Validating the Joint Sentiment Topic Model and the reversed Joint Sentiment Topic Model on Political Text*

Max Boiten
Christian Pipal
Martijn Schoonvelde
Gijs Schumacher

July 10, 2019

Abstract

Sentiment is an important concept in social science research to study issues such as public opinion, media tone, or campaign dynamics. While automated content analysis methods such as the application of sentiment dictionaries provide reasonably good results overall, these methods do not take into account that the sentiment dimension of a word can depend on its context. Adopting novel methods from computational linguistics, we present a semi-supervised sentiment-topic model (JST and reversed JST) to estimate topics and sentiment simultaneously in political text. Validating our results with a crowd coded data set of UK prime minister speeches, we show that taking topic specific sentiment into account improves the accuracy of the generated sentiment scores compared to commonly used dictionaries. Our findings highlight the connection between topics and sentiment, and have important implications for the analysis of emotional content in political texts.

*For conducting this research, Schoonvelde & Schumacher received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No 649281, EUENGAGE

Introduction

It is quite often claimed that public life, including political life, is increasingly sentimental. Picking up on this theme several recent papers analyze sentiment in media (Atteveldt et al. 2008; Young and Soroka 2012), in legislative debate (Proksch et al. 2019; Rheault et al. 2016), in electoral campaigns (Crabtree et al. 2019; Kosmidis et al. 2018), in prime minister speeches (Traber, Schoonvelde, and Schumacher 2019) and in party congress speeches (Schumacher et al. 2019). There is plenty of evidence that sentimental (or emotional) political rhetoric has the power to persuade and mobilize. For example, enthusiastic rhetoric increase turnout among the public, while angry rhetoric make people stronger in their convictions and may strengthen the persuasiveness of a message (Brader 2005).

This paper is concerned with the identification of sentiment in political rhetoric. Until now, the literature has mostly relied on sentiment dictionaries. These provide lists of words with a positive, negative and sometimes neutral connotation. The positive (or negative) sentiment of a text is then simply the (weighted) sum of matches with the positive sentiment dictionary. There is much to like about this approach. Oftentimes, these dictionaries have been extensively validated, and their application to textual data is highly scalable and comes at almost no additional cost. Still, the dictionary method has major limitations.

In particular, dictionary methods may conflate topics and emotions. The reason for this is that depending on the topic, a word may mean something positive or negative. In other words, its semantic orientation may depend on its context. The word “low” for instance carries positive sentiment when talking about unemployment, but means something negative when talking about economic growth. The same thing goes for an “increase” in political trust among the public versus an “increase” in political distrust. This makes valid estimates of sentiment difficult to obtain using a context-free dictionary.

In response to such issues, this paper distinguishes between topics and emotions in political texts. Drawing on work in computational linguistics, we introduce a semi-supervised sentiment-topic model to estimate topics and sentiment simultaneously. Contextual meaning is a well-known issue in computational linguistics. That field has developed a number of tools to measure sentiment across topics. However, these methods have not been applied yet to political texts—rather they have been validated using product or movie reviews (Lin and He 2009; Jo and Oh 2011; Lin et al. 2012). But such reviews usually serve one purpose only: to express a positive

or negative opinion about the product. Political texts like speeches and manifestos are much more complex, posing additional challenges to the models.

In this paper we validate the Joint Sentiment Topic model (JST) and the reversed Joint Sentiment Topic model (rJST) (Lin and He 2009; Lin et al. 2012) on political texts. Our analysis offers several insights. First, we demonstrate the effectiveness of models measuring topics and sentiment simultaneously in the setting of political speech. Both models show consistent separation of positive and negative words in their sentiment-topics, while also providing coherent topics. Furthermore, this paper offers a discussion on the roles of sentiment and topics in political texts which come with specific peculiarities. Empirically, our results emphasize the connection between topics and sentiment in political text. Comparing our results with different dictionaries used in political sentiment analysis, taking topic specific sentiment into account improves the accuracy of emotional polarity estimates substantially. Further, we show how JST can be used to assess the overall emotional tone of a political text, and how rJST provides an easy to use method to study topic specific sentiment.

The paper is structured as follows. First, we introduce sentiment analysis making the connection between methods in computer linguistics and political science. We then introduce the JST and reverse JST models. Next, we validate the JST and reverse JST results on speeches by UK prime ministers between 2007 and 2015 from the EUspeech dataset (Schumacher et al. 2016) and compare the obtained sentiment estimates with crowd coded data. Finally, we illustrate an empirical application measuring the development of emotional polarity in the UK House of Commons between 1988 and 2015.

Measuring sentiment in text

Because of their easy to use and highly scaleable nature, the use of sentiment dictionaries such as LIWC (Tausczik and Pennebaker 2010) or Lexicoder (Young and Soroka 2012) is a very popular tool to study the emotional content of political text. However, this approach comes with drawbacks.

First, dictionaries used for these automated approaches are predominately developed and validated on English texts, making them difficult to use across different languages resulting in lower quality sentiment estimates (Mohammad 2016).

Second, since human language is domain specific, dictionaries should ideally be developed

with a specific application in mind (e.g. Haselmayer and Jenny 2017; Rheault et al. 2016; Rudkowsky et al. 2018; Young and Soroka 2012). The costs associated with the development and validation of dictionaries prevents this in many cases. When such domain specific language dictionaries are not available, one can resort to supervised machine learning methods to classify text along a latent scale¹ or human crowd coding (Lind, Gruber, and Boomgaarden 2017). Since both approaches are associated with a manual process of labeling data, these methods come with a hefty price tag.

Third, sentiment dictionaries are ignorant to the topic of a text, average sentiment scores differ for each topic. For example, Young and Soroka (2012), who introduced the Lexicoder Sentiment Dictionary, show in their validations stark contrasts between topics. A somewhat positive crime article has an average sentiment of about as negative as a somewhat negative or neutral article on the economy. This issue can be solved by using domain-specific sentiment lexicons or by generating such lexicons for a specific corpus (see Rheault et al. (2016) for a political science application). These applications are useful when a corpus contains one (broad) topic only, but their comparability across domains remains doubtful.

In a corpus that contains multiple topics, a better approach would be one that adapts to different phrasing in different topics.² In what follows we provide a short overview of solutions from the computational linguistics literature. While reading one should keep in mind that the target of methods developed in computational linguistics is often different from social science tasks (Hopkins and King 2010). Most methods are developed for and validated on either movie or product reviews. The advantage of this type of data is that (i) the texts express a clear opinion or polarity and (ii) the texts are accompanied by the author’s rating and therefore validity is easily testable. In many political texts we lose both of those as politicians express opinions about multiple issues without providing an overall rating of any kind.

