



## UvA-DARE (Digital Academic Repository)

### Blueprints and fingerprints

*Politicians' use of emotional appeals in European democracies*

Pipal, C.

### Publication date

2024

[Link to publication](#)

### Citation for published version (APA):

Pipal, C. (2024). *Blueprints and fingerprints: Politicians' use of emotional appeals in European democracies*. [Thesis, fully internal, Universiteit van Amsterdam].

### General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

### Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

## Chapter 2

# Taking Context Seriously: Joint Estimation of Sentiment and Topics in Textual Data

### *Abstract*

Our understanding of media tone or campaign dynamics relies in important part on our ability to measure sentiment in texts. The workhorse computational instruments for measuring sentiment are sentiment dictionaries. While such dictionaries can provide reasonably good results, they do not consider that the sentiment a word expresses can depend on its topical context. As a solution to this problem, we demonstrate the benefits of jointly estimating sentiment and topics using semi-supervised joint sentiment-topic models (JST and rJST). We validate the JST model with a multilingual hand-coded data set of parliamentary speeches, and show that taking topic-specific sentiment into account improves the accuracy of sentiment estimates over commonly used off-the-shelf dictionaries. It achieves this without any additional costs related to human annotation. We also show that the reversed JST model can identify and track meaningful topic-specific sentiments in parliamentary debates over time. To facilitate the adoption of JST/rJST we have developed the R-package *sentitopics*, fully compatible with *quanteda*, *tm*, and *tidytext*. By accounting for context in sentiment, JST and rJST improve dictionary applications in a way that is feasible for most researchers.

---

This chapter is an article co-authored with Martijn Schoonvelde, Gijs Schumacher, and Max Boiten. The manuscript based on a revised version of this chapter has received a revise and resubmit at *Communication Methods and Measures*.

The toxicity of political debate in many countries has stimulated interest in emotions across the social sciences. More than before, the rhetoric of politicians in social media posts (Eberl et al., 2020; Heidenreich et al., 2020; Heidenreich et al., 2022; Widmann, 2021), campaigns (Cho, 2013; Haselmayer & Jenny, 2017; Ridout & Searles, 2011), debates (Boussalis et al., 2021; Rhodes & Vayo, 2019), and legislative and leader speeches (Osnabrügge et al., 2021; Rheault et al., 2016; Traber et al., 2019) is analyzed for sentiment or tone. Scholars of communication also study it in news coverage of politicians and elections (Dunaway et al., 2015; Hopmann et al., 2011; Kleinnijenhuis et al., 2019; Vargo et al., 2014). Beyond the political domain, sentiment is studied in a wide range of topics such as the sharing of online news (Pipal et al., 2022; Valenzuela et al., 2017), health news coverage (Kim, 2015), media reporting of immigration (Eberl et al., 2018; Lawlor, 2015), and user comments (Muddiman & Stroud, 2017).

The workhorse computational instrument for measuring sentiment in such texts is the sentiment dictionary (Baden et al., 2021). Sentiment dictionaries are lists of words with fixed positive or negative meaning and sentiment analysis typically consists of counting how often these words occur in a text. They are easy to use on large quantities of text and – once developed – cost-free (but see Rice & Zorn, 2021). A problem, however, is that a word can have a positive meaning in the context of one topic and negative or neutral meaning in another topic. This so-called domain-specificity problem (Chan et al., 2021) limits the applicability of sentiment dictionaries in different institutional settings (e.g. parliaments vs social media), across languages, and across different substantive topics (see, e.g. González-Bailón & Paltoglou, 2015). Recent studies have made progress on this issue by developing domain-specific dictionaries that rely on machine translation, word embeddings or human coding (Müller, 2022; Proksch et al., 2019; Rauh, 2018; Rheault et al., 2016; van Atteveldt et al., 2008; Widmann, 2021). The approach we introduce in this paper is different. It does not assume that the meaning of sentiment words is the same across the entire domain. We demonstrate and validate the utility of jointly estimating topics and sentiment in social text with the Joint Sentiment Topic model (JST) and the reversed Joint Sentiment Topic model (rJST) (Lin & He, 2009; Lin et al., 2012).

