NOTE: submitted for peer review December 18, 2019

What Drives U.S. Congressional Members' Policy Attention on Twitter?

Libby Hemphill, University of Michigan

Annelise Russell, University of Kentucky

Angela M. Schöpke-Gonzalez, University of Michigan

Training and Evaluating a Supervised Classifier

Training Data

Since not-policy tweets (topic "0") were overrepresented in our training dataset, in order to ensure that our supervised model could effectively learn patterns from policy-labeled tweets, we trained our best performing model (LR) on modified training sets from which we incrementally removed between 10% and 100% of randomly selected not-policy tweets (see Table 1). The best performing model according to F1 score, is that for which we removed 90% of all not-policy tweets. This proportion of not-policy tweets reflects approximately the median proportion of tweets represented by any given topic, supporting a nearly-balanced training set. This Tweet set amounted to 41,716 tweets total (39,704 policy tweets and 2,012 not policy tweets).

Preprocessing

In preparing text for use in a machine learning model, the text is divided into sequences of characters called *tokens* that are then used for analysis. Often, tokens are words, but in some cases they are multi-word phrases or parts of words such as word stems. Classifiers then look for associations between tokens and classes. How the text is processed into tokens impacts how classifiers make decisions, and here we describe the decisions we made during preparing the text for classification (or *preprocessing*).

We left words intact and did not reduce them to stems or lemmas. In the case of tweets, we expect that misspellings may often indicate different semantic meanings among terms with the same stemmed root, and thus potential association of different spellings with different topics. Stemming in these instances can remove the nuance in potential semantic meaning achieved by misspellings (Schofield & Mimno, 2016).

Given the prevalence of both English and Spanish language tweets, we removed English and Spanish stopwords using Python's Natural Language Toolkit (NLTK) English and Spanish stopword lists.

We employed NLTK's (Bird, Klein, & Loper, 2009) TweetTokenizer with parameters set to render all text lower case, to strip all Twitter username handles, and to replace repeated character sequences of length three or greater with sequences of length three. We complemented initial tokenization with removal of punctuation (including emojis), URLs, words smaller than two letters, and words that contain numbers.

Vectorization is the process of turning texts into numerical vectors that indicate the presence or absence of various tokens (or features) in a given text. We tested each of three vectorization approaches: simple one-hot encoding approach using Scikit-learn's (Pedregosa et al., 2011) DictVectorizer, simple bag-of-words approach with Scikit-learn's CountVectorizer, and bag-of-words term frequency inverse document frequency approach with Scikit-learn's TfidfVectorizer. Of these vectorization approaches, we found the simple bag-of-words approach using unigrams to result in the best performing models. This means that we represented tweets as unordered collections (or bags) of tokens (or words) using vectors that indicate, for each word in the entire collection of tweets, which are present in a given tweet.

Model Selection

This project trained and tested each of four types of classification models: a random guessing baseline dummy model (D) using stratified samples that respect the training data's class distribution, a Naive Bayes (NB) model, a Logistic Regression (LR) model,

and a Support Vector Machine (SVM) model. In each case, we used a 90-10 split for train-test data meaning that 90% of labeled tweets were used as training instances, and the models then predicted labels for the remaining 10%. We then compared the models' predictions with the human labels to evaluate their performance. After initial testing, for each of our top two performing models (LR and SVM), we subsequently evaluated whether the addition of Word2Vec (W2V) (Mikolov, Chen, Corrado, & Dean, 2013) word embedding features or Linguistic Inquiry and Word Count (LIWC) (Pennebaker, Booth, & Francis, 2015) features could improve their performance.

W2V models turn text corpora into numerical vectors (word embeddings), and groups these vectors of similar words together. Given sufficient data, these groupings can infer a word's meaning. For W2V feature embeddings, we attempted including each of W2V feature embeddings from Godin et al.'s pre-trained model (Godin, Vandersmissen, De Neve, & Van de Walle, 2015), as well as from a W2V model we trained on our Tweet set.

