212,457	65,240
Total	218,427	190,950	654,347	422,110

Table 2: Number of tweets by party, chamber, and gender. Two independent senators who caucus with Democrats have been grouped with them. Cell values are the count of tweets posted by the intersection of the row and column. Because of space constraints, we cannot include row totals: men posted 1,113,365 tweets, and women posted 372,469.

States project[1], George Washington Libraries[2], and the Sunlight Foundation[3]. Throughout 2017 and 2018, we periodically used the Twitter API to search for the user names in this composite list and retrieved the accounts' most recent tweets. Our final search occurred on January 3, 2019, shortly after the 115th Congress ended. In all, we collected 1,485,834 original tweets (i.e., we excluded retweets) from 524 accounts. The accounts differ from the total size of Congress because we included tweet data for MCs who resigned (e.g., Ryan Zinke) and those who joined off cycle (e.g., Rep. Conor Lamb); we were also not able to confirm accounts for every state and district. We summarize the accounts present in our dataset in Table 1 and the number of tweets posted by chamber, gender, and party in Table 2.

We used this data to train our unsupervised topic model, and to compare our supervised model's performance against our unsupervised model's performance.

Training Data for Supervised Classifier

We used a selection of Russell's human-labeled dataset (Russell 2017; 2018), which contains 45,395 tweets labeled with codes from the CAP codebook. We removed retweets from this set to limit our classification to original tweets, resulting in a total set of 39,696 labeled tweets used to train and test our models.

[1] https://github.com/unitedstates/congress-legislators

[2] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/UIVHQR

[3] https://sunlightlabs.github.io/congress/index.html\#legislator-spreadsheet

Methodology

Here we describe the processes of building and training the models and applying them to the 115th Congress tweets.

Preprocessing

We used the same preprocessing steps for both the supervised and unsupervised models and describe them below.

Stemming We evaluated whether or not to use stemming or lemmatization in training our unsupervised model by reviewing the relative interpretability of topics generated by models trained with each of stemmed texts and unstemmed texts. We found unstemmed texts render significantly more interpretable generated topics, likely reflecting syntactical associations with semantic meanings, which is consistent with prior work(Schofield and Mimno 2016). This pattern is likely especially relevant for tweet data given that unique linguistic patterns and intentional misspellings are used to communicate different semantic meanings. Stemming in these instances may remove the nuance in potential semantic meaning achieved by tweets' unique linguistic features including misspellings. Therefore, we did not use stemming or lemmatization in data preprocessing for the final models.

Stop lists, tokenization, and n-grams Given the prevalence of both English- and Spanish-language tweets, this project removed both English and Spanish stop words included in Python's Natural Language Toolkit (NLTK) default stop word lists.

We also used a combination of two tokenization approaches to prepare our data for modeling. First, we used Python's NLTK tweetTokenizer with parameters set to render all text lower case, strip Twitter user name handles, and replace repeated character sequences of length three or greater with sequences of length three (i.e. "Heeeeeeeey" and "Heeeey" both become "Heeey"). Second, we removed punctuation (including emojis), URLs, words smaller than two letters, and words that contain numbers.

In the early stages of model development, we evaluated the comparative intepretability and relevance of topics generated for document sets tokenized into unigrams, bigrams, and a combination of unigrams and bigrams. We found model results for document sets tokenized into unigrams to be most interpretable and relevant, and thus present only these results in this paper.

Specifying Models

Supervised Classifier We developed a supervised machine learning classifier on a collection of human-labeled tweets from the 113th Congress based on the 113th human-labeled data specified in our "Data" section (Russell 2017; 2018). Our training set comprised 90 percent of our total dataset, and our test set comprised 10 percent of the total dataset. We evaluated the performance of each of NLTK's[4] implementations of Nave Bayes and Logistic Regression classifier against a baseline of NLTK's Dummy Classifier performance, which provides a classification success rate expected via random guessing.

Unsupervised Topic Model We evaluated the performance of two unsupervised model implementations for relevance and interpretability: gensim's Latent Dirichlet Allocation (LDA) model and MALLET's LDA model wrapper (Řehůřek and Sojka 2010).

We began testing with a computationally inexpensive gensim LDA model. We found that the results returned were insufficiently interpretable to effectively answer our research questions. We thus tested an alternative, MALLET's LDA model wrapper, and found it relatively computationally inexpensive while rendering significantly more intepretable results than gensim's LDA implementation. We also considered Moody's lda2vec(Moody 2016), and though preliminary results did return more nuanced topics, we found its setup and computational inefficiency to be too cumbersome and costly for practical use. We present the results of the unsupervised model generated using MALLET's LDA model wrapper.

