A BETTER WORDSCORES: SCALING TEXT WITH THE CLASS AFFINITY MODEL

BY PATRICK O. PERRY[†] AND KENNETH BENOIT[*,‡]

Oscar Health[†] and London School of Economics and Political Science[‡]

Abstract

Probabilistic methods for classifying text form a rich tradition in machine learning and natural language processing. For many important problems, however, class prediction is uninteresting because the class is known, and instead the focus shifts to estimating latent quantities related to the text, such as affect or ideology. We focus on one such problem of interest, estimating the ideological positions of 55 Irish legislators in the 1991 *Dáil* confidence vote, a challenge brought by opposition party leaders against the then-governing *Fianna Fáil* party in response to corruption scandals. In this application, we clearly observe support or opposition from the known positions of party leaders, but have only information from speeches from which to estimate the relative degree of support from other legislators. To solve this scaling problem and others like it, we develop a text modeling framework that allows actors to take latent positions on a "gray" spectrum between "black" and "white" polar opposites. We are able to validate results from this model by measuring the influences exhibited by individual words, and we are able to quantify the uncertainty in the scaling estimates by using a sentence-level block bootstrap. Applying our method to the *Dáil* debate, we are able to scale the legislators between extreme pro-government and pro-opposition in a way that reveals nuances in their speeches not captured by their votes or party affiliations.

[*]Prepared for presentation at the 2022 COMPTEXT Workshop, Dublin, 6–7 May. This is a work in progress and while we would love you to cite it, please wait a few months for a more definitive (and hopefully published) version. This research was supported by the European Research Council grant ERC-2011-StG 283794-QUANTESS. We thank Jouni Kuha for comments.

1. Introduction. Text classification, where the goal is to infer a discrete class label from observed text, is a core activity in statistical and machine learning and natural language processing. Instances of this problem include inferring authorship (Mosteller and Wallace, 1963) or genre (Kessler et al., 1997), detecting deception (Newman, Pennebaker and Berry, 2003), classifying email as "spam" (Heckerman et al., 1998), or categorizing sentiment (Pang, Lee and Vaithyanathan, 2002). The huge appeal of the methods developed for these applications is that, from a small training set, it is possible to classify a large number of unlabelled documents to reasonable accuracy without costly human intervention.

In many applications, however, classification is an uninteresting goal, since the correct identification of the class is obvious and costless. It is fundamentally uninteresting, for example, to attempt to predict the political party of a speaker or the identity of a Supreme Court justice. Furthermore, in many social and political settings with observed discrete outcomes, institutions may cause predicted and observed class membership to diverge in significant ways. In parliamentary democracies where party discipline is enforced, for instance, voting may follow party lines even if the best predictions from observable features indicate more heterogeneous outcomes. In such cases, it is trivial to predict class (a legislator's vote) from observable covariates (political party). In the presence of these covariates, the text of a speech is ancillary to the goal of class label prediction.

Even when observing text does not improve prediction performance, it is not the case that text is uninformative. In legislative debates, the text that legislators generate through floor speeches may provide a direct opportunity for them to express their contrary and divergent preferences (see for instance Benoit and Herzog, 2012). With legal briefs, to take another example, it is trivial to classify opinions as majority or dissenting but using the observed text and other information it is possible to place the briefs on a spectrum between the two extremes (Clark and Lauderdale, 2010). Simply attempting to predict the category of opinion—for instance classifying *amicus curiae* briefs as pro-petitioner or pro-respondent (e.g. Evans et al., 2007), is of less direct interest since these categories are already known. The text of a document can reveal nuances that are not captured by and sometimes in disagreement with its class label.

Here, we focus on an application that is ill-suited to text classification but where text is nonetheless informative. We analyze the 1991 Irish *Dáil* confidence debate, previously studied by Laver and Benoit (2002) who used the debate speeches to demonstrate their "Wordscores" scaling method. The context is that in 1991, as the country was coming out of a recession, a series of corruption scandals surfaced involving improper property deals made between the government and certain private

TABLE 1
Irish Dáil debate speech statistics.

Government party members		*Opposition party members*	
Fianna Fáil (FF)	24	Democratic Left (DL)	3
Progressive Dems. (PD)	1	Fine Gael (FG)	22
		Green	1
		Labour (Lab)	7
Speech text			
Median length (leaders)	6,348 tokens		
Median length (others)	2,210 tokens		
Vocabulary size	9,731 word types		

companies. The public backlash precipitated a confidence vote in the government, on which the legislators (each called a *Teachta Dála*, or TD) debated and then voted to decide whether the current government would remain or be forced constitutionally to resign. Table 1 summarizes the composition of the Dáil in 1991 and provides some descriptive statistics about the speech texts, including the number of total words ("tokens") and unique words ("types"). We can use the debate as a chance to learn the legislators' ideological positions.

Because the Irish parliamentary context is characterized by strict party discipline, the move was largely symbolic and each legislator voted strictly with his or her party: all members of the governing parties (*Fianna Fáil* and the Progressive Democrats) voted to support the government, and all members of the opposition parties (the Democratic Left, *Fine Gael*, Green, and Labour) voted against. If we wanted to estimate different degrees of support, for instance to identify reluctant supporters of a vote, then we would need more than the uninformative voting that occurs entirely on party lines. In political science, this has long been a core challenge in testing theories of intra-party politics, because in parliamentary systems with strong party discipline legislators may "vote with their party possibly not because of their policy preferences, but rather in spite of them" (Schwarz, Traber and Benoit, 2017, 379). What they say, however, is typically not subject to party discipline and provides far more sincere information about their relative preferences.

Take, for example, the following excerpt from Noel Davern, a moderate from the *Fianna Fáil* party:

> It is not that the financial scandals have not occurred. They have occurred and the Government have taken quick action on them. In fact, we are not fully qualified to speak on them until we see the results of the full and independent inquiry.

Davern supports the government, but at the same time does not excuse them from all culpability. Contrast this with a typical opposition speech, calling for a vote against the confidence motion, from Labour TD Michael Ferris:

> Our decision to oppose this motion of confidence is a positive assertion of the disapproval of the ordinary people of the actions of this discredited Government. The people have watched with amazement the unfolding of scandals which

have tainted this Government. The Government cannot now be said to deserve the confidence of the people.

Both legislators express views that place them somewhere between the two extremes of absolute government support and absolute opposition support.

Where do Davern, Ferris, and the other 56 TDs that participated in the debate lie on this ideological spectrum? This is the essential question that we attack in this manuscript. In answering the question, we have at our disposal the speech texts, along with some additional information. We know that the leader of the government (Haughey, the *Fianna Fáil Taoiseach*) will give a speech at one extreme of the pro-government spectrum, and we know that the heads of the two major opposition parties (Spring and De Rossa, the Labour the Democratic Left leaders) will be at the extreme of the other end. We will use these three texts as reference points by which to scale the other 55 ambiguous texts whose positions are unknown and must be estimated.[1]

To solve our particular problem, we develop a new text scaling method that is broadly applicable to situations where most documents are unlabelled but we have a few examples of documents at the extremes of a hypothesized ideological or stylistic spectrum. Instead of predicting class membership, our objective in such problems is to *scale a continuous characteristic*, through measuring the fit of a text to a set of known classes based on its degree of similarity to typical texts from these classes. The novelty of this approach is that it can scale an unlimited number of texts whose latent positions – what we term *class affinities* – are unknown, from a small pair of archetypical, extreme reference texts. Scaled positions are always relative to these anchors, making the resulting estimates directly interpretable, unlike unsupervised scaling methods. While ours is not the first method to implement this form of supervised scaling, ours is the first to provide an explicit statistical foundation for scaling (unlike for instance Laver, Benoit and Garry (2003)) or to focus directly on estimating a latent attribute rather than adapting a machine learning technique designed for predicting a discrete class.

In what follows, we develop the *class affinity model* and demonstrate its use in scaling the degree of support or opposition expressed in the speeches made during the confidence debate. We start by outlining the foundations of our scaling model, contrasting it first to similar approaches designed for classification (Section 2), and then to lexicographical association methods in the form of sentiment dictionaries (Section 3). Section 4 then sets out the model, comparing this to related methods, high-

[1] In this example and generally for the applications to which our model applies, texts known to be extreme must be identified as reference texts. This is the same problem as in any machine learning application requiring the appropriate choice of training data, but in our class of problems is generally known *ex ante*, since the objective is to leverage a small number of known extreme positions to estimate a much larger number of unknown positions, relative to these extremes. See Laver, Benoit and Garry (2003) for a discussion of this problem.

lighting the differences through on statistical principles but also using our application. Sections 5.1 and 5.2 detail how this model and its reference distributions are estimated, while Section 9 relates the affinity model to related methods. In Section 6, we show how to measure the influence of individual words, provide recommendations for removing common terms that might skew the resultsm and describe how we applied this procedure to choose a tailored vocabulary for our application. Section 7 demonstrates how to estimate uncertainty for the class affinity scaled estimates. Finally, we summarize the results the results of fitting the class affinity model to our application (Section 8), and offer some concluding remarks.

2. Scaling with a classification method. We have stated repeatedly that classification is not our objective in this problem, but nonetheless there is a long tradition of fitting classification methods to text, and we might try applying one of those methods here. We have a "training set" of the three leadership speeches, one of which we can label as *Government* and two as *Opposition*. We can fit a supervised classification method to this training set and then use it to make predictions for the other 55 legislators.

Using the Naive Bayes text classification method popularized by Sahami et al. (1998), we would model the tokens in each speech text as independent draws from a label-dependent distribution estimated from the reference texts. Letting label $k = 1$ denote *Government* and label $k = 2$ denote *Opposition*, for each label $k \in \{1,2\}$ and word type v in our vocabulary $\mathcal{V}$, we would estimate p_{kv}, the probability that a random token drawn from a text with label k is equal to v. Typically we use the empirical word occurrence frequencies in the reference documents or some smoothed version thereof. Here and throughout the text, unless otherwise noted we will take our vocabulary to be the set of word types that appear at least twice in the leadership speeches, excluding common function words from the modified Snowball stop word list distributed with the `quanteda` software package (Porter, 2006; Benoit et al., 2018); we ignore words outside this set.

Under the "naive" assumption that tokens in a text are independent draws from the same distribution, assuming equal prior odds for each label, the log-odds that the label is *Government* given the word counts $x = (x_v)_{v \in \mathcal{V}}$ is

$$\eta(x) = \sum_{v \in \mathcal{V}} x_v \log(p_{1v}/p_{2v}),$$

where x_v denotes the number of times that word type v appears in the text, and $\mathcal{V}$ represents the total set of word types. The expression for $\eta(x)$ arises as the log ratio of two multinomial likelihoods with probability vectors p_1 and p_2. Using Naive Bayes classification for this two-class prediction problem, we would predict the label as *Government* when $\eta(x) > 0$, and we would predict the label

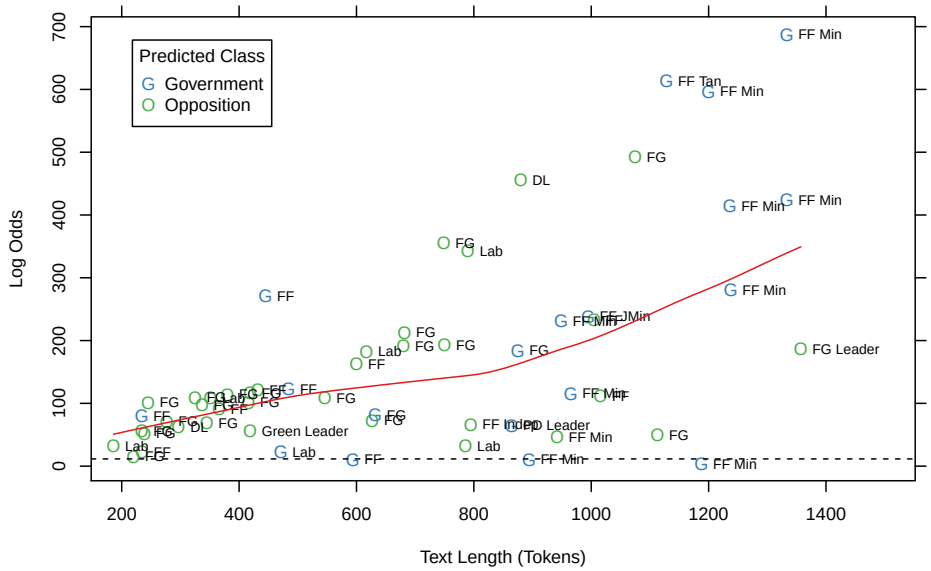


Fig 1: Odds of class membership for the debate speeches as predicted by a Naive Bayes model. Points above the dashed lines have predicted class probabilities exceeding 99.999%.

as *Opposition* when $\eta(x) < 0$.

