RESEARCH ARTICLE

The Automatic Analysis of Emotion in Political Speech Based on Transcripts

Christopher Cochrane[a], Ludovic Rheault[a], Jean-François Godbout[b], Tanya Whyte[a], Michael W.-C. Wong[a], and Sophie Borwein[a]

[a]Department of Political Science, University of Toronto, Ontario, Canada; [b]Department of Political Science, Université de Montréal, Quebec, Canada

ARTICLE HISTORY
Compiled June 18, 2023

ABSTRACT
Automatic sentiment analysis is used extensively in political science. The digitization of legislative transcripts has increased the potential application of established tools for the automated analyses of emotion in text. Unlike in writing, however, expressing emotion in speech involves intonation, facial expressions, and body language. Drawing on a new dataset of annotated texts and videos from the Canadian House of Commons, this paper does three things. First, we examine whether transcripts capture the emotional content of speeches. We find that transcripts capture sentiment, but not emotional arousal. Second, we compare strategies for the automated analysis of sentiment in text. We find that leading approaches performed reasonably well, but sentiment dictionaries generated using word embeddings surpassed these other approaches. Finally, we test the robustness of the approach based on word embeddings. Although the methodology is reasonably robust to alternative specifications, we find that dictionaries created using word embeddings are sensitive to the choice of seed words and to training corpus size. We conclude by discussing the implications for analyses of political speech.

KEYWORDS
Text-as-Data; Sentiment Analysis; Legislatures; Word Embeddings.

1. Introduction

Emotion in speech provides a window into a politician's public orientation toward an issue that goes beyond what we can infer from their vote in a legislature (Diermeier, Godbout, Yu, & Kaufmann, 2012; Proksch & Slapin, 2015; Schwarz, Traber, & Benoit, 2017). It is one thing to oppose a bill; it is another thing to be "disgusted" by it. Transcripts of legislative debates are important sources of data for the study of political speech (Soroka, Penner, & Blidook, 2009). The magnitude of these data has spawned a computational turn in the study of parliamentary discourse (Beelen et al., 2017; Grimmer & Stewart, 2013a; Hopkins & King, 2010; Laver & Benoit, 2002; Lowe & Benoit, 2013; Monroe & Schrodt, 2018; Quinn, Monroe, Colaresi, Crespin, & Radev, 2010; Rheault, Beelen, Cochrane, & Hirst, 2016; Rheault & Cochrane, 2020). A key component of this turn involves the automatic detection of emotion (Mohammad &

CONTACT Christopher Cochrane. Email: christopher.cochrane@utoronto.ca. Data and code available at `https://github.com/ccochrane/emotionTranscripts`.

Turney, 2012; Turney & Littman, 2003; Zhang & Liu, 2017). Despite the existence of established tools for detecting emotion in writing, there are at least two reasons to be skeptical of their accuracy for applications involving transcripts of legislative speech. First, most of these tools have been developed for analyses of non-political text, and we do not know if automatic tools for analyzing emotion in other domains—such as Twitter, movie reviews, or news stories—will work for legislatures. The domain specificity of legislative discourse has frustrated automated language analyses in the past (Hirst, Riabinin, Graham, Boizot-Roche, & Morris, 2014). Second, and more generally, we do not know whether a transcript captures the emotion of a speech. Inwardly, emotions are "irruptive motivational states" that "...interfere with the smooth unfolding of plans designed to secure our long-term goals" (Griffiths, 1997, 247). They trigger physiological responses such as heightened heart rate, blushing, watering eyes, increased salivation, cracking voices, and trembling. Emotions also involve outward cultural performances, such as shouting in anger, wailing in grief, or throwing a shoe (Hall, 2015). Unlike writing, the expression of emotion in speech is not confined to word-choice and syntax, and instead relies heavily on intonation (Bänziger & Scherer, 2005), facial expressions (Ekman, Davidson, & Friesen, 1990; Ekman & Friesen, 1969), and body language (Van den Stock, Righart, & de Gelder, 2007), which legislative transcripts do not record (Knox & Lucas, 2019). For these reasons, we expected transcripts to perform relatively poorly as sources of information about the emotional content of political speech.

We ask two central questions about measuring emotion from transcripts. First, do transcripts capture the emotion originally expressed in the corresponding video records of political speeches? And second, how well do existing tools for text-based sentiment analysis perform relative to human judgment? For each question, we provide concrete empirical evidence that can inform researchers interested in the study of emotion using speech transcripts. Our results make use of an original collection of televised parliamentary speeches from the Canadian House of Commons between 2015 and 2017, which we aligned with their official transcripts and supplemented with human annotations. The transcripts come from the written Hansard, the official record of parliamentary debates in Canada. The video record is maintained by the *Canadian Parliamentary Access Channel*.

Our analysis proceeds as follows. In the next section, we examine whether coders watching video clips of parliamentary debates perceive the same emotional content as do coders reading the corresponding entries in Hansard. We find that coders reliably capture the same sentiment (negative vs positive) regardless of whether they code from the video or text record. We also find, however, that judgements about the level of emotional activation (aroused vs subdued) are not consistent across text and video. In short, the sentiment of the speech is in the transcript, but the emotional arousal is not. We turn in the third section to analyze different tools for the automatic detection of sentiment in text. We find that although leading dictionary and supervised approaches predicted human judgments reasonably well, automatically generated sentiment dictionaries based on word embeddings surpassed these other approaches. This methodology leverages the ability of word embeddings to identify semantically related words, and we elaborate on the conditions needed for the induction of dictionaries. The penultimate section of the paper scrutinizes the sensitivity of this method. We demonstrate that sentiment dictionaries based on word embeddings can sometimes work across domains and are robust to different parameters. We do find, however, that the model is sensitive to very small corpus sizes and to choices of seed words. We conclude by discussing the implications of our findings for political communication

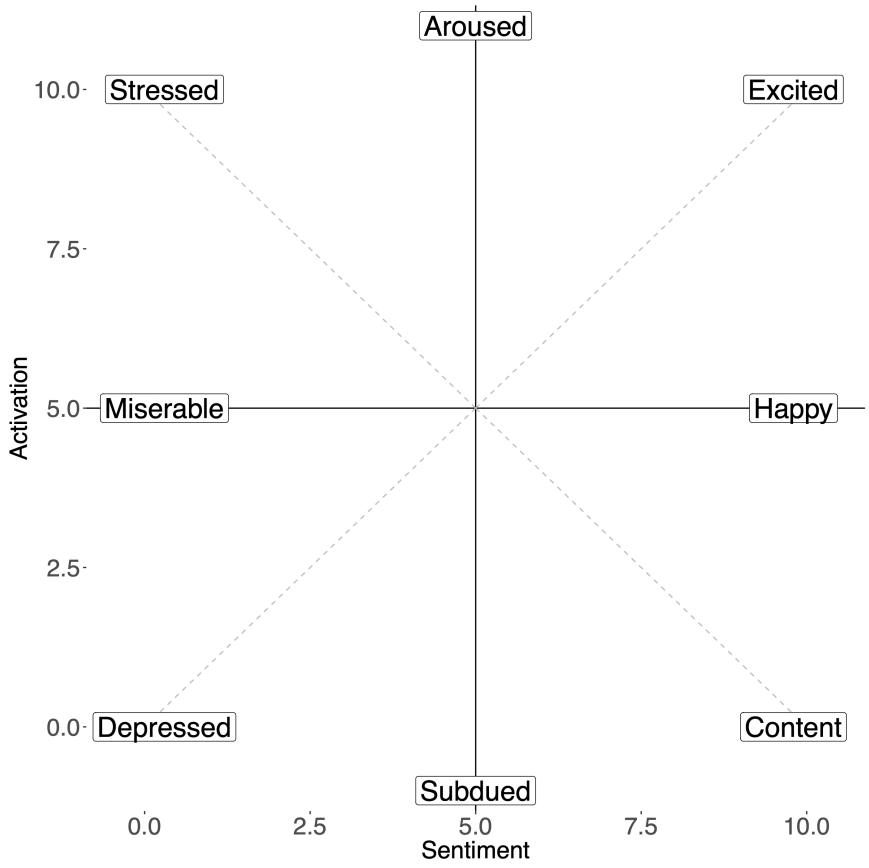
research.

2. Human Judgments of Text and Video Records

Studying emotion in detail is highly complex and undoubtedly extends beyond what we can capture with transcripts of speeches. But can we use these data to capture even basic aspects of emotional expression? The simplest model of emotion is the two-dimensional "core affect" model represented in Figure 2 (J. Russell, 1980; J. Russell & Barrett, 1999). The horizontal dimension captures the "valence" or "sentiment" of the speech—i.e., whether the speech expresses a positive or a negative feeling. The vertical dimension captures "arousal" or "activation" (Schlosberg, 1954)—i.e., whether the speech is animated or subdued. A person who expresses a negative feeling and is extremely animated about it—as in anger, for example—would be positioned in the upper left corner of Figure 2. A person expressing a positive feeling in a very calm way—as in contentment, for instance—would be in the bottom right corner. The bottom left captures negative/subdued, or "depressed," speech; the top right is positive/animated, or "excited," speech. Frustration, sadness, disappointment, and other types of emotional expression occupy middle-of-the-road positions on these dimensions. This is a parsimonious model of emotion. The key point for present purposes is that identifying emotion in speech implies capturing the positions of utterances on at least these two dimensions.