The MALLET LDA wrapper requires that we specify a number of topics to find in advance. We evaluated the performance of models generating between 5 and 70 topics in increments of five topics within this range. We found 50 topics to yield the most interpretable and relevant results without excessive redundancy in topics generated.

Evaluating Models

We used common approaches to evaluating the two models, and then compared the results from each with one another. The following subsections describe our evaluation processes in greater detail.

Evaluating the Supervised Model Using a logistic regression classifier achieved an F1 score of 0.791. The F1 score balances the precision and recall of the classifier to provide a measure of performance. We compared our classifier to a dummy classifier using the same training data, and it achieved only F1 = 0.10. Given the difference between our classifier's score and the dummy, we argue the supervised classifier achieved reasonable accuracy. We also compared our classifier's labels with the human labels and achieved a Cohen's kappa of 0.769, suggesting moderate agreement (McHugh 2012). We present the results only from our highest performing classifier (logistic regression) in this paper.

Evaluating the Unsupervised Model We used a Grounded Theory approach to interpret and label topics returned by our final unsupervised model (Glaser and Strauss 1967). This approach involved first labeling the baseline model's topics according to our initial interpretations, allowing us to discovery semantically important topics that arise from the data rather than from a predetermined topic set. Human interpretation and label assignment was performed by two domain experts and confirmed by a third. One expert has prior experience on legislative staff having served in a senior senator's office, and in public affairs for political organizations. She was able to determine whether the topics identified by the classifier were interpretable

[4]All algorithms used by the supervised model are available in NLTK (Bird, Klein, and Loper 2009), a Python library.

and were actually capturing political topics as members of Congress understand them. The second expert has spent 12 years working with political communication theory, allowing her to understand topics returned within the context of political behavior. The two labelers discussed disagreements until both agreed on the labels applied to all fifty topics. Following this first labeling process, we performed the second step of the Grounded Theory method, determining patterns across labels and grouping labels together that indicated topical similarity based on their highest-weighted features. For instance, one topic's top unigram features included *health, care, Americans, Trumpcare* and another included *health, care, access, women*. We bundled both of these topics under the umbrella *healthcare*. Finally, we reviewed our topic labels and features with a senior member of a U.S. policy thinktank to confirm the validity of our labels and interpretability of our topics.

We also compared our set of labels to the Comparative Agendas Project (CAP) Master Codebook (Bevan 2017), a codebook commonly used in social, political, and communications science to understand topics in different types of political discourse worldwide. Where possible, we assigned CAP codes to the related codes that resulted from our inductive labeling. A complete list of our topics, their associated CAP codes, and highly-weighted features is available in Table 3.[5] Not all of our codes had ready analogues in the CAP taxonomy. For instance, our unsupervised model identified *media appearance* as a separate and frequent topic. The CAP codebook includes only policy areas and not relationship-building or constituent service activities of government officials, and so no CAP label directly applies here. Further, we determined that some nuanced aspects of the topics returned by our unsupervised model are not captured by the policy focus of the CAP codebook. For example, our unsupervised model returned several topics clearly interpretable as related to veteran affairs. The CAP Codebook includes aspects of veteran affairs as a sub-topic under both *Housing* and *Defense*, but given that our unsupervised model detected veteran-related issues outside of these sup-topic areas, we determined that a new topic solely devoted to veteran affairs would better describe the thematic content of those topics. A similar situation applied to the unsupervised topic we call legislative process, some but not all aspects of which may have fallen under CAP cod *Government Operations*. The codes we identified that were not captured by the CAP codebook are marked by "-" in the Code Number column of Table 3.

Comparing Models' Output

The supervised model provides a single topic label for each tweet, and the unsupervised model provides probabilities for each tweet-class pair. In order to compare the labels between classifiers, we assigned tweets the highest-probability topic indicated by the unsupervised model. We measured interannotator agreement between our supervised and unsupervised models using Cohen's kappa (McHugh 2012). We opted to use Cohen's kappa as a measure that is widely used and well understood. Since we hope to compare agreement between only two annotators (supervised and unsupervised model), our data is not incomplete, and our topic codes are all of the same unit type, we determined that more robust measures such as Krippendorf's Alpha were not necessary (Hellgren 2012). Because our supervised model assigns topic labels only to policy-related tweets, a comparison between supervised and unsupervised models is only meaningful for those tweets both models labeled, or policy-related tweets. Therefore, we calculated a Cohen's kappa score between our supervised and unsupervised model for only policy-related tweets with label assignments corresponding to CAP codes.

The two classifiers achieved a Cohen's kappa of 0.262. We argue that the classifiers likely measure different things, and the low agreement suggests an opportunity for researchers to select the tool that matches their analytic goals. Low agreement suggests that the two models are well-differentiated; we return to this discussion in detail below under "What the Different Models Capture".