Lexicon-based methods treat text as a bag of words, neglecting the contextual meaning of words (Liu 2015). In practice, bag-of-words based methods tend to perform well, but in the case of political texts – in which a word’s meaning may depend on the context – there are good reasons to prefer a more fine-grained analysis. To distinguish between contexts, one has to separate the different topics that are discussed and the opinions expressed about these

¹For an application see Theocharis et al. 2016.

²A parallel to this task can be found in Lauderdale and Herzog (2016), who define the *wordshoal* model as an aggregation of separate *wordfish* models (Slapin and Proksch 2008) in order to estimate political positions across topics.

topics. We highlight two existing approaches, because they clearly illustrate the challenges of topic-specific sentiment analysis.

On the one hand, some approaches first identify topics in a text and then discern for every topic whether a reference is positive or negative. For example Somasundaran (2010) combines a dataset with coded pieces of opinionated text with a discourse approach to identify the object of every opinion in a text. In an application of these methods on online political discussions she presents very intuitive results, exactly pinpointing what topics authors are positive or negative about. In debates about gay marriage, for example, she finds that proponents refer positively to ‘love’ and ‘equality’, while opponents refer positively to ‘Jesus’, ‘Christianity’ and ‘Corynthians’. Such an analysis is extremely fine-grained, but to be useful comparatively it needs some form of aggregation to different topics (Liu 2015, p. 157-159).

On the other hand, there are models such as the Joint Sentiment Topic (JST) (Lin and He 2009; Lin et al. 2012) and the Aspect Sentiment Unification Model (ASUM) (Jo and Oh 2011). These models take some sentiment words as input and use these to estimate a combination of topic and sentiment within a text. The strength of these methods is that they assign topic and sentiment to words and are therefore able to find positive and negative connotations in different contexts at the same time. This way, the word ‘low’ can show as positive for the context of unemployment and negative for the context of wages. In product reviews, the models outperformed supervised models for the same task. These models, however, have not been used to trace sentiment in political texts. Given their promise, we evaluate JST for political speeches.

Joint Sentiment-Topic Model

The Joint Sentiment Topic Model (Lin and He 2009; Lin et al. 2012) is among the first unsupervised models to sentiment and topical content simultaneously. It builds on LDA, or Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003). Topic models such as LDA are a clustering method for analysing large amounts of textual data, treating documents as mixtures of topics. They take all documents in the corpus as input and then – assuming that these documents are generated by a fixed number of topics – estimates the prevalence of these topics as well as their content. Despite its simplicity LDA can be successful at identifying and estimating meaningful topics.

JST adds sentiment to this clustering approach. Instead of assuming that a document consists of a mixture of topics alone, JST assumes that a document consists of a mixture of

sentiments first, and a mixture of topics second (usually, positive, negative and neutral). That is, each word in a document is assigned a sentiment first and a topic second. As mentioned before, most sentiment models are built for reviews, a sort of text designed to communicate an overall sentiment. In this context, the assumptions of JST make sense, since the core goal of the writer is to convey positive or negative sentiment. However, there is usually a balance of aspects which you are positive or negative about. For example, if you are only positive about a phone’s screen, the topic distribution for the positive sentiment of your review is skewed towards mentioning the screen. After all, if you discuss something positively, it is likely to be the screen.

Reversed JST (rJST) turns these assumptions on their head: each document is assumed to consist of a mixture of topics first and a mixture of sentiments second. Each topic has a distribution over different sentiments. This reflects the generative process of a writer who first determines what to speak about and then for each topic what sentiment to convey. Reversed JST turned out to be ineffective for product reviews (Lin et al. 2012), but may theoretically come much closer to political texts. After all, politicians express opinions on different topics. In spatial or directional voting theory, parties communicate their positions on issues in order to shape voter’s perceived positions of those parties in the relevant political space(s) (e.g. Downs 1957; Davis, Hinich, and Ordeshook 1970; Rabinowitz and Macdonald 1989). Therefore, despite its ineffectiveness in product reviews, we include reverse JST in this study.

Table 1 shows the generative processes of LDA, JST and rJST. In the latter two, sentiment is added to the generative process. In addition to a topic distribution, every text also contains a distribution over sentiments (e.g. neutral, positive and negative).

The two models thus differ fundamentally. JST is developed to find one sentiment distribution for an entire text, allowing us to compare the overall expressed sentiment between texts. On the other hand, rJST’s primary goal is to find a sentiment distribution for each topic within a text. This allows for a comparison of topic specific sentiment between texts, but might not work well to analyse the overall expressed sentiment of texts containing multiple topics.

Data and methods

EUspeech data

In order to evaluate JST and rJST we use the EUspeech dataset (Schumacher et al. 2016), specifically the subset of British speeches (Blair, Cameron and Brown; $N = 787$). This selection

Table 1: Generative processes of LDA and JST

Latent Dirichlet Allocation	Joint Sentiment Topic model	reversed Joint Sentiment Topic model
1. For each document d , choose a topic distribution $\theta_d \sim Dir(\alpha)$ 2. For each word w_i in document d <ul style="list-style-type: none"> choose a topic $z_i \sim \theta_d$ choose a word w_i from the distribution over words defined by the topic z_i (parameter ϕ_{z_i}) 	1. For each document d , choose a sentiment distribution $\pi_d \sim Dir(\gamma)$ 2. For each sentiment label l in document d , choose a topic distribution $\theta_{d,l} \sim Dir(\alpha)$ 3. For each word w_i in document d <ul style="list-style-type: none"> choose a sentiment $l_i \sim \pi_d$ choose a topic $z_i \sim \theta_{d,l_i}$ choose a word w_i from the distribution over words defined by l_i and z_i (parameter $\phi_{z_i}^{l_i}$) 	1. For each document d , choose a topic distribution $\theta_d \sim Dir(\alpha)$ 2. For each topic z in document d , choose a sentiment distribution $\pi_{d,z} \sim Dir(\gamma)$ 3. For each word w_i in document d <ul style="list-style-type: none"> choose a topic $z_i \sim \theta_d$ choose a sentiment $l_i \sim \pi_{d,z_i}$ choose a word w_i from the distribution over words defined by l_i and z_i (parameter $\phi_{z_i}^{l_i}$)

Source: Blei, Ng, and Jordan (2003), Lin and He (2009) and Lin et al. (2012)

has several reasons. First, the speeches are all in one language (English), avoiding any problems in comparing translated texts and texts that were English originally. Second, the speeches are likely to cover multiple topics. In this, they are typical of the type of documents prevalent in political science. Finally, they discuss topics which are easily tested for face validity. We can expect the Prime Ministers to comment on major topics, therefore it does not require very deep contextual knowledge to assess the strength of the model. The top part of Figure 1 shows how many speeches by Blair, Brown and Cameron we have over time. The middle panel demonstrates the distribution of speech length.