What explains the position of an utterance on these dimensions? Certainly, words play a role. In a meta-analysis, Brooks et al. (2017) show that language, and specifically the naming of particular emotions such as disgust and anger, causes brain activity associated with the retrieval of memories about those emotions, which helps people to more quickly resolve ambiguous affective states. But non-verbal signals also matter. Scherer, Ladd, and Silverman (1984) studied how intonation affects people's perceptions of speech by experimentally altering the audio signals to preserve different aspects of its acoustic profile. They found that intonation conveys some emotion independently of word choice, magnifies other emotions in parallel with word choice, and conveys still other emotions in interacting with word choice. Facial expression matters as well (Ekman, 1992; Ekman et al., 1990; Ekman, O'Sullivan, Friesen, & Scherer, 1991), although a study by Meeren, van Heijnsbergen, and de Gelder (2005) of how people perceived emotion in hybrid images of face-body compounds found that judgments about emotional expressions in discrepant images were heavily biased toward the emotion conveyed in body language rather than the face. Perceptions of emotion in body language may also occur immediately, prior to conscious consideration (de Gelder, de Borst, & Watson, 2015). Van den Stock et al. (2007) find that emotional signals are clearer when facial expression, body language, and intonation all run in the same direction, as they normally do, and are distorted when these cues run in different directions. The general question of precisely how tone, face, body, and words interact to produce emotional signals is a complicated question beyond the scope of this paper. The salient role of non-verbal signals, however, is clear.

Given these findings, we expect the transcripts of parliamentary debates to perform relatively poorly as conveyors of emotional expression compared to human judgments based on full observation of the speeches. The gold standard for our analysis is human judgment of the video of speech fragments from the record of Question Period maintained by the *Canadian Parliamentary Access Channel*. We use sentences as the unit of analysis because it is the smallest natural unit to convey meaning in speech,



Source: James Russell. 1980. "A Circumplex Model of Affect." *Journal of Personality and Social Psychology* 39(6): 1161-1178.

Figure 1. Core Affect Model of Emotion

and we want to measure the performance of coders when they are focused as much as possible on the same content. The length of full speeches is too variable. A common challenge in studies based on large segments of text is that they often report low levels of intercoder reliability, or report none at all, which is understandable because coders may choose to focus on different parts of a longer text. Although collections of a fixed small number of sentences work well (Proksch, Lowe, Wäckerle, & Soroka, 2019), these would sometimes cut across entire speeches in our case, and there is significant variability in the length of paragraphs recorded in Hansard. Unlike in writing, moreover, the delineation of paragraphs as they appear in Hansard is decided by someone recording the speech rather than someone delivering it. Although this is true for sentences as well, the sentence is by far the most consistent natural unit of analysis in Hansard.

We recorded every third Question Period between January 2015 and December 2017, covering the last 10 months of the Conservative government of Stephen Harper and the first 23 months of the Liberal government of Justin Trudeau. These videos are all of the same (1080p) quality and in identical formats. The videos were precisely trimmed from the start of the first question to the end of the last answer, yielding 102 videos of about 45 minutes in length, on average. We figured 1020 speech fragments would be sufficient for measuring intercoder reliability at a reasonable level of precision. From each of these

videos, ten time-points ($mm : ss$) were selected at random. The sentence beginning just prior to each selected time-point was then extracted to form its own video clip. We used the punctuation in the official written record of the original language to determine the precise start and end of each sentence. In cases where the same sentence overlapped two of the randomly generated time-points, the previous sentence was also used.[1] The average length of an extracted video clip was just under nine seconds long, and the average clip contained 23 words. These clips were uploaded to *YouTube* and added to a *Qualtrics* survey instrument administered to three independent, bilingual coders. The ordering of the clips was randomized. For each video, the coders were asked to score on eleven-point scales (0-10) the sentiment (negative-positive) and activation (subdued-aroused) of the speech fragment (See Appendix A). The randomized presentation means that the same video was often presented to the same coder at different times.

We were able to identify the official Hansard record for all of the video speech excerpts, except one.[2] For speeches in French, we used the official English translation, which is the corpus on which all computational analysis to date has been conducted (but see Duval & Pétry, 2016). The texts of these speech fragments were then presented in random order to three different independent coders, who were given precisely the same instructions as the video coders, and were asked to indicate the sentiment and activation for the speech fragments on the same eleven-point scales. The text coders were also asked to come into the lab and code a randomly re-ordered version of the same instrument, three months after they had submitted their initial scores. This permitted analysis of the cross-time consistency of the text coders' judgments.

All of the coders were doctoral students, but none of the coders were aware of the hypotheses or topic of study. The coders worked independently and there was no calibration exercise of any kind. For both the text and video coders, two of the coders were male and one female. We provided identical instructions to both groups of coders, which are presented in the Appendix. For a detailed discussion of the implications of different methods of human coded content-analysis, see Weber et al. (2018).

Table 1. Summary of Coder Judgments

	Sentiment		Activation		
Coder	$\bar{x}$	σ	$\bar{x}$	σ	N
V1	4.3	1.5	5.5	1.3	1045
V2	4.8	1.8	5.9	1.2	1009
V3	4.7	1.1	5.5	1.1	519
T1	4.7	2.7	6.5	0.7	2040
T2	4.9	2.2	5.7	1.3	2039
T3	4.6	1.6	5.9	1.0	2033

Table 1 summarizes the judgments of text and video coders. Notice that the means for sentiment and activation were consistently below and above five, respectively. The coders perceived the sentiment of the speech fragments to be somewhat more negative than positive and more aroused than subdued. The last column is the number of

[1] In three cases, there were minor distortions in the indicated videos (email dings). The previous sentence was used for those cases as well.

[2] The missing speech implied that government ministers were taking bribes, and we suspect it was withdrawn from Hansard by the Member.

coded videos/transcripts received from each coder. Recall that the text coders coded from randomly ordered lists of transcripts for all of the selected videos, whereas the video coders coded clips drawn at random from the database of all videos. The coders were also able to move much more efficiently through text than through video. The turnaround times from the video coders were longer than for the text coders, and we received twice as many codes from the text coders.

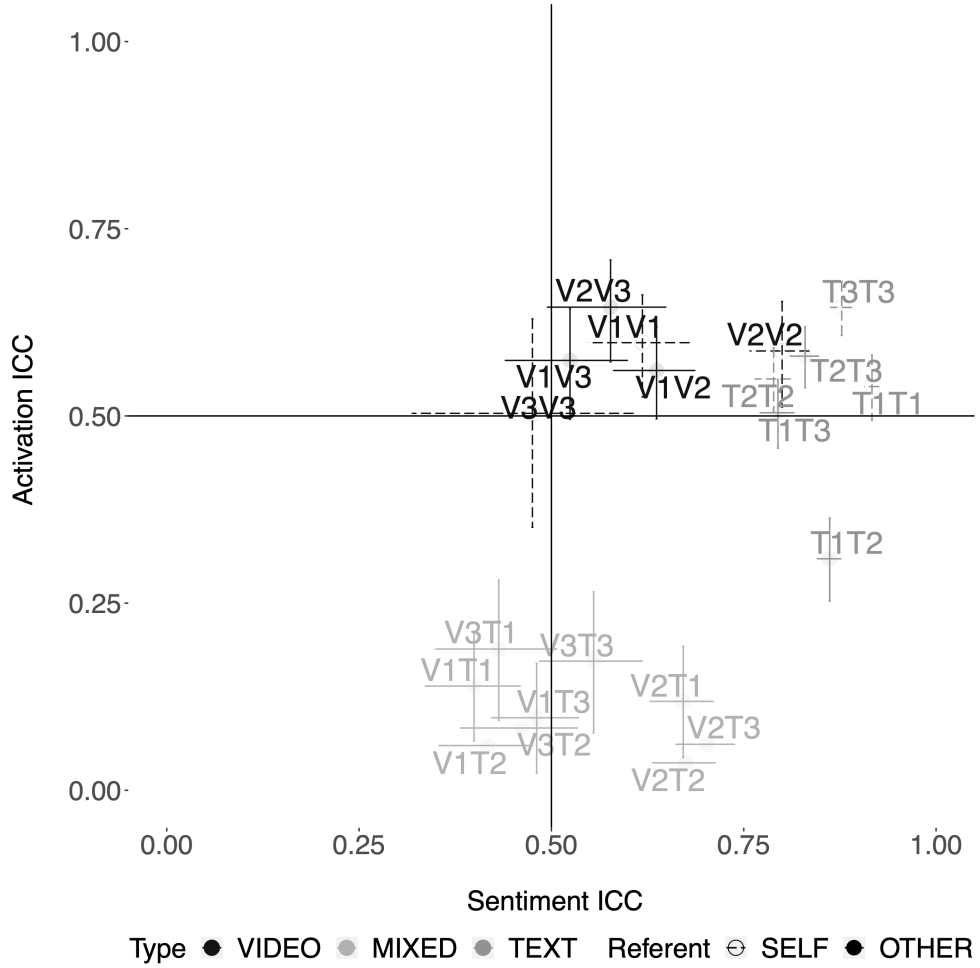


Figure 2. Coder Reliability of Text and Video Coders

Figure 2 summarizes the intraclass correlation coefficients between the average of the first two scores given to each sentence by each pair of coders, and also between the first and second score given to each video/transcript by the same coder at different points in time.[3] The horizontal axis captures the correlation coefficients for sentiment scores, and the vertical axis captures the correlation coefficients for activation scores. We consider intraclass correlation coefficients greater than $r \approx 0.4$ to be decent indicators

[3]Although video coders may have coded the same video more than twice, we restricted the analysis to their first two scores because the text coders coded each snippet only twice.

of reliability. Anything below .4 is normally considered to be poor, and anything above .6 is considered to be good (Cicchetti, 1994, 286).

The labels in Figure 2 indicate the specific pair of scores in the analysis—e.g., V2V3 implies that the comparison is between the scores of the second and third video coder. For ease of interpretation, when the 95% confidence intervals are represented as dashed lines it means that the correlation is between the scores given by the same coder at different times, whereas confidence intervals represented as solid lines convey that the correlation is between different coders. The shading in the graph corresponds to whether the correlation is between the scores of two video coders (VIDEO), two text coders (TEXT), or between a video coder and a text coder (MIXED). Thus, for example, the position of V2V2 in the top right quadrant indicates that the second video coder was highly reliable in her first and second scoring of each video on both the sentiment ($r = .80$) and activation ($r = .59$) dimensions. The scores of V1T2 in the bottom left quadrant indicate that the first video coder and second text coder were somewhat reliable in their sentiment scores ($r = .42$), but not at all reliable in their activation scores ($r = .06$).