Applying Models to 115th Congress' Tweets

We applied both the supervised and unsupervised models to all of the original tweets posted by the 115th Congress and use those results to illustrate the analytic potential of these new methodological tools. Table 3 summarizes the results by providing the topic description, the corresponding CAP codebook number (if applicable), the proportion of tweets that fell in that topic according to the supervised (SU) and unsupervised (UN1 and UN2) models. For each tweet, the unsupervised model returns the probability that the tweet belongs in each of the 50 topics identified. We provide both the most likely (UN1) and second-most likely topic (UN2) for each tweet.

Topic Distribution In Table 3, each of topics 1-23 correspond to the CAP macro-level code numbers. Topic 23, *cultural affairs* was not detected by either of our models, and as such receives a "-" value across all columns. Each topic 24-35 corresponds to expert interpretations of unsupervised topic model results; we assigned "0" to tweets whose topics we uninterpretable by either model. Since these latter topics apply only to the unsupervised model's results, each of these topics receives a "-" value in column "SU". Related, each of topics *energy* (#8), *housing* (#14), *foreign trade* (#18), and *international affairs* (#19) received "-" values in both columns "UN1" and "UN2)". These "-" values indicate that our unsupervised model did not detect topics that human interpretation would assign any of these four CAP Codebook topics.

Discussion

Examining the output from the two models on the same dataset, tweets from the 115th Congress, helps explain the models' differences, trade-offs, and clarifies the new insights available from the unsupervised model.

[5]The CAP Codebook does not include a codes numbered 11 or 22.

Topic Description	CAP #	SU	UN1	UN2
macroeconomics	1	0.060	0.074	0.042
civil rights	2	0.037	0.094	0.040
health	3	0.088	0.318	0.043
agriculture	4	0.010	0.019	0.014
labor	5	0.022	0.021	0.016
education	6	0.021	0.021	0.016
environment	7	0.018	0.045	0.012
energy	8	0.013	-	-
immigration	9	0.021	0.017	0.012
transportation	10	0.011	-	0.020
law and crime	12	0.043	0.026	0.018
social welfare	13	0.009	0.020	0.014
housing	14	0.003	-	-
domestic commerce	15	0.027	0.023	0.018
defense	16	0.066	-	0.019
technology	17	0.006	-	0.005
foreign trade	18	0.002	-	-
international affairs	19	0.027	-	-
government operations	20	0.037	0.072	0.059
public lands	21	0.009	-	0.010
cultural affairs	23	-	-	-
veterans	-	-	0.019	0.026
sports	-	-	0.020	0.011
district affairs	-	-	0.071	0.063
holidays	-	-	0.009	0.027
awards	-	-	-	0.009
politicking	-	-	-	0.042
self promotion	-	-	0.050	0.052
sympathy	-	-	-	0.010
emergency response	-	-	0.018	0.013
legislative process	-	-	-	0.059
constituent relations	-	-	-	0.022
power relations	-	-	0.026	0.024
uninterpretable	-	-	0.037	0.284

Table 3: Topic Distribution Across the 115th Congress. Contains proportional distributions across all topics for both supervised (SU) and unsupervised (UN1 and UN2) classifiers. UN1 indicates the highest probability category assigned, and UN2 indicates the second highest.

Comparing Model Labels

Remember that the resulting Cohen's kappa between supervised and unsupervised model results was only 0.262. Common practice suggests that a Cohen's kappa of .21-.40 indicates fair interannotator agreement, with 0 indicating interannotator agreement that approximates random choice (McHugh 2012). In this sense 0.262 does not represent particularly high interannotator agreement between the supervised and unsupervised models. In order to understand why this was, we examined individual cases of disagreement and agreement. We discuss example cases below, how these disagreements contribute to the Cohen's kappa score returned, and what they reveal about the utility of each of the modeling approaches.

First, we examined the proportion of supervised labels matching the first and second most probable unsupervised labels for policy-related tweets. Table 3 shows that both models were able to detect a topic in nearly all tweets (i.e., the proportion of *uninterpretable* appears relatively infrequently). This topic appears more frequently in the second set of labels from the unsupervised classifier (UN2). The *uninterpretable* topic's features were comprised largely of prepositions, conjunctions, and other words with little semantic meaning. This suggests that the most significant semantic meaning of a tweet labeled by the unsupervised model can likely be understood by using its maximum probability topic.

In Figure 1 (a), we include the supervised model and most-likely topic from the unsupervised model because those two topics are policy-related. In (b) and (c), we report the supervised classifier label and second-most likely unsupervised topic label to illustrate how it's possible to use each classifier for different analytic purposes. The supervised classifier, (b), shows us differences in political attention by party. The unsupervised classifier's second-most likely category enables us to compare the style by party. We can see that Republicans give less attention to *civil rights*, *environment*, and *social welfare* and more attention to *energy* and *defense* than Democrats. None of these differences are especially surprising given the parties' priorities, but our models show empirically that the differences are observable even in social media. Among the unsupervised model's style codes, Republicans exhibit more *self promotion* than Democrats, but the parties are nearly equal on those style codes. Our model affords similar comparisons among other groups such as gender or chamber that may provide insight into how political attention changes or how commuication styles differ among groups.