The sentiment score in the bottom panel of figure 1 is calculated using the sentiment dictionary of Rheault et al. (2016). Their dictionary is trained using word embeddings on UK parliamentary speeches and is therefore the most comprehensive and applicable dictionary we could use for this context. For this calculation, negations have not been taken into account. The positive score is calculated by dividing the number of positive words by the sum of positive and negative words. We show this score for two reasons. First, as we will use this dictionary as input for JST and reversed JST, it is important to know roughly what distribution it has

Figure 1: Data description of 787 UK PM speeches

in our data. Second, one crucial element we can learn from computational linguistics about training data is that they should ideally be balanced. In our case, the speeches are clustered towards the positive end of the scale as seen in the bottom panel of figure 1. This is a downside of political speeches: there is no clear-cut labeling or easily identifiable sentiment and therefore no way to preselect texts unless we have a manually selected training set. Then again, this is highly context sensitive, so it might not work very well.

This distribution is reminiscent of the findings of Somasundaran (2010), who finds that on both sides of online political debates there are mostly positive references, just to different concepts. If this also extends to PM speeches, it poses multiple problems for sentiment analysis and our models. If sentiment and topic are correlated and differences are as small as seen above, a different topic might well have a completely different sentiment score. This makes comparison of speeches without knowledge of topics even more challenging. For our models, it means that it will be harder to pick up negative sentiment-topics, because there are fewer negative words to structure the topics around. This is on the one hand a significant drawback of our data, on the other hand it is natural for the type of speeches we use. If our models can fit this well, it

means they are likely to be suited for political texts more generally, and show that this method is better at measuring sentiment compared to sentiment dictionaries.

Estimation of JST and rJST

There are multiple ways to estimate JST and reverse JST. In this paper, we follow the implementation of Lin and He (2009)³ using a Gibbs Sampler. Gibbs samplers iteratively approximate the posterior of a model which cannot easily be optimized linearly. Our implementation fits with the R package *quanteda* (Benoit 2017), to make it more easily applicable for social scientists. The implementation can be downloaded from CRAN.⁴ The downside of using a Gibbs sampler is that it is not easily parallelized but for this study it suffices.⁵

JST and rJST are able to estimate sentiment on a two-point (positive/negative) or three-point (positive/neutral/negative) scale, producing respective word lists associated with sentiment-topics based on this choice. In our application of JST/rJST models on British prime minister speeches, including the neutral category improves the face validity of the results, giving more intuitively coherent word lists for the positive and negative sentiment categories. This is in line with Lin and He (2009) who included the neutral category in the original model.

Both models are semi-supervised, which means that they take some coded input to base the model training on. This way the model requires very little input, yet can offer enough guidance to the learning process to meaningfully identify sentiment. As an input, JST requires a sentiment dictionary, coding positive and negative sentiments. The dictionary informs the sampler in two ways. First, rather than assigning a random sentiment label in the initialization phase, it puts words in their respective sentiment categories if they match with a dictionary entry. Second, this serves as a basis for the following iterations of the algorithm. There, the dictionary informs the sampler that words from the dictionary are more likely to appear in their respective sentiment category, while still allowing for deviation if found in the text.

As an input dictionary, we take the dictionary provided by Rheault et al. (2016)⁶. This is a dictionary constructed using word embeddings (Pennington, Socher, and Manning 2014) trained on debates from the British House of Commons. It is a well-validated, context-specific lexicon for British political speech. Therefore, it serves as an excellent basis for our algorithm.

³<https://github.com/linron84/JST>

⁴<https://cran.r-project.org/web/packages/rJST/index.html>

⁵For readers unfamiliar with Gibbs sampling, Resnik and Hardisty (2010) provide an introduction.

⁶For the source, see <https://github.com/lrheault/emotion>

Furthermore, it allows for a comparison between their measure of political sentiment based purely on the sentiment lexicon and the semi-supervised approach of JST and reverse JST.

Results

Expectations

Evaluating machine learning methods is a difficult task as we have several benchmarks to compare our results to. Before we turn to the results, let us quickly recap our expectations. First, we expect JST to perform better on the identification of a global sentiment, and reverse JST to perform better on identifying topic specific sentiment. Second, if topics and sentiment are correlated, we expect to see many topic words in the positive and negative categories, rather than in the neutral categories. If topic and sentiment are uncorrelated, as arguably is the case in product reviews, we expect topic words to cluster in the neutral category, because they co-occur with both negative and positive words and thus do not cluster well with either. If they are spread more around the positive and negative sentiment categories, it means that there is a correlation between some topic words and sentiment categories.

We test the first expectation in two ways. First, we compare model performance to the calculated lexicon score. Since our models are trained with the lexicon, this indicates mostly to what extent our models can incorporate prior information and structure topic-sentiment combinations. If our model is better at untangling the structure of political text, its performance will rightly deviate from the dictionary approach. We extend this by checking the words in sentiment categories to sentiment words from the dictionary. Second, we evaluate topic and sentiment performance through concurrent validity, comparing our models' estimated topic parameters with what we would expect from the topics given world events.

Our second expectation concerns face validity. We compare lists of words that our models strongly associate with topics and sentiments to judge to what extent they are mixed. This is not conclusive but can give a strong indication of how the two behave.

Finally, we compare the overall sentiment score of JST and the result obtained from various dictionaries with with a crowd coded data set of political speeches. This allows us to tests whether the model improves the accuracy of sentiment estimation when using automated methods. We follow the three steps described above and wrap up with some overall thoughts on what JST and rJST measure.

Face validity

Having concluded that our models function as we expect them to the next step is to evaluate if – and to what extent – they detect sentiments. As argued above, we move to an analysis of face validity since that is the best indicator we have at this point.