First, notice from the position of the points with dashed confidence intervals that each coder consistently adjudged the sentiment and activation of each speech fragment. The first score that each coder gave to a speech fragment was consistent with the subsequent score that they gave to that fragment, and this is true for both the text and video coders, and for both sentiment and activation scores. All but one of the self-referential dyads are in the top right quadrant, indicating high reliability on both dimensions. The second finding is that text coders were highly consistent with one another when it comes to their sentiment and, with one exception (T1T2), their activation scores, and so were the video coders. The text coders were especially consistent in their sentiment scores. Finally, the third and most important finding is that while the sentiment scores of video and text coders tended to align, their activation scores did not align at all. Notice the position of the light grey labels, which invariably occupy much higher scores along the horizontal (sentiment) axis than the vertical (activation) axis. There is no approximately acceptable level of inter-coder reliability between the activation scores of any text coder and any video coder.

3. The Automated Detection of Sentiment in Text Transcripts

We now turn to the efficacy of tools for the automatic detection of sentiment in text. Dictionaries—or "lexicons," as they are typically called in other disciplines—and supervised machine learning are two broad classes of tools for automating the detection of sentiment in text.[4] A sentiment dictionary is a list of positive and negative words or n-grams (sequences of words), often curated manually by researchers. The sentiment of a text can be calculated as a function of the balance of its n-grams that match the positive and negative lists in the dictionary. Although the initial creation of a dictionary is time-consuming and labor-intensive, dictionaries are easy to implement and rely on highly informed human judgment. We test five widely used sentiment dictionaries in our analysis: Lexicoder 3.0 (Daku, Soroka, & Young, 2015; Young & Soroka, 2012), Sentiwordnet 3.0 (Baccianella, Esuli, & Sebastiani, 2010; Esuli & Sebastiani,

[4]For general summaries of these methods and their applications, see for instance Quinn et al. (2010), Cambria, Schuller, Xia, and Havasi (2013), Grimmer and Stewart (2013a), Wilkerson and Casas (2017), and Benoit (2019).

2006), the Hu-Liu Lexicon (Hu & Liu, 2004), Jockers-Rinker's Lexicon (Jockers, 2015; Rinker et al., 2019), and VADER (Hutto & Gilbert, 2014).

Supervised machine learning induces a function from the patterns of features (e.g., words) in data already classified by humans, and then tests that function on other data for which only the features, and not the classifications, are known to the algorithm (Pang & Lee, 2008; Pang, Lee, & Vaithyanathan, 2002). For sentiment analysis, the weight of an n-gram is the degree to which it signals a positive rather than a negative classification for the texts in a corpus, within the constraints of some model. These weights are tested by predicting classifications of out-of-sample texts for which the true classification is known to the researcher. Applying these weights to words in as-of-yet unclassified texts facilitates predictions about whether that text is generally positive or negative. Support Vector Machines (SVM) are among the most popular models in this category (Vapnik, 1998). The SVM's optimization problem is to find the hyperplane in a high-dimensional vector space that maximizes the width of the margin between the hardest-to-classify cases (i.e., the support vectors) with the option to modify the shape of the classification boundary (i.e., the kernel). The weight of a feature—for example, an n-gram—captures the direction of its effect and its salience for separating cases along the maximum margin hyperplane (Winston, 2014).

Supervised learners are not predetermined by the initial judgments of researchers to nearly the same degree as manually curated dictionaries. Nonetheless, the advantage of dictionary approaches over supervised learning is that dictionaries are suitable to small corpora. The number of syntactically allowable combinations of words in a language means that establishing with any degree of certainty that an n-gram appears more commonly in one class of text rather than another requires a corpus large enough for key words to occur many times. The most common words in a language, moreover, are usually the least discriminating (Manning, Ragahvan, & Schutze, 2009). Although the corpus of parliamentary speech is more than large enough to satisfy the size requirements of machine learning, the problem for sentiment analysis is that, unlike for movie reviews (Chaovalit & Zhou, 2005; Ennedy & Nkpen, 2006; Pang & Lee, 2008), customer feedback forms (Lak & Turetken, 2014) or tweets with emojis (Kralj Novak, Smailović, Sluban, & Mozetič, 2015), there is no extensive human classification of parliamentary speech with which to train the models. Nor is there any guarantee that models trained on established corpora will work for the parliamentary domain. Hand-annotating parliamentary data would also defeat the purposes of using machine learning in the first place; it would be as time-consuming as building a dictionary, predetermine the results to the same extent, and be just as inflexible across time and domains. To be sure, the rise of language models relying on deep neural networks may soon eclipse traditional classifiers for sentiment analysis. Transformers such as the BERT model (Devlin, Chang, Lee, & Toutanova, 2018) have been used intensively after achieving state-of-the-art performance on a variety of tasks in computational linguistics, while offering portability across domains by fine-tuning classifiers on local corpora. The downside, however, is that the proper interpretation of influential features requires a strong familiarity with the field of deep learning. In disciplines where substantive interpretation matters, such as political science, we believe that transparent approaches like dictionaries will remain a desirable option.

We also test a third and increasingly common approach which involves creating dictionaries automatically, without the need for hand-made curation. We use the word2vec algorithm to generate distributed vector space representations of words—commonly known as "word embeddings" (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). Next, we take advantage of a property of these embeddings, namely their ability to

identify semantically related words, and generate a comprehensive list of words to populate a sentiment dictionary. As we illustrate below, this approach outperforms alternatives, is robust across some domains, and it can be very easily implemented at low cost.

The word2vec algorithm uses a shallow layer neural net to predict the occurrence of words on the basis of surrounding words (Mikolov, Sutskever, et al., 2013).[5] The weights (or coefficients) from this model become the word embeddings, representations with the convenient property that words commonly used in the same contexts are close to each other in the vector space. This means that distance metrics such as cosine similarity—based on the angles between vectors—can be used to uncover semantic associations between words. In this way, we can say that the algorithm embeds the contextual likeness of words in the angular distance between their corresponding vectors. Consistent with the word-use theory of meaning (Wittgenstein, 2009), mathematical operations on these vectors yield results that correspond remarkably well with meaningful semantic patterns in language; for instance, analogies. Canonically, models trained on large corpora are able to solve that the vector for "king" minus the vector for "man" plus the vector for "woman" approximates the vector for "queen," hence capturing the semantic relatedness between pairs of words (Mikolov, Corrado, Chen, & Dean, 2013; Pennington, Socher, & Manning, 2014).

The corpus for this stage of the analysis is the 77,730,436 tokens spoken in the Canadian House of Commons from the beginning of the 39th Parliament (January 29, 2006) and into the 42nd Parliament (April 19, 2018). The 39th Parliament is when the House of Commons first provided a structured machine-readable Hansard that can be scraped by researchers using code for replication. Prior to this, it is necessary to use Optical Character Recognition and extensive curating to extract the text and metadata of the Debates. Algorithms for generating word embeddings have hyperparameters that researchers can adjust for the purpose at hand. These hyperparameters include the number of dimensions in the vector space in which words are to be positioned, the size of the word window used to capture word co-occurrences, and a minimum threshold for dropping idiosyncratic words that appear very rarely in a corpus. Researchers can also specify the number of iterations (epochs), which specifies the number of times the algorithm passes through the corpus to optimize and re-optimize the position of words in the vector space. As we have no *a priori* expectations about the optimal settings of these parameters, and do not want to tune an algorithm on data that will be used for testing, we left all parameters at their default values. We thus use 300 dimensions, a word window of six words, a minimum word count of 10, and used five iterations. We also remove stopwords. After these adjustments, we have an effective corpus size of 49,713,429 tokens and a vocabulary of 40,597 unique words.

3.1. Dictionary Induction Using Word Embeddings

After training the word embedding model on parliamentary speech, we use positive and negative "seed words" to quickly induce a domain-specific sentiment lexicon. The remainder of this section traces the origins of this approach to the creation of dictionaries, before discussing its theoretical underpinnings.

[5]There are two popular variants. The Continuous Bag of Words (CBOW) algorithm assigns vectors to maximize the likelihood of a word appearing, given its context. The Skip Gram algorithm assigns vectors to maximize the likelihood of contexts appearing, given each word. We use the CBOW algorithm.

Using seed words for creating sentiment dictionaries began with Turney (2002), and was later elaborated upon in Turney and Littman (2003). Turney's objective was to quantify the semantic orientation of phrases with respect to reference words representing positive and negative sentiment, inspired by earlier work from Hatzivassiloglou and McKeown (1997) on predicting the valence of adjectives.[6] Turney and Littman (2003) considered two canonical methods in information retrieval for measuring relationships between words, namely pointwise mutual information (PMI) and latent semantic analysis (LSA). Whereas PMI measures the probability of word co-occurrences, LSA relies on singular value decomposition (SVD) to produce a low-rank approximation of a corpus matrix (the document-feature matrix). The reduced SVD matrix contains numerical vectors for both documents and words: document vectors will be similar to each other when they contain similar words, and word vectors will be similar to each other when they appear in the same documents. Turney and Littman relied on these properties to rank words from a corpus vocabulary as a function of their similarity with a manually curated list of seed word pairs. The SVD variant of their approach was recently replicated by Hamilton, Clark, Leskovec, and Jurafsky (2016).

New distributed representations of words such as Mikolov, Corrado, et al. (2013)'s *word2vec*, discussed above, and Pennington et al. (2014)'s *GloVe* vectors, were natural candidates for extending Turney (2002)'s approach. To our knowledge, the first attempt to adapt Turney's model of dictionary induction using word embeddings came from political scientists, who used the methodology for studying legislative speeches (Rheault, 2016; Rheault et al., 2016).[7] Rheault et al. (2016) relied on a customized implementation of the GloVe vectors fitted on a corpus of lemmas (the canonical form of words) and their associated parts of speech, with the objective of disambiguating sentiment dictionaries when applied to political texts. A similar technique was proposed independently in Fast, Chen, and Bernstein (2016) for the induction of dictionaries beyond sentiment. Many related approaches—using seed words and embeddings—have been used for the unsupervised induction of multilingual dictionaries (Ruder, Vulić, & Søgaard, 2019, for an overview, see), and a similar method was proposed by Rice and Zorn (2019) for studying Court opinions.

