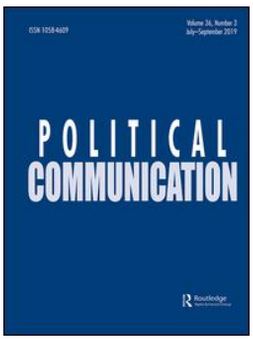

Political Communication

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/upcp20

Dictionaries, Supervised Learning, and Media Coverage of Public Policy

Lindsay Dun , Stuart Soroka & Christopher Wlezien

To cite this article: Lindsay Dun , Stuart Soroka & Christopher Wlezien (2020): Dictionaries, Supervised Learning, and Media Coverage of Public Policy, Political Communication, DOI: 10.1080/10584609.2020.1763529

To link to this article: https://doi.org/10.1080/10584609.2020.1763529

POLITICAL COMMUNICATION
https://doi.org/10.1080/10584609.2020.1763529

Routledge
Taylor & Francis Group

Dictionaries, Supervised Learning, and Media Coverage of Public Policy

Lindsay Dun [a], Stuart Soroka [b], and Christopher Wlezien [a]

[a]Department of Government, University of Texas at Austin; [b]Department of Communication and Media, University of Michigan

ABSTRACT

There are many different approaches to automated content analysis. This paper focuses on dictionaries and supervised learning; in addition to comparing the effectiveness of the two, we argue for the advantages of using them in combination. We do so in a research area in which we have an independent objective referent: government spending. With an eye toward capturing the accuracy of media coverage on public policy, we apply both hierarchical dictionary counts and supervised learning to measure mass media coverage of change in US defense spending. Both approaches appear to do well at capturing a media "policy signal" in the area, which provides an important test of convergent validity. While the results highlight the value of both dictionary and machine learning methods used independently, they also illustrate ways in which the two can be used in combination.

KEYWORDS

automated content analysis; machine learning; content-analytic dictionaries; policy feedback; public responsiveness

The earliest applications of large-scale automated content analysis relied almost exclusively on dictionaries. In the early days of readily-available digital text, dictionaries such as the General Inquirer (Stone et al., 1966) and Linguistic Inquiry and Word Count (LIWC; Pennebaker et al., 2001), alongside more specialized approaches such as Diction (Hart, 2000), were the standard approaches to extracting signals from large bodies of text. Over time, however, supervised-based approaches have proliferated. (See Grimmer & Stewart, 2013 for a particularly valuable review of the field.)

Many scholars suppose that supervised-learning approaches dominate those relying on dictionaries, predicated mainly on the supposition that computers are better able to capture the quantities of interest. Because supervised methods are able to take into account relationships between multiple words, quite possibly at different places in a document, they may be better able than simple dictionary-based word count approaches to capture the context in which words occur. This belief is reasonable and also supported by recent research.

This paper considers a third way, namely, the combination of dictionary-based and supervised-learning approaches. Indeed, supervised-learning applications almost always use dictionaries, if only implicitly (Hopkins & King, 2010; Monroe et al., 2008; also see Wilkerson & Casas, 2017), and we attempt here to demonstrate the advantages of a more explicit *dictionary-plus-supervised-learning* approach. We assess the success of the

CONTACT Christopher Wlezien Wlezien@austin.utexas.edu Department of Government, University of Texas at Austin, Austin, TX 78712-1704 USA.

Supplementary data for this article can be accessed on publisher's website at https://doi.org/10.1080/10584609.2020.1763529.

methodology by comparing content analyses of media coverage of public policy with an independent objective referent – government spending. The substantive importance of this test case is discussed in the next section.

Below, we apply hierarchical dictionary counts and supervised learning to media coverage of change in US defense spending, a publicly salient domain in which there is strong evidence of thermostatic public responsiveness (Wlezien, 1995, 1996; Soroka & Wlezien, 2010). We rely on a corpus of roughly 2 million articles from 17 newspapers over the period between 1980 and 2018, applying a combination of hierarchical dictionary searches as well as a random forest machine learning algorithm trained on sentences extracted using dictionaries and coded by humans – an approach which we label the *dictionary-plus-supervised-learning* approach. Results using this method perform only slightly better than do the hierarchical-dictionary results in terms of capturing a media "policy signal" that moves alongside spending change. Whether the improvement is enough to justify the use of supervised learning rather than hierarchical dictionaries in this instance is not entirely clear. We nevertheless see supervised learning as a way to augment and extend dictionary-based approaches, and *vice versa*. We thus consider several ways in which the two approaches can be used in tandem. Substantively speaking, we regard our results as highlighting the accuracy of US media coverage of defense spending; methodologically speaking, our emphasis is on the potential value of combining dictionary- and machine-learning-based methods of automated content analysis.

Background

The initial motivation for our research is to better understand the role of mass media as a mechanism(s) of policy feedback. Some research demonstrates negative feedback, where the public adjusts its public preference "inputs" thermostatically in response to policy "outputs" (Wlezien, 1995; Erikson et al., 2002; Eichenberg & Stoll, 2003; Jennings, 2009; Soroka & Wlezien, 2010; Ellis & Faricy, 2011; Ura & Ellis, 2012; Wlezien & Soroka, 2012; Pacheco, 2013). Other research finds positive feedback, where an increase in policy leads people to want more spending in that domain (see the excellent review of the large and diverse literature in Beland & Schlager, 2019). Both relationships, and perhaps especially the thermostatic one, require that the public receives information about policy outputs. Our substantive interest is in assessing the role of the mass media in communicating this information; and doing so requires a method of capturing, across large bodies of media content, measures of mass-mediated policy information.

Previous research on media coverage of policy has typically not focused on policy outputs. Scholars have tended to concentrate on media priorities (and frames) and their impact on policy decision-making (Baumgartner & Jones, 1993; Boydstun, 2013; Card et al., 2015; McCombs & Shaw, 1972). While important and relevant to policymaking (and the public), this work does not help us understand whether mass media content reflects what policymakers actually do.

Some other recent research does address media coverage of policy, focusing particularly on defense spending (Neuner et al., 2019; Soroka & Wlezien, 2019). That work proposes a measure of media coverage that focuses on policy *change*, not the actual levels of policy. There are two main reasons for this. First, it may be that the change in policy is what the

media coverage reflects, much as is the case for the economy (see Soroka et al., 2015; Wlezien et al., 2017). Second, we can directly measure media coverage of change using relatively simple dictionaries. That research shows that the resulting media signal ($^{\Delta}M_t$) closely follows spending change (ΔP_t) itself, per the following equation:

$$^{\Delta}\mathrm{M_t} = \mathrm{f}\{\Delta \mathrm{P_t}\}. \quad (1)$$

Prior research concentrates on a single domain in a single country, spending on defense in the US. We follow that lead in the analysis that follows. We thus regard this paper as (a) building directly on previous work relying solely on dictionary-based content analysis of defense, but also (b) adding to a growing body of work addressing the role that mass media play in public opinion formation and political representation.