The quantity $\eta(x)$ measures the strength of the evidence that the label of a text is *Government* or *Opposition*, and we can use this quantity to scale the 55 virgin texts. Unfortunately, the Naive Bayes scaling method has serious drawbacks. First, the estimated log odds tend to be absurdly high. On our example, the median absolute log odds is 197.8, corresponding to an unrealistically high probability of class membership exceeding $1 - 10^{-85}$. Second, because $\eta(x)$ is measuring the strength of the evidence, longer texts will tend to have higher absolute log odds. We illustrate both of these defects in Fig. 1, where we plot the absolute odds of class membership as a function of text length.

Related methods suffer from versions of this same problem. Multinomial inverse regression (Taddy, 2013) regularizes the probability vector estimates p_1 and p_2 and adds a calibration step to the log-odds, but it still suffers from the same drawbacks as Naive Bayes. Discriminative methods, like those used by Joachims (1998) and Jia et al. (2014), are affected to a degree depending on their choice of features. With logistic regression, for example, when the features are linear functions of the counts x, then it will still be the case that longer documents have more extreme counts and hence more extreme predictions.

One could potentially fix the issue of longer documents getting more extreme predicted probabilities by using relative frequencies (x_v/n) as features instead of absolute counts (x_v). However, even with this choice of features there is still a fundamental disconnect between the classification philosophy and the goals of scaling. In the classification world, a document is either "black" or "white;"

for an unlabelled document, the method will tell you the probability that the label is black. In reality, though, a text is "gray," a mixture of black and white. This is a fundamental difference in perspective that precludes using a classification method for our task. We expand on this metaphor below.

3. Scaling with dictionaries. Not all text scaling methods take the black-and-white classification view of the world. One of the most successful alternatives is dictionary-based scaling (Stone, Dunphy and Smith, 1966; Pennebaker, Francis and Booth, 2001; Hu and Liu, 2004). In their simplest forms, dictionary methods conceive as each text as a mixture of two contrasting poles, such as positive and negative. Neutral words get discarded from the vocabulary. The scaling of a text is determined by the average orientation of its tokens.

There are many variations of dictionary-based scaling but for concreteness we will focus on Grimmer and Stewart's (2013) formulation. To apply that scaling to the problem at hand—scaling debate speeches—we would need two non-overlapping lists: one of words associated with *Government* and one of words associated with *Opposition*. Given these lists, we would assign a score $s_v = +1$ to each word type v in the *Government* list, and a score $s_v = -1$ to each word type v in the *Opposition* list. The dictionary-based scaling of a text with token count vector x would be

$$t(x) = \frac{1}{n} \sum_{v \in \mathcal{V}} x_v s_v,$$

where $n = \sum_{v \in \mathcal{V}} x_v$; this quantity is equal to the difference in word type occurrence rates between the *Government* and *Opposition* lists.

It is labor-intensive and error-prone to build a custom dictionary for each application, so often when practitioners apply dictionary scaling methods, they use off-the-shelf dictionaries instead of building their own. For our application, the Lexicoder sentiment dictionary (LSD, 2015 version), "a broad lexicon scored for positive and negative tone and tailored primarily to political texts," would be a natural choice (Young and Soroka, 2012, 211). However, as those authors note, applying an off-the-shelf dictionary to a new domain often leads to undesirable results. Table 2 illustrates this point in the context of our application by comparing the word orientations as determined by the LSD with their empirical associations with *Government* and *Opposition* as observed in the leadership speeches. The rows indicate the LSD-assigned orientations of the words, taking negations into account as recommended by Young and Soroka (2012). The columns are the associations of each word with the government/opposition status of the speaker, using the "keyness" G^2 likelihood ratio score, where associations with Government speaker usage of terms where $p <= 0.05$ classed as "Government", and similarly for associations with Opposition speaker usage, and terms outside of that range classed

TABLE 2
Comparing government and opposition words to Lexicoder sentiment dictionary matches.

	Government/Opposition		
Sentiment	Government	Neutral	Opposition
Positive	11 partners, progress, balance, achieved, legitimate, best, forward, better, improvement, improvements	377 confidence, like, great, well, ensure, hope good, opportunity, normal, responsible	2 wealth, creation
Neutral	66 public, now, economic, per, economy, cent, growth, new, way, community	2,329 government, country, business, irish, made, many, us, can, years, must	54 people, political, house, mr, one, taoiseach, minister, deputy, time, questions
Negative	8 problems, ireland's, debt, difficulties, deficit, deterioration, opposite, implications	346 scandals, ireland, difficult, allegations, failed, concern, scandal, unfortunately, innuendo, loss	0 —

as "Neutral." We display the number of word types in each cell, along with the most common words.

If the dictionary were appropriate for our application, we should observe positive words associated with government usage, and negative words associated with opposition usage. The patterns in Table 2, however, show a very different result. Only 11 "positive" words have high usage in the government leadership speech, and no "negative" words have high usage in the opposition leadership speeches. Most "positive" and "negative" words do not have a clear association with either *Government* or *Opposition*. Furthermore, there are some worrying cases where the dictionary orientation is counter to the association between the classes. For example, while the LSD declares the word to be negative, in the context of the debate *deficit* refers simply to a fiscal outcome; likewise, *confidence* is related to the question of the debate, and not intended to convey positive valence. Despite being designed to detect political valence, the dictionary fails here since it has not been tailored for this particular debate. Terms that are associated with one type of affect generally are used differently in the context of the no-confidence debate.

Beyond the problem of domain adaptation, the more fundamental issue with dictionary methods is that their basic premise—that each word has a clear orientation—is inappropriate in our domain. Most words in our application do not clearly either belong in one category or the other. We can seen this in Table 2, where over 95% of the word types do not have statistically significantly different usage rates between the government and opposition leadership speeches. The vast majority of words get used by both government and opposition, and thus have mixed associations with both classes. Some dictionaries try to adjust for this by giving non-binary scores to the words (Bradley and Lang, 1999), but these adjustments are often *ad hoc*, and they suffer from the same domain adaption problems. In the sequel, we present an alternative method that allows for mixed word association while simultaneously adapting to the domain.

TABLE 3
Word- and document-level assumptions from three scaling methods.

		Documents: *Gray*	Documents: *Black/White*
Words: *Gray*		Affinity Model	Classification
Words: *B/W*		Dictionaries	

4. The affinity model. Classification methods assume that each text is a member a well-defined category. Dictionary methods do not make this strong assumption, but they too take an unrealistic view of the world by supposing that each word has a well-defined orientation. Table 3 highlights this difference, and makes clear that there is room for a third worldview allowing both texts and words to be gray. We will formalize this intuition in a statistical model that we refer to as the "affinity model."

Our basic conceptual model is that over the course of a speech, a speaker's orientation switches back and forth between *Government* mode and *Opposition* mode. When she is in *Government* mode, she chooses words in the same manner as the government leadership. Likewise, when she is *Opposition* mode, she chooses words in the same manner as the opposition leadership. We should place the speaker on the spectrum between the two extremes of pro-government and pro-opposition according to what proportion of time she spends in each mode. Our perspective is that documents do not have "true" classes, but instead they are mixtures of classes. This perspective is related to but differs from that of Biecek et al. (2012), who instead assume that items have true classes but uncertain class labels.

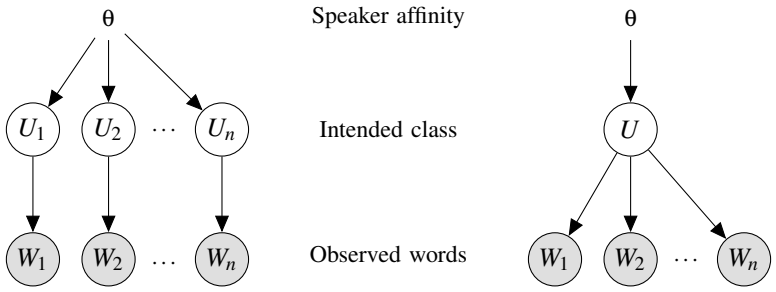
Formally, let $\mathcal{V}$ denote the vocabulary of word types, a set with cardinality $|\mathcal{V}| = V$. Encode the text of a speech as a sequence of tokens $W = (W_1, W_2, \ldots, W_n)$, with each token W_i belonging to $\mathcal{V}$. In our model, the speaker's underlying orientation evolves in parallel to the text and can be represented as discrete latent random variable $U = (U_1, U_2, \ldots, U_n)$ taking the values $i = 1, \ldots, k$, where the value U_i denotes the speaker's underlying orientation while uttering token W_i. We will in general suppose that there are K possible orientations, identified with the labels $1, \ldots, K$.

In our conceptual framework, a speech and the corresponding underlying orientation sequence are realizations of some speaker-specific random process. For $k = 1, \ldots, K$, we define a speaker's affinity toward orientation k as θ_k, the expected proportion of time that her underlying orientation is k:

$$\theta_k = \mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n} I(U_i = k)\right\},$$

where $I(\cdot)$ denotes the indicator function. Each speaker has an underlying affinity vector $\theta = (\theta_1, \ldots, \theta_K)$.

In our specific application, there are $K = 2$ orientations. Each debate speaker has a separate affinity vector $\theta = (\theta_1, \theta_2)$. We will scale each speaker by estimating his or her affinities for *Government* (θ_1) and *Opposition* (θ_2).



(a) Class affinity model (b) Classification Model

Fig 2: Generative model for the underlying orientation U and the token sequence W, contrasting the class affinity model to the classification model.

We will impose two simplifying assumptions to make inference under our model tractable. First, we will suppose that $U_1, U_2, \ldots, U_n$ are independent and identically distributed. This forces that for every label k, and position i, the underlying orientation is randomly distributed with $\Pr(U_i = k) = \theta_k$. Second, we will suppose that $W_1, W_2, \ldots, W_n$ are independent conditional on U, and that the distribution of $W_i \mid U$ depends only on U_i and is the same for all positions i. This positional invariance allows us to define for each label k and word type v the probability

$$p_{kv} = \Pr(W_i = v \mid U_i = k)$$

and it allows us to define the reference probability vector $p_k = (p_{kv})_{v \in \mathcal{V}}$. Our two simplifying assumptions result in a generative model: for each position $i = 1, \ldots, n$, the speaker picks an underlying orientation with probabilities determined by θ; given that the underlying orientation is $U_i = k$, the speaker picks token W_i according to distribution p_k. Fig. 2(a) summarizes this generative process.