First, we turn to the words the models most strongly associate with sentiment/topic combinations. Table 2 shows two examples of such word vectors. Both discuss terrorism. However, JST seems to conflate different responses with the war in Syria and economic themes around migration. Since JST determines sentiment by global association and topic by association within the sentiment category, its associations within topic were expected to be weaker. rJST, on the other hand, consistently selects words regarding fighting terrorism in Afghanistan and Pakistan. Across all topics, rJST outperforms JST in topic consistency. See tables A2 and A3 for more examples.

Table 2: The 20 words with strongest association to two topics related to terrorism

Neutral	JST		Neutral	Reverse JST	
	Positive	Negative		Positive	Negative
immigr	polic	terror	afghanistan	forc	pakistan
migrat	forc	terrorist	afghan	secur	terrorist
come	secur	threat	troop	defenc	terror
peopl	train	peopl	british	nato	afghanistan
migrant	support	extremist	armi	train	action
control	last	face	polic	britain	border
system	govern	syria	countri	oper	area
uk	local	must	helmand	arm	threat
illeg	clear	muslim	taliban	militari	fight
worker	work	attack	al	nation	britain
net	month	ideolog	karzai	take	strategi
student	intern	isil	elect	threat	attack
economi	servic	defeat	year	protect	countri
employ	respons	fight	qaeda	alli	world
econom	communiti	islam	number	equip	take
busi	effort	young	civilian	clear	pakistani
border	strategi	clear	increas	capabl	govern
eu	oper	iraq	helicopt	effort	deal
reduc	arm	say	taleban	mission	zardari
visa	contin	islamist	month	home	million

Note: Both are drawn from the 25-topic models

Overall, this performance is as expected. At this point, we believe it safe to conclude that rJST is better suited for our application than JST. The more appropriate generative process

yields positive results both in selecting the right sentiment words, as shown in table A1 and in identifying meaningful topics. Given our approach at identifying a more aspect-level sentiment, rJST performs best, both theoretically and empirically. With this in mind, the rest of this section focuses on evaluation of rJST.

Given that our topics seem consistent as judged by the words associated with them, we dig a bit deeper to evaluate if the estimated topic and topic-sentiment parameters carry meaningful information. Figure 2 shows monthly average topic parameters for two topics, the Afghanistan war topic from table 2 and a topic on the financial crisis (see table A4). The financial crisis topic shows a surge after September 2008 when Lehman Brothers fell. The topic spikes again October 2009 when the European sovereign debt crisis began. The topic shortly peaks again in May-June 2012, around the Greek parliamentary elections. The topic clearly responds to world events with political salience. The war in Afghanistan works a bit differently. The two high peaks occur because of a speech Gordon Brown gave to the troops in Afghanistan 30-08-2009 and a David Cameron speech on Afghanistan at the North Atlantic Council Summit 04-09-2014. Both concern planned events rather than the more unplanned events of the financial crisis.

Figure 2: Monthly mean topic parameters for two topics (reverse JST, 25 topics)

Not all topics seem as consistent as the topics above. There are several things to be said about this. First, not every pattern is meaningful to us, even if it is a statistically meaningful latent factor for our corpus. This is not a failure of the model, but rather something we must accept and filter ourselves. Second, it is possible to improve our judgement by calculating term scores (Blei and Lafferty 2009). Term scores distinguish terms more unique for a sentiment

category, similar to tf-idf weighting. This should give a more accurate view of the words around which topic-sentiment combinations cluster.

Finally, to assess the topic-sentiment scores of reverse JST, we looked at how the sentiment of the financial crisis topic developed over time in figure 3. For this we defined a cutoff point of 0.05 for the topic scores of a speech to be included in this test since, since topics with little presence are expected to have bad sentiment estimates. The figure shows an initial negative reaction to the fall of Lehman Brothers, followed by a positive response. Then, when the extent of the Eurocrisis becomes clear, the sentiment moves down again. Finally, there is a positive peak around the agreement on the first Greek bailout package.

Figure 3: Reverse JST crisis topic positive sentiment scores over time

Note: The scores in this graph are monthly averages. Given the selection of cases with 5% or more on the crisis topic the topic scores are also higher than in figure 2.

While these results look encouraging, the overall high scores are notable. This can be related to the relative prominence of positive terms in our speeches. In any case, the score can be very difficult to interpret in absolute terms. Within the same topic it has merit, but as so far we have no frame of reference, a score of 0.75 is more positive than 0.5, but does not necessarily imply that the underlying text has to be positive.

But does the topic sentiment estimation of rJST differ from using a sentiment dictionary? To investigate this, we look at the sentiment of speeches on the Afghanistan war. For this we selected all speeches with a topic score above 0.2, meaning that the model identified more than 20 percent of the text being associated with the war in Afghanistan, resulting in a total

of 45 speeches. We further checked if their titles were indicative of the chosen topic. To estimate the sentiment with a dictionary method, we applied the LIWC dictionary to these selected speeches, calculated their emotional polarity, and compared the results with the topic-sentiment scores from the rJST model. Figure 4 shows how these scores differ across the 45 speeches on Afghanistan.

Figure 4: Emotional polarity of speeches on Afghanistan, rJST vs. dictionary. $n = 45$.

Interpreting words in context

Now we have established that the determination of sentiments is reliable, we still need to cover the question whether rJST actually improves the interpretation of words in various contexts. To do so, we have selected the 100 most strongly associated words per sentiment-topic combination and per unique word created a list of the sentitopics with which is it associated. The model used is the same 25-topic model used for the other evaluations, with three sentiment categories in each topic (neutral, positive, negative).

Among these 7500 words (25 topics * 3 sentiment categories * 100 most strongly associated words), there are 2906 unique words. It is logical that there are far fewer unique words in this than words in general, since many words carrying sentiment are used across contexts. Out of these 2906 words, 739 (25.4%) are only associated with neutral sentiment categories. Figure 5 shows the to what extent the words that are associated to a sentiment category are linked to different sentiment categories. It shows that most words are either negative or positive, but that there is a rather sizeable number of words that occurs in the top 100 most strongly associated

words of both negative and positive sentitopics.

Figure 5: Occurrence of words in different sentiment categories

Note: For the proportion of positive topic assignments, the neutral category was left out of consideration. A score of 0 thus means presence solely in negative categories.

It is these words we are interested most in, because they show where topics and sentiment are connected. And not only does it show they are connected, it also shows rJST is able to discern these different meanings. We pick a number of words to illustrate these differences.