Additional similarities and differences in the proportions of SU and UN2 in Figure 1 may help us to understand the utility of each model. We can see that our unsupervised model's maximum probability topic predictions did not include topic labels *transporation*, *defense*, *technology*, *public lands* for any policy-related tweets. However, we can also see that each of these topics are included among the unsupervised model's second most probable topic predictions. This suggests that in some policy topics' cases, each of our supervised and unsupervised models may be able to predict approximately similar ideas if both most probable and second most probable topics are taken into account.

Interestingly, we see that the *government operations* topic is predicted with about the same frequency by the supervised model and each of the unsupervised model's first most probable topic assignment and second most probable topic assignment. This again suggests that the unsupervised model and supervised model both may be able to predict approximately similar ideas, but also begs the question, which was the first most probable topic assignment for each of those tweets those that were labeled *government operations* with second-highest probability?

In general, we notice that among the second-most probable topics for the unsupervised model, topics 24-35, or all those topics that are not featured in the CAP Codebook, oc-

tweet Text	UN1	UN2
Sorghum market is very dependent on export. China is main market w/ 75% of it, then Mexico. #FarmBill	*agriculture*	*uninterpretable*
NYC is proud to send 120 #FirstResponders to help Texans in the wake of #Harvey https://t.co/vdumEKFnQA	*emergency response*	*district affairs*
@realDonaldTrump @RepMcEachin @RepJoeKennedy Any time someone is brave enough to serve, we must respect that choice, not stomp on their rights and liberties. #TransRightsAreHumanRights	*civil rights*	*veterans*

Table 4: Sample tweets labeled *International Affairs* by Supervised Classifier. UN1 refers to the maximum probability class according to the unsupervised classifier, and UN2 indicates the second-highest probability class.

cur more frequently. This suggests that the topic of highest probability more effectively captures a policy focus, and the topic of second-highest probability captures the way in which, the reason for, and with whom MCs are discussing these policy issues.

Rare Topics

That none of the CAP Codebook topics *energy*, *housing*, *foreign trade*, *international affairs* are detected by the unsupervised model indicates that these topics are occurring rarely enough that the unsupervised model does not have enough data to detect these topics as distinct. In reviewing Figure 1, it is possible to see that a very low proportion of tweets were labeled *housing* (#14) or *foreign trade* (#18) by the supervised classifier, for example. That these topics are not frequently occurring according to the supervised model either suggests that that MCs do not spend a lot of time tweeting about each of topics *housing* or *foreign trade*. These omissions are likely an artifact of the U.S. federal government's structure. Though housing, for example, is in part an issue addressed by the U.S. Department of Housing and Urban Development and thus an appropriations issue for the U.S. Congress, many housing issues are addressed at state and local levels. It is possible that given this distribution of significant responsibility to the state and local levels, MCs spend less time talking about housing at the federal level.

Where the Models Diverge

It is also possible that those terms associated with these topics for the supervised model are being associated with or grouped together with terms corresponding to different topics. For example, given that *international affairs*, *defense*, and *immigration* topic areas have some overlapping topical relevance, it is possible that certain tweets were labeled by the unsupervised classifier as *defense* or *immigration* rather than *international affairs* as by the supervised classifier. By examining the features associated with each topic in each classifier (see Tables 6 and 7), we can see some of these differences. For instance, the *defense* topic (#16) has features "iran", "nuclear", and "strategy" in the unsupervised model and "veteransday", "drones", and "stolenvalor" in the supervised model. In the unsupervised model, topics about *veterans* emerged that were distinct from *defense* and *housing*, where veterans occur in the CAP codebook. So we see that similar terms are associated with different topics in the two types of models, and that explains some of their differences. However, each set of associations is reasonable and interpretable, and that suggests that the models capture different latent properties with their feature-class associations.

Table 4 describes several sample tweets that were labeled *international affairs* by the supervised model, and indicates how the unsupervised model labeled the same tweet. For example, the unsupervised model labeled the first tweet featured as *agriculture* because of the presence of the hashtag #FarmBill and mention of sorghum. At the same time, the tweet also reflects international affairs topical focus by mentioning the international trade market and naming other countries in that market. In this case, each of the supervised and unsupervised models captured different thematic focuses of the tweet both perhaps of comparable relevance.

