The Ghost in the Machine?

Dictionaries, Supervised Learning, and Media Coverage of Public Policy*

Lindsay Dun
University of Texas at Austin
ldun@austin.utexas.edu

Stuart Soroka
University of Michigan
ssoroka@umich.edu

Christopher Wlezien
University of Texas at Austin
wlezien@austin.utexas.edu

Abstract: There are many different approaches to automated content analysis. They come in several main varieties, including (1) one that relies on dictionaries and (2) another that employs machine learning. Many scholars suppose that the machine-based approaches dominate those relying on dictionaries, seemingly predicated on the supposition that computers can do things that people cannot. This is reasonable and also is supported by some research, most of which relies on human-coded test sets. In this paper, we offer a somewhat different argument. Rather than test the effectiveness of either dictionaries or supervised learning at capturing a signal, we argue for and demonstrate the advantages of using the two in combination. We also do so in a research area in which we have an independent objective referent: government spending. We apply hierarchical dictionary counts and supervised learning to measure mass media coverage of change in US defense spending. The dictionary approach does reasonably well at capturing a media "policy signal;" but adding supervised methods appears to increase the accuracy of that signal, albeit to a relatively small degree. Results highlight the value of different methods of automated content analysis and also how the methods can be used in combination.

Keywords: Policy feedback, public responsiveness, thermostatic model, political communication, automated content analysis

* Prepared for presentation at the Texas Methods Conference, Houston, 2019. A previous version was presented at the Policy Agendas Workshop at the University of Texas at Austin. For research assistance, we thank Connor Dye; for helpful comments, we thank Ross Buchanan, Maraam Dwidar, EJ Fagan, Bryan Jones, Heike Kluver, Yphtach Lelkes, Katherine Madel, Anne Rasmussen, Jochen Rehmert, Jeroen Romeijn, Sean Theriault, Dimiter Toshkov, and Rens Vliegenthart. The research is supported by National Science Foundation Grants SES-1728792 and SES-1728558.

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 and LIWC, alongside more specialized approaches such as Rod Hart's Diction software, were the standard approaches to extracting signals from large bodies of text. Over time, however, supervised-based approaches have proliferated. (See Grimmer and Stewart 2013 for a particularly valuable review of the field.)

Many scholars suppose that supervised-learning approaches dominate those relying on dictionaries, seemingly predicated on the supposition that computers are better able to capture the quantities of interest. For instance, because supervised methods are able to take into account the co-existence of 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 illustrated by recent research.

This paper considers a third way, namely, the combination of dictionary-based and supervised-learning approaches. Indeed, supervised-learning applications often use dictionaries, if only implicitly (Hopkins

and King 2010; Monroe, Colaresi, and Quinn 2008; also see Wilkerson and Casas 2017), and we attempt here to demonstrate the advantages of a more explicit *dictionary-plus-supervised-learning* approach. We assess the success of our methodology by comparing content analyses of media coverage of public policy with an independent objective referent – government spending. This is motivated by research demonstrating public responsiveness to policy; the idea is to see whether media coverage is the mechanism of responsiveness, providing information to the public about what policymakers do.[1]

Below, we apply hierarchical dictionary counts and then add random forest 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 and Wlezien 2010). We rely on a corpus of nearly 2.2 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. The coding that is augmented by machine-learning performs slightly better than does the dictionary-only results, in terms of capturing a media "policy signal" that moves alongside spending change. Whether the improvement is enough to justify the addition of machine learning to the dictionary-based analysis is not entirely clear. We nevertheless regard the results as highlighting: (a) the rather striking accuracy of media coverage of defense spending; and (b) the potential value of combining dictionary- and machine-learning-based methods of automated content analysis.

Background

We begin with some substantive background for the analysis below – an account of public responsiveness to policy, and a conceptualization of media coverage of policy. (This explication borrows in part from Soroka and Wlezien N.d.)