For each position $i = 1, \ldots, n$, the chance that word v appears in position i is

$$\Pr(W_i = v) = \sum_{k=1}^{K} \Pr(U_i = k) \Pr(W_i = v \mid U_i = k) = \sum_{k=1}^{K} \theta_k \, p_{kv}.$$

Further, $W_1, W_2, \ldots, W_n$ are independent conditional on U, so that the probability of observing the token sequence $w = (w_1, \ldots, w_n)$ is

$$\Pr(W = w) = \prod_{i=1}^{n} \Big(\sum_{k=1}^{K} \theta_k \, p_{k w_i} \Big) = \prod_{v \in \mathcal{V}} \Big(\sum_{k=1}^{K} \theta_k \, p_{kv} \Big)^{x_v}, \tag{1}$$

where x_v is the number of times word v appears in the text. At a high level, this is the same generative model as that used for a topic model (Blei, Ng and Jordan, 2003). The main difference between these models is that topic models are typically unsupervised, but the affinity model uses supervision to estimate $p_1, p_2, \ldots, p_K$. We elaborate more on the connection to topic models in Section 9.4.

We note also that the affinity model can be seen as a generalization of the Naive Bayes model depicted in Fig. 2(b). In the Naive Bayes model, each document has a single underlying orientation, U. All words in the document share the same underlying orientation. The parameter θ can be seen as the prior distribution for U. In Naive Bayes, we do not estimate θ, but instead we estimate $\Pr(U = k \mid X_1, \ldots, X_n)$ for each class k. In Naive Bayes, each document has just one underlying orientation. The power of the affinity model is that it allows the underlying orientation to vary with the word position.

In our application, and indeed in most applications involving natural text, our two simplifying assumptions are unlikely to be true. Speakers do not order their words arbitrary, but do so in a way that respects grammatical and other structure, so in reality $W_1, \ldots, W_n$ are not conditionally independent given U. Further, speakers are not likely to alternate between orientations within sentences, so $U_1, \ldots, U_n$ are likely not independent in real speech. Nonetheless, it is still plausible that one could fit the affinity model to a real text to get an informative estimate $\hat{\theta}$. In assessing the uncertainty of $\hat{\theta}$, though, one may not want to lean too heavily on the independence assumptions. We will return to this point in Sec. 7.

5. Estimating affinities.

5.1. *Estimating affinity vectors.* The affinity model described in Section 4 lends itself naturally to likelihood-based estimation. We first consider the problem of estimating the affinity vector θ for a particular text, when we are given the reference distributions $p_1, \ldots, p_K$. We will return to the issue of estimating the reference distributions in Section 5.2.

The parameter space for the affinity vector is the simplex $\Theta \subset \mathbb{R}^K$ consisting of all vectors θ with non-negative components satisfying the equality constraint $\sum_{k=1}^K \theta_k = 1$. One implication of the equality constraint is that the model is over-parametrized, which makes estimating θ directly awkward. To handle this constraint, we will reparametrize the model in terms of a $(K-1)$-dimensional contrast vector β.

In the $K = 2$ case, we set $\beta = (\theta_2 - \theta_1)/2$, so that $\theta_1 = 1/2 - \beta$ and $\theta_2 = 1/2 + \beta$; the parameter space for β is $\mathcal{B} = [-1/2, 1/2]$. In the general case we let β be defined by the relation

$$\theta = \theta_0 + C\beta, \tag{2}$$

where θ_0 is any point in the interior of the parameter space and the contrast matrix $C \in \mathbb{R}^{K \times (K-1)}$ has full rank and satisfies $C^{\mathrm{T}} 1 = 0$. In principle θ_0 and C can be arbitrary, but for concreteness we will take θ_0 to be the center of the parameter space $\theta_0 = (1/K, 1/K, \ldots, 1/K)$, and we will take C to be the Helmert matrix. The parameter space for the contrast vector, then, is $\mathcal{B} = \{\beta \in \mathbb{R}^{K-1} : \theta_0 + C\beta \succeq 0\}$,

where $\succeq$ denotes component-wise partial order. With this particular choice of θ_0 and C, the general case agrees with the special case when $K = 2$.

Following equation (1), the log-likelihood function for the contrast vector is

$$l(\beta) = \sum_{v \in \mathcal{V}} x_v \log \mu_v, \tag{3}$$

where $\mu_v = \sum_{k=1}^{K} \theta_k p_{kv}$ and $\theta = \theta(\beta)$. We will estimate β by maximizing $l(\beta)$ or a penalized version thereof.

In the special case when $K = 2$, the score and observed information functions gotten from differentiating the log likelihood are

$$u(\beta) = l'(\beta) = \sum_{v \in \mathcal{V}} \frac{p_{2v} - p_{1v}}{\mu_v} x_v,$$
$$I(\beta) = -l''(\beta) = \sum_{v \in \mathcal{V}} \frac{(p_{2v} - p_{1v})^2}{\mu_v^2} x_v.$$

The expected information is

$$i(\beta) = \mathrm{E}\{I(\beta)\} = n \sum_{v \in \mathcal{V}} \frac{(p_{2v} - p_{1v})^2}{\mu_v}.$$

To define the analogous functions in the general case, define the matrix-valued function $Q = Q(\beta) \in \mathbb{R}^{K \times \mathcal{V}}$ with $Q_{kv} = p_{kv}/\mu_v$. In the general case, the analogous functions are

$$u(\beta) = C^{\mathrm{T}} Q x, \tag{4}$$

$$I(\beta) = C^{\mathrm{T}} Q X Q^{\mathrm{T}} C, \tag{5}$$

where $X \in \mathbb{R}^{\mathcal{V} \times \mathcal{V}}$ is the diagonal matrix with $X_{vv} = x_v$ for $v \in \mathcal{V}$. The expected information is

$$i(\beta) = n C^{\mathrm{T}} Q P^{\mathrm{T}} C = n C^{\mathrm{T}} P Q^{\mathrm{T}} C,$$

where $P \in \mathbb{R}^{K \times \mathcal{V}}$ is the matrix with kth row equal to p_k^{T} for $k = 1, \ldots, K$.

The observed information function $I(\beta)$ is positive semidefinite, indicating that the log likelihood function $l(\beta)$ is concave. We can estimate β by maximizing the log likelihood using the Newton-Raphson iterative method. The expensive part of this maximization procedure is computing $I(\beta)$, which takes time $O(VK^2)$, or faster if the count vector x is sparse. In our experience on the *Dáil* speeches, the method typically converges after about five iterations. The difficult part of the optimization is that we must restrict the search to the parameter space $\mathcal{B}$; we accomplish this using an interior-point barrier method (Boyd and Vandenberghe, 2004, Ch. 11).[2]

[2] An alternative estimation procedure is to use an expectation-Maximization algorithm, which produces similar results

In exchange for adding a small bias to the estimates, we can reduce the variance and remove the explicit inequality constraints on the parameter space. In particular, Firth (1993) shows that in the asymptotic regime where n tends to infinity, adding a penalty of order $O(1)$ to a log likelihood adds a term of size $O(1/n)$ to the bias of the estimator (sometimes reducing the estimator's bias, but not necessarily doing so in our setting). In our case, we choose a positive scalar λ and define the penalty function

$$\psi_\lambda(\theta) = \lambda \sum_{k=1}^{K} \log \theta_k.$$

Then, we estimate the affinities by maximizing the penalized log likelihood $\tilde{l}_\lambda(\beta) = l(\beta) + \psi_\lambda(\theta)$, where $\theta = \theta(\beta)$. The penalty ensures that $\tilde{l}_\lambda$ is strictly concave, and further that the maximizer $\hat{\beta}_\lambda$ is unique and belongs to the interior of the parameter space. For the analyses in this manuscript, we use the penalty value $\lambda = 0.5$. Section 5.2 provides some theoretical justification for this penalty value in a related context. (In our Appendix C we explore a range of these values, confirming the choice of 0.5.)

5.2. *Estimating reference distributions.* The reference distributions $p_1, p_2, \ldots, p_K$ themselves need to be estimated from data. In our framework, this learning step requires not large volumes of training data, but rather texts that are clearly polar examples of each reference class, to form benchmarks for estimating the other texts' affinities to these classes. In the context of our specific application, the 1991 Irish *Dáil* confidence debate, recall that the contrasting $K = 2$ classes represent *Government* ($k = 1$) and *Opposition* ($k = 2$). We will use the leaders of the government and opposition respectively to represent the archetype texts for each class. *Taoiseach* (Prime Minister) Charles Haughey's speech forms the government reference text for estimating p_1, and the speeches from the two opposition party leaders (Spring and de Rossa) form the reference texts for estimating p_2.

To estimate a particular reference distribution p, we will suppose in general that we have at our disposal m texts drawn from this distribution of lengths $n_1, n_2, \ldots, n_m$. We denote the vectors of word counts for these texts by $x_1, x_2, \ldots, x_m$. In our application, $m = 1$ for estimating the *Government* reference, and $m = 2$ for estimating the *Opposition* reference. We will use smoothed empirical frequencies to estimate p_v as advocated by Lidstone (1920). We choose a nonnegative smoothing constant α and

but requires many more iterations. Here, we prefer the simplicity of the Newton-Raphson approach and the fact that that the derivation of the Newton-Raphson step permits us to work out the relationship more directly with other measures (such as Wordscores) and to derive the influence measures. In Appendix B, we show how EM can be used as an alternative fitting procedure.

estimate the probability of word type v as

$$\hat{p}_v = \left(\alpha + \sum_{j=1}^{m} x_{jv}\right) \Big/ \left(V\alpha + \sum_{j=1}^{m} n_j\right).$$

Specifically, we will set $\alpha = 0.5$. It is not essential to smooth the estimates of p, but doing so reduces estimation variability.

There are many reasonable choices for the smoothing constant α, including choosing α adaptively (Fienberg and Holland, 1972). In natural language processing, it is common to take $\alpha = 1$ so that $\hat{p}$ is the maximum *a posteriori* estimator under a uniform prior (Jurafsky and Martin, 2009, Sec. 4.5.1). From a frequentist standpoint, the value $\alpha = 0.5$—which corresponds to using a Jeffreys prior for p—is slightly more defensible. In the regime where V is fixed and n tends to infinity, using the results from Firth (1993) one can show that using $\alpha = 0.5$ results in an expected Kullback-Leibler divergence from $\hat{p}$ to p of order $O(n^{-3/2})$ instead of $O(n^{-1})$ for other choices of α.

Once we have estimates $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_K$ of the reference distributions, to get an estimate of the class affinity vector θ for a text, we use the methods from Section 5.1, using the estimated class distributions in place of their true values. This plug-in procedure allows us to get point estimates of θ. There are two sources of uncertainty in each estimate $\hat{\theta}$: randomness in the vector of counts, x, and randomness in the reference distribution estimates $\hat{p}_1, \hat{p}_2, \ldots, p_K$. For a full uncertainty estimate, we need to account for both sources. We will return to this point in Section 7.

One limitation of our estimation scheme is that it does not use information from the non-extreme speeches. We could potentially try to incorporate these speeches into our estimates $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_K$ using a semi-supervised approach like that of Murphy, Dean and Raftery (2010), for example. However, there are dangers to doing so, namely the possibility of fitting word probabilities from lexicon that does not relate to the primary axis of class affinity. In work applying unsupervised scaling to Irish budget speeches in the parliament, for instance, Lowe and Benoit (2013, 308-309) found that the unusual position of *Sinn Féin* introduced an alternative dimension to the debate to an otherwise mainly government-opposition divide, causing unsupervised estimates of the positions from *Sinn Féin* TDs to be estimated wrongly when compared to human coding. By selecting extreme texts on a known dimension, the scaling is based on affinities to classes known to the analyst to be represented clearly by these extreme texts, rather than the myriad of other possible language patterns found in texts that are not so clearly representative of the class extremes.

6. Vocabulary diagnostics and selection. An additional advantage of the simple analytic form of the affinity is how it facilitates computationally efficient diagnostic checking for the model fit.