The word ‘hard’, for instance, has a general negative connotation, but is associated in two positive sentitopics. Both share that they are about solving or dealing with difficult situations. One about helping people in difficult situations, the other about facing such situations. Similarly, ‘success’ and ‘unemployment’ occur in both sentiment categories, depending on the context they are used in. The word ‘danger’ has double polarities for the same reasons. ‘Being in danger’ is negative, while ‘saving people from danger’ is identified as positive by rJST in a topic on military interventions.

‘Justice’, equally, exists in different categories. The context of cracking down hard on crime to serve justice is identified as negative, while it is positive in a topic of commemoration. Between the commemoration sentitopics, however, the word ‘love’ also takes an ambiguous position. The word is associated with three negative sentitopics and one positive sentitopic. In all three negative categories, the phrase ‘loved ones’ has the strongest presence. When ‘loved ones’ is used, this generally implies a loss, hence the presence of ‘love’ in negative categories. Stemming, however, produces the same token for ‘love’ as for ‘loved’, leaving the two to be

conflated easily. Regardless, rJST identifies them in their respective contexts.

Validation with coded data

To test whether our results provide a benefit over other methods, we compare emotional polarity estimates against a crowd coded data set of European leader speeches. This speech level comparison includes the results obtained from the JST model (since JST is more useful for document level sentiment estimates), a highly domain specific dictionary trained on word embeddings (Rheault et al. 2016), and more general dictionaries such as Lexicoder (Young and Soroka 2012), LIWC (Tausczik and Pennebaker 2010), and Bing (Hu and Liu 2004).

To calculate the emotional polarity of these speeches using the introduced sentiment-topic model, we estimate a JST model with 45 topics⁷ and use the seed dictionary of Rheault et al. 2016 as an input. To calculate the polarity score, we calculate the difference between positive and negative scores over the sum of positive and negative scores in a text. Similarly, we calculate the results based on the different dictionaries.

To calculate an emotion polarity score based on human coding, we crowd coded 17 speeches from UK prime ministers using the Crowdfunder platform. Coders were asked to judge a sentence on a 5-point-scale (very negative, somewhat negative, neutral, somewhat positive, very positive). Each sentence from these speeches was coded by at least 3 coders, and only includes judgments from coders who passed a 0.8 threshold on Gold judgment tasks. Since a simple mean of means aggregation has been proven to produce valid results (Benoit 2017), we calculate the overall sentiment score of each speech by taking the mean of all sentiment judgements.

We then compare the correlations between our different sentiment estimates with the speech sentiment obtained from the coded data. Judging our model performance based on all speeches we have coded data for in figure 6, the JST model performs poorly compared to all other dictionary methods. However, something unusual seems to be going on here. While we would expect the Rheault et al. 2016 dictionary to perform very well in this task since it was trained using word embeddings and the specific domain in question, the even better results obtained by the Bing, Lexicoder, and LIWC dictionary is striking. Comparing this to previous applications, the correlation between dictionary scores and the coded data is unusually high. What’s even more suprising, the small seed dictionary we used when estimating the model even performs

⁷This number of topics has been shown to be accurate for UK prime minister speeches in earlier work by Traber, Schoonvelde, and Schumacher 2019 and, based on a qualitative inspection of the resulting senti-topics, provides a sensible set of topics.

Figure 6: Sentiment estimates vs. coded sentiment, all speeches ($n = 17$)

slightly better than the much more complex full Rheault dictionary. When looking at the speeches in detail, this seems to be an artifact of the small and unbalanced test set we have. For instance, the two shortest and most positive speeches don't contain a single negative word and only talk about the English football team and the World Cup. Given our limited set of coded data, we cannot be sure at this point if these strange results are an artefact of the short nature of some of the speeches in our set (since the models tries to fit many different senti-topics to each speech), or if their extreme positivity drives our model off.

We try to balance our comparison by excluding speeches with less than 50 sentences, thus excluding short and extremely positive speeches from the comparison. This leaves us with 9 speeches in the comparison, and as figure figure 7 shows, the results are much more sensible now. While the results obtained by Rheault's dictionary are still exceptionally good for a dictionary application, all other dictionaries now show performances one would expect from

Figure 7: Sentiment estimates vs. coded sentiment, min. 50 sentences ($n = 9$)

non-domain specific dictionaries. Most interestingly, our JST model outperforms the high bar set by Rheault’s dictionary substantially. These results show that even when having a highly domain specific dictionary at hand, taking the relationship between topics and sentiment into account improves the accuracy of automatically estimated sentiment scores. Since the dictionary we used to feed the model is a small seed dictionary, these results mean that we can get equally good (or even better) results by employing a sentiment-topic model compared to creating a much more costly domain specific dictionary.

Application: Measuring emotion in parliamentary debates

We now use the methods described in this paper in an empirical application. To highlight the substantial benefit of topic specific sentiment analysis, we measure the emotional polarity of

parliamentary debates in the House of Commons between 1988 and 2013. Previous work showed a steady increase in positivity since the 1980s (Rheault et al. 2016), contrasting pundits’ and media accounts of political debate becoming more negative over the last decade.

For this application we use the data provided by the ParlSpeech Corpus (Rauh, De Wilde, and Schwalbach) which includes all speeches from the British House of Commons between November 1988 and January 2015. In total, our corpus consists of 1,540,065 speeches from 1994 speakers with an average speech length of 195 words.

For the following analysis we processed the data in the following way: First, we removed a set of procedural phrases required by the rules of the House of Commons such as ”right honorable friend” or ”right honorable member” from the corpus.⁸ We further removed all, current and historic, names of British constituencies. This is necessary to prevent the model from clustering topics and sentiments around these commonly occurring phrases and words. Second, we drew a random sample of 10 percent of the data to increase computational efficiency, and aggregated all speeches by speaker and year. Finally, we converted the corpus to a document feature matrix containing 16,921 documents and 44,539 features.⁹

To identify the optimal number of topics, we estimated a series of LDA topic models and calculated their harmonic mean fit. Based on these results, we estimated a JST model with 105 topics and used the seed dictionary of Rheault et al. 2016 as a sentiment dictionary input. We chose JST over rJST since we are mainly interested in the overall expressed sentiment in parliamentary speeches. To compare our results to those reported in Rheault et al. 2016, we applied their dictionary to the same corpus as well. To measure emotional polarity, we divided the difference between positive and negative scores by the sum of positive and negative scores for each speaker-year entry, calculated a weighted average per year, and standardized the final score since we are interested in the development of emotional polarity over time rather than its absolute value. This procedure applies to both JST and dictionary results.