Mediating Public Responsiveness to Policy

That policy can feed back on public opinion is well known, and research highlights two general relations. The first is *negative feedback*, where the public adjusts its public preference "inputs" thermostatically based on policy "outputs" (Wlezien 1995; Soroka and Wlezien 2010). Here, the public's preference for more policy – its relative preference, *R* – represents the difference between the public's preferred level of policy (*P**) and the level it actually gets (*P*):

$$R_t = P^*_t - P_t, \quad (1)$$

where *t* represents time. In the model, relative preferences change if either the preferred level of policy *or* policy itself changes. This equation is straightforward in theory, but less so in practice, as we rarely observe *P**. Because of this and the fact that all of the variables have (or would have) different metrics, we need to rewrite the model as follows:

$$R_t = a_0 + \beta_1 O_t + \beta_2 P_t + e_t, \quad (2)$$

where *a* and e_t represent the intercept and the error term respectively, and *O* designates a variety of other determinants of relative preferences, e.g., economic and national security. If people respond thermostatically, β_2 will be less than 0.

The public may not respond thermostatically to policy change, of course, and it even may be that policy feeds back positively on preferences – an increase in spending could lead people to want more

[1] Although the expectation may seem obvious, little research demonstrates that media content contains an adequate number of informative policy cues and much works laments the failures. For a review, see Neuner, et al (N.d.).

spending in that domain. This *positive feedback* thus is between P and P^* in equation 1 above, where the public's preferred levels of policy are conceived to be a function of policy itself:

$$P^*_t = f\{P_t\}. \quad (3)$$

To be absolutely clear, there would be positive feedback if this relation is positive.[2] Most of the research on positive feedback conceives of individuals as responsive to their own personal consumption of policy, rather than the collective policy per se (e.g., Soss and Schram 2007). This is of relevance to our representation of positive feedback, as we might observe that individuals' P^*_{it} would reflect the micro-level P_{it} instead of the macro P_t. Of course, individuals could respond to both the micro and macro levels, as follows:

$$P^*_{it} = f\{P_{it}, P_t\}. \quad (4)$$

For example, people may observe an increase in policy that works and favor more.

Much like equation 1 above, equation 4 is more straightforward in theory than in practice, as we often do not directly observe P^*_{it} and cannot fully account for it using proxies. This means that we sometimes only can detect positive feedback indirectly, from the coefficient β_2 in equation 2. The coefficient thus would encompass both negative and positive feedback, and actually capture their net effect, i.e., if $\beta_2 < 0$, we cannot conclude that there is no positive feedback, just that negative feedback exceeds any positive effect. By implication, the coefficient provides a minimal estimate of each – positive and negative – type of effect.

Regardless of our expectations or results, the mechanism of feedback, thermostatic or otherwise, is of critical importance – we are interested in how the public receives information about policy outputs. Our research is motivated by an interest in determining whether mass media serve this function, focusing in this paper on measurement.

Conceptualizing Media Coverage of Public Policy

Previous media research relating to policy has focused almost exclusively on *inputs* into the policymaking process, not policy *outputs* themselves. To be clear: measures of media coverage have not been concentrated on what policymakers actually do, but on media priorities (and frames) and their impact on policy activity (McCombs and Shaw 1972; Baumgartner and Jones 1993; Boydstun 2014; Card, et al 2015). Given our interest in public responsiveness, we want to know whether and how mass media content reflects what is happening to policy.

We might suppose that the news media reflect the level of policy that has been adopted. The media signal (M) at a point in time t then would be a function of the size of the policy (P), as in the following very basic equation:

$$M_t = g\{P_t\}. \quad (5)$$

If we could conceive of the two variables being measured on the same scale, we might even posit that the relationship between them would be an identity function. Here media coverage would perfectly capture policy, perhaps with some random error.[3] While this conceptualization may be tempting,

---

[2] Just as the coefficient of feedback in equation 2 need not be negative, the relationship in equation 3 need not be positive, and a negative effect is possible if policy increases cause people to want less than they did at the previous point in time. This is what we might predict if policy does not improve conditions or makes things worse. Also see Soroka and Wlezien (2010: 115-119).

[3] In terms of the equation relating them, the coefficient would be 1.0 and the intercept 0.0.

there is reason to expect that the media report on policy change. This is what we observed for the economy, where coverage reflects its direction, not its level (Soroka et al 2015; Wlezien, et al 2017).