Ideally, our fit should exhibit two characteristics. First, it should not be driven by a small number of word types, but instead it should be determined by an accumulation of information from many different word types. Second, the word types that show the most influence in determining the fit should be ones that make sense from a subject matter perspective. To check whether our scaling results satisfy these properties, and to better understand them generally, we will develop an influence measure to characterize the impact of each word type in determining the overall fit.

Our strategy for assessing influence stems from Cook (1977), who, in the context of linear regression, assesses the influence of each observation by measuring the change that results from deleting the observation. Proceeding analogously, we will measure the influence of a word type $v \in \mathcal{V}$ by setting the corresponding token count x_v to zero and observing the change in the class affinity estimate $\hat{\theta}$. Ideally, we would do this by computing the maximizer $\hat{\theta}^{(-v)}$ of the log likelihood (or, when regularizing, the penalized log likelihood) gotten after setting x_v to zero, but the large number of word types makes this impractical. In fact, a more direct analogy with Cook's distance would delete v from all documents, not just the reference text, but doing so would be more computationally intensive and give similar results. We will settle for finding a computationally simple closed-form approximation to $\hat{\theta}^{(-v)}$.

Suppose that x is a vector of token counts for the particular text of interest, and that $\hat{\theta} = \theta_0 + C\hat{\beta}$ is the affinity vector estimate gotten from $\hat{\beta}$, the maximizer of the corresponding log likelihood $l(\beta)$ defined in (3). Making the dependence on x explicit, the score and observed information functions are

$$u(\beta;x) = C^{\mathrm{T}}Qx, \qquad I(\beta;x) = C^{\mathrm{T}}QXQ^{\mathrm{T}}C,$$

where $X \in \mathbb{R}^{\mathcal{V}\times\mathcal{V}}$ is a diagonal matrix with $X_{vv} = x_v$ for $v \in \mathcal{V}$ and $Q = Q(\beta)$ is as defined in Section 5.1.

For an arbitrary word type $v \in \mathcal{V}$, consider the effect of setting $x_v = 0$. This defines a new vector of token counts $x^{(-v)}$ defined by $x_v^{(-v)} = 0$ and $x_w^{(-v)} = x_w$ for all $w \neq v$. Let e_v denote the vth standard basis vector in $\mathbb{R}^{\mathcal{V}}$ and define $h_v = C^{\mathrm{T}}\hat{Q}e_v$, where $\hat{Q} = Q(\hat{\beta})$. Note that $x = x^{(-v)} + x_v e_v$, so that

$$u(\hat{\beta};x) = u(\hat{\beta};x^{(-v)}) + x_v h_v, \qquad I(\hat{\beta};x) = I(\hat{\beta};x^{(-v)}) + x_v h_v h_v^{\mathrm{T}}.$$

Since $u(\hat{\beta};x) = 0$, this implies that evaluating the score function with the new data at the old estimate gives

$$u(\hat{\beta};x^{(-v)}) = -x_v h_v. \tag{6}$$

The maximizer $\hat{\beta}^{(-v)}$ of the new log likelihood is roughly equal to the first Newton scoring step from $\hat{\beta}$. We can compute this step explicitly by first computing the inverse of the observed information matrix:

$$\begin{aligned} \{I(\hat{\beta};x^{(-v)})\}^{-1} &= \{I(\hat{\beta};x) - x_v h_v h_v^{\mathrm{T}}\}^{-1} \\ &= \{I(\hat{\beta};x)\}^{-1} + (x_v^{-1} - \tilde{h}_v^{\mathrm{T}} h_v)^{-1} \tilde{h}_v \tilde{h}_v^{\mathrm{T}} \end{aligned} \tag{7}$$

where $\tilde{h}_v = \{I(\hat{\beta};x)\}^{-1} h_v$.

Approximating the maximizer by the first Newton step from $\hat{\beta}$ gives

$$\begin{aligned} \hat{\beta}^{(-v)} &\approx \hat{\beta} + \{I(\hat{\beta};x^{(-v)})\}^{-1} u(\hat{\beta};x^{(-v)}) \\ &= \hat{\beta} - (x_v^{-1} - \tilde{h}_v^{\mathrm{T}} h_v)^{-1} \tilde{h}_v, \end{aligned}$$

where we have used (6) and (7) to simplify the expression. Using this approximation for $\hat{\beta}^{(-v)}$ gives us an approximation for the change in the estimated affinities:

$$\begin{aligned} \hat{\theta} - \hat{\theta}^{(-v)} &= C\hat{\beta} - C\hat{\beta}^{(-v)} \\ &\approx (x_v^{-1} - \tilde{h}_v^{\mathrm{T}} h_v)^{-1} C\tilde{h}_v. \end{aligned}$$

Motivated by this approximation, we define our influence measure as

$$d_v = (1/2)\|(x_v^{-1} - \tilde{h}_v^{\mathrm{T}} h_v)^{-1} C\tilde{h}_v\|_1, \tag{8}$$

where $\|\cdot\|_1$ denoteds 1-norm. When we are regularizing the estimates, using a penalized log likelihood $\tilde{l}(\beta;x)$ in place of $l(\beta;x)$, we define the influence similarly, using the negative Hessian $-\nabla^2_\beta \tilde{l}(\beta;x)$ in place of $I(\beta;x)$.

Using a 1-norm instead of a Euclidean norm in the definition of d_v allows us to interpret d_v as the total amount of positive change to the components of $\hat{\theta}$. Given that $1^{\mathrm{T}}(\hat{\theta} - \hat{\theta}^{(-v)}) = 0$, this is also equal to the total amount of negative change.

In the our results (including those presented in Fig. 4) we excluded words appearing only once and words on the English Snowball "stop" word list. Why did we exclude these words?

After fitting affinity model to the complete vocabulary and using it to scale the 55 non-leadership speeches, we computed the influence measures as defined in (8) for each speech word count vector x and word type v,as well as the direction of influence (whether the appearance of the word pushes the fit towards *Government* or *Opposition*). This gave us a 55×9731 matrix of (speech, word) influence measures. Most of the entries of this matrix are zero since most count vectors x are sparse and words

that do not appear in a speech have no influence on its affinity estimate. For each word type, we recorded the count of nonzero speech influence entries, along with the median and maximum of the nonzero entries. (We report these values in the Appendix, where we also compare the influence measures for each word to alternative selection scores such as entropy or a G^2 association measure.)

The influence of a word is determined by its usage rate and the degree to which is usage is imbalanced across the reference classes. The most influential words are those that appear frequently and exhibit a small imbalance between *Government* and *Opposition*, or else appear moderately and exhibit a large imbalance between the two classes. This holds generally: influential words tend to either be highly imbalanced, or moderately imbalanced with high usage rates. For example *social*, *nation*, and *economic* influence the affinity fit towards *Government*, and *people* and *taoiseach* influence the affinity fit towards *Opposition*. However, we also discovered that certain function words like *and* and *the* exerted a big influence on the fit. These function words have slightly imbalanced usage rates in the reference texts, which, compounded with a high usage rate, results in a large net influence. This sensitivity to stylistic differences is a manifestation of a common critique of the related Wordscores scaling method (Beauchamp, 2012; Grimmer and Stewart, 2013). To reduce sensitivity to stylistic differences, we eliminated function words (the Snowball English "stop" words) from our analysis.

We also saw that a few rare words like *attribute* and *proof* have large influence. These words are not meaningful discriminators on substantive grounds, but they show up as influential because they only appear once in the reference speeches. The estimated probabilities for these words are unreliable. Their influence is determined purely by estimation variability. To get around this, in our final analysis we choose to exclude these words—the *hapax legomena*—that only appear once in the reference speeches.

After excluding stop words and *hapax legomena*, we were left with a reduced vocabulary $\mathcal{V}$ of 1321 word types. We re-fit the model and re-scaled the speeches, computing the influences of the word types in the reduced-vocabulary model. Table 4 shows the most influential *Government* and *Opposition* words, computed as before.

7. Uncertainty quantification. In principle, it is possible to get standard errors for an affinity estimate $\hat{\theta}$ directly from the expected or observed information function (5). However, this likelihood-based standard error is likely too narrow, because it ignores uncertainty in the estimates of the reference distributions $(\hat{p}_1, \ldots, \hat{p}_K)$, and it relies on the independence assumptions in the model. Ignoring uncertainty in the reference distribution estimates is inappropriate when the reference set is small, as it is here (three leadership speeches). Similarly, the independence assumption—that word tokens

TABLE 4
Influential words after feature selection. Median and maximum influence (×100) exerted by the most influential words, grouped by direction of influence. Medians are computed over texts containing the word.

Government				Opposition			
Word	Count	Median	Max	Word	Count	Median	Max
deasy	3	0.9	1.8	people	54	1.3	4.9
nation	12	0.8	1.8	taoiseach	43	0.8	3.1
cent	26	0.8	3.4	democrats	23	0.7	1.9
social	30	0.7	10.4	minister	44	0.6	2.5
corresponding	1	0.7	0.7	system	37	0.6	2.7
1990	17	0.7	1.9	house	54	0.5	1.9
union	9	0.7	1.0	o'kennedy	5	0.5	0.9
per	31	0.7	3.1	progressive	24	0.5	1.4
belief	3	0.7	1.0	say	39	0.5	1.3
reform	19	0.6	2.4	issue	27	0.5	1.4
1987	20	0.6	4.0	million	26	0.5	1.5
economic	33	0.6	2.8	headings	2	0.5	0.5
roads	6	0.6	2.6	wealth	6	0.5	1.4
development	29	0.6	2.7	printed	2	0.4	0.7
new	38	0.6	1.6	said	41	0.4	1.6

in different positions of a text are independent of each other—simplifies the analysis, but it is likely violated in real-world data. To accurately assess the uncertainty in our estimates, we need a method that accounts for the uncertainty in the reference distribution estimates and the dependence between nearby words in text.

To estimate the sampling distribution of the scaling estimates under dependence between word tokens, we will use a block bootstrap that respects the natural linguistic structure of the text, by following Lowe and Benoit (2013)'s recommendation to resample texts at the sentence level to simulate sampling variation but also to capture meaningful dependencies among words within natural syntactic units. To properly account for uncertainty in the reference distribution estimates, we will also construct sentence-level bootstrapped reference speeches. To describe the procedure, we let $t_1, \ldots, t_R$ denote the reference texts, and we let s denote the scaled text. The procedure is as follows:

1. For bootstrap replicates $b = 1, \ldots, B$:
   (a) Construct bootstrapped reference texts $t_1^{*b}, \ldots, t_R^{*b}$, where t_i^{*b} has sentences drawn with replacement from t_i, with the same total number of sentences.
   (b) Use the bootstrapped reference texts $t_1^{*b}, \ldots, t_R^{*b}$ to estimate the reference distributions $\hat{p}_1^{*b}, \ldots, \hat{p}_K^{*b}$ as described in Sec. 5.2.
   (c) Construct a bootstrap version of the scaled text, s^{*b} by resampling sentences from s, with replacement.
   (d) Treating the reference distribution estimates $\hat{p}_1^{*b}, \ldots, \hat{p}_K^{*b}$ as fixed, construct an affinity-scaling estimate $\hat{\theta}^{*b}$ from s^{*b}.
2. Use the sample standard deviation of $\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*B}$ as the bootstrapped estimate of the standard

error of the affinity scaling estimate $\hat{\theta}$ for s.

We performed this procedure for all of 55 non-leadership speeches, getting a separate bootstrap standard error for each. For comparison, we computed likelihood-based standard error for the estimates from the Fisher information conditional on the reference estimates. Unsurprisingly, the bootstrap standard errors are generally wider than the likelihood-based estimates. The two uncertainty estimates are both on the same order of magnitude, with the bootstrap standard error being less than 1.5 times as large as the likelihood-based standard error for most of the speeches (87%); the median ratio of the two standard errors is 1.3. In the sequel, we use bootstrap standard errors to quantify the uncertainty in the affinity estimates.