Figure 8 shows the development of emotional polarity in the British House of Commons between 1988 and 2013, comparing the results of our model with those of Rheault et al. 2016 dictionary and the Lexicoder dictionary.¹⁰ There is an upwards trend in both cases, showing that the rhetoric of debates in the House of Commons got more positive throughout that period.

⁸A full list of excluded phrases can be found in the appendix.

⁹Text preprocessing included: removal of numbers and punctuation, lowercasing, removing stopwords, Porter stemming, and trimming (min. 3)

¹⁰We show a comparison with other sentiment dictionaries in the appendix.

Figure 8: Emotional Polarity in the British House of Commons, 1988-2013.

For the most part, the estimates from the JST model are below those of the Rheault et al. 2016 dictionary. An interesting finding is that the two methods show a diverging development from 2010 onwards: While the JST scores keep increasing, the dictionary method picks up a negative development.

Next we are interested in the difference in emotional expression between government and opposition parties. Figure 9 and 10 show the emotional polarity of government and opposition in the British House of Commons between 1988 and January 2013. Both JST and the dictionary show a difference in emotional polarity between government and opposition throughout the time period.

Finally, to pick up on the claim that political life is becoming more sentimental, we investigate whether parliamentary debate in Britain became more emotional over the last two decades. For this we calculate a measure of overall arousal, that is the sum of positive and negative scores from the JST model. Figure 11 shows this development between 1988 and 2013. The results show that overall arousal increased between 1988 and the early 2000s, and remained on a quite steady level since 2004. This suggests that political debate indeed became more emotional since the end of the 1980s in the Britain.

Figure 9: Emotional Polarity of Government and Opposition in Britain. 1988 - 2013. JST.

Figure 10: Emotional Polarity of Government and Opposition in Britain. 1988 - 2013. Rheault et al. dictionary

Conclusion

In this paper, we set out to explore the relation between topics and sentiment in political texts. While in computer linguistics measures of sentiment already differentiate between contexts, this idea has not yet been applied to political science. We study this relation by applying the JST and the rJST. We found that political texts have a number of extra challenges. First, finding evenly distributed training data is complicated (see also Hopkins and King 2010). Second, the texts do not have a central focus like most of the training data from computational linguistics. Nevertheless, our models performed reasonably well. rJST in particular was able to distinguish between topics and sentiment categories.

Figure 11: Arousal in the British House of Commons, 1988-2013. Estimates from JST model.

Furthermore, we found that topics and sentiment are very much related. Our models picked up sentiment and topics in the same categories, suggesting that there is some correlation. This finding is further strengthened by the fact that rJST placed words carrying sentiment in different categories depending on the context, coming closer to their real polarity than a normal dictionary approach would.

Empirically, our comparison with a set of crowd coded speeches showed that using a sentiment-topic model improves the accuracy of sentiment estimates substantially compared to other, even highly domain specific, dictionaries. These benefits might be particularly strong when texts are long enough to contain several topics. This might be especially useful when analysing aggregated texts, for instance when making comparisons over time.

These findings echo those of Somasundaran (2010), who showed in online political debates that different sides chose to make references to different concepts, rather than one and the same concept. This complicates the study of sentiment in political texts. We suggest that for further studies into sentiment that some measure of controlling for topics is implemented. rJST shows promise and could be developed further, for example by comparing it to a reversed version of ASUM (Jo and Oh 2011).

Finally, this study does not go down completely to the aspect level. To entangle the relation between topic and sentiment, it might be necessary to dig down completely to the aspect level in a similar fashion to Somasundaran (2010). Such an approach would be less informative about sentiment, but can teach us a lot about how politicians intertwine topics and sentiments.

References

- Atteveldt, Wouter van, Jan Kleinnijenhuis, Nel Ruigrok, and Stefan Schlobach. 2008. “Good News or Bad News? Conducting Sentiment Analysis on Dutch Text to Distinguish Between Positive and Negative Relations”. *Journal of Information Technology & Politics* 5, no. 1 (): 73–94. ISSN: 1933-1681. doi:10.1080/19331680802154145. <http://www.tandfonline.com/doi/abs/10.1080/19331680802154145>.
- Benoit, Kenneth. 2017. *quanteda: Quantitative Analysis of Textual Data*. R package version 0.99.9001. <http://quanteda.io>.
- Blei, David M, and John D Lafferty. 2009. “Topic models”. In *Text mining: classification, clustering, and applications*, ed. by Akash Srivastava and Mehran Sahami, 10:34. 71. Chapman & Hall/CRC Data Mining / Knowledge Discovery Series.
- Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. “Latent dirichlet allocation”. *Journal of machine Learning research* 3 (Jan): 993–1022.
- Brader, Ted. 2005. “Striking a responsive chord: How political ads motivate and persuade voters by appealing to emotions”. *American Journal of Political Science* 49 (2): 388–405.
- Crabtree, Charles, Matt Golder, Thomas Gschwend, and Indridi H. Indridason. 2019. “It’s Not Only What you Say, It’s Also How You Say It: The Strategic Use of Campaign Sentiment”. *Journal of Politics*, no. Forthcoming. http://mattgolder.com/files/research/campaign%7B%5C_%7Dsentiment.pdf.
- Davis, Otto A., Melvin J. Hinich, and Peter C. Ordeshook. 1970. “An Expository Development of a Mathematical Model of the Electoral Process”. *American Political Science Review* 64 (2): 426–448. doi:10.2307/1953842.
- Downs, Anthony. 1957. *An economic theory of democracy*. New York: Harper & Row.
- Haselmayer, Martin, and Marcelo Jenny. 2017. “Sentiment analysis of political communication: combining a dictionary approach with crowdcoding”. *Quality and Quantity* 51 (6): 2623–2646. ISSN: 15737845. doi:10.1007/s11135-016-0412-4. [https://homepage.univie.ac.at/martin.haselmayer/..](https://homepage.univie.ac.at/martin.haselmayer/)
- Hopkins, Daniel J, and Gary King. 2010. “A method of automated nonparametric content analysis for social science”. *American Journal of Political Science* 54 (1): 229–247.
- Hu, Mingqing, and Bing Liu. 2004. “Mining and summarizing customer reviews”. In *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*, 168. New York, New York, USA: ACM Press. ISBN: 1581138889. doi:10.1145/1014052.1014073. <http://portal.acm.org/citation.cfm?doid=1014052.1014073>.
- Jo, Yohan, and Alice H Oh. 2011. “Aspect and sentiment unification model for online review analysis”. In *Proceedings of the fourth ACM international conference on Web search and data mining*, 815–824. ACM.
- Kosmidis, Spyros, Sara B Hobolt, Eamonn Molloy, and Stephen Whitefield. 2018. “Party Competition and Emotive Rhetoric”. *Comparative Political Studies*. ISSN: 15523829. doi:10.1177/0010414018797942. <https://doi.org/10.1177/0010414018797942>.
- Lauderdale, Benjamin E, and Alexander Herzog. 2016. “Measuring political positions from legislative speech”. *Political Analysis* 24 (3): 374–394.
- Lin, Chenghua, and Yulan He. 2009. “Joint sentiment/topic model for sentiment analysis”. In *Proceedings of the 18th ACM conference on Information and knowledge management*, 375–384. ACM.
- Lin, Chenghua, Yulan He, Richard Everson, and Stefan Ruger. 2012. “Weakly supervised joint sentiment-topic detection from text”. *IEEE Transactions on Knowledge and Data engineering* 24 (6): 1134–1145.