The second example where the models diverge was labeled by the unsupervised model as *emergency response* (highest probability) or *district affairs* (second highest probability). It is possible that the supervised model associated words like "proud", "send", and "help" with the international affairs topic, but the thematic focus of the tweet does indeed appear to be more related to emergency response and district affairs via discussion of "#FirstResponders", hurricane Harvey, and mention of local communities New York City and Texas. In this case, the unsupervised model captures more relevant themes than the supervised model. These disagreements illustrate that the fixed parameters of the CAP Codebook may not capture all types of conversation by MCs on Twitter. Since the unsupervised model represents a bottom-up Grounded Theory approach, it is able to capture a new topic not featured in the CAP Codebook but that provides additional nuance to the analysis of these tweets.

New Insights from the Unsupervised Model

The final example in Table 4 raises another important difference between our supervised and unsupervised models. Where our supervised model is limited by the policy-focused labels defined by the CAP Codebook, our unsupervised model is not. We see that approximately 21.32% and 35.76% of all maximum probability and second maximum probability topics detected fall in topic codes that are not captured in the CAP Codebook. These topics (a) reveal that MCs spend a significant portion of their tweets posting about issues other than policy and (b) identify the topics and behaviors beyond policy that they exhibit.

The additional topics the unsupervised classifier identified largely reference activities related to personal relationship building. The ability to distinguish these activities enables analyses of behavior beyond policy debates such as home style. For instance, to differentiate Fenno's (Fenno 1978) "person-to-person" style from "servicing the district". Relationship building topics include *sports*, *holidays*, *awards*, *sympathy*, *constituent relations*, and *power relations*. Service-related topics include *district affairs*, *emergency response*, and *legislative process*. Prior research has conducted similar style analysis on smaller sets of tweets around general elections (Evans, Cordova, and Sipole 2014), and our models enable this analysis at scale and throughout the legislative and election calendars.

That somewhere between 20 and 40 percent of all tweets are labeled with these topics tells us that the CAP Codebook alone is unable to capture how and why MCs use Twitter. The supervised model does a good job of capturing the policy topics present in the CAP codebook and facilitating the comparative analysis that codebook is designed to support. The unsupervised model enables us to see that MCs invest significant attention in personal relationship building and constituent service on Twitter and to analyze those topics and behaviors that are not clearly policy-oriented.

Recommended Uses of Each Model

Given the performance and capabilities of each model, we recommend the supervised model for conducting comparative studies across political systems, and we recommend the unsupervised model for understanding the nuances of the U.S. Congress (e.g., their topics, their non-policy content and behavior). Though useful for comparative studies, our unsupervised model helps us understand that Congress' communications on Twitter do not easily map to the Comparative Agendas Codebook and that there are different ways to classify MCs Twitter behavior that help us understand more about Congress' attention.

To further understand the usefulness of each model, we assessed each of their Cohen's Kappa agreement with labels assigned by a human annotator to 13,943 policy-related tweets made available by Russell (as used in (Russell 2018; 2017)). As expected, since our supervised model was trained on a subset of this data, we found that our supervised model and human-labeled tweets receive a high Cohen's kappa of 0.96, close to nearly complete agreement. The low Cohen's kappa between manual labels and the unsupervised classifier suggest that the taxonomies are mostly independent—these coding approaches measure and capture different aspects of the tweets.

Model 1	Model 2	Cohen's Kappa Score
Unsupervised	Supervised	0.012
Unsupervised	Hand-Labeled	0.013
Supervised	Hand-Labeled	0.960