From a theoretical standpoint, the reliance on *pairs* of seed words to generate sentiment dictionaries is essential to overcome the problem of opposite words (loosely speaking, antonyms) occurring around the same context words. Expressions of sentiment with an opposite valence, for instance the words "support" and "oppose", can often be found in identical contexts. As an illustration, consider the following sentences:

I urge my colleagues to *support* this bill.
I urge my colleagues to *oppose* this bill.

A strict adherence to the distributional hypothesis—"[w]ords that occur in similar contexts tend to have similar meanings" (Turney & Pantel, 2010, 153)—may lead to the incorrect conclusion that "support" and "oppose" share the same meaning. Unsurprisingly, a well documented limitation of word embeddings is that searching for expressions most similar to "good" will often yield opposite words entangled with synonyms (see e.g. Tang et al., 2014). The problem plagues most models that rely on word co-occurrences. Scholars studying sentiment are not merely interested in finding

[6]Pang and Lee (2008) summarize the history of this development.
[7]Amir et al. (2015) also used word embeddings to predict the sentiment of Twitter terms using a labeled set of words and phrases, but using a regression-based approach.

expressions that relate to sentiment, but also to organize them along a scale that reflects polarity. As a result, a necessary step for sentiment analysis is to characterize each pole of the dimension of interest.

The solution in Turney's approach is to rely on seed word *pairs* to resolve the underlying relation between opposite words. By identifying two words that represent opposites on a dimension of interest—sentiment—the displacement vector between the pair of seed words measures the difference between two antithetical points on that dimension. A similar problem is involved in the resolution of analogies with word embeddings. Just like the subtraction of the vector for "queen" from "king" captures an opposition in terms of gender, the difference between word vectors for "good" and "bad" measures opposite poles reflecting sentiment.[8] How similar are the word vectors relative to the displacement vector—and not to each seed individually—is what allows the researcher to reveal whether they are closer to either end of the dimension under study. Fast et al. (2016) faced a similar problem when creating generic topic dictionaries from word embeddings. Because they rely on a single list of seed words, they had to hire crowdsource workers to remove antonyms and other misclassified words. In contrast, our approach does not require human interventions, beyond the selection of seed word pairs, but it is applicable only to theoretical dimensions for which opposite words are available. When validating this approach, we find that simply taking the difference between cosine similarities relative to each pair of seed words performs at least as well—if not better—than projecting word vectors onto the displacement vector per se. Since a single pair of seed words (say, "good" and "bad") may not capture the sentiment dimension perfectly due to corpus noise, we also examine how many seed words are sufficient to produce a stable sentiment dictionary. We return to this question when exploring the robustness of dictionaries generated using word embeddings.

In geometric terms, we build on the assumption that, for any given word, if the angles formed between the vectors for positive words are, on average, smaller than the angles formed between the vectors for negative words, then the word is semantically closer to positive than to negative words, and we infer from that a positive sentiment for the word. The opposite is true if the word's vector forms smaller angles with the vectors for negative words than for positive words. The dot product of two unit vectors is equal to the cosine of the angle between the vectors. For vectors $\vec{\mathbf{v}}$ and $\vec{\mathbf{w}}$, the cosine of the angle θ between them—the cosine similarity—is given by:

$$\cos\theta = \frac{\vec{\mathbf{v}} \cdot \vec{\mathbf{w}}}{\|\vec{\mathbf{v}}\|\|\vec{\mathbf{w}}\|} = \frac{\sum_{i=1}^{n} v_i w_i}{\sqrt{\sum_{i=1}^{n} v_i^2}\sqrt{\sum_{i=1}^{n} w_i^2}} \tag{1}$$

The range of $\cos\theta$ is [-1,1], with -1 for angles of 180° (i.e., when the vectors point in opposite directions, indicating dissimilarity) to 1 for angles of 0° (i.e., when the vectors point in identical directions, indicating semantic identity). The value of $\cos\theta$ is 0 when the vectors are orthogonal to each other at 90°, indicating semantic independence. Taken together, larger cosine similarities indicate that two words are more semantically alike than words with smaller cosine similarities. With a list of $p = 1, 2, \ldots, a$ positive seed word vectors $\vec{\mathbf{v}}_p$, and a list of $q = 1, 2, \ldots, b$ negative seed word vectors $\vec{\mathbf{u}}_q$, we can therefore define a function S for the sentiment of any word $\vec{w}$ as:

[8]For a detailed examination of the relevance of arithmetic operations performed on words embeddings, see Ethayarajh, Duvenaud, and Hirst (2018).

$$S(\vec{\mathbf{w}}) = \sum_{p=1}^{a} \frac{\vec{\mathbf{w}} \cdot \vec{\mathbf{v}}_p}{\|\vec{\mathbf{w}}\| \|\vec{\mathbf{v}}_p\|} - \sum_{q=1}^{b} \frac{\vec{\mathbf{w}} \cdot \vec{\mathbf{u}}_q}{\|\vec{\mathbf{w}}\| \, \|\vec{\mathbf{u}}_q\|} \tag{2}$$

The sentiment of a word is the sum of its cosine similarities to a set of positive seed words minus the sum of its cosine similarities to a set of negative seed words.

For seed words, we adapted a list of positive and negative words from Turney and Littman (2003). Our positive seeds are "good", "excellent", "correct", "best", "happy", "positive", and "fortunate"; our negative seeds are "bad," "terrible," "wrong," "worst," "disappointed," "negative," and "unfortunate." From Equation 2, we generated a domain specific sentiment lexicon by calculating the sentiment of each word uttered in parliament, $S(\vec{\mathbf{w}})$, as the sum of its cosine distances from the set of positive seeds minus the sum of its cosine distances from the negative seeds. To restate, words with positive values for $S(\vec{\mathbf{w}})$ have positive valence and words with negative values have negative valence. Table 2 summarizes the 30 most positively and negatively valenced words in our lexicon. The words in the left-hand column are those, on balance, most likely to appear alongside other positive words and least likely to appear alongside negative words. Words in the right-hand column reflect the opposite. We sum the valences of the words in the transcripts of video clips to generate a sentiment score for each sentence. Accounting for negation words (valence shifters) and modifiers did not affect performance, and so we exclude them for reasons of parsimony. Dividing the sentiment of each sentence by the number of words also did not matter, though we would normally do so with longer speech fragments.

3.2. Validating Methods for Sentiment Analysis

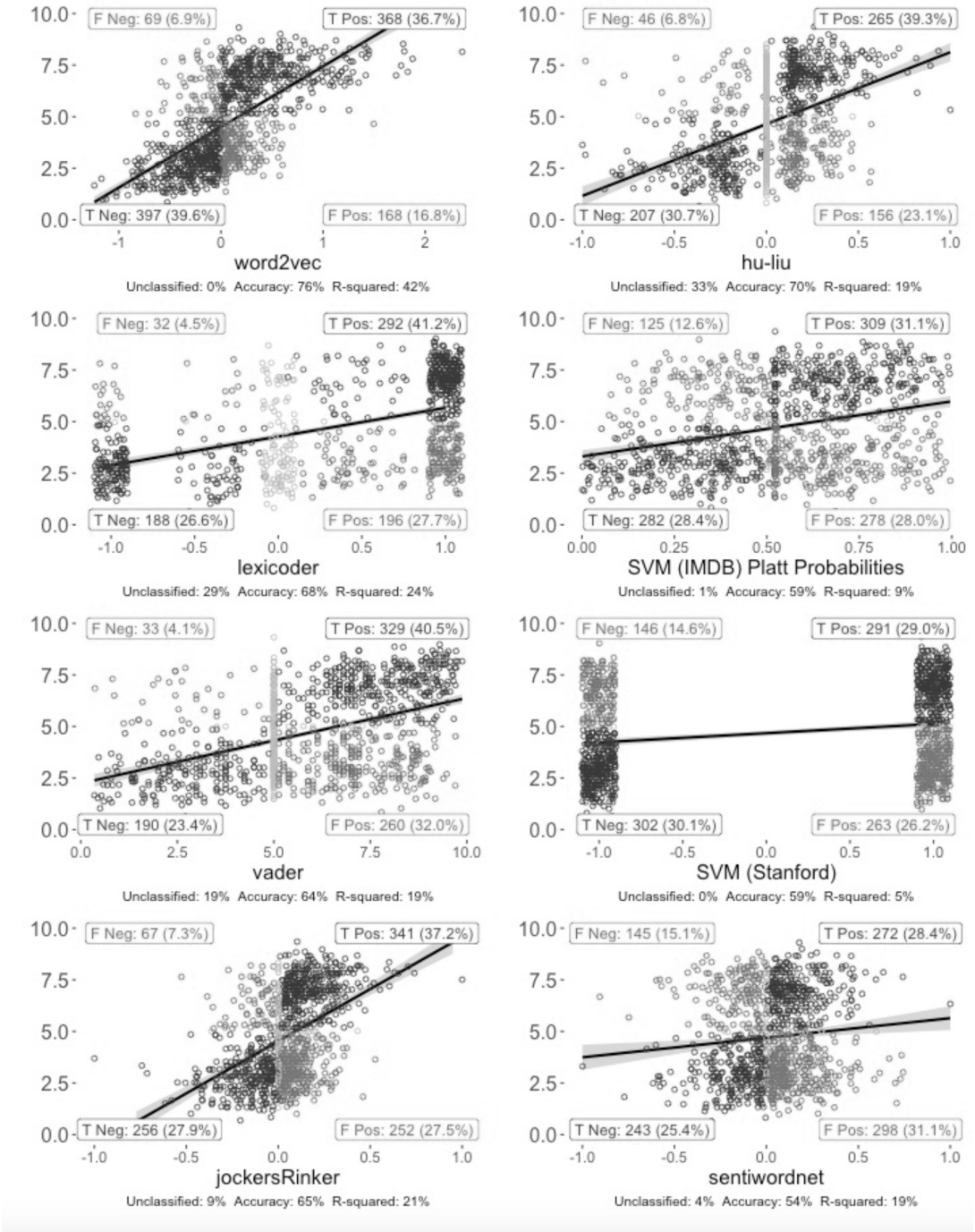
Figure 3 compares tools that we tested for predicting the sentiment scores of our human text coders, including the dictionary induced using word2vec embeddings described above. For each measure, the confusion matrix overlays a jitterplot to show the proportional reduction in error alongside the classification accuracy of the different tools. For the measure of accuracy, we consider a classification successful if the method produces a score in the positive range—past the midway point—while the human coders also coded a text as positive on average; the opposite must be true for negative predictions. This corresponds to the percent correctly predicted commonly used in binary classification tasks.