LIWC dictionaries are collections of words categorized and subcategorized into semantic and syntactic categories (e.g. emotional tone, percentage of words in the text that are pronouns, affect, biological processes, leisure activities, swear words, etc.). We used the LIWC 2015 Dictionary composed of approximately "6,400 words, word stems, and select emoticons" grouped into approximately 90 categories and subcategories (Pennebaker et al., 2015). Using this dictionary, we extracted LIWC categorical groupings for each token (word) in our Tweet set.

Evaluation Measures

We implemented each of D, NB, LR, SVM models with W2V and LIWC features using Scikit-learn with modifications to model configurations to improve predictive performance. To evaluate the performance of our models, we calculated F1 scores for each model. F1 scores are essentially the weighted average of precision and recall and

are a common performance measure for general classification models. There is no threshold or ladder of F1 performance, but higher is generally better.

Supervised Model Performance

Using a logistic regression (LR) classifier, we achieved the highest F1 score (0.79) in our ensemble. We compared our classifier to a dummy classifier using the same training data, and it achieved only F1 = 0.07. Given the difference between our classifier's score, the dummy, and the more complicated models and feature additions we evaluated (NB, SVM, W2V, and LIWC), we argue we have achieved high accuracy and that the LR classifier is the best available. Table 1 displays our models' results according to F1.

Table S1: Classifier Performance Summary

Classifier	**F1 Score**
Dummy	0.07
Naïve Bayes	0.70
Logistic Regression (LR)	0.79
LR + pre-trained word2vec features	0.78
LR + original word2vec features	0.77
LR + LIWC	0.78
Support Vector Machine (SVM)	0.78
SVM + pre-trained word2vec features	0.77
SVM + original w2word2vecv features	0.77
SVM + LIWC	0.78

What Drives U.S. Congressional Members' Policy Attention on Twitter? -

SUPPLEMENTARY DOCS

Additional Tables

Table S2. Topic Distribution and Top Term Features Associated with Each Policy Code

CAP #	CAP Label	Freq.	Prop.	Associated terms
1	Macro-economics	115,258	10.9%	budgetconference, unemployment, fiscalcliff, budgetdeal, renewui, manufacturing, budget, taxreform, sequestration, fiscal, debt, debtceiling, debtcrisis, sequester, dontdoublemyrate
2	Civil rights	80,958	7.7%	enda, passenda, nsa, marriageequality, surveillance, paycheckfairness, vra, abortion, lgbt, stopthebans, txlege, marchonwashington, marriage, snowden, talkpay
3	Health	155,638	14.7%	defundobamacare, obamacare, healthcare, obamacares, medicare, mentalhealth, nih, makedclisten, obamacareinthreewords, ocare, cancer, autism, stoprxdrugabusewv, flood, nsa
4	Agriculture	18,380	1.7%	farmbill, gmo, freedomtofish, fisheries, farm, sugar, agricultural, catfish, fishermen, crop, monsanto, fishing, nominee, stamp
5	Labor	41,067	3.9%	fmla, minimumwage, laborday, wia, raisethewage, familyact, righttowork, minimum, miners, pensions, cojobs, nlrb, jobs, obamacare
6	Education	43,748	4.1%	talkhighered, studentloan, tuition, studentloans, dontdoublemyrate, prekforall, investinkids, headstart, erate, restoreta, educational, stem, tribal, walmart
7	Environment	31,176	2.9%	actonclimate, climate, leahysummit, chemsafetyact, climatechange, chesbay, oilspill, keeptahoeblue, pollution, riograndedelnorte, tsca, carbontax, americarecyclesday, brownfields, nsa
8	Energy	22,401	2.1%	energyefficiency, hydropower, reca, helium, coal, energyindependence, biofuels, tva, americanenergy, keystonepipeline, cleanenergy, whitehouses, shovel, nebraska
9	Immigration	35,416	3.3%	immigration, immigrationreform, cirmarkup, cir, immigrants, momento, amnesty, immigrant, dreamers, hoevencorker, daca, econ, farmbill, nsa, acted
10	Transportation	17,903	1.7%	skagitbridge, obamaflightdelays, bridgeact, faa, airport,