Table 5: Comparing Model Output to Manual Labels

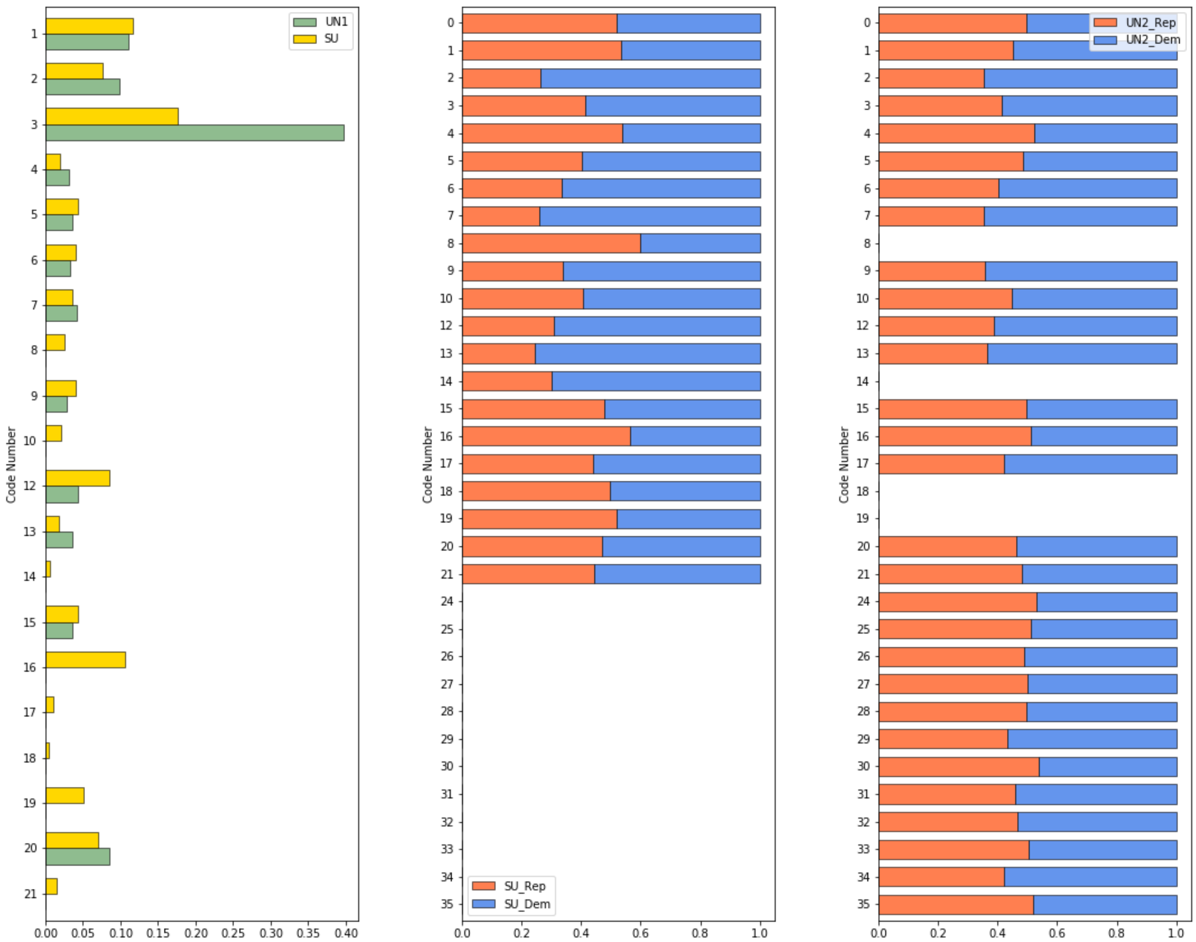
Our unsupervised model's findings suggest potential extensions to the CAP Codebook to reflect tweet-specific U.S. legislative discourse (see codes 24-35 in Table 3. It is worth noting that the proposed additional topics largely do not reflect policy topics, but rather different forms of communicating with constituents. These codes describe topics that embody certain communication styles more than topical ideas such as Public Lands or Health. The emergence of these topics suggests that unsupervised topic modeling supports different kinds of analysis than manual and codebook-oriented labeling. Unsupervised topic modeling can also detect communication styles and intents. Further investigation about this potential could yield interesting insights concerning the utility of unsupervised topic modeling to understanding the political communications space.

Future Work

Political Attention The most exciting next steps for work in this area involve using the models to study political attention and social media use in political communication. Our models dramatically reduce the costs of obtaining labeled data for comparative analysis and provide a mechanism for identifying additional behavior beyond policy discussions. For example, by using our supervised model, we are able to study the complete 115th Congress and their relative attention to policies over time. We can compare attention between subgroups (e.g., parties, chambers, regions of the U.S.) and over time (e.g., around primaries, during recess). When combined with similar advances at labeling and accessing news content (Saito and Uchida 2018; Gupta et al. 2018), our models also facilitate analysis of the political agenda, enabling researchers to test the paths through which topics reach the mainstream or legislation.

International Comparison Our supervised model directly allows for comparative analyses by using a standard codebook designed for such studies. Whether our models work on tweets in languages other than English (and marginally Spanish) is an open question. We are currently training models on German parliamentary tweets from 2017 in order to evaluate the potential utility of our approach for understanding politicians' public-facing rhetoric across different languages and country contexts.

Evaluating Non-Policy Related Topics The dataset we used to train the supervised model includes additional human-labeled binary tags indicating personal relationship



(a) (b) (c)

Figure 1: Topic Distributions. (a) shows the overall topic distribution according to the supervised classifier. (b) contains a 100% stacked bar chart indicating the proportion of tweets of that type that came from Republicans (red) and Democrats (blue) as labeled by the supervised classifier. (c) shows the proportion of tweets of that type that came from Republicans (red) and Democrats (blue) as indicated by the second-highest topic according to the unsupervised classifier.

or service-related content in the 113th Congress' tweets. Interestingly, many of these manual tags reflected similar personal relationship or service-related tags that our unsupervised model independently detected. Future work could compare these results—manual labels and unsupervised labels—for this relational content and potentially inform another useful computational tool.