Some of the dictionary tools performed well overall, but results were invariably mixed. Lexicoder was attuned very effectively to negative sentiment, but not as well to positive sentiment. Notably, many of the video transcripts contained no words in the Lexicoder, Hu-Liu, and Vader dictionaries. In total, 33 percent of the sentences were unclassified by Hu-Liu, 29 percent by Lexicoder, and 19 percent by Vader. Among the dictionaries, Sentiwordnet classified the largest proportion of the sentences, but it was also the least accurate.[9]

[9]For each approach, unclassified sentences are not included in the calculations of accuracy and R-squared. We exclude them because their inclusion as "neutral" classifications substantially reduces the performance of these dictionaries, and, for the purposes of comparing established dictionaries to the approach based on word embeddings, we wanted to represent the established dictionaries in the strongest possible

Table 2. Most Valenced Words in Induced Sentiment Lexicon

	Positive Words		Negative Words	
1.	excellent	.2630	horrible	-.3066
2.	mentorship	.2218	terrible	-.2936
3.	highquality	.2189	panic	-.2915
4.	outstanding	.2128	horrendous	-.2842
5.	invaluable	.2071	outrageous	-.2829
6.	innovative	.2012	disastrous	-.2819
7.	midwives	.2001	despicable	-.2758
8.	talents	.1978	scandalous	-.2742
9.	topnotch	.1942	horrific	-.2724
10.	collaboratively	.1917	inexcusable	-.2691
11.	fortunate	.1903	horribly	-.2635
12.	secure	.1899	indiscriminate	-.2586
13.	productive	.1894	immoral	-.2565
14.	develop	.1892	devastating	-.2559
15.	welcome	.1880	senseless	-.2551
16.	excellence	.1877	callous	-.2551
17.	ethic	.1877	deception	-.2545
18.	continue	.1872	unspeakable	-.2539
19.	worldclass	.1872	unfortunate	-.2524
20.	strengthen	.1869	unjustified	-.2510
21.	cooperative	.1865	inhumane	-.2500
22.	constructive	.1863	inexplicable	-.2493
23.	dedicated	.1839	atrocious	-.2486
24.	thoughtful	.1820	shameful	-.2483
25.	tirelessly	.1780	irresponsible	-.2474
26.	cooperatively	.1767	disgraceful	-.2468
27.	collaborative	.1763	reprehensible	-.2466
28.	build	.1762	disgusting	-.2463
29.	happy	.1752	pernicious	-.2463
30.	constructively	.1731	meanspirited	-.2459



Notes: Jitterplots and predictive accuracy scores for different sentiment analysis methods, validated against human-based annotations. "T Pos" stands for true positives, "T Neg" for true negatives, "F Neg" for false negatives, and "F Pos" for false positives. The "word2vec" label indicates the sentiment dictionary generated using word embeddings. See text for details.

Figure 3. Predicting Human Judgments of Sentiment in Text

As we discussed earlier, there is no pre-classified or human annotated Hansard on which to train supervised learners, and manually annotating Hansard would be time-consuming, costly, and potentially predetermine the results to the same extent as human curated dictionaries. Nonetheless, it might be possible to leverage models trained on large annotated corpora from other domains. Thus, we also tested a number of supervised learners trained on tweets from the annotated Stanford Twitter corpus and movie reviews from the IMDB. The top performing tool in this class was the Support Vector Machine trained on the IMDB reviews with the prediction transformed to Platt probabilities, which is intensive computationally. The SVM classified all but one percent of the sentences. Despite the high accuracy in the domains in which these learners were trained, this approach was less accurate than most of the dictionaries that we tested. Given this, we find little evidence to support the strategy of classifying the sentiment of parliamentary speech with a SVM model trained on annotated corpora from either of these other domains.[10]

The top left panel of Figure 3 summarizes the effectiveness of the sentiment lexicon induced from word embeddings. When it came to predicting human judgments, the lexicon's classifications were more precise, accurate, and specific than any of the other measures we tested. The accuracy rate is 74%, and the measure explained nearly twice as much of the variation in human judgments as any other tool, except Lexicoder. It also classified all of the sentences in our corpus. The approach based on word embeddings was as comprehensive as the supervised learners tested and more accurate than the best dictionaries.

4. Sensitivity Tests

We now explore how the performance of sentiment dictionaries based on word embeddings is affected by domain specificity, choice of seed words, chance, and corpus size.

Figure 4 summarizes results comparing alternative specifications for predicting the judgments of human coders about the sentiment of a text. The vertical axis represents the R-squared associated with each variant. The bars correspond to the different model specifications. For example, the sentiment lexicon created with all of the seed words we outlined earlier, using word embeddings trained on the Hansard corpus, is the best performing method, explaining over 41 percent of the variation in human judgments. When relying on word embeddings trained on the Google News corpus instead (the

---

way. In response to a helpful suggestion, we also experimented by including in the Lexicoder analysis the entire paragraph surrounding the sentences that we extracted. In effect, this meant that a greater number of words would align with the Lexicoder dictionary, which could conceivably result in a better classification of the context of the sentence in our analysis. In following this approach, we found a small decrease in the accuracy of Lexicoder's classification and in the amount of variance that it explained, but an appreciable decrease, from 31% to 10%, in the proportion of our sentences that Lexicoder was unable to classify.

[10]We are confident that supervised models trained on annotated parliamentary text would represent an excellent strategy for analyzing sentiment in parliamentary corpora, provided that there was enough annotated data with which to train the models.. Parliamentary data are not normally annotated for sentiment, however, and the process of annotating them is time consuming and costly.

third bar from the left), the same approach performed nearly as well, at just under 40 percent of explained variance. Substantively speaking, we may interpret this last result by saying that the manner in which sentiment is expressed in the news is probably not drastically different from the manner in which it is expressed in parliament. Thus, we could have relied on embeddings trained on the Google News corpus to create sentiment lexicons with the proposed approach, even though the final domain of application differed. This supports the finding of Spirling and Rodriguez (in press) that models pretrained on large, similar corpora can sometimes be used as substitutes to those trained specifically on smaller local corpora.

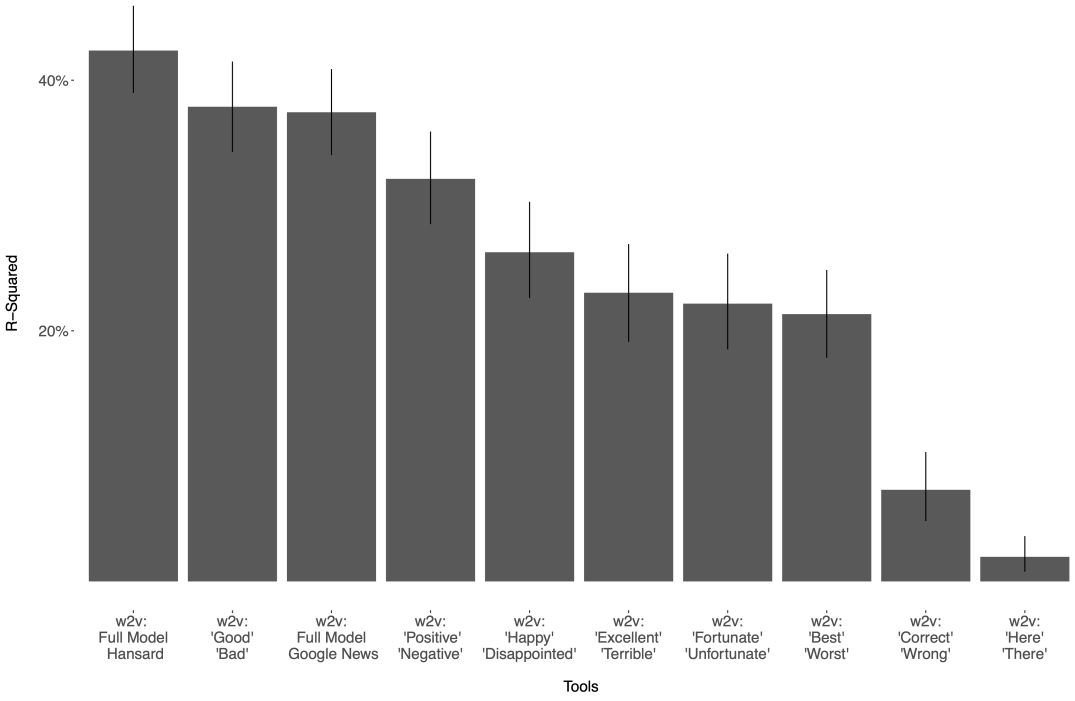


Figure 4. Comparing Alternative Specifications

We also examined the consequences of using different seed words. As it turns out, the full list of seven juxtaposed dyads of seed words performed better than any single dyad on its own. Nonetheless, using only the seed words 'good' and 'bad' performed nearly as well as the full list of seeds. By contrast, the seed words 'correct' and 'wrong' contributed little to the performance of the model. Significantly, neutral words, like 'Here' and 'There', perform very poorly as predictors of sentiment.