We think we can more directly measure media coverage of change ($^{\Delta}M$). That is, we can envision and implement an analysis that identifies media content characterizing the direction in which policy is moving (Soroka and Wlezien N.d.). Even relatively simple dictionaries allow us to do so, as we will see shortly. It is less clear how we would measure coverage of levels of policy, particularly spending, which happens to be the focus of our empirical analysis. Even accepting that the media reports on policy levels, an appropriate content analytic dictionary seems elusive, or at least less effective.[4]

In sum, it seems likely that media coverage focuses on policy change and that it is possible to design a content analysis that could reliably extract this "signal." Armed with a measure of the media policy signal, we could assess whether it follows actual policy change over time. That is, we could estimate the following:

$$^{\Delta}M_t = h\{\Delta P_t\}. \qquad (6)$$

For this exploratory analysis, we concentrate on a single domain in a single country, spending on defense in the US.

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 newspapers 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).[5] 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, but 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

[4] That said, we can construct a proxy for coverage of levels based on the cumulation of coverage of spending change (see Neuner, et al N.d.).

[5] Note that full-text search terms are searched as phrases, i.e., "national security," not "national" and "security" separately.

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 newspapers start in 1980 – most enter the dataset in the early 1990s. All are gathered up to the end of the 2018 fiscal year.

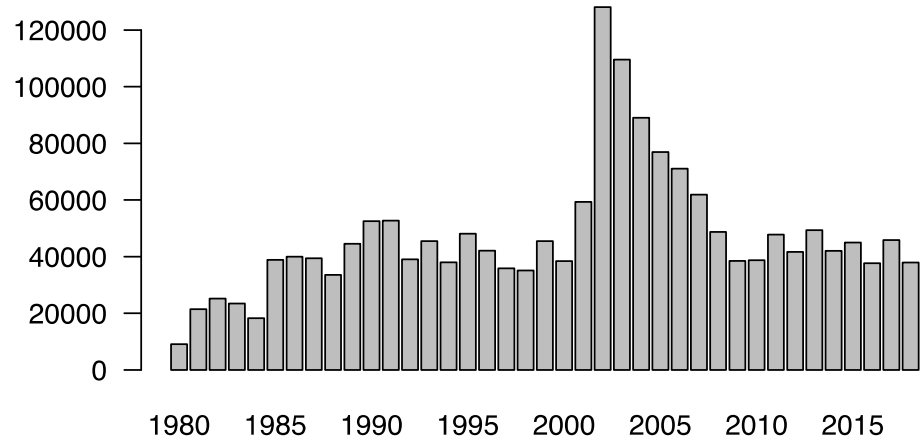
Our selection of newspapers is based on availability, alongside circulation, with some consideration given to regional coverage. In the end, we have 17 of the highest-circulation newspapers in the US, three of which aim for national audiences, and seven of which cover considerably large regions in the 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 1, 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 of this text, then, but rather all *sentences* within this corpus that focus on spending. 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 Ken Benoit's *quanteda* package, in R. Note that the dictionary has been subjected to testing in Soroka and Wlezien (N.d.), and was constructed from our own reading of keyword-in-context (kwic) retrievals, augmented by thesaurus searches. Our dictionary-building procedure, in a nutshell, is 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 of our dictionary words, 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 1,841,780 sentences. Figure 1 plots the number of spending sentences in our database by fiscal year.

Figure 1. Defense Spending Sentences, by Fiscal Year



This is the raw material for the analyses that follow.

Measuring the "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 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 already 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, 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 596,186 "spending" sentences that also include "direction" keywords.