				thud, transportation, maritime, highway, airports, harbormaintenancetax, rail, amndmnt, alleviate
12	Law and crime	88,135	8.3%	vawa, gun, voicesofvictims, guncontrol, gunsense, gunviolence, guns, passmjia, mjia, adoption, msa, sexualassault, nra, firearms, amndmnt
13	Social welfare	16,172	1.5%	snap, foodstamps, hungry, freecellphones, poverty, nutrition, seniors, older, socialsecurity, nationalservice, protectseniors, disappointment, welfare, hunger
14	Housing	5,514	0.5%	gsereform, fha, housing, fhfa, homeless, homelessness, liheap, mortgage, gse, revitalization, walkable, homeowners, affordablehousing, ndhfa
15	Domestic commerce	52,471	5.0%	fixflood, safechemicalsact, detroitbankruptcy, sandy, sandyrecovery, flood, fixfloodinsurancenow, patent, smallbiz, fema, smallbusiness, tourism, sandyaid, startupact, appears
16	Defense	148,929	14.1%	stolenvalor, drones, veterans, drone, ndaa, backlog, veteransday, endthevabacklog, missiletonowhere, nomination, assault, obamacare, immigration, energy
17	Technology	11,700	1.1%	marketplacefairness, nonettax, broadband, cyber, internetsalestax, cable, mfa, marketplacefairnessact, fcc, cybersecurity, internet, nasa, bolden, commissioners
18	Foreign trade	3,964	0.4%	trade, export, drywall, exports, olympic, overseas, concerning, automakers, currency, event, plants, connecticut, buyamerican, arms
19	International affairs	58,220	5.5%	benghazi, syria, egypt, waterstrategy, alqaeda, foreignrelations, israel, libya, ukraine, standwithisrael, nuclear, immigration, obamacare, nsa
20	Government operations	92,833	8.8%	nomination, irsscandal, postal, filibusterreform, irs, nominations, inform, nominee, gopshutdown, endgridlock, perez, confirmation, judges, nuclearoption, nominees
21	Public lands	17,807	1.7%	commissiononnativechildren, indiancountry, native, wildfires, monument, indian, wrda, wildfire, park, fundourparks, forest, parks, grazing, trashed

What Drives U.S. Congressional Members' Policy Attention on Twitter? -

SUPPLEMENTARY DOCS

Table S3. Analysis of tweet content for each of four sampled time periods, for both most prolific MCs in the *health* policy area and MCs that paid most attention to *health* relative to other policy areas.

MC	February 2017	March 2017	July 2017	January 2018
Patty Murray (Senate, D-WA)	Largely discussing nomination of Tom Price to U.S. Secretary of Health and Human Services, and against GOP's counter proposal to ACA.	Largely against GOP's counter proposal to ACA.	Largely against GOP's proposal to repeal ACA.	Largely against GOP's efforts in healthcare. Some discussion of children's healthcare provision, community health center funding, and the U.S. opioid epidemic.
Rob Portman (Senate, R-OH)	Largely discussing the U.S. opioid epidemic and related policy proposals. Minimal discussion of children's health.	Largely discussing the U.S. opioid epidemic and related policy proposals. Some discussion against ACA.	Largely against ACA. Minimal discussion of U.S. opioid epidemic.	Largely discussion of U.S. opioid epidemic. Some discussion of children's healthcare and mine worker healthcare.
Richard J. Durbin (Senate, D-IL)	9 tweets. Largely discussion against GOP's counter proposal to ACA and U.S. opioid epidemic. Minimal discussion of biomedical research funding, heart health, and airline staff health.	Largely against GOP's counter proposal to ACA. Some discussion of mental health and medical research issues.	Largely against GOP's proposal to repeal ACA.	2 tweets. Discussion of medical research and Legionnaire's disease disclosure policies.
Frank Pallone, Jr. (House of Reps., D-NJ)	Largely against GOP's counter proposal to ACA. Minimal discussing nomination of Tom Price to U.S. Secretary of Health and Human Services.	Largely against GOP's counter proposal to ACA.	Largely against GOP's proposal to repeal ACA. Minimal provision of information about Medicaid policies.	Discussion of children's healthcare, women's healthcare funding, community health center funding, and the impact of environmental pollution on public health. Minimal