The Media Corpus

Examining the possibility that media coverage is a mechanism for public responsiveness requires a reliable measure of the media policy signal. What is the best way to identify this signal? Our examination focuses on a combination of dictionary- and supervised-learning-based techniques applied to a massive corpus of newspaper stories. This corpus can be drawn from a number of full-text resources, of course; we rely on Lexis-Nexis due to access to the Web Services Kit (WSK), which facilitates the downloading of several hundred thousand stories, formatted in xml, in a single search request. A search request can be based on either pre-coded subjects, or full-text keywords, or both. We use a combination, as follows: STX001996 or BODY(national defense) or BODY(national security) or BODY(defense spending) or BODY(military spending) or BODY(military procurement) or body (weapons spending).[1] STX001996 is the "National Security" index term, one of five sub-topics with the "International Relations and National Security" topic. It captures the lion's share of articles on defense policy, spending and otherwise. Of course, Lexis-Nexis' assignment of topics is most likely a function of their own dictionary-based word search, and our assumption is that their search is more developed than ours would be. Even so, in order to not miss other spending-related articles, we add the full-text (BODY) search terms identified above.

We arrive at the above search terms based on some preliminary tests, exploring the reliability with which different searches capture relevant articles and avoid too many irrelevant ones. Even so, we invariably miss some articles relevant to spending, and our analyses identify a considerable volume of irrelevant material as well. We suspect that using the "National Security" index term means that we err on the side of Type I rather than Type II errors, i.e., we are more likely to include items that we shouldn't than exclude ones we should include. That said, we expect that most irrelevant articles do not factor into our measure of the net media signal, since we use a combination of layered dictionaries to identify the instances of spending mentions most likely to pertain to change in defense spending. Diagnostic analyses support this expectation, as we will see.

Our working database relies on the following newspapers: *Arizona Republic, Arkansas Democrat-Gazette, Atlanta Journal-Constitution, Boston Globe, Chicago Tribune, Denver*

Post, Houston Chronicle, LA Times, Minneapolis Star-Tribune, New York Times, Orange County Register, Philadelphia Inquirer, Seattle Times, St. Louis Post-Dispatch, Tampa Bay Tribune, USA Today, and *Washington Post*. Not all newspaper databases start in 1980 – most enter the dataset in the early 1990s. All are gathered up to the end of the 2018 fiscal year, i.e., September 30.

Our selection of newspapers is based on availability, alongside circulation, with some consideration given to regional coverage. In the end, we have 17 of the highest-circulation newspapers in the US, three of which aim for national audiences, and seven of which cover considerably large regions in the northeastern, southern, midwestern, and western parts of the country. Combining these newspapers offers, we think, a reasonable representation of the national news stream, at least where newspapers are concerned. Using a relatively wide range of newspapers has an additional advantage: to the extent that the language and/or focus of defense stories varies across outlets, there are advantages to building both dictionaries and supervised learning models across a corpus that is relatively broad. The total database includes 2,171,189 stories, albeit with more from the mid-1990s onwards, when all of our 17 newspapers are in the database. This can be seen in Appendix Figure A1, which plots the number of articles by year across each of our 17 newspapers.

Not all of this content is focused on defense spending. The analyses that follow are not based on this text, then, but rather all *sentences* within this corpus that focus on spending, as defense spending change is our concept of interest. To be clear: our working database in the analyses that follow is at the sentence level, where sentences are extracted from the larger database using a simple keyword-in-context (KWIC) search identifying all sentences with a keyword related to government spending. We do this using a SPEND dictionary, which includes the following words:

SPEND: allocat*, appropriation*, budget*, cost*, earmark*, expend*, fund*, grant*, outlay*, resourc*, spend*

This dictionary search (and all subsequent dictionary searches) is implemented in the *quanteda* package in R (Benoit et al., 2018). Note that the dictionary has been subjected to testing in Soroka and Wlezien (2019), and was constructed from a reading of keyword-in-context (kwic) retrievals, augmented by thesaurus searches. The dictionary-building procedure, in a nutshell, was as follows: (1) read a random draw of articles extracted using Lexis-Nexis keywords, and establish a simple set of words that seem to capture "spending," (2) augment that list using a thesaurus, and (3) search for each dictionary word, extracting kwic entries and reviewing those entries to ensure that every word is used, most of the time, in the way in which we anticipate – in this instance, in the context of a sentence about spending. Applying this dictionary to our news-story corpus results in a database of 1,775,008 sentences. Figure 1 plots the number of spending sentences in our database by fiscal year. This is the raw material for the analyses that follow.

Capturing a "Media Policy Signal"

Having produced the corpus, we now need to implement our content analyses. As noted above, there are many approaches to computer-automated content analysis. The earliest ones employed dictionaries created by investigators, and this approach still is widely used

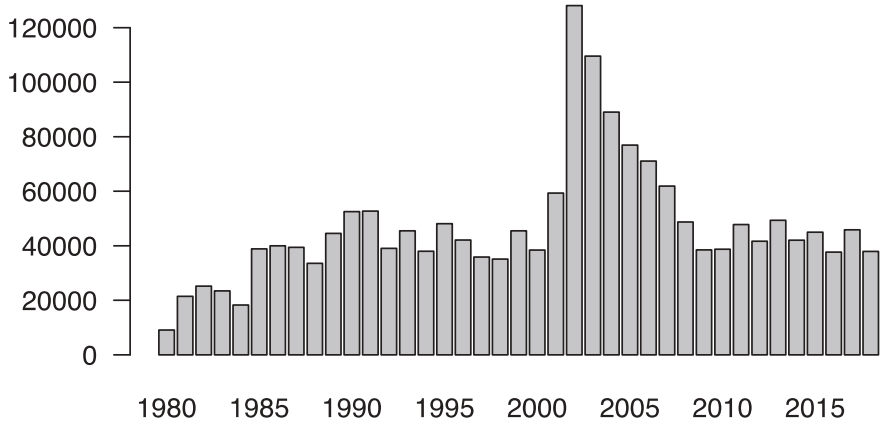

Figure 1. Defense spending sentences, by fiscal year.

today. Increasingly, however, data analysts rely on machine learning approaches, some of which are entirely unsupervised by humans and others supervised. (There also is a combined "semi-supervised" approach.) For our purposes, dictionaries and supervised methods are most appropriate, because we know our classification categories.