Finally, we examine the sensitivity of the model to the size of the corpus. Here, we generated derivative corpora by subtracting at random, and repeatedly, an increasingly large percentage of the days covered in our analysis. The days were extracted at random from across the entire period covered by our corpus. We removed days in increments of one percent. Thus, we begin with the full corpus, then derive a second corpus from the full corpus by removing one percent of days at random, then derive a third corpus by removing two percent of the days at random from the full corpus, and so on to the smallest corpus, which removed 99 percent of the days in the original corpus. We trained a new word2vec model on each of these progressively smaller corpora, and used

these models to generate sentiment scores for each of the sentences coded by human coders. We then compared the sentiment scores generated by word2vec dictionaries to the sentiment scores assigned by the human coders. Our expectation is that models trained on smaller corpora will perform less well at predicting human judgements.

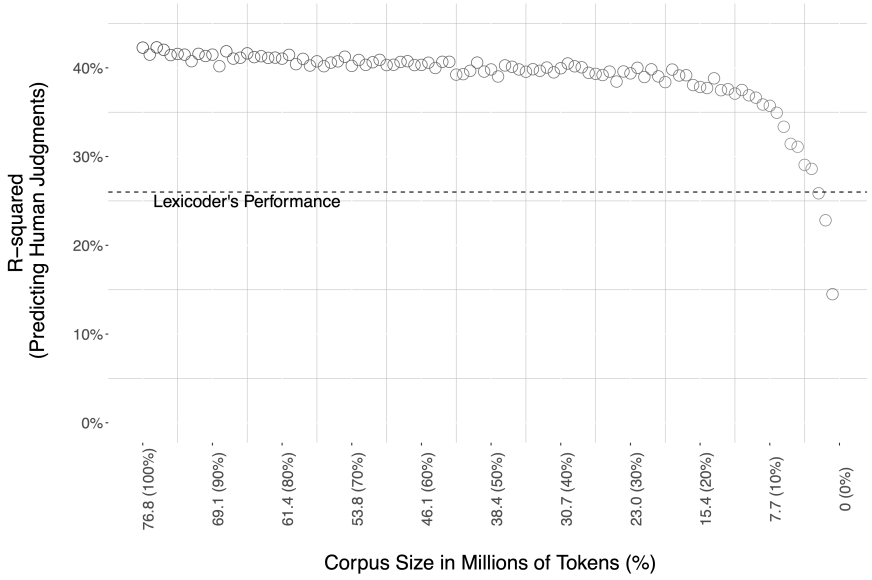


Figure 5. Impact of Corpus Size

The results of this analysis are summarized in Figure 4. The horizontal axis captures the size of the corpus that we used to train each model, ranging from 100 percent of the full corpus (comprised of about 77 million tokens) to just 1 percent of that corpus. The vertical axis summarizes the performance of each model at predicting the average sentiment score assigned to sentences by the human coders. Thus, as discussed earlier, the point in the top left corner of Figure 4 indicates that the full model explains more than 40 percent of the variation in human judgements, which, recall, is a 60 percent increase in performance compared to the next best tool that we tested, Lexicoder. As we see in the Figure, the performance of word embeddings for this task is robust to repeated applications of the model, although performance begins to decline precipitously at about 8 million words, and it no longer performs better than Lexicoder when trained on a corpus of less than two million words.

5. Discussion

Emotion is intrinsic to political life. As Crabtree, Golder, Gschwend, and Indriason (2020, 5) recently put it, "parties can influence perceptions of the world and [...] vote choice not only through the substantive content of their campaign messages but also through the emotive content of their campaigns." Understanding how emotion can be reliably measured from different modalities of communication—text, audio, images—could spur advances in the study of political campaigns, in particular strategies relying on emotional appeals. In legislative settings, analysis of emotion in communication may help us detect the intensity of support and opposition to legislative proposals and programs (Lupia, Soroka, & Beatty, 2020; Proksch et al., 2019), which is not obvious

from voting records alone, and especially not from voting records in contexts with strict party discipline. Although legislative transcripts are increasingly available, it is unclear whether they are capable of capturing the emotion of a speech. Indeed, there are good reasons to expect that they do not.

We began with the simplifying assumption that sentiment and arousal are core dimensions of emotion. Existing studies find that emotional expression involves a complex combination of verbal and non-verbal signals. Given that transcripts capture only verbal signals, this raises questions about the degree to which a transcript captures the emotional content of a speech. To examine this question, we compared human judgements about the emotional content of video clips and corresponding transcripts of parliamentary debate in the Canadian House of Commons. We report mixed results. We found that video and text coders perceived the same sentiment in the speech fragments, but coders annotating the same content across different media did not perceive the same level of emotional intensity. We infer that transcripts capture sentiment and not arousal. This finding suggests a limitation to using transcripts of speeches to capture emotion, and reinforces the importance of emerging work that focuses on analyzing audio and visual data in political communication (e.g. Casas & Webb Williams, 2019; Dietrich, Enos, & Sen, 2019; Dietrich, Hayes, & O'Brien, 2019; Hwang, Imai, & Tarr, 2019; Knox & Lucas, 2019).

Nonetheless, we do find strong evidence that transcripts capture the *sentiment* of a speech. When coding from transcripts of videos, there was a high degree of alignment between human coders in their assessment of the sentiment in the speech fragments. The finding that sentiment appears in the transcript of speeches has implications for a much broader research agenda in political communication. Sentiment itself, or what is often called tone, has proven to be a useful concept in the study of political communication, especially written communication (Stockmann, 2011; Young & Soroka, 2012). Studies on the tone of media coverage (Soroka, Young, & Balmas, 2015; Wlezien & Soroka, 2019) and press releases (Meyer-Gutbrod & Woolley, 2020) have been a staple of communication research. A growing stream of research in political communication also relies on social media data to investigate public sentiment (Dang-Xuan, Stieglitz, Wladarsch, & Neuberger, 2013; Flores, 2017; Oliveira, Bermejo, & dos Santos, 2017; Yarchi, Baden, & Kligler-Vilenchik, 2020) and the sentiment of public officials (Eberl, Tolochko, Jost, Heidenreich, & Boomgaarden, 2020; Rauh, Bes, & Schoonvelde, 2020; Zavattaro, French, & Mohanty, 2015). Our findings suggest that automated measures of sentiment work for transcripts of political speeches as well. In legislative settings, we suspect that sentiment is often the dimension of interest for substantive research. If we can know that a politician is consistently and strongly opposed to a bill, for example, then that is probably sufficient for most purposes. More generally, however, measures capturing the detailed emotional disposition of a speaker—such as anxiety vs. confidence, anger vs. sadness, and so on—would open avenues of research about political speech that extend beyond what we find can be captured in transcripts.

We then turned to test the efficacy of tools for the automatic detection of sentiment in text. We tested leading tools for automated sentiment analysis in terms of their effectiveness at predicting human judgments about the sentiment in transcripts of parliamentary speech fragments. We examined tools from three broad classes: dictionaries, supervised machine learning, and a method of dictionary induction based on word embeddings. Some dictionaries performed well, but they were not able to classify many of the sentences in our corpus, and they were occasionally unbalanced in terms of their success at predicting negative and positive classifications. Unlike dictionaries, supervised learners trained on established corpora classified all of the sentences in our

corpus, but they were less accurate than leading dictionaries. Lastly, we generated word embeddings using Mikolov, Corrado, et al. (2013)'s word2vec, and used small lists of positive and negative seed words to induce a domain specific sentiment dictionary from these embeddings. That approach outperformed on every indicator the other tools that we tested.

Finally, we tested the sensitivity of sentiment dictionaries based on word embeddings to alternative specifications, a different domain, and varying corpus sizes. We found that the performance of this approach is affected by alternative choices of seed words. We also find that word embeddings perform well for this task even when trained on just a small fraction of our corpus. As we expected, however, the automatic sentiment dictionaries become less reliable when based on word embeddings fitted on smaller corpora of just a few million words. In contrast, the same approach using word embeddings generated from a large out-of-domain corpus—the Google News corpus—outperformed even the most effective human-generated dictionary that we tested. Thus, although we find that word embeddings trained on smaller corpora are less effective for this task, they can in some cases be substituted with embeddings trained externally.

In sum, the results in this paper uncover opportunities and limitations of applying established tools for the automatic analysis of emotion in text to transcripts of legislative debates and other speeches. We also provide and validate an important use-case for word embeddings in the automatic generation of sentiment lexicons for political text. The utility of word embeddings, however, extends beyond sentiment analysis. Indeed, word embeddings have recently been used to study semantic change (Rodman, 2020) and ideology (Rheault & Cochrane, 2020). In validating downstream indicators based on word embeddings against human judgement, and in finding that the methodology is robust to re-initialized applications on subtly different corpora, we strengthen the case for using word embeddings in political science.

Acknowledgement(s)

This paper was improved by feedback received at the Centre for the Study of Democratic Citizenship at McGill, the Department of Political Science at Université Laval, the Department of Political Science at Western University, the 2nd Annual Politics and Computational Social Science Conference, the 115th Conference of the American Political Science Association, the 2019 Conference of the Canadian Political Science Association, and, especially, from comments by Jacob Montgomery, Bryce Dietrich, J. Scott Matthews, Sven-Oliver Proksch, François Pétry, Yannick Dufresne, and David Armstrong. We are also grateful for the exceptionally detailed and constructive criticism from both of the Journal's anonymous reviewers. We also thank Meghan Snider, Pierre-Oliver Bonin, Jason Vandenbeukel, Katie Moez, Stefan Ferraro, and Justin Savoie for their excellent assistance coding. We are responsible for any remaining errors.

Funding

This research was funded by the Social Sciences and Humanities Research Council of Canada.

References

Abercrombie, G., & Batista-Navarro, R. (2020). Sentiment and Position-Taking Analysis of Parliamentary Debates: A Systematic Literature Review. *Journal of Computational Social Science*, 1–26.