				discussion of government shutdown's effect on healthcare services.
Brett Guthrie (House of Reps., R-KY)	0 tweets	Largely supporting GOP's counter proposal to ACA. Some provision of information about Medicaid policies.	0 tweets.	2 tweets. Discussion of health-related legislative action and support for Medicaid.
Michael C. Burgess (House of Reps., R-TX)	Largely supporting GOP's counter proposal to ACA. Minimal discussion of emergency medical care support, Medicaid improvement, and heart health.	Largely supporting GOP's counter proposal to ACA. Minimal discussion of health care center support and of U.S. opioid epidemic.	4 tweets. Largely in support of GOP's proposal to repeal ACA. Minimal discussion of electronic health record legislation.	9 tweets. Discussion in support of flu shot practices, children's healthcare, human trafficking awareness and prevention, U.S. opioid epidemic, and cancer treatment.
Grace F. Napolitano (House of Reps., D-CA)	1 tweet. Discussion of mental health.	Largely against GOP's counter proposal to ACA. Minimal discussion of mental health support.	5 tweets. Discussion of mental health service provision, Medicare funding, and against GOP's proposal to repeal ACA.	1 tweet. Discussion of support for mental health services.
Diane Black (House of Reps., R-TN)	9 tweets. Largely supporting GOP's counter proposal to ACA. Minimal discussing nomination of Tom Price to U.S. Secretary of Health and Human Services.	Largely supporting GOP's counter proposal to ACA.	7 tweets. Largely discussion of lowering Medicare costs, veterans' healthcare, and telehealth support. Some discussion in support of GOP's proposal to repeal ACA.	7 tweets. Largely discussion of government shutdown's impact on healthcare services. Minimal discussion of children's healthcare.

What Drives U.S. Congressional Members' Policy Attention on Twitter? -

SUPPLEMENTARY DOCS

Overall	Discussion of healthcare generally, debate surrounding GOP's proposal to repeal ACA, information about children's dentistry, and debate surrounding legalization of marijuana.	Discussion of mental health services, healthcare provider behaviors, plans to address the U.S. opioid epidemic, infant health, different health issues' awareness promotion, and environmental pollution's public health effects.	Largely debate surrounding GOP's proposal to repeal ACA.	Largely debate surrounding GOP's efforts in healthcare. Some discussion of children's healthcare.

References

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Sebastapol, CA: O'Reilly Media, Inc.

Godin, F., Vandersmissen, B., De Neve, W., & Van de Walle, R. (2015). Multimedia Lab @ ACL WNUT NER Shared Task: Named Entity Recognition for Twitter Microposts using Distributed Word Representations. *Proceedings of the Workshop on Noisy User-Generated Text*, 146–153. Retrieved from http://www.aclweb.org/anthology/W15-4322

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. Retrieved from http://arxiv.org/abs/1301.3781

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research: JMLR*, *12*(Oct), 2825–2830. Retrieved from http://www.jmlr.org/papers/v12/pedregosa11a.html

Pennebaker, J. W., Booth, J., & Francis, M. E. (2015). *LIWC2015*. Retrieved from https://liwc.wpengine.com/

Schofield, A., & 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. Retrieved from https://www.transacl.org/ojs/index.php/tacl/article/view/868