Finally, 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:

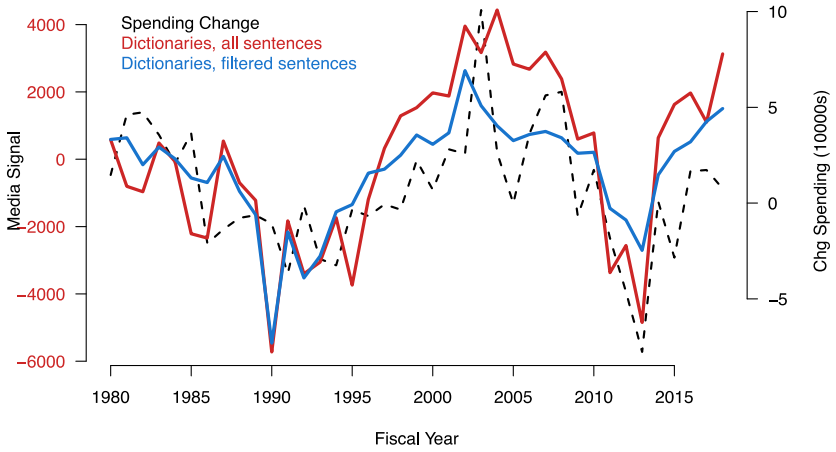
> DEFENSE: army, navy, naval, air force, marines, defense, military, soldier, war, cia, homeland, weapon, terror, security, pentagon, submarine, warship, battleship, destroyer, airplane, aircraft, helicopter, bomb, missile, plane, service men, base, corps, iraq, afghanistan, nato, naval, cruiser, intelligence

Doing so isolates 436,129 "spending" sentences that contain "direction" keywords and words clearly related to defense. We may still have irrelevant sentences in our database, but layering dictionaries on top of each other will, we think, produce an increasingly reliable measure. Indeed, our own recent work suggests that this is the case (Soroka and Wlezien, N.d.).

Note that our approach here is identical to the use of hierarchical dictionary counts as implemented in Young and Soroka (2012) and Belanger and 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; see Grimmer and Stewart 2013 for an especially helpful review.) There sometimes is a perception that dictionaries are simple word lists, concocted based only on a thesaurus, where individual words are not subjected to testing, and where results are thus likely to either capture, or miss, 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.

Our previous research details how the application of successive dictionaries works (Soroka and Wlezien N.d.); another paper further highlights the strong connection between this dictionary-based and human coding (Neuner et al. N.d.). Both prior papers 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;" 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.

Figure 2. The Dictionary Media Signal and Spending Change



How closely does our dictionary-based measure of the media policy signal track actual government spending on defense? Figure 2 displays the aggregate signal across all 17 newspapers alongside changes in defense appropriations (budget authority), in FY2000 US dollars, drawn from the *Historical Tables* distributed by the OMB. The correlation between spending change and the signal based on all kwic retrievals with UP/DOWN keywords but without a DEFENSE keyword is 0.59; the correlation between spending change and a signal based on kwic retrievals that also include a DEFENSE keyword is an only slightly larger 0.61.[6]

Dictionary-plus-Supervised-Learning

The dictionary-plus-supervised-learning approach relies on the same body of sentence-level data as the purely dictionary-based approach. In this instance, however, we base the supervised learning on a subset of human-coded data. We proceeded with human coding in several stages. In the first two stages, 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 are used in part to test the relevance of our dictionary keywords (see Soroka and Wlezien N.d.), so a third round then draws 40 sentences from seven of the more 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 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 input 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 is another point at which the supervised-learning is aided by a dictionary-based analysis.) We then follow this approach for two additional samples of coding.

In sum, we draw a sample of 5,623 sentences from the roughly 1.8 million sentences that include spending cues, with a particular focus on sentences that also include a defense keyword. We then use 80% (4,498) of the sample to train the computer and the other 20% (1,125) to test the algorithm. The training relies on human codes from Amazon Mechanical Turk (MTurk). We collect 5 codes for each sentence; and we ask for coding on both the relevance and direction of each sentence: (1) whether each sentence is about spending, and then, if so, (2) the direction of spending. We collapse the resulting codes into 4 categories: 0 = not about defense spending; 1 = about defense spending but not change; 2 = about a spending decrease; 3 = about a spending increase. To code sentences, we require 60% of the 5 coders to agree a sentence is relevant and at least 60% of MTurkers to agree on a direction code. In exploratory analyses, we found that these choices optimized algorithm performance and produced the highest average class accuracy in our test set.