8. Results. Fig. 3 displays the estimated government affinities for all 55 speeches using the reduced vocabulary as discussed in Sec. 6. Within each party grouping, which separates the *Fianna Fáil* ministers from the non-ministerial backbench speakers, we plot the median affinity position as a vertical bar. The figure includes 95% confidence intervals, computed using the sentence-level bootstrap from the previous section.

At both the level of the government versus opposition and inter-party levels, the results are entirely in line with expectations: not only are the parties arrayed in an order that would be consistent with expectations, with opposition parties on the *Opposition* side, and the governing parties on the other, but also we see that speeches from the different parties align with the extremity of their positions in regards to the establishment. The speeches of most centrist opposition party, *Fine Gael*, express a more moderate anti-*Government* positions than either the left party Labour or the far-left Democratic Left party. This median difference emerges clearly even though we considered the speeches of the Labour and Democratic Left leaders as equivalent for the purposes of training the *Opposition* class.

The more interesting distinctions emerge when we examine *intra*-party differences in expressed position. Among the government ministers, it is not surprising to see that John Wilson, the FF Deputy Prime Minister (*Tánaiste*, or "FF Tan" in the plot), and Gerard Collins, the Foreign Minister and a senior *Fianna Fáil* minister had extreme *Government*-oriented estimated positions exceeded only by the *Taoiseach* Charles Haughey himself. What is more interesting is that the next minister in the estimated ranking, Albert Reynolds, would later become the next *Taoiseach*. At the other extreme, among the most *Opposition*-oriented government minister we see notable examples in Raphael (Ray) Burke, who was removed from his ministerial position the following year, and Mary O'Rourke, who months later would challenge Albert Reynolds for the party leadership.

The "back-bench" FF members voted with the government but generally gave speeches that were

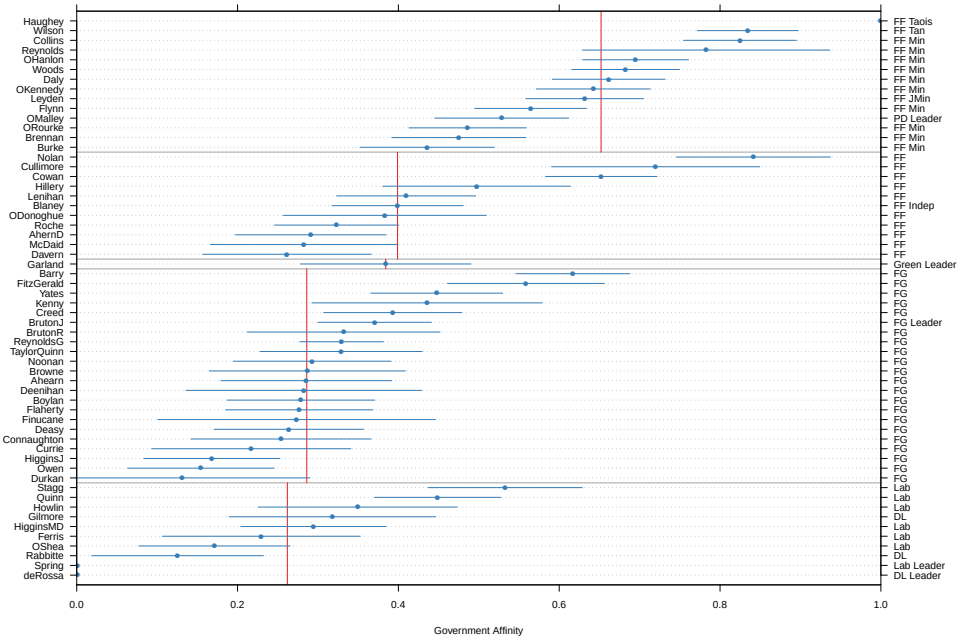


Fig 3: Affinity scaling estimates ($\hat{\theta}_1$) with bootstrap 95% confidence intervals

far more lukewarm than the FF ministers. Correspondingly, we see that the estimated estimated *Government* affinities for the back-benchers are generally lower than those of the minsters. There were three exceptions, members with extreme estimated *Government*-oriented affinities: Nolan, Cullimore, and Cowan. One of these members, Brian Cowen, became Minister for Labour the following year, and occupied senior positions include Prime Minister for the next two decades.

On the opposition side, we see a similar set of heterogeneous estimated affinities. Two salient examples of extreme estimated *Government*-oriented affinities are *Fine Gael* TD Garret FitzGerald, a former and future Prime Minister, and TD Peter Barry, who had fought Fitzgerald in 1987 for party leadership. Both emphasized fairly standard economic concerns, attacking the government's poor economic performance rather than its corrupt behavior. It is notable that the member with the highest estimated pro-opposition affinity, DL member Pat Rabbitte who would later become leader of the Labour Party; in his speech, he engaged in a personal set of attacks against the *Taoiseach* and specifically attacking his character and judgment.

Overall, the balance of the positions expressed in the debate favored the opposition position, although this reflected the fact that the majority of the speeches (36 out of 55) were from opposition TDs. More interestingly, half of the government speakers, predominantly backbenchers, expressed positions in their speeches closer to the opposition than to the government side—such as Neil Blaney whom we quoted above, with a position of under 0.40 and right at the median of the FF backbenchers—reinforcing our point that speeches are far more informative than voting when it

comes to revealing preferences.

The results of applying the class affinity scaling model to the confidence debate speeches provides a results consistent with expectations and with previous scholarly investigations of this episode (Laver and Benoit, 2002). Using only the texts of the speeches, we have succeeded at revealing differences between the speakers that were not apparent from their party affiliations.

9. Connections to other methods.

9.1. *Dictionary methods.* In the special case that the reference distributions $p_1, p_2, \ldots, p_K$ have disjoint supports—that is, when no two classes k and l are such that both $p_k(v) > 0$ and $p_l(v) > 0$ for some word type v—affinity scaling is exactly equivalent to dictionary scaling.

To make this equivalence clear, suppose that for each word type $v \in \mathcal{V}$, at most one of the reference probabilities $p_{1v}, p_{2v}, \ldots, p_{Kv}$ is nonzero. When this is the case, we can partition the vocabulary as a union of disjoint sets, $\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2 \cup \cdots \cup \mathcal{V}_K$, where

$$\mathcal{V}_k = \{v \in \mathcal{V} : p_{kv} > 0\}.$$

Here, $\mathcal{V}_k$ is the set of word types associated with label k. The disjoint support condition ensures that each word type v is associated with exactly one label.

Under the disjoint support condition, when we observe the ith token w_i, we can immediately infer the underlying orientation u_i to be the only class with this word in its support. The log-likelihood simplifies to

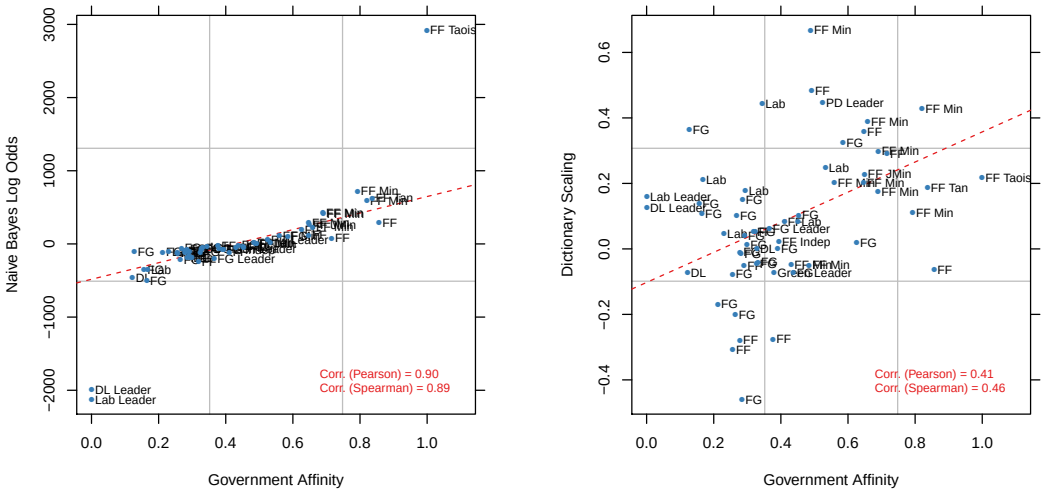
$$\begin{aligned} l(\theta) &= \sum_{v \in \mathcal{V}} x_v \log\Big(\sum_{k=1}^{K} \theta_k p_{kv}\Big) \\ &= \sum_{k=1}^{K} \sum_{v \in \mathcal{V}_k} x_v \log(\theta_k p_{kv}) \\ &= \sum_{k=1}^{K} n_k \log \theta_k + (\text{constant}), \end{aligned}$$

where $n_k = \sum_{v \in \mathcal{V}_k} x_v$ and the constant does not depend on θ. In this case, the maximum likelihood estimate of the class affinity vector is

$$\hat{\theta} = \Big(\frac{n_1}{n}, \frac{n_2}{n}, \ldots, \frac{n_K}{n}\Big).$$

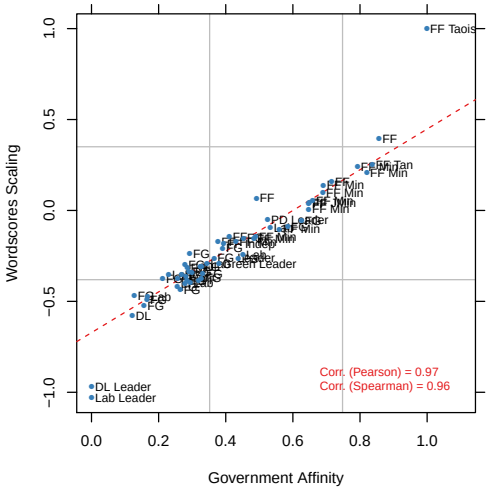
That is, the estimated class affinities are the token occurrence rates in the support sets $\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_K$.