Model Improvements Our next steps involve evaluating whether alternative topic modeling techniques can achieve even more nuanced topic results. One such approach may include implementing Moody's lda2vec approach (Moody 2016). Moody's approach builds upon word2vec's ability to determine word to word relationships in order to augment LDA's ability to determine word to document relationships by building document-level abstractions. Nikita compared the lda2vec's returned topical output with a genism LDA model's returned output (Nikita, 2016). His replication attempt provided a helpful example of comparative model im-

Code #	Associated Terms
1	budgetconference fiscalcliff budgetdeal
2	passenda enda nsa
3	defundobamacare obamacare healthcare
4	farmbill gmo sugar
5	fmla minimumwage laborday
6	talkhighered dontdoublemyrate studentloans
7	actonclimate leahysummit climate
8	energyefficiency energyindependence
9	immigrationreform immigration cirmarkup
10	skagitbridge obamaflightdelays faa
12	vawa guncontrol gunviolence
13	snap nutrition hungry
14	gsereform fha housing
15	fema sandyrelief sandy
16	veteransday drones stolenvalor
17	marketplacefairness nonettax broadband
18	trade exports export
19	benghazi standwithisrael
20	nomination irs nominations
21	commissiononnativechildren tribalnations

Table 6: Features Associated with Topics in the Supervised Classifier

plementation and evaluation methods of both lda2vec and gensim LDA models.

Niu's work with topic2vec modeling explores ways to address the issue of uninterpretable output from LDA models (Niu et al., 2015); he notes that because LDA assigns high probabilities to words occurring frequently, those words with lower frequencies of occurrences are less represented in the derivation of topics, even if they have more distinguishable semantic meaning than the more frequently occurring words. In a similar approach to Moody, Niu tries to combine elements of word2vec and LDA modeling techniques to address the issue of under-specific topics resulting from pure LDA methods. His topic2vec approach yields more specific terms (i.e. where LDA yields patients and medical in each of two topics when run on a particular dataset, topic2vec yields aricept, memantine, and enbrel in one topic and anesthesiologists, anesthesia and comatose in a second topic with the same dataset). We did not pursue topic2vec in our tests due to replicability issues encountered with the methodologies proposed, but they offer an interesting avenue to explore in attempts to improve the model.

Conclusion

We provide computational models to facilitate research on political attention in social media. The supervised model classifies tweets according to the CAP codebook, enabling comparative analyses across political systems and reducing the labor required to label data according to this common codebook. The unsupervised model labels tweets according to their policy topic, social function, and behavior. It enables nuanced analyses of U.S. Congress, especially the intersection of their policy discussions and relationship buildling. Together, these two models provide methodological tools for understanding the impact of political speech on Twitter and comparing political attention among groups and over time.

Code #	Associated Terms
1	budget government house bill
1	tax families cuts
1	tax jobs reform
2	trump president american
2	today life rights
2	women rights equal
3	health care americans
3	opioid help health
3	health getcovered enrollment
4	energy jobs farmers
5	obamacare jobs year
6	students education student
7	climate change water
9	children border families
10	will infrastructure funding
12	sexual human trafficking
12	gun violence congress
13	families workers care
15	small jobs businesses
16	iran nuclear security
17	internet netneutrality open
20	act bill protect
20	court judge senate
20	trump investigation russia
20	obama rule president
21	national protect public
24	veterans service honor
24	women service men
25	nan game team
26	office help hours
26	today great new
26	service academy fair
26	hall town meeting
27	day happy today
27	family will friend
27	happy year celebrating
28	school congressional art
29	make work keep
29	forward work working
30	read week facebook
30	tune watch live
30	hearing watch committee
31	families victims prayers
32	help disaster hurricane
33	bill house act
33	today discuss issues
33	vote voting voter
34	work community thank
35	today president mayor

Table 7: Features Associated with Topics in the Unsupervised Model

Acknowledgments

We are especially grateful to Annelise Russell for sharing her data and enabling us to train the supervised classifiers. This material is based upon work supported by the National Science Foundation under Grant No. 1525662.

References

Bevan, S. 2017. Gone fishing: The creation of the comparative agendas project master codebook. Technical report.

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. "O'Reilly Media, Inc.".

Bossetta, M. 2018. The digital architectures of social media: Comparing political campaigning on facebook, twitter, instagram, and snapchat in the 2016 U.S. election. *Journal. Mass Commun. Q.* 1077699018763307.

Evans, H. K.; Cordova, V.; and Sipole, S. 2014. Twitter style: An analysis of how house candidates used twitter in their 2012 campaigns. *PS Polit. Sci. Polit.* 47(2):454–462.