Dictionaries

Our working corpus is already premised on two dictionary searches: first, in the use of Lexis-Nexis topics (derived from proprietary dictionaries) and additional full-text search terms relating to national security, and second, in the application of the SPEND dictionary to extract sentences related to spending. Even this database will include content not directly related to spending change, however. Our aim is thus to narrow our focus further.

First, we narrow to sentences that mention both spending *and* direction. We identify direction using UP and DOWN dictionaries, built and implemented using the same process described above. The dictionaries are as follows:

UP: accelerat*, accession, accru*, accumulat*, arise*, arose, ascen*, augment*, boom*, boost, climb*, elevat*, exceed*, expand*, expansion, extend*, gain*, grow*, heighten*, higher, increas*, increment*, jump*, leap*, more, multiply*, peak*, rais*, resurg*, rise*, rising, rose, skyrocket*, soar*, surg*, escalat*, up, upraise, upsurge, upward

DOWN: collaps*, contract*, cut*, decay*, declin*, decompos*, decreas*, deflat*, deplet*, depreciat*, descend*, diminish*, dip*, drop*, dwindl*, fall*, fell, fewer, less, lose, losing, loss, lost, lower*, minimiz*, plung*, reced*, reduc*, sank, sink*, scarcit*, shrank, shrink*, shrivel*, shrunk, slash*, slid*, slip*, slow*, slump*, sunk*, toppl*, trim*, tumbl*, wane, waning, wither*

Applying these dictionaries to our sentence-level corpus identifies 575,602 "spending" sentences that also include "direction" keywords.

Second, in order to reduce the number of false positives, i.e., sentences that are captured in our search but are in fact not directly related to defense spending, we run

a simple dictionary search over the set of spending change sentences to confirm that each includes at least one of the following DEFENSE words:

DEFENSE: army, navy, naval, air force, marines, defense, military, soldier, war, cia, homeland, weapon, terror, security, pentagon, submarine, warship, battleship, destroyer, airplane, aircraft, helicopter, bomb, missile, plane, service men, base, corps, iraq, afghanistan, nato, naval, cruiser, intelligence

Doing so isolates 206,426 "spending" sentences that contain "direction" keywords and words clearly related to defense.[2]

To summarize the hierarchy of dictionaries employed, we begin with *articles* from Lexis-Nexis that relate to "national security," from which we extract *sentences* relating to "spending" and then isolate those that also indicate "direction" and explicitly relate to "defense." Note that the ordering of the sentence-level dictionaries does not matter, as they technically are not nested but rather are applied jointly. There is reason to think that the dictionaries will produce an increasingly reliable measure (see Soroka & Wlezien, 2019). This is not assured, however, and even if successful, there still may be irrelevant sentences in our database.

Note that our approach here is identical to the use of hierarchical dictionary counts as implemented in past work (e.g., Bélanger & Soroka 2012; Young & Soroka 2012). We also regard this application of dictionaries as very similar to the "learning" inherent in supervised learning methods used for large-N content analysis (e.g., Jurka et al., 2012.) There sometimes is a perception that dictionaries are necessarily general, not specific to particular domains of interest, and that they are simple word lists, concocted based only on a thesaurus, where words are not subjected to testing, and where results are thus likely to either miss much relevant material or capture a good deal of irrelevant material. This certainly can be true, but the use of several iterations of testing during the dictionary-building stage, and the subsequent use of multiple dictionaries that essentially removes false positives, i.e., a spending word that is in fact not related to defense, makes for a rather different dictionary-based analysis – one that has used a corpus and human coding to "learn" about the terms most relevant to the analysis. Admittedly, this is not as easy and cheap as using generic off-the-shelf dictionaries but it also is not as intensive (and expensive) as supervised learning approaches. The analytical benefits of context-specific dictionaries have been explored in other work (Muddiman et al., 2019), and it is worth noting that just because a dictionary is domain specific does not mean it cannot be used by other scholars for related research. The dictionaries used here may be useful for analyses of different types of media, and the SPEND and UP/DOWN dictionaries may be useful for analysis of spending information across other policy domains. Of course, these possibilities would require testing these dictionaries on other datasets.

Previous research details how the application of successive dictionaries works (Soroka & Wlezien, 2019); another paper further highlights the strong connection between this dictionary-based approach and human coding (Neuner et al., 2019). Both articles also demonstrate a means by which to create a measure of the "media policy signal," based on this sentence-level coding, and compare it with actual spending change. We use a slightly different (but functionally equivalent) approach here: we code sentences in which there are more UP words than DOWN words as "1" and sentences in which there are more DOWN

words than UP words are coded as "-1," while other sentences are coded "0." (Note that the vast majority of sentences have just one direction keyword.) We then sum these values across sentences, by fiscal year (October-September). The resulting measure captures both direction and magnitude, and can be calculated for all newspapers or single newspapers. The resulting signal is shown below, in Figure 3, but before we turn to that, we review our dictionary-plus-supervised learning approach.

Dictionary-plus-Supervised-Learning

Figure 2 illustrates the steps taken for the layered dictionary approach alongside the steps taken for the dictionary-plus-supervised-learning approach. The latter relies on the same body of sentence-level data as the purely dictionary-based approach, that is, the sentences identified using the "spending" dictionaries on national security articles from Lexis-Nexis. In this case, however, we supervise learning based on a subset of human-coded data to identify defense spending change.

It is of course possible to use a machine-learning approach mostly independent of the dictionaries, although the initial identification of the corpus would in most cases still require some kind of dictionary-based search. As noted in Figure 2, however, our combined approach relies on dictionaries to identify the sentences that humans code in order to train our machine learning algorithm. Doing so allows us to focus human coding on a more balanced corpus, i.e., a corpus with a sufficiently high rate of spending mentions.

We proceeded with human coding in four rounds. In the first two rounds, in order to ensure at least a minimal amount of human coding across all keywords in our SPEND dictionary, we draw a random sample of 25 sentences from each of our 11 keywords. These rounds were used in part to test the relevance of dictionary keywords (see Soroka & Wlezien, 2019), so a third round then draws 40 sentences from seven of the most common and reliable spending keywords. The cumulative results of these three rounds of coding produce a somewhat imbalanced coding set, where a majority of sentences are still not

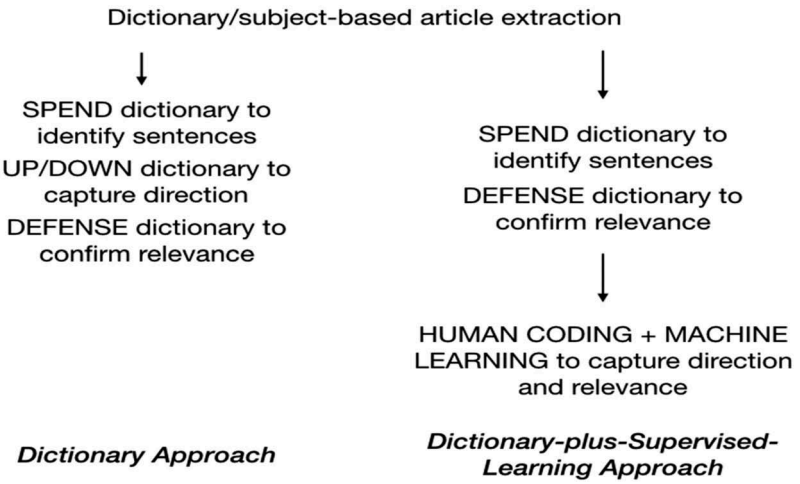

Figure 2. Dictionary and dictionary-plus-supervised-learning approaches.