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

[6] Media coverage in a particular year also may reflect spending decisions from the previous year or else effectively anticipate spending decisions in subsequent years. It thus is worth noting that the correlation between the media signal in year t is 0.45 with spending in fiscal year t-1 and 0.64 with spending in fiscal year t+1. Also see Neuner, et al (N.d.).

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 application of the UP and DOWN (direction) and DEFENSE (relevance) dictionaries.

To use the data in the supervised-learning model, we first transform our entire corpus into a document frequency matrix in R. We retain unigrams and bigrams in the document frequency matrix and remove all words/word pairs that occur less than two times in the training dataset. Doing so retains 20,480 features to use in prediction; this is known as a "bag of words" feature set. We subset that document frequency matrix into the training and test sets with human codes and then the uncoded sentences we will use in prediction. We then transform our training subset, test subset, and remaining uncoded sentences into standard matrix format in R, which discards all document variables (but keeps each sentence's unique id number). The random forest model needs a factor variable for prediction and testing, so we turn the human coded values associated with each sentence in the training and testing matrices into a factor object in R. In the end, we have five components:

a. One matrix of sentence words for training;
b. One factor object of training values associated with each training sentence;
c. One matrix of sentence words for testing;
d. One factor object of training values associated with each test sentence; and
e. One matrix of sentence words in all uncoded sentences for predictions.

All random forest analyses are carried out using the *randomForest* and *rfUtilities* packages in R. We ran a tuning function to determine the optimal number of "tree" splits in the random forest algorithm—the tuning function suggested 210 trees. Using the information in our training data matrix and categories, we instruct the random forest model in R 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.

After the model is trained, we run a cross-validation function to determine how well our model generalizes to data the model has not seen, and to make sure we do not overfit our model to specific training data. Using the trained model to predict codes in our test set of 1,125 sentences suggests a proportional match of .77, which is slightly higher than what we find (.65) using coding from the hierarchical dictionaries alone. (To make the comparison, we collapse the 4-category supervised learning codes into a 3-category variable to better comport with the way in which the results from dictionary-based methods are stored; specifically, we recode 0 and 1 values to 0, 2 to -1, and 3 to 1.) The two approaches produce identical codes for 70% of the sentences. We now have a trained algorithm that can be used to predict uncoded sentences.

To generate supervised learning predictions, we code the entire 1.8 million sentences (excluding those 5,623 sentences used in training and testing) using the trained model. (We do so in batches, since the matrix for all news stories is too large to transform.) We then sum the codes, transformed into three categories as described above, by fiscal year. Figure 4 plots the resulting supervised learning media signal alongside the one produced using dictionaries, both of which are unstandardized. The correspondence between the two media signals is striking, and the Pearson's correlation (.94) confirms a very tight relationship. How does each measure compare with spending change? The correlation with spending change is slightly larger for the supervised learning signal (.67) than for the dictionary signal (.59), but the difference is not highly reliable (p=.099). In sum, the layering of the supervised

learning on top of the dictionaries yields a relatively small gain in the accuracy of our media policy signal.

Figure 3. The Two Media Signals Compared

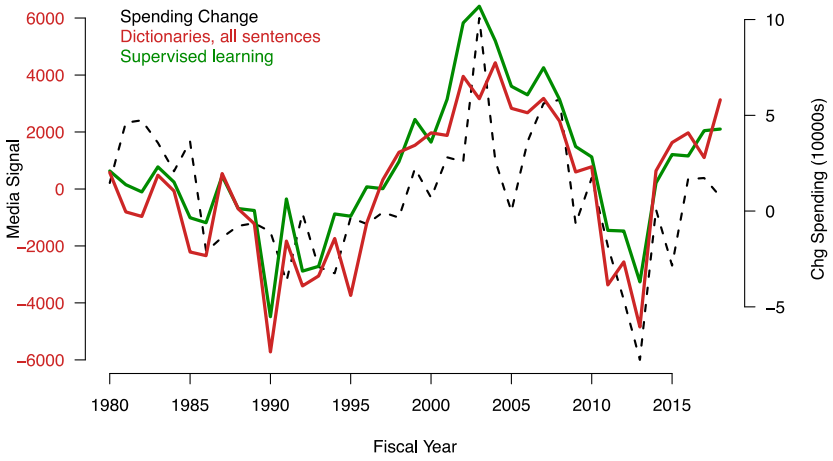


Table 1.