Fenno, R. F. 1978. *Home Style: House Members in Their Districts*. Longman.

Frechette, C., and Ancu, M. 2017. A content analysis of twitter use by 2016 US congressional candidates. *The Internet and the 2016 Presidential Campaign* 25.

Gabel, M. J., and Huber, J. D. 2000. Putting parties in their place: Inferring party Left-Right ideological positions from party manifestos data. *Am. J. Pol. Sci.* 44(1):94–103.

Glaser, B. G., and Strauss, A. L. 1967. The discovery of grounded theory: Strategies for qualitative research. *Social Forces* 46(4):555.

Golbeck, J.; Auxier, B.; Bickford, A.; Cabrera, L.; Conte McHugh, M.; Moore, S.; Hart, J.; Resti, J.; Rogers, A.; and Zimmerman, J. 2018. Congressional twitter use revisited on the platform's 10-year anniversary. *Journal of the Association for Information Science and Technology* 69(8):1067–1070.

Gupta, S.; Sodhani, S.; Patel, D.; and Banerjee, B. 2018. News category network based approach for news source recommendations. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 133–138.

Hellgren, K. A. 2012. Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in quantitative methods for psychology* 8(1):23–34.

Hemphill, L.; Culotta, A.; and Heston, M. 2016. #polar scores: Measuring partisanship using social media content. *Journal of Information Technology and Politics* 13(4):365–377.

Hemphill, L.; Otterbacher, J.; and Shapiro, M. 2013. What's congress doing on twitter? In *Proceedings of the 2013 conference on Computer supported cooperative work*, CSCW '13, 877–886. New York, NY, USA: ACM.

Hong, S., and Kim, S. H. 2016. Political polarization on twitter: Implications for the use of social media in digital governments. *Gov. Inf. Q.* 33(4):777–782.

Karlsen, R. 2011. A platform for individualized campaigning? social media and parliamentary candidates in the 2009 norwegian election campaign. *Policy & Internet* 3(4):1–25.

LaMarre, H. L., and Suzuki-Lambrecht, Y. 2013. Tweeting democracy? examining twitter as an online public relations strategy for congressional campaigns'. *Public Relat. Rev.* 39:360–368.

Laver, M.; Benoit, K.; and Garry, J. 2003. Extracting policy positions from political texts using words as data. *Am. Polit. Sci. Rev.* 97(2):311–331.

McHugh, M. L. 2012. Interrater reliability: the kappa statistic. *Biochem. Med.* 22(3):276–282.

Moody, C. E. 2016. Mixing dirichlet topic models and word embeddings to make lda2vec.

Moon, S. J., and Hadley, P. 2014. Routinizing a new technology in the newsroom: Twitter as a news source in mainstream media. *J. Broadcast. Electron. Media* 58(2):289–305.

Murthy, D. 2015. Twitter and elections: are tweets, predictive, reactive, or a form of buzz? *Inf. Commun. Soc.* 18(7):816–831.

Oliver, J. E., and Rahn, W. M. 2016. Rise of the trumpenvolk: Populism in the 2016 election. *Ann. Am. Acad. Pol. Soc. Sci.* 667(1):189–206.

Quinn, K. M.; Monroe, B. L.; Colaresi, M.; Crespin, M. H.; and Radev, D. R. 2010. How to analyze political attention with minimal assumptions and costs. *Am. J. Pol. Sci.* 54(1):209–228.

Řehůřek, R., and Sojka, P. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA. `http://is.muni.cz/publication/884893/en`.

Russell, A. 2017. U.S. senators on twitter: Asymmetric party rhetoric in 140 characters. *American Politics Research* 1532673X17715619.

Russell, A. 2018. The politics of prioritization: Senators' attention in 140 characters. *The Forum* 16(2):331–356.

Saito, T., and Uchida, O. 2018. Automatic labeling to classify news articles based on paragraph vector. *International Journal of Computers* 3.

Schofield, A., and Mimno, D. 2016. Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics* 4(0):287–300.

Shapiro, M. A., and Hemphill, L. 2017. Politicians and the policy agenda: Does use of twitter by the U.S. congress direct new york times content? *Policy & Internet* 9(1):109–132.

Slapin, J. B., and Proksch, S.-O. 2008. A scaling model for estimating Time-Series party positions from texts. *Am. J. Pol. Sci.* 52(3):705–722.

Straus, J. R.; Glassman, M. E.; Shogan, C. J.; and Smelcer, S. N. 2013. Communicating in 140 characters or less: Congressional adoption of twitter in the 111th congress. *PS Polit. Sci. Polit.* 46(1):60–66.

Yu, B.; Kaufmann, S.; and Diermeier, D. 2008. Classifying party affiliation from political speech. *Journal of Information Technology & Politics* 5(1):33–48.