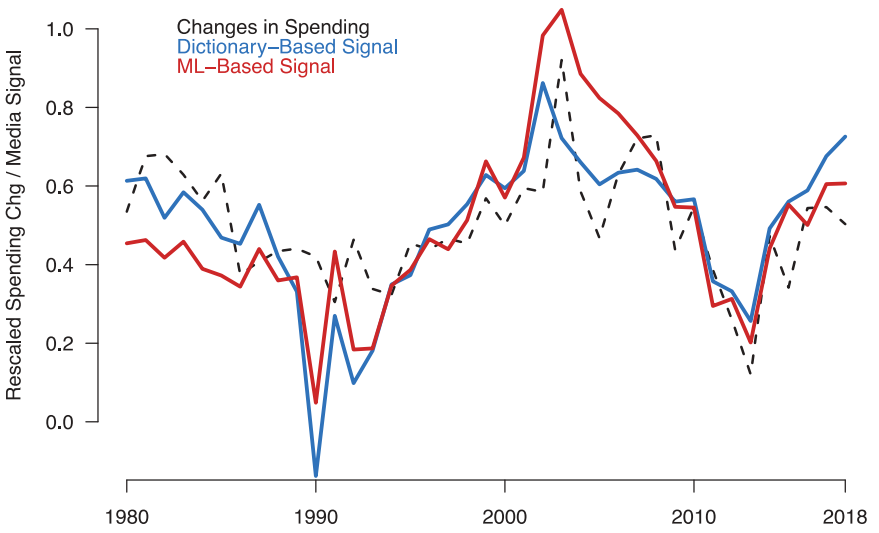

Figure 3. The media signals and spending change.

coded as indicating changes in defense spending (see coding details below), so a fourth round draws 100 sentences from each of our seven keywords, but this time also including at least one word from the DEFENSE dictionary. This is done to improve balance in the human-coded data, which is critical to the training of an effective model, i.e., one that discriminates among sentences instead of simply predicting the most common category in the training set for the overwhelming majority of sentences.

Note that this approach may increase the baseline correspondence between our dictionary-based and supervised-learning approaches – the latter begins with a similar set of dictionary-derived content, and is trained on dictionary-derived sentences as well. But this partial reliance on a dictionary will produce more balanced data, which is critical to effective machine-learning. Because our objective is to see if we can improve on dictionary-only results with a dictionary-plus-supervised-learning approach, we explicitly use the dictionary to assist the machine learning algorithm in this way.

We follow this approach for two additional samples of coding. In total, we draw 5,441 sentences from the sentences that include spending cues, with a particular focus on sentences that also include a defense keyword. We randomly selected 80% (4,353) of the sample to train the computer, and leave the remaining 20% (1,088) to test the resulting algorithm.

Human coding is done online through Amazon Mechanical Turk (MTurk). We collect 5 codes for each sentence; and we ask for codes on both the relevance and direction of each sentence. We ask (1) whether a sentence is about defense spending, and then, if so, (2) the direction of spending change. We collapse the resulting codes into 4 categories: 0 = not about defense spending; 1 = about defense spending but with no information about the direction of change; 2 = about a spending decrease; 3 = about a spending increase. We then aggregate human codes based on the following decision rule: we require at least 60%, i.e., 3 or more, of the 5 MTurk coders to agree a sentence is relevant *and* at

least 60% to agree on a direction code. In exploratory analyses, we found that our algorithm performed slightly better in our test set (measured both in overall accuracy and average class precision/recall) when we assigned classes based on 60% agreement rather than 80% agreement.[3]

In sum, the MTurk data is transformed into sentence-level codes as follows:

(a) If <60% of MTurkers say a sentence is relevant, the sentence gets a 0;
(b) If ≥60% of MTurkers say a sentence is relevant but <60% of MTurkers agree on a direction, the sentence gets a 1;
(c) If ≥60% of MTurkers say a sentence is relevant and ≥60% of MTurkers say spending is going down, the sentence gets a 2; and
(d) If ≥60% of MTurkers say a sentence is relevant and ≥60% of MTurkers say spending is going up, the sentence gets a 3.

Note that this is similar to the dictionary approach, taking into account the combined application of the SPEND, DIRECTION and DEFENSE dictionaries.

The supervised-learning model requires a document frequency matrix, in which we retain unigrams and bigrams,[4] and stem all text features. We remove English stop words as well as all words/word pairs that occur less than two times in the training dataset, and then discard all features in the uncoded sentences that are not in the training set. Doing so retains 10,834 features to use in prediction; this is known as a "bag of words" feature set. We subset that document frequency matrix into the training (80% of the sample) and test (20% of the sample) sets, as described above.

Random forest analyses are carried out using the *randomForest* and *rfUtilities* packages in R. We selected the random forest classifier (Breiman, 2001) because it is generally considered one of the more accurate and efficient multi-class algorithms.[5] We trained our random forest model using 250 trees, which we found offered a good balance between prediction accuracy and computing time.[6] To further improve model performance, we ran a tuning function,[7] assuming a 250-tree model, to determine the optimal number of variables available to the model for splitting at each tree node – this is known in random forest analyses as the "mtry" parameter. The tuning function suggested that setting mtry = 416 minimized prediction error in data not used in model training.[8]

After determining the appropriate number of trees for the model to "grow," as well as the optimal value of mtry to program in model specifications, we instruct the random forest model to weight samples drawn from each training category (0, 1, 2 and 3) in inverse proportion to how often they show up in the training set. We do this because the training set is unbalanced – we still have too many 0's – and inverse weighting should lead the random forest model to produce more accurate predictions in uncategorized data. (For more detail on random forest models, see, e.g., Montgomery & Olivella, 2018; Siroky, 2009.) We do not run k-fold cross validation to test the effectiveness of our final model, as random forest models generally are not prone to overfitting.[9]

We use two indicators to judge model performance in our test set: precision and recall. Recall is the proportion of correct predictions made by the model out of all human coded sentences in a class (correct predictions divided by column sums in our confusion matrices presented in Table 1). Precision is the proportion of correct predictions made

by the model out of all model predictions made in a class (correct predictions divided by row sums in our confusion matrices).