	Spending Change	RFC Mturk	Dictionary
Spending Change	1	0.67	0.59
RFC MTurk	0.67	1	0.94
Dictionary	0.59	0.94	1

Table 2.

	Proportion
Dictionary = Human code	0.65
Machine Learning = Human Code	0.77
Machine Learning = Dictionary	0.70

Is the 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, it seems very likely that we would arrive at the same conclusion with both measures; either way, we find a strong correspondence between media coverage and defense spending change. That said, there may be other environments in which there is a clearer difference between the dictionary- and dictionary-plus-supervised learning approaches. And the slight increase in accuracy seen here does signal the potential for supervised-learning to find (meaningful) variation that is missed by dictionaries alone. We discuss this further below.

Conclusion and Discussion

A carefully constructed measure of media coverage of policy *outputs* is of real significance for those interested in policy responsiveness and representation. As noted above, most 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. N.d.). A means of identifying the media policy signal offers not just a measure of media accuracy, it speaks to the potential for representative democracy, policy domain by policy domain, across time and space.

That said, 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 it's "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, and we tested the "improvement" by comparing each signal to actual spending change. Our analysis reveals that the dictionary-only approach does about as well as the dictionary-plus-supervised-learning methods. We are not staking a claim on the success of dictionary-based approaches in all instances. 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 modelling (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.

We also have not explored the applicability of these approaches to domains other than defense. Our long-term objective is to do exactly this. The hope is that the dictionaries will require only minimal revision for other domains; at least, that has been our objective here. This is a testable proposition, left for future work. So too is the possibility of developing a rather different set of dictionaries for use in policy domains that are less characterized by spending. Defense may be an easy test for our measure; environment would be much tougher. Even so, it may be possible to develop a set of dictionaries that capture change in environmental regulation, and this would be useful in understanding responsiveness and representation in that domain. And there are supervised learning alternatives.

The methods to automated content analysis need not compete, however, and can be used in combination. 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 are subjects for future research. For the time

being, we have seen that we can capture the coverage of actual policy change in the media using both traditional methods and modern data-intensive ones.

References

Althaus, Scott. 2003. *Collective Preferences in Democratic Politics*. Cambridge: Cambridge University Press.

Altheide, David L. 1997. "The News Media, the Problem Frame, and the Production of Fear." *Sociological Quarterly* 38(4): 647–68.

Barabas, Jason. 2009. "Not the Next IRA: How Health Savings Accounts Shape Public Opinion. *Journal of Health Policy, Politics and Law* 34:181-217.

Barabas, Jason and Jennifer Jerit. 2009. "Estimating the Causal Effects of Media Coverage on Policy Specific Knowledge." *American Journal of Political Science* 53(1): 73-89.

Bartle, John, Sebastian Dellepiane-Avellaneda, and James Stimson. 2011. "The Moving Centre: Preferences for Government Activity in Britain, 1950–2005." *British Journal of Political Science* 41(2): 259-285.

Baumgartner, Frank, and Bryan D. Jones. 1993. *Agendas and Instability in American Politics*. Chicago: University of Chicago Press.

Beland, Daniel. N.d. "Policy Feedback and the Politics of the Affordable Care Act." *Policy Studies Journal*, forthcoming.

———. 2010. "Reconsidering Policy Feedbacks: How Policies Affect Politics." *Administration and Society* 42(2): 568-590.

Bennett, Stephen. 1988. "'Know-Nothings' Revisited: The Meaning of Political Ignorance Today. *Social Science Quarterly* 69:476-490.

Berelson, Bernard R., Paul F. Lazarsfeld, and William N. McPhee. 1954. *Voting*. Chicago: University of Chicago Press.

Blei, David, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning and Research* 3:993–1022.

Boydstun, Amber. 2013. *Making the News*. Chicago: University of Chicago Press.