9.2. *Wordscores.* The "Wordscores" scaling method developed by Laver, Benoit and Garry (2003) turns out to be closely related to class affinity scaling. That method, which is primarily used to scale

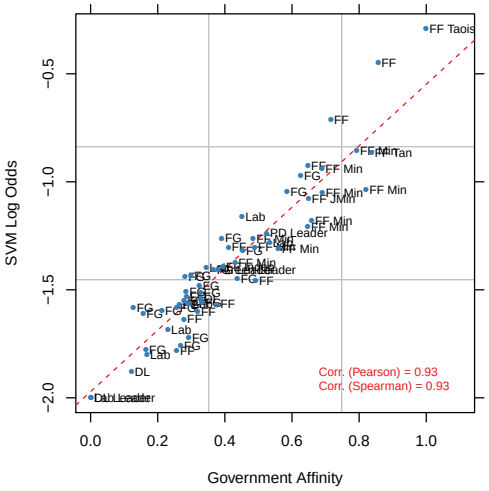


(a) Naive Bayes

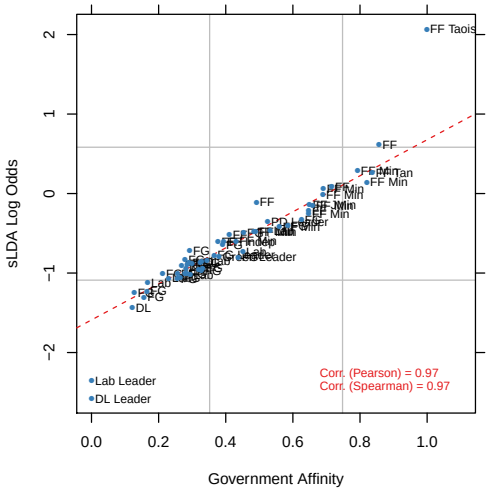
(b) Dictionary



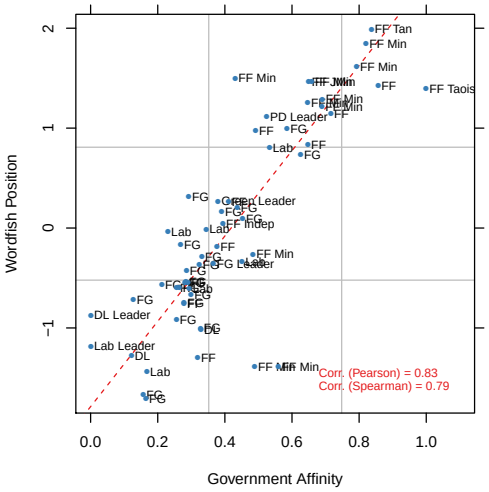
(c) Wordscores



(d) Support Vector Machine



(e) Supervised LDA



(f) Wordfish

Fig 4: Comparisons between scaling methods

documents between $K = 2$ reference classes works well in practice but has been criticized for having *ad hoc* theoretical foundations (Lowe, 2008). We can show, however, that Wordscores scaling is closely related to affinity scaling, and gives highly correlated results for texts that are not close to the extremes (represented by the reference text positions). We elaborate on this connection below.

In its simplest form, Wordscores takes as given reference distributions for each class, denoted p_1 and p_2. The method defines the wordscore of a word type $v \in \mathcal{V}$ as

$$s_v = \frac{p_{2v} - p_{1v}}{p_{1v} + p_{2v}}. \tag{9}$$

Word types that only appear in class 2 have scores of $+1$, while types that only appear in class 1 have scores of -1. Other types have intermediate values indicating the relative degrees of association with the two classes. The unnormalized "text score" of a length-n text with token count vector x is then the average wordscore of its tokens:

$$t(x) = \frac{1}{n} \sum_{v \in \mathcal{V}} \frac{p_{2v} - p_{1v}}{p_{1v} + p_{2v}} x_v, \tag{10}$$

Texts with positive $t(x)$ values tend to be more like class 2, while texts with negative $t(x)$ values tend to be more like class 1.

The magnitude of the unnormalized score $t(x)$ is not directly interpretable. To fix this, Martin and Vanberg (2007) advocate rescaling the score to ensure that average reference texts from the two classes have scores of -1 and $+1$. To realize the Martin–Vanberg scaling, for $k = 1, 2$ define

$$t_k = \sum_{v \in \mathcal{V}} \frac{p_{2v} - p_{1v}}{p_{1v} + p_{2v}} p_{kv}.$$

An average text of length n from class k has token counts satisfying $x_v/n = p_{kv}$, so that its score is $t(x) = t_k$. Using the relation $p_{1v}/(p_{1v} + p_{2v}) = 1 - p_{2v}/(p_{1v} + p_{2v})$ termwise in the sum, one can verify that $t_1 = -t_2$. The Martin–Vanberg wordscore scaling is

$$\tilde{t}(x) = -\frac{t_2 + t_1}{t_2 - t_1} + t(x) \cdot \frac{2}{t_2 - t_1} = t(x)/t_2.$$

An average text x from class 1 satisfies $\tilde{t}(x) = -1$; an average text x' from class 2 satisfies $\tilde{t}(x') = +1$.

The wordscore scaling $\tilde{t}(x)$ turns out to be deeply connected to affinity scaling. To see this connection, note that using the parameterization from Section 5.1, the score and observed information functions for the affinity model evaluated at $\beta = 0$ are

$$u(0) = 2 \sum_{v \in \mathcal{V}} \frac{p_{2v} - p_{1v}}{p_{1v} + p_{2v}} x_v = 2n t(x),$$

$$i(0) = 2n \sum_{v \in \mathcal{V}} \frac{(p_{2v} - p_{1v})^2}{p_{1v} + p_{2v}} = 2n (t_2 - t_1).$$

There is a striking relationship between the scaled text score and the derivatives of the mixture model log likelihood:

$$\tilde{t}(x)/2 = \{i(0)\}^{-1}u(0).$$

The right hand side of this expression is equal to the first Fisher scoring iterate computed while maximizing $l(\beta)$ starting from the initial value $\beta = 0$. When the maximizer $\hat{\beta}$ is close to 0, it will be approximately equal to this first iterate. Thus, when a text is roughly balanced between the two reference classes ($\hat{\beta} \approx 0$), it will also be the case that

$$\tilde{t}(x) \approx 2\hat{\beta} = \hat{\theta}_2 - \hat{\theta}_1;$$

In these cases, the wordscore scaling is a linear transformation of the estimated class affinities.

We demonstrate the quality of this approximation in Fig. 4c, where we plot the wordscore scaling versus the estimated government affinity for the debate speeches. We can see that there is very good agreement between the two scalings, and that when $\tilde{t}(x) \approx 0$, the two scalings are almost identical.

9.3. *Support vector machines and logistic regression.* We have just shown analytically that affinity scaling gives similar results to Wordscores. It turns out that, when the number of reference documents is small, up to scaling, both methods are approximately equivalent to classifying with a support vector machine or linear regression.

Suppose that we are in the two-class ($K = 2$) case, and that there is one reference document for each class. Imagine fitting a linear classifier that tries to predict class using a document's word frequencies as features. With a vocabulary size V greater than the number of training documents, the two classes can be perfectly separated as long as the two reference distributions p_1 and p_2 corresponding to the training documents are not identical. In this case, the support vector machine fit and the logistic regression fit are identical, up to differences that arise from regularizing the coefficients.

Given a document with length n and word count vector x, its feature vector is its vector of word frequencies, $n^{-1}x$. The feature vectors for the two training documents are p_1 and p_2. Up to a constant of proportionality, the maximum margin predictor, expressed as a function of x is

$$\begin{aligned} \eta(x) &= (p_2 - p_1)^{\mathrm{T}}\{n^{-1}x - (1/2)(p_1 + p_2)\} \\ &= \frac{1}{n}\sum_{v \in \mathcal{V}} (p_{2v} - p_{1v})x_v + (\text{const.}) \end{aligned} \tag{11}$$

Since the classes are perfectly separated, and multiple of this predictor gives the same classification performance on the training set; the precise scaling chosen by the fitting procedure will depend on the regularization parameters.

Comparing the support vector machine scaling (11) with the unnormalized wordscores scaling (10), we can see that the only substantive difference is the denominator $p_{1v} + p_{2v}$ in the coefficient on x_v. Thus, up to a constant shift and scale, if $p_{1v} + p_{2v}$ is roughly constant relative to $p_{2v} - p_{1v}$, then the two methods will give similar results. In light of the connection between Wordscores and affinity scaling developed in Sec. 9.2, this implies that in these situations, the support vector machine results will be highly correlated with the affinity scaling results.

We verified the connection between the two methods empirically, using the `SVMlight` software with the default tuning parameters (Joachims, 1999). Fig. 4d shows the support vector machine estimated log odds plotted against the affinity scaling results. Both scalings give similar results (correlation 0.92). The main distinction is that the numerical value of the support vector machine log odds is determined completely by the regularization parameter and is thus uninterpretable. The affinity scaling of a document, by contrast, can be interpreted directly.

9.4. *Topic models.* Topic models share a similar perspective with the affinity model in that both represent texts as mixtures of topics, with each topic having an associated word distribution. In our framework, the topics correspond to the reference classes, and the text-specific topic weights correspond to class affinities. We learn the class distributions from a set of labeled reference texts. This approach differs from that taken by unsupervised topic models (Blei, Ng and Jordan, 2003; Grimmer, 2010), where estimated topics may or may not correspond to scaling quantities of interest.

Supervised variants of topic models allow for associations between labels and topics, but these models all assume that class membership is discrete, not a continuous scale (McAuliffe and Blei, 2008; Ramage et al., 2009; Roberts, Stewart and Airoldi, 2016). These supervised models force clear associations between the topics and the scaling quantities of interest, but they assume that the texts have discrete labels indicating class membership. Even Ramage et al.'s (2009) Labeled LDA—the closest analogue to the affinity model—assumes that each document expresses a sparse subset of the reference topics. The fundamental assumption of discrete class membership places these methods in the same category as other classification methods like Naive Bayes, estimating the probability of class membership, not class affinity.

Despite their philosophical differences, in practice supervised topic models can give scalings that are highly correlated with the affinity model scaling. The connection to supervised topic models is easiest to understand in the case of McAuliffe and Blei's (2008) Supervised Latent Dirichlet Allocation (sLDA), which models a text-specific label as a random quantity linked to a linear function of the text-specific topic weights. Roughly speaking, the method works in two stages. In the first stage,

sLDA fits a topic model to the reference texts. In the second stage, sLDA fits a logistic regression model using the fitted topic weights as predictors and the class label as response. In practice, sLDA fits the topics and the logistic regression simultaneously, but when the number of topics is larger than the number of reference texts, any differences between sequential and simultaneous fitting are determined by the regularization parameters and the random initialization.

The connection between sLDA and affinity model scaling is closest with two topics and two reference texts. In this case, since the number of topics equals the number of reference texts, sLDA can get a perfect fit by allocating one topic to each reference text, and can separate the two classes perfectly given the topic weights $(\hat{\theta}_1, \hat{\theta}_2)$ by using a linear predictor for the odds of class membership of the form $\eta = b(\hat{\theta}_2 - \hat{\theta}_1)$, where the coefficient b gets determined by the regularization parameters. When the sLDA fit gets used for prediction on the unlabelled texts, the fitted topic weights $(\hat{\theta}_1, \hat{\theta}_2)$ will be the same as the values from a fitted affinity model (again, ignoring the effects of regularization regularization and initialization). The sLDA score will be highly correlated with the difference in estimated affinities.

In the case when there are more topics and more reference texts, the relationship between affinity scaling and sLDA is not as simple, but the same general intuition still holds and the two methods still give highly correlated results. Fig. 4e illustrates this with a model using 10 topics, where the correlation between the non-reference text scalings from the two methods is 0.98. Here, the sLDA method gives unreasonable results for the extremes. Furthermore, the interpretation of the scaling value if different: odds of class membership for sLDA, versus degree of membership for the affinity model.

9.5. *Unsupervised methods.* Some approaches to scaling texts, including Latent Semantic Indexing (Deerwester et al., 1990) and Slapin and Proksch (2008)'s "Wordfish" Poisson scaling method, estimate latent text-specific traits using unsupervised methods. Often, the estimated traits are correlated with recognizable attributes, and so they can be used to scale ideology. Letting x_{iv} denote the count of word type v in text i, the Slapin and Proksch (2008) Wordfish model specifies that x_{iv} is a Poisson random variable with mean λ_{iv}, where $\log \lambda_{iv} = \alpha_i + \psi_v + \theta_i \beta_v$ for some unknown text-specific parameters (α_i and θ_i) and word-specific parameters (ψ_v and β_v). Estimates of θ_i have been shown to provide valid estimates of latent positions expressed in speeches (Lowe and Benoit, 2013).

The drawback to unsupervised scaling of this sort, however, is that they provide no guarantee that the estimated latent trait corresponds to the quantity of interest. We demonstrate this behavior in Fig. 4f, where we plot the Wordfish scaling estimates of the debate speeches versus the affinity

scaling estimates. The two methods give similar results (correlation 0.82), but there are also some notable differences. The government and opposition leaders are not the most extreme examples as determined by Wordfish, indicating that even in this focused context—a debate over a confidence motion—the primary dimension of difference is something other than the government-opposition divide.