Panel 1a of Table 1 shows that our algorithm generalizes to out-of-sample data fairly well. Precision and recall in the test set (1a) are reasonably high, although reliably capturing spending change and direction (codes 2 and 3) appears here to be more difficult than identifying irrelevance (code 0). Panel 1b produces similar metrics using the dictionary results. Indeed, these metrics are strikingly close to the ones in 1a,[10] although the dictionary is slightly worse at predicting 0s and 3s than the random forest classifier.

In an ideal world, we could work to improve the random forest model's precision and recall. It seems likely that model performance is partly limited by the fact that humans who read these sentences don't always have the same interpretation of what is "relevant" to defense spending, and what sentences specify an "increase" or "decrease" in spending. The fact that the dictionary and algorithm both struggle the most with "3" sentences suggests that the words used to specify a spending increase may be either (a) ambiguous or (b) also used in ways that do not always indicate a spending increase. If this is correct, there is a ceiling on how accurate our model can be, given the training data we have.

In spite of these limitations, results in Table 1 give us some preliminary indication of how well a random forest model can predict defense-spending categories designated by humans. These averages do not necessarily indicate how well the algorithm will perform in classification of the entire dataset, of course. As noted in Hand (2006, p. 10), the supervised learning paradigm assumes that "there is some fixed clear external criterion which is used to produce class labels." We have designed our classification categories to fit a specific research task, but our categories do not have objective thresholds. This class ambiguity is, we suspect, reflected in the accuracy. Accuracy alone is thus not always the ideal benchmark for model performance. Luckily, we can compare out-of-sample predictions to an objective referent: change in defense spending. This is a relatively rare opportunity in machine learning analyses.

Table 1. Confusion matrices.

1a. Algorithm Performance in Test Set								
		Human Codes						
		0	1	2	3		Precision	Recall
Predictions	0	306	58	6	29	Class 0	0.77	0.83
	1	40	233	27	83	Class 1	0.61	0.61
	2	8	35	66	7	Class 2	0.57	0.60
	3	15	55	11	109	Class 3	0.57	0.48
Accuracy in Test Set: 65.63%								
1b. Dictionary Performance in Test Set								
		Human Codes						
		0	1	2	3		Precision	Recall
Predictions	0	220	28	1	16	Class 0	0.83	0.60
	1	94	247	25	110	Class 1	0.52	0.65
	2	26	59	78	17	Class 2	0.43	0.71
	3	29	47	6	85	Class 3	0.51	0.37
Accuracy in Test Set: 57.9%								

Comparing Aggregate Results

In order to compare the results from our two approaches, we generate a second "media signal" similar to the one described above based on dictionaries, but this time based on the supervised learning results. To do so, we first code the entire 1.7 million sentences (excluding those 5,441 sentences used in training and testing) using the trained model.[11] We then aggregate the resulting codes by fiscal year, to produce a media signal. Figure 3 plots the resulting media signals, for both dictionary and supervised-learning results, alongside spending change, based on defense appropriations (budget authority) in FY2000 US dollars, drawn from the *Historical Tables* distributed by the OMB. In order to plot these series on the same axis, each is standardized to have a range of 1, and then centered to have a mean of 0.5.

The correspondence between the two media signals is striking, as is their correspondence with spending. The two media signals – one based entirely on dictionaries, and the other based on supervised learning – have a Pearson's correlation of 0.85. How does each measure compare with spending change? The correlation with spending change is only marginally larger for the supervised learning signal (0.63) than for the dictionary signal (0.60), and the difference is not reliable (N = 39; for the difference in correlation coefficients, p = .84). In sum, dictionary-plus-supervised-learning approach yields only a very small gain in the accuracy of our media policy signal.

Is the marginal gain in accuracy enough to justify the additional costs of the supervised learning, in terms of both human coding and computational resources? In this particular case, we would arrive at the same conclusion with both measures; either way, we would find a relatively strong correspondence between media coverage and defense spending change. We take this as evidence of the measurement validity of both the dictionary and dictionary-plus-machine-learning approaches. It is reassuring to know that dictionaries and machine-learning produce, in this case at least, roughly similar results.

It is possible that machine learning methods would offer greater analytical benefits in other research domains. The hint of an increase in accuracy here does signal the potential for supervised-learning to find (meaningful) variation that is missed by dictionaries alone. Our take-away from this analysis, therefore, is not that supervised learning always adds little to what we can extract from media content using dictionaries. However, we do take our results to suggest that high quality, well-developed dictionaries remain a clearly valuable tool for content analysts. Given that they are more straightforward – and often less costly – to implement, this seems an important methodological finding.

Combining Approaches

Recall that our objective is only partly to consider the dictionary and dictionary-plus-supervised-learning approaches independently – we also want to the consider the two approaches used in tandem. We expect that there are several ways in which this might be fruitful.

The first way in which approaches can be combined is already widespread, although the dictionary portion often is implicit rather than explicit. In order to either identify the corpus, or to achieve balance in a training set used for machine learning, or both, a simple dictionary search can be critical. Calling our method "dictionary-plus-supervised-learning" is intended

to make the dictionary-based identification of the corpus especially clear – it recognizes more explicitly that dictionaries are regularly a precursor to machine learning.

Another straightforward combination of the two approaches is to use dictionaries as a check on supervised learning, and vice versa. This involves no more than what has been done above. The central quantity of interest – in this case, a fiscal-year measure of the media signal – is estimated using each approach independently, and the results are compared. Doing so offers a valuable test of the concurrent validity of each measure. Where the dictionary is concerned, we now know hierarchical dictionaries produce a result that is in line with supervised learning based on human coding. Where the dictionaries-plus-supervised-learning approach is concerned, confirming that results are similar to a set of dictionaries including words that are very clearly connected to the quantities of interest provides a valuable check on what is otherwise a somewhat opaque coding mechanism, i.e., one in which the words guiding the analysis are determined by the algorithm.

A third, more ambitious combined approach is to use supervised learning to improve the dictionary, and *vice versa*: using supervised learning to identify words not already included in the dictionary, and then using the improved dictionary to rebuild a corpus for supervised learning, and so on. Table 2 offers an illustration of the first step in this process. The table relies on the dictionary-plus-supervised-learning results from the preceding section. We take the resulting codes from the random forest coding, return to the corpus, and extract the words from our entire dataset that are highly correlated with each code. Doing so offers a glimpse of the words that are doing most of the work in the

Table 2. Words associated with machine-learning codes.