Bucchi, Massimiano, and Renato G. Mazzolini. 2003. "Big Science, Little News: Science Coverage in the Italian Daily Press, 1946-1997." *Public Understanding of Science* 12(1): 7-24.

Campbell, Andrea. 2012. "Policy Makes Mass Politics." *Annual Review of Political Science* 15: 333-351.

———. 2003. *How Politics Makes Citizens*. Princeton: Princeton University Press.

Card, Dallas, Amber E Boydstun, Justin H Gross, Philip Resnik, Noah A Smith. 2015. "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), pages 438–444, Beijing, China, July 26-31, 2015.

Converse, Philip E. 1970. "Attitudes and Non-Attitudes: Continuation of a Dialogue. In Edward R. Tufte (ed.), *The Quantitative Analysis of Social Problems*. Reading, Mass: Addison-Wesley.

-----. 1964. "The Nature of Belief Systems in Mass Publics." In David Apter (ed.), *Ideology and Discontent*. New York: Free Press.

Daku, Mark, Stuart Soroka and Lori Young. 2015. *Lexicoder*. Software available at lexicoder.com.

Davie, William R., and Jung Sook Lee. 1995. "Sex, Violence, and Consonance/Differentiation: An Analysis of Local TV News Values." *Journalism and Mass Communication Quarterly* 72(1): 128–38.

Delli Carpini, Michael and Scott Keeter. 1996. *What American Know about Politics and Why it Matters*. New Haven: Yale University Press.

Deutsch, K.W., 1966. *The Nerves of Government: Models of Political Communication and Control*. New York: Free Press.

Dunaway, Johanna. 2011. "Institutional Influences on the Quality of Campaign News Coverage." *Journalism Studies* 12(1):27-44.

Durr, Robert H. 1993. "What Moves Policy Sentiment?" *American Political Science Review* 87: 158-170.

Easton, David. 1965. *A Framework for Political Analysis*. Englewood Cliffs NJ: Prentice-Hall.

Eichenberg, Richard, and Richard Stoll. 2003. "Representing Defence: Democratic Control of the Defence Budget in the United States and Western Europe." *Journal of Conflict Resolution* 47: 399-423.
Ellis, Christopher, and Christopher Faricy. 2011. "Social Policy and Public Opinion: How the Ideological Direction of Spending Influences Public Mood." *The Journal of Politics* 73 (04): 1095–1110.
Erikson, Robert S., Michael B. MacKuen and James A. Stimson. 2002. *The Macro Polity*. Cambridge: Cambridge University Press.
Fording, Richard. N.d. "Medicaid Expansion and the Political Fate of Governors who Support it." *Policy Studies Journal*, forthcoming.
Friedman, Sharon H., Sharon Dunwoody and Carol L. Rogers, eds. 1999. *Communicating Uncertainty: Media Coverage of New and Controversial Science*. New York: Routledge.
Grimmer, Justin and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis for Political Texts." *Political Analysis* 21(3): 267-297.
Hakhverdian, A. 2012. "The Causal Flow between Public Opinion and Policy: Government Responsiveness, Leadership, or Counter Movement?" *West European Politics*, *35*(6): 1386-1406.
Hopkins, Daniel and Gary King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science* 54(1):229-247.
Iyengar, Shanto. 1991. *Is Anyone Responsible? How Television Frames Political Issues*. Chicago: University of Chicago Press.
Jennings, Will. 2009. "The Public Thermostat, Political Responsiveness and Error Correction: Border Control and Asylum in Britain, 1994-2007." *British Journal of Political Science* 39:847-870.
Jennings, Will and Christopher Wlezien. 2015. "Preferences, Problems, and Representation." *Political Science Research and Methods* 3(3): 659-681.
Jurka, Timothy P., Loren Collingwood, Amber Boydstun, Emiliano Grossman, and Wouter van Atteveldt. 2012. RTextTools: Automatic text classification via supervised learning. http://cran.r-project.org/web/packages/RTextTools/index.html.
Lawrence, Regina G. 2000. "Game-Framing the Issues: Tracking the Strategy Frame in Public Policy News." *Political Communication* 17(2): 93-114.
McCombs, Maxwell W., and Donald L. Shaw. 1972. "The Agenda-Setting Function of Mass Media." *Public Opinion Quarterly* 36(2):176-187.
Mettler, Suzanne. 2005. *Soldiers to Citizens: Thee G.I. Bill and the Making of the Greatest Generation*. New York: Oxford University Press.
Mettler, Suzanne, and Joe Soss. 2004. "The Consequences of Public Policy for Democratic Citizenship: Bridging Policy Studies and Mass Politics." *Perspectives on Politics* 2(1):55-73.
Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn. 2008. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." *Political Analysis* 16 (4):372-403.
Morgan, Stephen L. and Minhyoung Kang. 2015. "A New Conservative Cold Front? Democrat and Republican Responsiveness to the Passage of the Affordable Care Act. *Sociological Science* 2:502-526.
Neuner, Fabian, Stuart Soroka and Christopher Wlezien. N.d. "Mass Media as a Source of Public Responsiveness to Policy." *International Journal of Press/Politics*, forthcoming.
Pacheco, Julianna. 2013. "Attitudinal Policy Feedback and Public Opinion." *Public Opinion Quarterly* 77:714-734.
Page, Benjamin I. and Robert Y. Shapiro. 1992. *The Rational Public: Fifty Years of Trends in Americans' Policy Preferences*. Chicago: University of Chicago Press.
Popkin, Samuel and Michael Dimock. 1999. Political Knowledge and Citizen Competence. In Stephen Elkin and Karol Salton (eds), *Citizen Competence and Democratic Institutions*. University Park: Pennsylvania State University.