- Lind, Fabienne, Maria Gruber, and Hajo G. Boomgaarden. 2017. "Content Analysis by the Crowd: Assessing the Usability of Crowdsourcing for Coding Latent Constructs". *Communication Methods and Measures* 11, no. 3 (): 191–209. ISSN: 1931-2458. doi:10.1080/19312458.2017.1317338. <https://www.tandfonline.com/doi/full/10.1080/19312458.2017.1317338>.
- Liu, Bing. 2015. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.
- Mohammad, Saif M. 2016. "Sentiment Analysis: Detecting Valence, Emotions, and Other Affective States from Text". *Emotion Measurement* (): 201–237. doi:10.1016/B978-0-08-100508-8.00009-6. <https://www.sciencedirect.com/science/article/pii/B9780081005088000096?via%7B%5C%7D3Dihub>.
- Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. "Glove: Global vectors for word representation". In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.
- Proksch, Sven-Oliver, Will Lowe, Jens Wäckerle, and Stuart Soroka. 2019. "Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches". *Legislative Studies Quarterly* 44, no. 1 (): 97–131. ISSN: 03629805. doi:10.1111/lsq.12218. <http://doi.wiley.com/10.1111/lsq.12218>.
- Rabinowitz, George, and Stuart Elaine Macdonald. 1989. "A directional theory of issue voting". *American Political Science Review* 83 (1): 93–121.
- Rauh, Christian, Pieter De Wilde, and Jan Schwalbach. *The ParlSpeech data set: Annotated full-text vectors of 3.9 million plenary speeches in the key legislative chambers of seven European states*. doi:doi:10.7910/DVN/E4RSP9. <https://doi.org/10.7910/DVN/E4RSP9>.
- Resnik, Philip, and Eric Hardisty. 2010. *Gibbs sampling for the uninitiated*. Tech. rep. Maryland University College Park Institute for Advanced Computer Studies.
- Rheault, Ludovic, Kaspar Beelen, Christopher Cochrane, and Graeme Hirst. 2016. "Measuring Emotion in Parliamentary Debates with Automated Textual Analysis". *PloS one* 11 (12): e0168843.
- Rudkowsky, Elena, Martin Haselmayer, Matthias Wastian, Marcelo Jenny, Štefan Emrich, and Michael Sedlmair. 2018. "More than Bags of Words: Sentiment Analysis with Word Embeddings". *Communication Methods and Measures* 12 (2-3): 140–157. ISSN: 19312466. doi:10.1080/19312458.2018.1455817. <https://www.tandfonline.com/action/journalInformation?journalCode=hcms20>.
- Schumacher, Gijs, Daniel Hansen, Mariken A.C.G. van der Velden, and Sander Kunst. 2019. "A new dataset of Dutch and Danish party congress speeches". *Research & Politics* 6, no. 2 (): 205316801983835. ISSN: 2053-1680. doi:10.1177/2053168019838352. <http://journals.sagepub.com/doi/10.1177/2053168019838352>.
- Schumacher, Gijs, Martijn Schoonvelde, Tanushree Goyal, and Erik de Vries. 2016. *EUSpeech: A New Dataset of EU Elite Speeches*. doi:10.7910/DVN/XPCVEI.
- 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.
- Somasundaran, Swapna. 2010. "Discourse-level relations for Opinion Analysis". PhD thesis, University of Pittsburgh.
- Tausczik, Yla R, and James W Pennebaker. 2010. "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods". *Journal of Language and Social Psychology* 29 (1): 24–54. doi:10.1177/0261927X09351676. <http://jls.sagepub.com>.

- Theocharis, Yannis, Pablo Barberá, Zoltán Fazekas, Sebastian Adrian Popa, and Olivier Parnet. 2016. "A Bad Workman Blames His Tweets: The Consequences of Citizens' Uncivil Twitter Use When Interacting With Party Candidates". *Journal of Communication* 66, no. 6 (): 1007–1031. ISSN: 00219916. doi:10.1111/jcom.12259. <https://academic.oup.com/joc/article/66/6/1007-1031/4082403>.
- Traber, Denise, Martijn Schoonvelde, and Gijs Schumacher. 2019. "Errors have been made, others will be blamed: Issue engagement and blame shifting in prime minister speeches during the economic crisis in Europe". *European Journal of Political Research* (). ISSN: 03044130. doi:10.1111/1475-6765.12340. <http://doi.wiley.com/10.1111/1475-6765.12340>.
- Young, Lori, and Stuart Soroka. 2012. "Affective news: The automated coding of sentiment in political texts". *Political Communication* 29 (2): 205–231.

Appendix

Comparison with dictionary results

Figure 1 shows a comparison between the positive sentiment score from JST and the dictionary score for each text. Based on this, we can conclude that JST is capable of estimating sentiment scores that are similar to those calculated using the sentiment dictionary. As discussed above, there is a significant bias in the dictionary sentiment scores on our corpus. The scores JST assigns span a wider range than the dictionary scores.

Figure 1: JST positive sentiment score versus dictionary score

Note: The JST sentiment scores show the estimated probability of words being assigned to the positive sentiment categories for each text.

Reverse JST is not developed to measure a document-level sentiment parameter. With fewer words attributed to a topic, the measure for sentiment becomes less reliable. Topic parameters measure most topics in every article, although some only very little. Nonetheless, this means that we cannot aggregate a score to compare to the document-level dictionary measure. To evaluate the performance on prior information of rJST we need to assess to what extent it links sentiment words to sentiment categories. Table A1 shows the overlap between words lists of a sentiment in the trained model and the prior dictionary. The table shows consistently good performance on identifying positive words with positive sentiments across models. We could argue that rJST performs better on this measure, because the more topics it has, the better it identifies words that co-occur with negative sentiment words from our dictionary.