Relevant, Spending Increase		Relevant, Spending Decrease	
Word	Correlation	Word	Correlation
Billion	0.42	Cuts	0.39
Homeland	0.29	Reductions	0.32
Increase	0.21	Cut	0.27
Billions	0.19	Reduction	0.27
Defense	0.17	Defense	0.24
Security	0.16	Spending	0.24
Military	0.16	Military	0.16
Spending	0.14	Cutting	0.14
Increases	0.12	Reduce	0.12
Pentagon	0.12	Automatic	0.11
Increased	0.11	Deficit	0.1
Iraq	0.1	Budget	0.1
Afghanistan	0.09	Domestic	0.1
Year	0.09	Billion	0.09
Dollars	0.09	Reduced	0.09
Department	0.08	Cutbacks	0.09
Increasing	0.08	Acrosstheboard	0.09
Next	0.08	Tax	0.08
War	0.07	Reducing	0.07
Bill	0.07	Deep	0.07
Inflation	0.06	Proposed	0.06
Fiscal	0.06	Programs	0.06
Wars	0.06	Taxes	0.06
Request	0.06	Pentagon	0.06
Weapons	0.05	Reagan	0.06
Missile	0.05	Deeper	0.06
Nearly	0.05	Savings	0.05
Pentagons	0.05	Sequestration	0.05

automated coding (and, by implication, the words that humans interpret most clearly as indicating our quantities of interest). Doing so also allows us to consider words that were not, but maybe should have been, part of the dictionaries.

Table 2 shows all words that correlate at r = 0.05 or higher with codes 2 (relevant, and indicating a spending decrease) and 3 (relevant, and indicating a spending increase) from the supervised learning. For the most part, words are as we should expect – they clearly signify the quantities of interest, and they are for the most part already included in the dictionaries. Consider some of the top words in the increase column, i.e., "billion" and "increase"; and some of the top words in the decrease column, i.e., "cuts" and "reductions". Some words appear in both columns, such as "spending"; these are likely the words that identify relevance. And there are some words that seem to indicate spending that are not already in our spending dictionary, i.e., "supplemental" and "request."

Generally speaking, neither of the latter two words – supplemental and request – are necessarily spending words. But given that we are working with a corpus that includes only defense-related articles, it may be that these words are reliably about spending. If that is the case, then the sentence-based corpus that we extract using the SPEND dictionary may be missing some relevant data. What might results look like if we tested and then added the word supplemental to our spending dictionary, added some human coding based on sentences with the new keyword, and then re-estimated the supervised-learning model? Our estimates of the media signal may improve slightly. And further examination of the words correlated with the supervised-learning codes might reveal yet additional revisions to the dictionaries.

We do not go down that path for the time being, but we want to highlight it as one potentially fruitful outcome of combining the approaches illustrated above. The hierarchical-dictionary and dictionary-plus-supervised-learning approaches may each be useful in certain circumstances; in others, they may *both* be useful.

Discussion

A carefully constructed measure of media coverage of policy outputs is of real importance for those interested in policy responsiveness and representation. As noted above, many citizens learn about most policies indirectly, often through mass media. The opinion-policy link thus depends heavily not just on the volume but on the accuracy of media coverage of policy. Where media provide accurate policy cues, there are good reasons to expect public responsiveness and policy representation. Where media cues are systematically different from actual policy, the potential for responsiveness and representation is limited (see Neuner et al., 2019). Identifying an accurate media policy signal allows us to assess the potential for representative democracy, policy domain by policy domain, across time and space.

If the aim of a media policy signal is to assess the accuracy of the information citizens receive, then the degree to which the measure reflects (a) accurate versus inaccurate coverage, or (b) valid measurement versus invalid measurement, is of critical importance. Put differently, if we find that media do not reflect spending change, we would like to know that the finding reflects biases in media coverage, not simply the methodological difficulties of large-scale content analysis. This is the main motivation for this paper. We would like to be confident that we have given media coverage its "best shot" at reflecting spending change.

We considered here the possibility that adding a supervised-learning approach to dictionary-based analyses would lead to an improved measure of the media policy signal, focusing specifically on defense spending. Our analysis reveals that the hierarchical-dictionary approach does about as well as the dictionary-plus-supervised-learning method in capturing the signal in newspaper coverage. This may be due to the fact that there is a ceiling on how well media content reflects policy change, and we are reaching that ceiling in both the dictionary and machine learning approaches. It also may be that that the dictionary approach actually does not fully capture the underlying signal and machine learning just adds little. Both may be at work.

We are not staking a claim on the success of dictionary-based approaches in all instances, of course. As Grimmer and Stewart (2013) note, the best content-analytic approach will vary widely depending on the requirements of the data and theory. Dictionaries may be especially effective in this case because of the limited and readily-identifiable words referring to both spending and direction and the need to identify explicit cues that would be clear to media consumers. The former likely reduces the gains from supervised learning methods, since human coding augmented by computational methods may offer little gain in reliability. The latter reduces the advantages of automated clustering methods such as latent Dirichlet allocation (LDA; e.g., Blei et al., 2003) or structural topic modeling (STM; e.g., Roberts et al., 2014), which capture correlations between words that need not be proximate or related in ways that would be meaningful for the average reader. Each of these other approaches obviously are of real value in other types of content analysis.

Approaches to automated content analysis need not compete, however, as they can be used in combination. That has been our primary emphasis in the final sections of this paper. We already have seen benefits of dictionaries in creating a more relevant corpus for use in supervised learning. It also may be that supervised learning can help improve dictionaries, by identifying relevant and irrelevant words. These seem to be important subjects for future research in communications studies and political science.

Notes

1. Note that full-text search terms are searched as phrases, e.g., "national security," not "national" and "security" separately.
2. Note that for our analyses below that compare the dictionary and machine learning approaches on uncoded sentences, all sentences used in machine learning model training also had to be removed from the dictionary dataset. This slightly reduces the total number of dictionary sentences with a spending word, direction word, and defense word to 204,565.
3. We do not provide detailed results based on somewhat different ways of aggregating human coding (e.g., requiring 80% agreement), but results do not differ significantly from what is reported here, and are available upon request. Since articles are assigned direction codes only when they are relevant, and there are instances in which not all coders agree on relevance, there are a small number of relevant articles for which there are only 3–4 direction codes. This is one reason why the 3-coder cutoff used here is advantageous. Syntax and data to replicate reported results are available at https://doi.org/DOI:10.17605/OSF.IO/TPA6U
4. We also tested the model including trigrams in the feature set. The performance of this model was so similar to that of the unigram-bigram model that we decided the additional computational resources required to estimate a unigram-bigram-trigram model were not justifiably offset by model improvements.
5. A multi-class SVM model leads in this instance to less accurate predictions than the random forest approach.