Roberts, Stewart, Tingley, Lucas, Leder-Luis, Gadarian, Albertson, and Rand. 2014. "Structural topic models for open-ended survey responses." *American Journal of Political Science* 58(4): 1064-1082.

Soroka, Stuart. 2014. "Reliability and Validity in Automated Content Analysis," in *Communication and Language Analysis in the Corporate World*, Roderick P. Hart, ed., Hershey PA: CGI Global.

Soroka, Stuart, Dominic Stecula, and Christopher Wlezien. 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.

Soroka, Stuart N. and Christopher Wlezien. N.d. "Tracking the Coverage of Public Policy in Mass Media." *Policy Studies Journal*, forthcoming.

Soroka, Stuart N. and Christopher Wlezien. 2010. *Degrees of Democracy: Politics, Public Opinion and Policy*. Cambridge University Press.

Soss, Joe and Sanford Schram. 2007. "A Public Transformed? Welfare Reform as Policy Feedback." *American Political Science Review* 101:111-127.

Stimson, James. 1999. *Public Opinion in American: Moods Cycles and Swings,* 2nd ed. Boulder CO: Westview Press.Ura, Joseph. 2014. "Backlash and Legitimation: Macro Political Responses to Supreme Court Decisions." *American Journal of Political Science* 58:1100-126.

Ura, Joseph Daniel, and Christopher R Ellis. 2012. "Partisan Moods: Polarization and the Dynamics of Mass Party Preferences." *The Journal of Politics* 74 (01): 277–91.

Weaver, Vesla and Amy Lerman. 2010. "The Political Consequences of the Carceral State." *American Political Science Review* 104:817-833.

Wilkerson, John and Andreu Casas. 2017. "Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges." *Annual Review of Political Science* 20:529-544.

Wlezien, Christopher. 2004. "Patterns of Representation: Dynamics of Public Preferences and Policy." *Journal of Politics* 66:1-24.

———. 1996. "Dynamics of Representation: The Case of US Spending on Defense." *British Journal of Political Science* 26:81-103.

———. 1995. "The Public as Thermostat: Dynamics of Preferences for Spending." *American Journal of Political Science* 39:981-1000.

Wlezien, Christopher and Stuart Soroka. 2012. "Political Institutions and the Opinion-Policy Link." *West European Politics* 35(6): 1407-1432.

Wlezien, Christopher, Stuart Soroka and Dominik Stecula. 2017. "A Cross-National Analysis of the Causes and Consequences of Economic News." *Social Science Quarterly* 98(3): 1010-1025.

Appendix Figure 1.