The performance on identifying negative sentiments does lag behind, most likely due to the shortage of negative words in our corpus. While it is easy to balance training data for reviews using their ratings, political scientists don't have such luxury. Given the excellent performance on identifying positive words, it is likely that a more balanced set of texts could yield better performance on both sentiments. Constructing such a training set is beyond the scope of this study. Nonetheless, this underperformance due to an unbalanced dataset does influence all our results. Given the numbers of false positives for the negative sentiment, rJST seems to perform better under these circumstances.

Overall, we can conclude that our models perform as we wanted them to. They incorporate prior information well and JST document-level scores resemble those from the dictionary. Reverse JST on the other hand did not, but by design that should be the model that performs better on the aspect level in our case given the potential big differences in terminology between topics.

Table A1: Overlap between likely words for sentiment and dictionary

		Positive		Negative	
	Topics	True Posi- tive	False Pos- itive	True Posi- tive	False Pos- itive
JST	10	62.9	2.3	15.5	9.5
	20	62.6	3.15	15	11.1
	25	69.16	2.88	15.6	10.68
reverse JST	10	62.6	2	16.2	8.3
	20	59.2	2.5	16.9	6.4
	25	59.92	2.84	17.68	7.92
	50	49.84	2.64	16.74	5.98
	100	38.64	2.91	14.44	3.37

Note: The above figures are the average number of words from the dictionary that occur in the top 100 words most strongly associated with a topic across all topics in the model. False positives are sentiment words from the ‘wrong’ category.

Table A2: Top 20 words for 4 random JST topics

Topic 8			Topic 17			Topic 22			Topic 25		
Neutral	Positive	Negative	Neutral	Positive	Negative	Neutral	Positive	Negative	Neutral	Positive	Negative
eurozon	right	bank	peopl	job	chang	tax	also	global	growth	invest	energi
europ	decis	lend	money	economi	countri	transpar	time	financi	deficit	new	oil
franc	servic	indistinct	tax	econom	let	trade	next	bank	britain	industri	price
european	govern	problem	pay	futur	system	compani	well	world	govern	technolog	food
british	clear	mortgag	spend	busi	problem	rule	look	crisi	economi	compani	produc
euro	secur	busi	pension	market	way	agenda	believ	system	debt	innov	invest
french	job	system	budget	help	rule	secur	make	economi	countri	growth	global
germani	polic	financi	rate	work	differ	m	necessari	countri	spend	develop	suppli
chancellor	case	money	save	invest	put	cooper	cours	action	take	open	market
britain	debat	hous	year	decis	yes	inform	meet	fiscal	deal	great	consum
merkel	made	deal	incom	last	simpli	ern	announc	stimulus	would	govern	demand
german	proper	rate	put	new	wrong	lough	new	monetari	budget	economi	renew
itali	respons	got	increas	creat	frank	uk	see	regul	tax	market	need
greec	public	billion	economi	interest	argument	pay	nation	problem	recoveri	success	fuel
sarkozi	review	compani	taxpay	confid	end	avoid	possibl	polic	public	competit	gas
currenc	believ	happen	cost	low	turn	lauren	prepar	america	rate	export	increas
singl	safe	england	allow	build	take	gas	move	europ	money	britain	rise
let	protect	bonus	parliament	us	doesn	indistinct	support	trade	pay	uk	product
angela	power	loan	basic	industri	without	eu	continu	rate	tough	product	electr
term	need	small	extra	plan	drive	growth	avail	imf	grow	univers	nuclear

Note: These are four topics picked by the authors for internal consistency.

Table A3: Top 20 words for 4 random JST topics

Topic 7			Topic 10			Topic 13			Topic 16		
Neutral	Positive	Negative	Neutral	Positive	Negative	Neutral	Positive	Negative	Neutral	Positive	Negative
year	servic	health	money	govern	economi	famili	govern	polic	oil	agre	trade
per	public	nhs	tax	econom	growth	children	local	crime	price	presid	china
cent	care	nurs	pay	make	deficit	help	help	communiti	food	g	g
billion	hospit	doctor	spend	job	britain	child	communiti	prison	suppli	agreement	countri
two	patient	dementia	cut	plan	debt	life	work	behaviour	demand	summit	deal
first	nation	year	budget	get	countri	parent	give	sentenc	increas	also	tax
five	peopl	treatment	go	need	deal	chang	power	crimin	fuel	meet	world
four	need	diseas	rate	decis	cut	relationship	area	street	consum	discuss	growth
three	want	drug	put	time	recoveri	feel	respons	peopl	economi	develop	chines
number	chang	healthcar	cost	last	year	kid	servic	justic	product	import	economi
week	reform	everi	taxpay	set	public	wellb	build	punish	global	world	compani
half	make	cancer	deficit	interest	difficult	polici	new	knife	gas	next	britain
seven	new	life	figur	back	sector	love	organis	antisoci	nuclear	make	global
across	choic	gp	increas	year	recess	evid	part	victim	electr	propos	uk
budget	better	prevent	less	deliv	take	coupl	public	action	sourc	believ	corrupt
spend	save	medic	incom	mean	spend	chris	enterpris	court	reduc	need	action
inde	best	heart	introduc	nation	unemploy	gdp	want	young	bill	forward	polit
month	improv	mother	extra	polici	still	practic	go	carri	high	togeth	issu
averag	person	treat	save	real	reduc	mother	author	parent	commod	growth	agenda
ten	live	scientist	low	creat	risk	societi	support	offend	car	econom	summit

Note: These are four topics picked by the authors for internal consistency.

Table A4: Top 20 words for 4 random JST topics

Neutral	Topic Financial crisis	
	Positive	Negative
countri	economi	global
world	intern	world
bank	system	crisi
fiscal	financi	action
trade	market	economi
stimulus	need	must
monetari	futur	act
regul	bank	financi
imf	institut	cannot
european	fund	recess
protection	help	globalis
g	polici	america
everi	interest	climat
europ	necessari	london
rate	togeth	face
tax	time	downturn
cut	econom	creat
obama	cooper	risk
financ	minist	common
billion	new	demand

Figure 2: Emotional Polarity in the British House of Commons, 1988-2015.