10. Discussion. In our application and in others like it, the correct prediction of a class is no longer a relevant benchmark because the process of producing political text is expected to produce heterogeneous text within each class. For us, the class—here, voting for or against the confidence motion, which was perfectly correlated with government or opposition status—is observed and uninteresting, while the heterogeneity is the primary interest. Despite what would seem obvious from a measurement model or scaling perspective, however, a standard approach in evaluating machine learning applications in political science has been predictive accuracy benchmarked against known classes (e.g. Evans et al., 2007; Yu, Kaufmann and Diermeier, 2008). This focus on estimating correct classes not only wrongly shifts attention away from the substantively interesting variation in latent traits, but also may ultimately impair classification generality by encouraging over-fitting to reduce predictive error.

Our proposed alternative, class affinity scaling, is based on a probability model similar to those underlying class predictive methods, but allows for mixed class membership. We have shifted focus from class prediction, something typically uninteresting in the social sciences, to a form of latent parameter estimation, while retaining the advantages of supervised learning approaches where the analyst controls the inputs that anchor the model. While there is a strong tradition in some disciplines, such as political science, of adapting machine learning to produce continuous scales, practitioners are often unaware of the differences in modeling assumptions between classification and scaling methods (e.g. Laver, Benoit and Garry, 2003), or they have not fully explored the implications of these assumptions (e.g. Beauchamp, 2012). We have highlighted the differences and similarities in a form that encourages future development.

The novelty of our approach is that it provides a statistical foundation for a method to scale an unlimited number of texts whose positions are unknown, from a small reference set whose positions are known, with direct quantitifaction of the uncertainty of these estimates. It both updates and extends the widely used Wordscores approach of Laver, Benoit and Garry (2003), and provides distinct advantages over related approaches adapted from methods for predicting classes. Relative to dictionary approaches, furthermore, it overcomes the limitations of static word associations by correctly

learning the relationships and weights of all words to known reference texts, providing a method that is more contextually appropriate and far more economical.

The relative simplicity of our method makes it amenable to direct mathematical analysis. This simplicity allowed us to draw connections between Naive Bayes classification, dictionary-based scaling, and a host of other methods. We were further able to exploit the analytic simplicity of the affinity scaling model to develop an influence measure assessing the sensitivity of the fit, which we then used to guide our vocabulary selection and to validate our fits to the *Dáil* debate.

Using our method to explore the nuances of the speeches in the 1991 *Dáil* confidence motion, we produced estimates for each speaker that accord with both a qualitative reading of the speech transcripts and an expert understanding of Irish politics. Our application is a hard domain problem, where no known lexicographical map exists to differentiate government versus opposition speech and dictionary-based scaling, even with a dictionary derived from political text, gives unsatisfactory results. With limited training from the leadership speeches, class affinity scaling is able to adapt to the context of the debate and give a meaningful scaling. The method has applications far beyond political text, however, and could be used to score more standard sentiment problems on a continuous scale, or applied to any other problem for which contrasting reference texts can be identified.

References.

Beauchamp, Nick. 2012. "Using text to scale legislatures with uninformative voting." http://nickbeauchamp.com/work/Beauchamp_scaling_current.pdf.

Benoit, Kenneth and Alexander Herzog. 2012. "Intra-Party Conflict Over Fiscal Austerity." LSE manuscript, October 29.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller and Akitaka Matsuo. 2018. "quanteda: An R package for the quantitative analysis of textual data." *Journal of Open Source Software* 3(30):774.

Biecek, P., E. Szczurek, M. Vingron and J. Tiuryn. 2012. "The R Package bgmm: Mixture Modeling with Uncertain Knowledge." *Journal of Statistical Software* 47:1–31.

Blei, D.M., A.Y. Ng and M.I. Jordan. 2003. "Latent Dirichlet allocation." *The Journal of Machine Learning Research* 3:993–1022.

Boyd, S. and L. Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press.

Bradley, M M and P J Lang. 1999. "Affective norms for English words (ANEW): Instruction manual and affective ratings.".

Clark, Tom S. and Benjamin Lauderdale. 2010. "Locating Supreme Court Opinions in Doctrine Space." *American Journal of Political Science* 54(4):871–890.

URL: *http://dx.doi.org/10.1111/j.1540-5907.2010.00470.x*

Cook, R Dennis. 1977. "Detection of influential observation in linear regression." *Technometrics* 19(1):15–18.

Deerwester, Scott, Susan T Dumais, George W Furnas, Thomas K Landauer and Richard Harshman. 1990. "Indexing by latent semantic analysis." *Journal of the American society for information science* 41(6):391–407.

Evans, Michael, Wayne McIntosh, Jimmy Lin and Cynthia Cates. 2007. "Recounting the Courts? Applying Automated Content Analysis to Enhance Empirical Legal Research." *Journal of Empirical Legal Studies* 4(4):1007–1039.

Fienberg, Stephen E. and Paul W. Holland. 1972. "On the Choice of Flattening Constants for Estimating Multinomial Probabilities." *Journal of Multivariate Analysis* 2:127–134.

Firth, D. 1993. "Bias reduction of maximum likelihood estimates." *Biometrika* 80(1):27–38.

Grimmer, J. 2010. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political Analysis* 18(1):1–35.

Grimmer, Justin and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21(3):267–297.

Heckerman, D., E. Horvitz, M. Sahami and S. S. Dumais. 1998. "A Bayesian approach to filtering junk e-mail." *Proceedings of the AAAI-98 Workshop on Learning for Text Categorization* pp. 55–62.

Hu, Minqing and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM pp. 168–177.

Jia, Jinzhu, Luke Miratrix, Bin Yu, Brian Gawalt, Laurent El Ghaoui, Luke Barnesmoore, Sophie Clavier et al. 2014. "Concise comparative summaries (CCS) of large text corpora with a human experiment." *The Annals of Applied Statistics* 8(1):499–529.

Joachims, T. 1999. Making large-Scale SVM Learning Practical. In *Advances in Kernel Methods - Support Vector Learning*, ed. B. Schölkopf, C. Burges and A. Smola. Cambridge, MA: MIT Press chapter 11, pp. 169–184.

Joachims, Thorsten. 1998. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*. Springer pp. 137–142.

Jurafsky, Daniel and James H. Martin. 2009. *Speech and Language Processing*. 2 ed. Upper Saddle River, NJ: Pearson.

Kessler, Brett, Geoffrey Numberg, Hinrich Schütze, Brett Kessler and Geoffrey Numberg. 1997. "Automatic detection of

text genre." *The eighth conference* pp. 32–38.

Laver, Michael and Kenneth Benoit. 2002. "Locating TDs in policy spaces: the computational text analysis of Dáil speeches." *Irish Political Studies* 17(1):59–73.

Laver, Michael, Kenneth Benoit and John Garry. 2003. "Estimating the policy positions of political actors using words as data." *American Political Science Review* 97(2):311–331.

Lidstone, G. J. 1920. "Note on the general case of the Bayes-Laplace formula for inductive or *a posteriori* probabilities." *Transactions of the Faculty of Actuaries* 8:182–192.

Lindstrom, M. J. and D. M. Bates. 1988. "Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data." *J. Am. Stat. Assoc.* 83:1014–1022.

Lowe, Will. 2008. "Understanding wordscores." *Political Analysis* 16(4):356–371.

Lowe, William and Kenneth Benoit. 2013. "Validating Estimates of Latent Traits From Textual Data Using Human Judgment as a Benchmark." *Political Analysis* 21(3):298–313.

Martin, L. W. and G. Vanberg. 2007. "A robust transformation procedure for interpreting political text." *Political Analysis* 16(1):93–100.

McAuliffe, Jon D and David M Blei. 2008. Supervised topic models. In *Advances in neural information processing systems*. pp. 121–128.

Mosteller, F. and D. L. Wallace. 1963. "Inference in an Authorship Problem." *J. Am. Stat. Assoc.* 58(302):275–309.

Murphy, T.B., N. Dean and A. E. Raftery. 2010. "Variable Selection and Updating In Model-Based Discriminant Analysis for High Dimensional Data with Food Authenticity Application." *Ann Appl Stat.* 4:396–421.

Newman, M L, James W Pennebaker and D S Berry. 2003. "Lying words: Predicting deception from linguistic styles." *Personality and social . . .* .

Pang, B., L. Lee and S. Vaithyanathan. 2002. "Thumbs up? Sentiment classification using machine learning techniques." *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* pp. 79–86.

Pennebaker, James W, Martha E Francis and Roger J Booth. 2001. *Linguistic inquiry and word count: LIWC 2001*. Mahway, NJ: Erlbaum Publishers.

Porter, Martin. 2006. "Snowball English stop word list." http://snowball.tartarus.org/algorithms/english/stop.txt.

Ramage, Daniel, David Hall, Ramesh Nallapati and Christopher D Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics pp. 248–256.

Roberts, Margaret E, Brandon M Stewart and Edoardo M Airoldi. 2016. "A model of text for experimentation in the social sciences." *J. Am. Stat. Assoc.* 111(515):988–1003.

Sahami, Mehran, Susan Dumais, David Heckerman and Eric Horvitz. 1998. "A Bayesian approach to filtering junk e-mail." 62:98–105.

Schwarz, Daniel, Denise Traber and Kenneth Benoit. 2017. "Estimating the Policy Preferences of Legislators in Parliamentary Systems: Comparing Speeches to Votes." *Political Science Research and Methods* 5(2):379–396.

Slapin, Jonathan B. and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52(3):705–722.

Stone, Philip J, Dexter C Dunphy and Marshall S Smith. 1966. *The general inquirer: A computer approach to content analysis.* Cambridge, MA: MIT press.

Taddy, Matt. 2013. "Multinomial inverse regression for text analysis." *J. Am. Stat. Assoc.* 108(503):755–770.

Young, Lori and Stuart Soroka. 2012. "Affective News: The Automated Coding of Sentiment in Political Texts." *Political Communication* 29(2):205–231.

Yu, Bei, Stefan Kaufmann and Daniel Diermeier. 2008. "Classifying Party Affiliation from Political Speech." *Journal of Information Technology & Politics* 5(1):33–48.

APPENDIX A: DETECTING INFLUENCE

Table A1 shows the results of applying our influence measure(8) to our pre-filtered document-term matrix for all 55 documents, grouped by the direction of influence.

TABLE A1
Median and maximum influence ($\times 100$) exerted by the most influential words, grouped by direction of influence. Medians are computed over texts containing the word.

Government				Opposition			
Word	Count	Median	Max	Word	Count	Median	Max
and	55	1.3	2.5	the	55	2.5	4.7
our	49	0.9	2.7	that	55	1.3	3.5
graduate	3	0.8	0.9	to	55	1.2	2.6
deasy	3	0.7	1.6	they	55	1.0	2.6
attribute	1	0.7	0.7	a	55	0.9	1.7
social	30	0.6	8.0	is	55	0.9	1.7
per cent	26	0.6	3.2	not	55	0.7	1.6
corresponding	1	0.6	0.6	people	54	0.7	3.0
nation	12	0.6	1.4	it	55	0.7	1.7
proof	2	0.6	1.0	he	42	0.6	2.0
1987	20	0.5	2.7	at	54	0.5	1.3
economic	33	0.5	2.1	his	43	0.5	1.4
will	55	0.5	1.5	taoiseach	43	0.5	1.3
international	18	0.5	1.1	by	55	0.4	0.7
union	9	0.5	0.9	as	55	0.4	1.2

We can see, for example, that the word type *social* exhibited influence on 30 speeches. For one of these speeches, deleting the word *social* has the affect of shifting the speech's affinity estimate away from *Government* by 0.08; the median shift for the 30 speeches is 0.006. Deleting *social* shifts the fit away from *Government*; equivalently, the appearances of *social* push the fit towards *Government*.