6. We also tested the model with 500 trees and found that this offered only a small improvement (measured in accuracy, precision, and recall) over 250 trees, but took substantially longer to train.
7. For this we used the tuneRF() function in the R package *randomForest*.
8. The tuning function also took into account class weights, or the inverse proportion of each prediction class in the total training sentences.
9. See, e.g., https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#ooberr.
10. Dictionary test set codes were generated as follows: if the dictionary counts indicated the sentence had at least one defense word but no direction word (or equal numbers of "up" and "down" words), the sentence got a "1". If the dictionary counts indicated that the sentence had ≥ 1 defense word, and the sentence had more spending "down" words than "up" words, the sentence got a "2". Lastly, if dictionary counts indicated a sentence had ≥ 1 defense word, and the sentence had more spending "up" words than "down" words, the sentence got a "3".
11. A total of 5,673 sentences were coded by either MTurkers or students. In initial model testing to determine how to maximize model performance, we used the student codes as well as the MTurker codes. The sentences we use for model prediction are all sentences not coded by either MTurkers or students. The remaining 1,769,335 uncoded sentences are what we use in prediction.

Acknowledgment

*Previous versions of this paper were presented at the Texas Methods Conference, Houston, 2019, and the Policy Agendas Project workshop at the University of Texas at Austin, also in 2019. For research assistance, we thank Connor Dye; for helpful comments, we thank Ross Buchanan, Francisco Cantu, Scott Cook, Maraam Dwidar, EJ Fagan, Jonathan Homola, Bryan Jones, Ryan Kennedy, Heike Kluver, Yphtach Lelkes, Katherine Madel, Anne Rasmussen, Jochen Rehmert, Jeroen Romeijn, Randy Stevenson, Sean Theriault, Dimiter Toshkov, Rens Vliegenthart, Guy Whitten, the special issue editors, and anonymous reviewers.

Disclosure Statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Science Foundation [SES-1728558,SES-1728792].

Data Availability Statement

The data described in this article are openly available in the Open Science Framework at https://doi.org/DOI:10.17605/OSF.IO/TPA6U.

Open Scholarship

This article has earned the Center for Open Science badge for Open Materials. The materials are openly accessible at https://doi.org/DOI:10.17605/OSF.IO/TPA6U.

Notes on contributors

Lindsay Dun is a Ph.D. student in the Department of Government at the University of Texas at Austin. Her research focuses on political communication, voting behavior, and content analysis methods.

Stuart Soroka is the Michael W. Traugott Collegiate Professor of Communication and Media & Political Science, and Research Professor in the Center for Political Studies at the Institute for Social Research, University of Michigan. His research focuses on political communication, the sources and/or structure of public preferences for policy, and the relationships between public policy, public opinion, and mass media.

Christopher Wlezien is the Hogg Professor of Government and Faculty Associate in the Policy Agendas Project at the University of Texas at Austin. His primary, ongoing research develops and tests a "thermostatic" model of public opinion and policy and his other major project assesses the evolution of voter preferences over the course of the election "timeline," both in the US and other countries.

ORCID

Lindsay Dun http://orcid.org/0000-0002-2982-5386
Stuart Soroka http://orcid.org/0000-0001-7524-0859
Christopher Wlezien http://orcid.org/0000-0002-0719-697X

References

Baumgartner, F., & Jones, B. D. (1993). *Agendas and instability in American politics*. University of Chicago Press.

Beland, D., & Schlager, E. (2019). Varieties of policy feedback: Looking backward and moving forward. *Policy Studies Journal*, *47*(2), 184–205. https://doi.org/10.1111/psj.12340

Bélanger, É., & Soroka, S. (2012). Campaigns and the prediction of election outcomes: Can historical and campaign-period prediction models be combined?. *Electoral Studies*, *31*, 702-714. https://doi.org/10.1016/j.electstud.2012.07.003

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). Quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774. https://doi.org/10.21105/joss.00774

Blei, D., Andrew, N., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning and Research*, *3*, 993–1022.

Boydstun, A. (2013). *Making the news*. University of Chicago Press.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Card, D., Boydstun, A. E., Gross, J. H., Resnik, P., & Smith, N. A. (2015, July 26-31). The media frames corpus: Annotations of frames across issues. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)* (pp. 438–444).

Eichenberg, R., & Stoll, R. (2003). Representing defence: Democratic control of the defence budget in the United States and Western Europe. *Journal of Conflict Resolution*, *47*(4), 399–423. https://doi.org/10.1177/0022002703254477

Ellis, C., & Faricy, C. (2011). Social policy and public opinion: How the ideological direction of spending influences public mood. *The Journal of Politics*, *73*(4), 1095–1110. https://doi.org/10.1017/S0022381611000806

Erikson, R. S., MacKuen, M. B., & Stimson, J. A. (2002). *The macro polity*. Cambridge University Press.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis for political texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statistical Science*, *21*(4), 1–14. https://doi.org/10.1214/088342306000000349

Hart, R. P. (2000). *DICTION 5.0: The text analysis program*. Sage-Scolari.

Hopkins, D., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, *54*(1), 229–247. https://doi.org/10.1111/j.1540-5907.2009.00428.x

Jennings, W. (2009). The public thermostat, political responsiveness and error correction: Border control and Asylum in Britain, 1994-2007. *British Journal of Political Science*, *39*(4), 847–870. https://doi.org/10.1017/S000712340900074X

Jurka, T. P., Collingwood, L., Boydstun, A., Grossman, E., & van Atteveldt, W. (2012). *RTextTools: Automatic text classification via supervised learning*. http://cran.r-project.org/web/packages/RTextTools/index.html

McCombs, M. W., & Shaw, D. L. (1972). The agenda-setting function of mass media. *Public Opinion Quarterly*, *36*(2), 176–187. https://doi.org/10.1086/267990

Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, *16*(4), 372–403. https://doi.org/10.1093/pan/mpn018

Montgomery, J. M., & Olivella, S. (2018). Tree-based models for political science data. *American Journal of Political Science*, *62*(3), 729–744. https://doi.org/10.1111/ajps.12361

Muddiman, A., McGregor, S. C., & Stroud, N. J. (2019). (Re)claiming our expertise: Parsing large text corpora with manually validated and organic dictionaries. *Political Communication*, *36*(2), 214–226. https://doi.org/10.1080/10584609.2018.1517843

Neuner, F., Soroka, S., & Wlezien, C. (2019). Mass media as a source of public responsiveness to policy. *International Journal of Press/Politics*, *24*(3), 269–292. https://doi.org/10.1177/1940161219832416

Pacheco, & Julianna. (2013). The Thermostatic Model of Responsiveness in the American States. *State Politics and Policy Quarterly*, *13*(3), 306-332.