Amir, S., Astudillo, R. F., Ling, W., Martins, B., Silva, M., & Trancoso, I. (2015). INESC-ID: A Regression Model for Large Scale Twitter Sentiment Lexicon Induction. In *Proceedings of the 40th annual meeting of the association for computational linguistics (acl 2015).*

Baccianella, S., Esuli, A., & Sebastiani, F. (2010, may). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In N. C. C. Chair) et al. (Eds.), *Proceedings of the seventh international conference on language resources and evaluation (lrec'10).* Valletta, Malta: European Language Resources Association (ELRA).

Bänziger, T., & Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech Communication*, *46*(3-4), 252–267.

Beelen, K., Thijm, T. A., Cochrane, C., Halvemaan, K., Hirst, G., Kimmins, M., ... Whyte, T. (2017). Digitization of the Canadian Parliamentary Debates. *Canadian Journal of Political Science*, *50*(3), 849–864.

Benoit, K. (2019). Text as data: An overview. In L. Curini & R. Franzese (Eds.), *The sage handbook of research methods in political science and international relations.* London, UK: SAGE Publishing.

Brooks, J. A., Shablack, H., Gendron, M., Satpute, A. B., Parrish, M. H., & Lindquist, K. A. (2017). The Role of Language in the Experience and Perception of Emotion: a Neuroimaging Meta-analysis. *Social Cognitive and Affective Neuroscience*, *12*(2), 169–183.

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, *28*(2), 15–21.

Casas, A., & Webb Williams, N. (2019). Images That Matter: Online Protests and the Mobilizing Role of Pictures. *Political Research Quarterly*, *72*(2), 360–375.

Chaovalit, P., & Zhou, L. (2005, 01). Movie review mining: A comparison between supervised and unsupervised classification approaches. In *Proceedings of the 38th hawaii international conference on system sciences.*

Cicchetti, D. V. (1994). Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology. *Psychological Assessment*, *6*(4), 284–290.

Cochrane, C., Greenaway, C., Whyte, T., & Rheault, L. (2019). Language, Neural Nets, and the Nature of Meaning. Vancouver, BC: Conference of the Canadian Political Science Association.

Crabtree, C., Golder, M., Gschwend, T., & Indriason, I. H. (2020). It Is Not Only What You Say, It Is Also How You Say It: The Strategic Use of Campaign Sentiment. *The Journal of Politics*, *82*(3), 000–000.

Daku, M., Soroka, S., & Young, L. (2015). *Lexicoder 3.0.* Retrieved from `http://lexicoder.com`

Dang-Xuan, L., Stieglitz, S., Wladarsch, J., & Neuberger, C. (2013). An Investigation of Influentials and the Role of Sentiment in Political Communication on Twitter During Election Periods. *Information, Communication & Society*, *16*(5), 795–825.

de Gelder, B., de Borst, A. W., & Watson, R. (2015). The Perception of Emotion in Body Expressions. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(2), 149-158.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Diermeier, D., Godbout, J.-F., Yu, B., & Kaufmann, S. (2012, jan). Language and Ideology in Congress. *British Journal of Political Science*, *42*(1), 31–55.

Dietrich, B. J., Enos, R. D., & Sen, M. (2019). Emotional Arousal Predicts Voting on the U.S. Supreme Court. *Political Analysis*, *27*(2), 237–243.

Dietrich, B. J., Hayes, M., & O'Brien, D. Z. (2019). Pitch perfect: Vocal pitch and the emotional intensity of congressional speech. *American Political Science Review*, *113*(4), 941–962.

Duval, D., & Pétry, F. (2016). L'analyse automatisée du ton médiatique: Construction et utilisation de la version française du Lexicoder Sentiment Dictionary. *Canadian Journal of Political Science*, *49*(2), 197–220.

Eberl, J.-M., Tolochko, P., Jost, P., Heidenreich, T., & Boomgaarden, H. G. (2020). What's in a Post? How Sentiment and Issue Salience Affect Users' Emotional Reactions on Facebook. *Journal of Information Technology & Politics*, *17*(1), 48-65.

Ekman, P. (1992). Facial Expression and Emotion. *American Psychologist*, *48*(4), 384–392.

Ekman, P., Davidson, R. J., & Friesen, W. V. (1990). The Duchenne Smile: Emotion Expression and Brain Physiology. *Journal of Personality and Social Psychology*, *58*(2), 342–353.

Ekman, P., & Friesen, W. V. (1969, feb). Nonverbal Leakage and Clues to Deception. *Psychiatry: Journal for the Study of Interpersonal Processes*, *32*(1), 88–106.

Ekman, P., O'Sullivan, M., Friesen, W. V., & Scherer, K. R. (1991). Face, Voice, and Body in Detecting Deceipt. *Journal of Nonverbal Behaviour*, *15*(2), 125–135.

Ennedy, A. L. K., & Nkpen, D. I. I. (2006). Sentiment Classification of Movie Reviews

Using Contextual Valence Shifters. *Computational Intelligence*, *22*(2), 1–23.

Esuli, A., & Sebastiani, F. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. *Proceedings of the 5th Conference on Language Resources and Evaluation*, 417–422. Retrieved from `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.61.7217`

Esuli, A., & Sebastiani, F. (2007). SENTIWORDNET: A high-coverage lexical resource for opinion mining. *Evaluation*, 1–26. Retrieved from `http://ontotext.fbk.eu/Publications/sentiWN-TR.pdf`

Ethayarajh, K., Duvenaud, D., & Hirst, G. (2018). Towards Understanding Linear Word Analogies. In *Proceedings of the 57th annual meeting of the association for computational linguistics (acl 2018)* (pp. 3253–3262).

Fast, E., Chen, B., & Bernstein, M. S. (2016). Empath: Understanding Topic Signals in Large-Scale Text. In *Proceedings of the 2016 chi conference on human factors in computing systems* (pp. 4647–4657).

Flores, R. D. (2017). Do Anti-Immigrant Laws Shape Public Sentiment? A Study of Arizona's SB 1070 Using Twitter Data. *The American Journal of Sociology*, *123*(2), 333.

Griffiths, P. E. (1997). *What Emotions Really Are.* Chicago, IL: University of Chicago Press.

Grimmer, J., & Stewart, B. M. (2013a). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297.

Grimmer, J., & Stewart, B. M. (2013b). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*, 267–297.

Hadash, G., Kermany, E., Carmeli, B., Lavi, O., Kour, G., & Jacovi, A. (2018). Estimate and replace: A novel approach to integrating deep neural networks with existing applications. *arXiv preprint arXiv:1804.09028*.

Hall, T. H. (2015). *Emotional Diplomacy: Official Emotion on the International Stage.* Cornell, NJ: Cornell University Press.

Hamilton, W. L., Clark, K., Leskovec, J., & Jurafsky, D. (2016). Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora. In *Proceedings of the 2016 conference on empirical methods in natural language processing (emnlp 2016)* (Vol. 2016, p. 595).

Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the Semantic Orientation of Adjectives. In *Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL* (p. 174-181).

Hinton, G. E. (1986). "distributed representations". In Rumelhart, D. E., J. Mcclelland, & J. L. (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (pp. 77–109). Cambridge, MA: MIT Press.

Hirst, G., Riabinin, Y., Graham, J., Boizot-Roche, M., & Morris, C. (2014). Text to Ideology or Text to Party Status? In B. Kaal, I. Maks, & A. van Elfrinkhof (Eds.), *From Text to Political Positions: Text Analysis Across Disciplines.* Amsterdam, NL: John Benjamins.

Hopkins, D. J., & King, G. (2010). A Method of Automated Nonparametric Content Analysis for Social Science. *American Journal of Political Science*, *54*(1), 229–247.

Hu, M., & Liu, B. (2004). Mining opinion features in customer reviews. In *Proceedings of the 19th national conference on artifical intelligence* (pp. 755–760). AAAI Press. Retrieved from `http://dl.acm.org/citation.cfm?id=1597148`

.1597269

Hutto, C. J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM-2014)*, 216–225.

Hwang, J., Imai, K., & Tarr, A. (2019). Automated Coding of Political Campaign Advertisement Videos: An Empirical Validation Study. In *2019 annual meeting of the society for political methodology (polmeth xxxvi)*.

Jockers, M. L. (2015). Extracts Sentiment and Sentiment-Derived Plot Arcs from Text. *Cran-R*, 1–12. Retrieved from `https://cran.r-project.org/web/packages/syuzhet/syuzhet.pdf`

Knox, D., & Lucas, C. (2019). A Dynamic Model of Speech for the Social Sciences. In *Available at ssrn: https://ssrn.com/abstract=3490753* (pp. 1–31).

Kosmidis, S., Hobolt, S. B., Molloy, E., & Whitefield, S. (2019). Party Competition and Emotive Rhetoric. *Comparative Political Studies*, *52*(6), 811–837.

Kour, G., & Saabne, R. (2014a). Fast classification of handwritten on-line arabic characters. In *Soft computing and pattern recognition (socpar), 2014 6th international conference of* (pp. 312–318).

Kour, G., & Saabne, R. (2014b). Real-time segmentation of on-line handwritten arabic script. In *Frontiers in handwriting recognition (icfhr), 2014 14th international conference on* (pp. 417–422).

Kralj Novak, P., Smailović, J., Sluban, B., & Mozetič, I. (2015, 12). Sentiment of emojis. *PLOS ONE*, *10*(12), 1-22. Retrieved from `https://doi.org/10.1371/journal.pone.0144296`

Lak, P., & Turetken, O. (2014, Jan). Star ratings versus sentiment analysis – a comparison of explicit and implicit measures of opinions. In *47th hawaii international conference on system sciences* (p. 796-805).

Laver, M., & Benoit, K. (2002). Locating TDs in Policy Spaces: The Computational Text Analysis of Dáil Speeches. *Irish Political Studies*, *17*(1), 59–73.