We can also see in Table A1 that there are words that a few rare words like *attribute* and *proof* have large influence. These words are not meaningful discriminators on substantive grounds, but they show up as influential because they only appear once in the reference speeches. The estimated probabilities for these words are unreliable. Their influence is determined purely by estimation variability. To get around this, in our final analysis we choose to exclude these words—the *hapax legomena*—that only appear once in the reference speeches.

It is possible that Snowball word list could have missed some influential function words, but inspecting the words in Table 4 and the other words further down in the order, we found that this was not the case for our application. The only suspicious words are *say* and *said*, but in the context of the debate, it makes sense that these words are pro-*Opposition*. When the word *said* gets used, it is typically used to quote the government ("they said" or "they continue to say"), usually by an opposition member criticizing the government. Likewise, at first glance it may seem suspicious that *per cent* is at the top of the *Government* list, but in fact this often used to cite national statistics about the economy and the GDP, using the state of the economy explain the unrest.

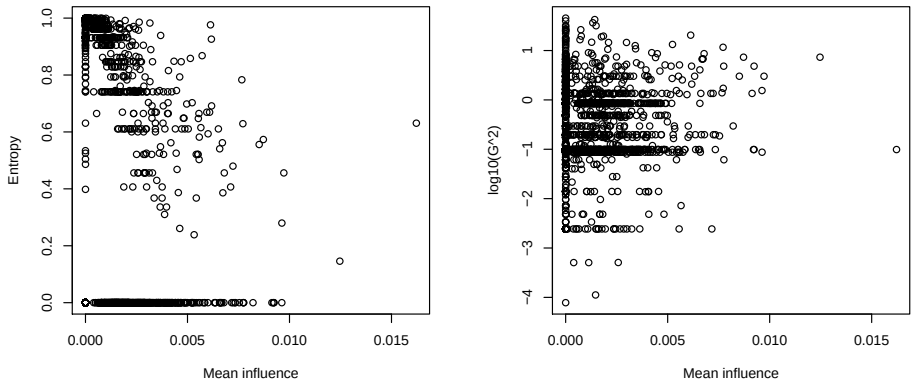


Fig A1: Comparison of mean word influence with word class association measures using the G^2 likelihood ratio statistic and Shannon entropy (with log base 2).

Overall, after looking at the most influential words in the two lists, we are comforted that most words have relatively low influence individually, and that the high-influence words make sense in the context of the application.

As a final check on our influence measure, we compared other approaches such as measuring the discrimination of words between our two classes using both likelihood-ratio and entropy measures, comparing whether our influence statistic simply reflects the uneven distribution of words across the government and opposition word frequncies in our corpus. Figure A1 plots both the word association with the government-opposition classes from the reference texts using the log base 2 Shannon entropy, as well as the G^2 likelihood ratio statistic test of association with the class label. There is no clear association between words that should be excluded because of their high entropy (near 1.0) or low association scores (χ^2 close to 0). Our measure of influence captures more than this because it is conditional not just on relative word frequencies but also on how the word's presence exerts leverage on the scaled results.

APPENDIX B: ALTERNATIVE FITTING METHOD

An anonymous reviewer suggested we explore fitting the affinity model using the EM-algorithm instead of the Newton-Raphson procedure described in Sec. 5.1. Indeed, the E- and M-steps for the model take simple forms. In that procedure, we are given text sequence $W \in \mathcal{V}^n$ and distributions $p_1, \ldots, p_K$. The goal is to estimate θ. For $i = 1, \ldots, n$ and $k = 1, \ldots, K$ we introduce the variables $u_{ik} = \Pr(U_i = k \mid W_i)$. In the E-step, we compute an estimate of each u_{ik} using the current estimate $\hat{\theta}^t$ of θ. This estimate is given by

$$\hat{u}^t_{ik} = \sum_{v \in \mathcal{V}} \frac{\hat{\theta}^t_k \, p_{kv}}{\sum_{l=1}^{K} \hat{\theta}^t_l \, p_{lv}} 1\{W_i = v\}$$

In the M-step, we maximize the expected log-likelihood to find an updated estimate $\hat{\theta}_{t+1}$. The components of this estimate are given by

$$\hat{\theta}_k^{t+1} = \frac{\sum_{i=1}^{n} \hat{u}_{ik}^t}{\sum_{l=1}^{K} \sum_{i=1}^{n} \hat{u}_{il}^t}$$

For penalized likelihood estimation with smoothing parameter λ, we modify $\hat{\theta}_k^{t+1}$ by adding λ to the numerator and $K\lambda$ to the denominator. The EM fitting procedure initializes $\hat{\theta}^t$ to some value—we chose $(1/K, \ldots, 1/K)$ in our simulations—then alternates between applying the E-step and the M-step until convergence.

The combination of one E-step and one M-step is an EM iteration. The time complexity of a single EM iteration is $O(VK)$, lower than the complexity for a Newton-Raphson iteration, $O(VK^2)$. This does not necessarily translate to a faster fitting procedure, though, because the EM algorithm can require more iterations to converge. In the context of fitting a mixture model, for example, Lindstrom and Bates (1988) observed that Newton-Raphson fitting was faster than the EM algorithm in certain settings.

We compared the two fitting procedures on our data set, using smoothed estimates with $\lambda = 0.5$. We used both EM and Newton-Raphson to scale all 58 *Dáil* speeches, recording for each the total computation time and the values of $\hat{\theta}^t$ as the algorithms progressed. Then, taking $\hat{\theta}$ as the final value of the Newton-Raphson procedure, we computed the Euclidean distance to the optimum $\|\hat{\theta}^t - \hat{\theta}\|_2$ for each legislator and iteration number. Finally, for each algorithm and iteration number, we averaged the distance values over all speeches.

On average, an Newton-Raphson iteration took about 7.6 times longer than an EM iteration (with a standard error of 0.2). For a fair comparison between EM and ER, we use "EM iteration equivalents" so that 1 Newton-Raphson iteration counts for 7.6 EM iteration equivalents.

In Fig. A2, we see that the two fitting procedures require about the same time to converge. The EM algorithm has faster individual iterations, but on average requires 30–50 steps to converge. The Newton-Raphson procedure has slower iterations, but only requires 4–5 steps. The total computation until convergence is comparable for both.

For our application, the two fitting procedures have comparable performance characteristics. However, the Newton-Raphson procedure computes more than just the estimate $\hat{\theta}$: it also computes the observed Fisher information matrix (the Hessian of the negative log-likelihood). We require this matrix to compute the word influence measure from Sec. 6, and if we were not using bootstrap-based confidence intervals for θ, we would also require this matrix to compute Wald intervals. In situations when either diagnostics or likelihood-based confidence intervals are required, then, the

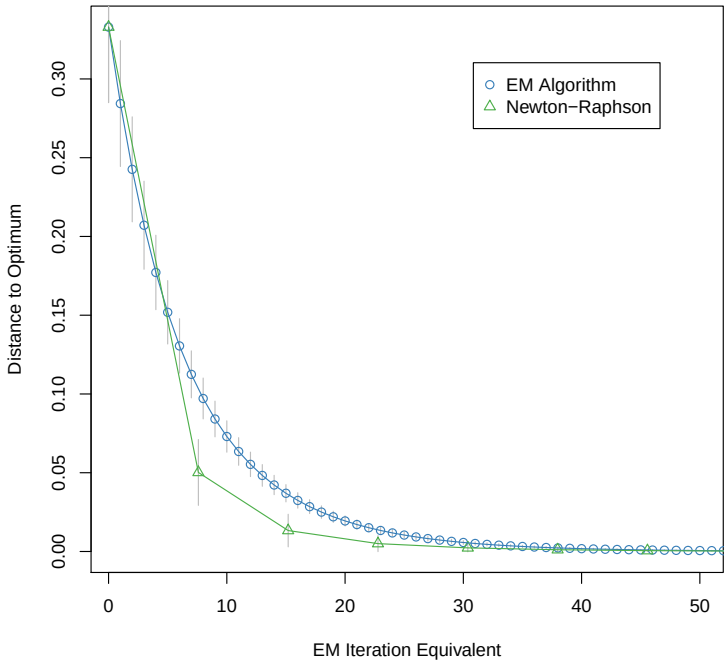


Fig A2: Average Newton-Raphson and EM algorithm performance comparison on the 58 *Dáil* speeches. Gray segments indicate 95% confidence intervals.

Newton-Raphson procedure has a clear advantage.

APPENDIX C: SENSITIVITY ANALYSIS

Our procedure has two tuning parameters: λ, for regularizing the affinity estimates; and α, for regularizing the reference distribution estimates. We set both of these parameters to 0.5, with this value chosen for its connections to the Jeffreys prior and to bias-reduced estimation (Firth, 1993). Despite the theoretical appeal of the value 0.5, we wanted to investigate the sensitivity of the estimates to the choice of its value. Ideally, the estimates should not depend to much on the tuning parameter values.

To investigate the sensitivity, we varied each tuning parameter λ and α separately, choosing 101 evenly-spaced values between 0 and 2. For each choice of the tuning parameter, we computed an affinity estimate $\hat{\theta}$ for each of the 58 *Dáil* speeches. Then, we computed the Spearman correlation between the vector of the 58 values of $\hat{\theta}_1$ and the corresponding vector computed with $\lambda = \alpha = 0.5$. Fig. A3 plots the results. In this figure, we can see that as λ and α varied, the Spearman correlations ranged between 0.996 and 1.000. The rank order of the legislator positions were nearly identical for all choices of the tuning parameters we tried. The results in our application are robust to tuning parameter selection.

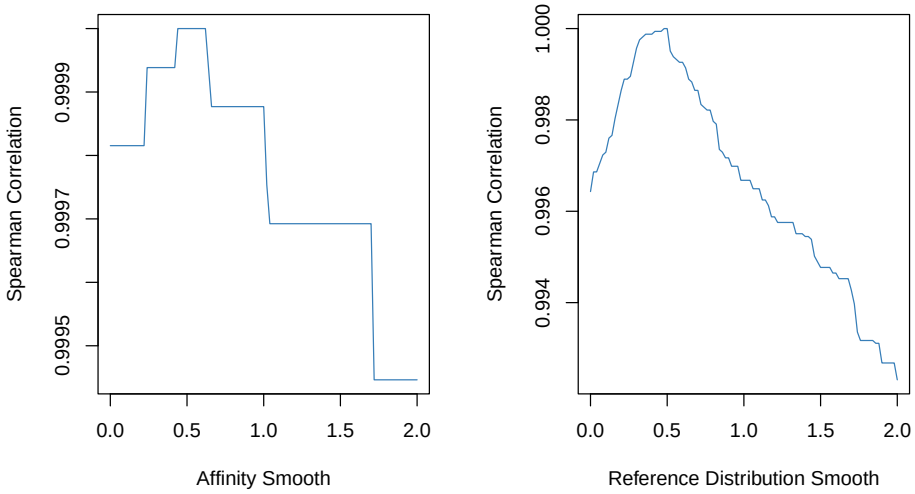


Fig A3: Spearman correlation of the estimates of $\hat{\theta}_1$ for the 58 *Dáil* speeches from tuning parameters $\lambda = 0.5$ and $\alpha = 0.5$ with the estimates when the affinity smooth λ and the reference distribution smooth α vary.

ADDRESSES:
OSCAR HEALTH, 295 LAFAYETTE ST, 6TH FLOOR, NEW YORK, NY 10012;
DEPARTMENT OF METHODOLOGY, LSE, LONDON WC2A 2AE, UK
E-MAIL: pperry@hioscar.com
kbenoit@lse.ac.uk