Pennebaker, J. W., Francis, M., & Booth, R. (2001). *Linguistic Inquiry and Word Count: LIWC 2001*. Erlbaum.

Roberts, S., Tingley, L., Leder-Luis, G., Albertson, B., & Rand, G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, *58*(4), 1064–1082. https://doi.org/10.1111/ajps.12103

Siroky, D. S. (2009). Navigating random forests and related advances in algorithmic modeling. *Statistics Surveys*, *3*, 147–163. https://doi.org/10.1214/07-SS033

Soroka, S., Stecula, D., & Wlezien, C. (2015). It's (Change in) the (Future) economy, stupid: Economic indicators, the media, and public opinion. *American Journal of Political Science*, *59*(2), 457–474. https://doi.org/10.1111/ajps.12145

Soroka, S. N., & Wlezien, C. (2010). *Degrees of democracy: Politics, public opinion and policy*. Cambridge University Press.

Soroka, S. N., & Wlezien, C. (2019). Tracking the coverage of public policy in mass media. *Policy Studies Journal*, *47*(1), 471–491. https://doi.org/10.1111/psj.12285

Stone, P. J., Dumphy, D. C., & Ogilvie, D. M. (1966). *The general inquirer: A computer approach to content analysis*. MIT Press.

Ura, J. D., & Ellis, C. R. (2012). Partisan moods: Polarization and the dynamics of mass party preferences. *The Journal of Politics*, *74*(1), 277–291. https://doi.org/10.1017/S0022381611001587

Wilkerson, J., & Casas, A. (2017). Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science*, *20*(1), 529–544. https://doi.org/10.1146/annurev-polisci-052615-025542

Wlezien, C. (1995). The public as thermostat: Dynamics of preferences for spending. *American Journal of Political Science*, *39*(4), 981–1000. https://doi.org/10.2307/2111666

Wlezien, C. (1996). Dynamics of representation: The case of US spending on defense. *British Journal of Political Science*, *26*(1), 81–103. https://doi.org/10.1017/S0007123400007420

Wlezien, C., & Soroka, S. (2012). Political institutions and the opinion–policy link. *West European Politics*, *35*(6), 1407–1432. https://doi.org/10.1080/01402382.2012.713752

Wlezien, C., Soroka, S., & Stecula, D. (2017). A cross-national analysis of the causes and consequences of economic news. *Social Science Quarterly*, *98*(3), 1010–1025. https://doi.org/10.1111/ssqu.12445

Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication*, *29*, 205-231.

Appendix

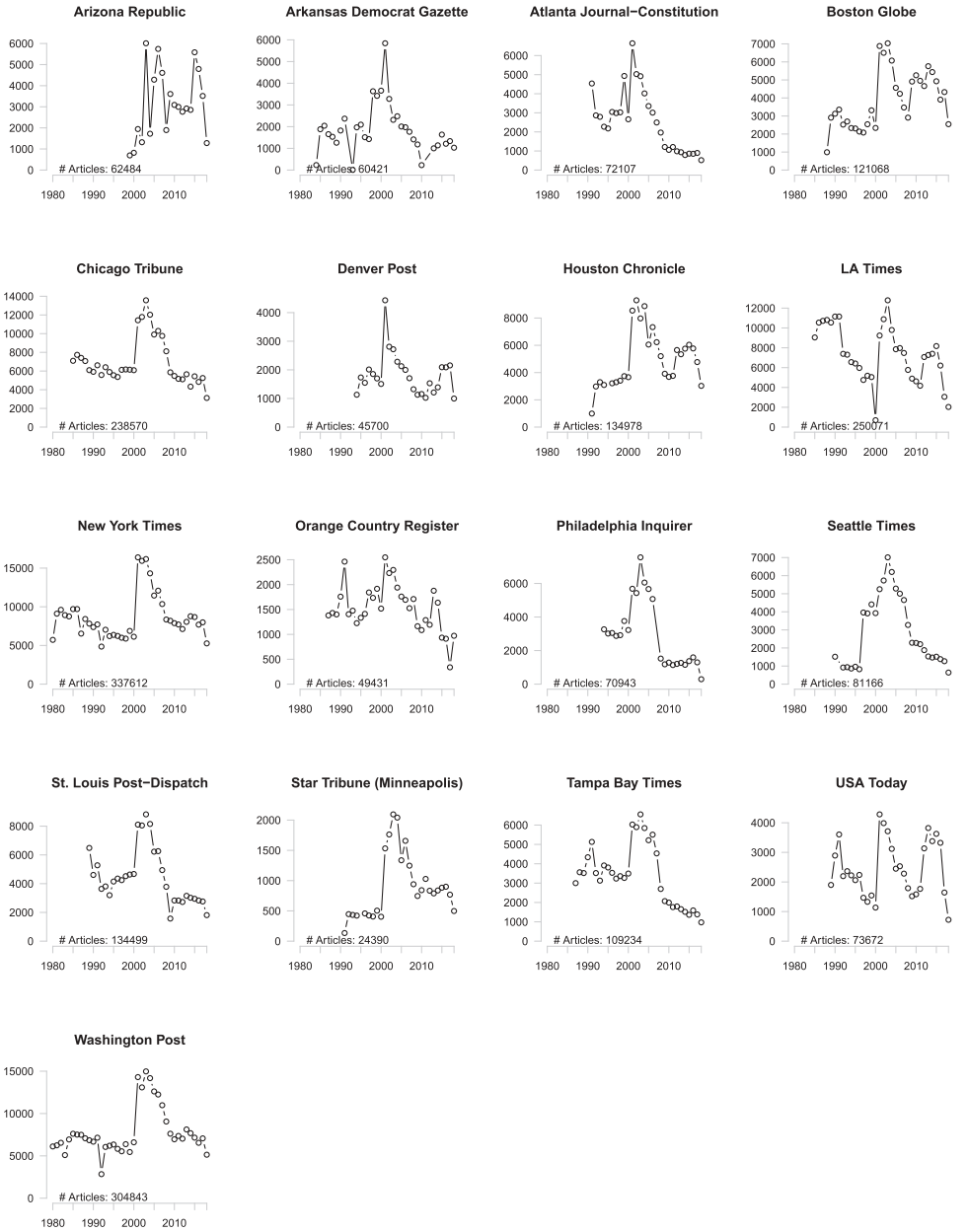

Figure A1. Article Counts by year across newspapers.