Lowe, W., & Benoit, K. (2013). Validating Estimates of Latent Traits from Textual Data Using Human Judgment as a Benchmark. *Political Analysis*, *21*(3), 298–313.

Lupia, A., Soroka, S., & Beatty, A. (2020). What Does Congress Want From the National Science Foundation? A Content Analysis of Remarks From 1995 to 2018. *Science Advances*, *6*(33), eaaz6300.

Manning, C. D., Ragahvan, P., & Schutze, H. (2009). *An introduction to information retrieval*. Cambridge, UK: Cambridge University Press.

Meeren, H. K. M., van Heijnsbergen, C. C. R. J., & de Gelder, B. (2005). Rapid Perceptual Integration of Facial Expression and Emotional Body Language. *Proceedings of the National Academy of Sciences*, *102*(45), 16518–16523.

Meyer-Gutbrod, J., & Woolley, J. (2020). New Conflicts in the Briefing Room: Using Sentiment Analysis to Evaluate Administration-Press Relations from Clinton through Trump. *Political Communication*, 1–19.

Mikolov, T., Corrado, G., Chen, K., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, *arXiV:1301*, 1–12. Retrieved from `http://arxiv.org/pdf/1301.3781v3.pdf`

Mikolov, T., Karafiat, M., Burget, L., Cernocky, J. H., & Khudanpur, S. (2010). Recurrent Neural Network Based Language Model. In *Interspeech 2010* (pp. 1045–1048). Makuhari, Chiba, Japan: ISCA.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). "dis-

tributed representations of words and phrases and their compositionality". In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 26* (pp. 3111–3119). Curran Associates, Inc.

Mohammad, S. M., & Turney, P. D. (2012). Crowdsourcing a Word-emotion Association Lexicon. *Computational Intelligence*, *29*(3), 436–465.

Monroe, B. L., & Schrodt, P. A. (2018). Introduction to the special issue: The statistical analysis of political text. *Political Analysis*, *16*(4), 351–355.

Nulty, P., Theocharis, Y., Popa, S. A., Parnet, O., & Benoit, K. (2016). Social Media and Political Communication in the 2014 Elections to the European Parliament. *Electoral studies*, *44*, 429–444.

Oliveira, D. J., Bermejo, P. H., & dos Santos, P. A. (2017). Can Social Media Reveal the Preferences of Voters? A Comparison Between Sentiment Analysis and Traditional Opinion Polls. *Journal of Information Technology & Politics*, *14*(1), 34-45.

Pan, J., & Chen, K. (2018). Concealing Corruption: How Chinese Officials Distort Upward Reporting of Online Grievances. *The American Political Science Review*, *112*(3), 602-620.

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2), 1–135.

Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 conference on empirical methods in natural language processing (EMNLP 2002)* (pp. 79–86). Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/W02-1011`

Park, B., Greene, K., & Colaresi, M. (2020). Human Rights are (Increasingly) Plural: Learning the Changing Taxonomy of Human Rights from Large-scale Text Reveals Information Effects. *The American Political Science Review*, *114*(3), 888-910.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/D14-1162`

Proksch, S.-O., Lowe, W., Wäckerle, J., & Soroka, S. (2019). Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches. *Legislative Studies Quarterly*, *44*(1), 97–131.

Proksch, S.-O., & Slapin, J. B. (2015). *The Politics of Parliamentary Debate: Parties, Rebels, and Representation.* Cambridge, UK: Cambridge University Press.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to Analyze Political Attention with Minimal Assumptions and Costs. *American Journal of Political Science*, *54*(1), 209–208.

Rauh, C., Bes, B. J., & Schoonvelde, M. J. (2020). Undermining, Defusing or Defending European Integration? Assessing Public Communication of European Executives in Times of EU Politicisation. *European Journal of Political Research*, *59*(2), 397-423.

Rheault, L. (2016). Expressions of Anxiety in Political Texts. In *Proceedings of 2016 emnlp workshop on natural language processing and computational social science* (pp. 92–101).

Rheault, L., Beelen, K., Cochrane, C., & Hirst, G. (2016). Measuring Emotion in

Parliamentary Debates with Automated Textual Analysis. *PLoS ONE*, *11*(12), 1–18.

Rheault, L., & Cochrane, C. (2020). Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis*, *28*(1), 1–22.

Rice, D. R., & Zorn, C. (2019). Corpus-Based Dictionaries for Sentiment Analysis of Specialized Vocabularies. *Political Science Research and Methods*, 1–16.

Rinker, T., Higgins, J., Ward, G., Possel, H., Mechura, M. B., Hu, M., ... Malaescu, I. (2019). *Lexicon: Lexicons for text analysis.* Retrieved from `https://cran.r-project.org/web/packages/lexicon/lexicon.pdf`

Rodman, E. (2020). A timely intervention: Tracking the changing meanings of political concepts with word vectors. *Political Analysis*, *28*(1), 87-111.

Ruder, S., Vulić, I., & Søgaard, A. (2019). A Survey of Cross-Lingual Word Embedding Models. *Journal of Artificial Intelligence Research*, *65*, 569–631.

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (2018). More Than Bags of Words: Sentiment Analysis With Word Embeddings. *Communication Methods and Measures*, *12*(2-3), 140–157.

Russell, J. (1980). A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178.

Russell, J., & Barrett, L. (1999). Core Affect, Prototypical Emotional Episodes, and Other Things Called Emotion: Dissecting the Elephant. *J Pers Soc Psychol*, *76*(5), 805–819.

Russell, J. A. (2003). Core Affect and the Psychological Construction of Emotion. *Psychological Review*, *110*(1), 145–172.

Scherer, K. R., Ladd, D. R., & Silverman, K. E. A. (1984). Vocal cues to speaker affect: Testing two models. *The Journal of the Acoustical Society of America*, *76*(5), 1346-1356.

Schlosberg, H. (1954). Three Dimensions of Emotion. *Psychological Review*, *61*(2), 81–88.

Schwarz, D., Traber, D., & Benoit, K. (2017). Estimating Intra-Party Preferences: Comparing Speeches to Votes. *Political Science Research and Methods*, *5*(2), 379–396.

Soroka, S., Penner, E., & Blidook, K. (2009). Constituency Influence in Parliament. *Canadian Journal of Political Science*, *42*(3), 563–591.

Soroka, S., Young, L., & Balmas, M. (2015). Bad News or Mad News? Sentiment Scoring of Negativity, Fear, and Anger in News Content. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 108–121.

Spirling, A., & Rodriguez, P. (in press). Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research. *The Journal of Politics*.

Stockmann, D. (2011). Race to the Bottom: Media Marketization and Increasing Negativity Toward the United States in China. *Political Communication*, *28*(3).

Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). *The General Inquirer: A Computer Approach to Content Analysis.* Cambridge, MA: MIT Press.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014). Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (acl 2014)* (pp. 1555–1565).

Turney, P. D. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th annual meeting of the association for computational linguistics (acl)* (pp. 417–424).

Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism. *ACM Transactions on Information Systems*, *21*(4), 315–346. Retrieved from https://dl.acm.org/doi/10.1145/944012.944013

Turney, P. D., & Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.

Van den Stock, J., Righart, R., & de Gelder, B. (2007). Body Expressions Influence Recognition of Emotions in the Face and Voice. *Emotion*, *7*(3), 487–494.

Vapnik, V. N. (1998). *Statistical Learning Theory.* New York, NY: John Wiley & Sons.

Weber, R., Mangus, J. M., Huskey, R., Hopp, F. R., Amir, O., Swanson, R., ... Tamborini, R. (2018). Extracting Latent Moral Information From Text Narratives: Relevance, Challenges, and Solutions. *Communication Methods and Measures*, *12*(2-3), 119–139.

Wilkerson, J., & Casas, A. (2017). Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science*, *20*(1), 529-544.

Winston, P. H. (2014). *Lecture 16: Learning Support Vector Machines.* MIT OpenCourseWare. Retrieved from https://www.youtube.com/watch?v=_PwhiWxHK8o

Wittgenstein, L. (2009). *Philosophical Investigations* (4th ed.). West Sussex, UK: Blackwell Publishing.

Wlezien, C., & Soroka, S. (2019). Mass Media and Electoral Preferences During the 2016 US Presidential Race. *Political Behavior*, *41*(4), 945-970.

Yarchi, M., Baden, C., & Kligler-Vilenchik, N. (2020). Political Polarization on the Digital Sphere: A Cross-Platform, Over-Time Analysis of Interactional, Positional, and Affective Polarization on Social Media. *Political Communication*, 1–42.

Young, L., & Soroka, S. (2012). Affective News: The Automated Coding of Sentiment in Political Texts. *Political Communication*, *29*(1), 205–231.

Zavattaro, S. M., French, P. E., & Mohanty, S. D. (2015). A Sentiment Analysis of U.S. Local Government Tweets: The Connection Between Tone and Citizen Involvement. *Government Information Quarterly*, *32*(3), 333-341.

Zhang, L., & Liu, B. (2017). Sentiment analysis and opinion mining. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning and data mining* (pp. 1152–1161). Boston, MA: Springer US. Retrieved from https://doi.org/10.1007/978-1-4899-7687-1_907 doi:

Appendix A. Coding Instructions

(1) On a scale from 0-10, where 0 indicates that the speaker was very subdued, 5 indicates that they were in a normal state of calm, and 10 indicates that the speaker was very animated, please indicate the emotional state of the speaker.
   - Here, emotional activation means level of emotional arousal, where a very low score indicates an unusually subdued or very low level of emotional arousal, and a very high score indicates an unusually animated or high level of emotional arousal.

(2) On a scale from 0-10, where 0 indicates that the speaker was expressing a very negative sentiment, 5 indicates that they were expressing a neutral sentiment,

and 10 indicates that they were expressing a very positive sentiment, please indicate the sentiment of the speech?

- The sentiment describes the valence of the speech fragment. A very low score indicates that the speaker is conveying a very negative (unfavorable) sentiment about something, whereas a very high score indicates that the speaker is conveying a very positive (favorable) sentiment about something.