The JST and rJST approach to estimating sentiment scores diverges from dictionary methods in two important ways. First, in assuming a fixed meaning of words dictionaries risk producing invalid estimates of sentiment in a text. The word ‘hard’, for instance, has a general negative connotation, but when it refers to overcoming difficult problems its meaning is clearly positive. Or consider ‘danger’: ‘saving people from danger’ is positive, while ‘being in danger’ is undoubtedly negative. In this paper we show an analysis of legislative rhetoric in which 30% of sentiment words have a positive meaning in some (political) topics and a negative meaning in others (see section 2.3). JST and rJST estimate topics and sentiment jointly, which allows the sentiment score of a word to vary with the topic in which it appears in a text. This mitigates the risk of biasing sentiment

scores. Second, JST and rJST can be used to generate uncertainty estimates of sentiment scores, which is not feasible with dictionary methods.

The core assumption underpinning JST and rJST is that a text is generated from a mixture of topics and sentiment but the order in which they do so is different. JST first assigns words to a sentiment category, and then assigns words to a topic. Hence, JST offers an efficient way to estimate the sentiment of a text taking into account that different topics are discussed. In contrast, rJST first assigns a word to a topic and then to a sentiment category. This way rJST allows researchers to describe topic-specific sentiment in a text. Importantly, it does so more efficiently than existing multi-step procedures that first identify the topic of a text, and then use (usually context-free) dictionaries to measure sentiment within these coded topics.

In this paper we detail the JST and rJST procedures and demonstrate their validity and utility for examining sentiment in social text. Our validation exercises offer several insights. First, we demonstrate that JST is a better predictor than dictionaries of human-coded sentiment in Dutch, English and German legislative speech. It achieves this without any laborious human coding. We also show that JST replicates the widely reported finding that government MPs use more positive sentiment than opposition MPs do. Second, we show that rJST produces valuable estimates of topic-specific sentiment. We illustrate this by showing that rJST output can be explained by government/opposition dynamics, and that it responds to high-profile events. Furthermore, we introduce the *sentitopics* R-package to estimate JST and reverse JST. *sentitopics* follows *tidy* principles (Wickham et al., 2019) and is fully compatible with widely-used text analysis packages such as *quanteda* (Benoit et al., 2018), *tm* (Feinerer et al., 2008) and *tidytext* (Silge & Robinson, 2016), making it easy for researchers to incorporate JST and rJST into their computational text analysis workflow.

In sum, we demonstrate that by jointly estimating topics and sentiment researchers can easily improve over plain dictionary applications. We demonstrate that with JST researchers can more precisely assess the overall emotional tone of a text than with dictionaries alone, without the need for coded training data. With rJST researchers can extract topic-specific sentiment, a useful quantity for many social science applications. Both tools allow for better measurement of sentiment, and help researchers to study new questions about emotional appeals in social text.

## 2.1 The advantages of jointly estimating sentiment and topics

Sentiment dictionaries such as LIWC (Tausczik & Pennebaker, 2010) or Lexicoder (Young & Soroka, 2012) are highly scalable, easy to use, and – once developed – their applica-

tion is cost-free. A problem with general sentiment dictionaries, however, is that human language is domain-specific (Chan et al., 2021). For example, dictionaries that have been developed on a news article corpus have been found to perform better than general sentiment dictionaries in analyzing sentiment in news articles (Boukes et al., 2019; van Atteveldt et al., 2021; Young & Soroka, 2012). Rauh (2018) demonstrates that a sentiment dictionary tailored to German legislative rhetoric and party manifestos outperforms a generalized dictionary. These findings imply that within the realms of politics and media tailored dictionaries are bound to perform better than generalized tools. But is this sufficient to mitigate bias? We think it is not. Because even within the realm of, say, politics, sentiment may strongly depend on the topics under discussion. A politician who talks about national security is bound to use language that is generally associated with negative sentiment (e.g. words related to war), but does not carry a specific sentiment within this context. There is no general answer to how much of a problem topic-specificity is in communication science research. But as we show in this paper, rJST specifically is a tool to examine its extent in particular applications, and both JST and rJST have procedures to describe sentiment while accounting for the issue of topic-specificity, allowing for more in-depth sentiment analysis.

JST and rJST have two further advantages. First, because they rely only on seed dictionaries, both models can easily be deployed in different languages. This is particularly useful because existing sentiment dictionaries are primarily developed for and validated on English texts with non-English dictionaries comparatively rare (see Baden et al., 2021). While recent work shows that machine-translated dictionaries can work, their correspondence with human judgments varies across languages (Proksch et al., 2019).

Second, other procedures to generate topic-specific sentiment exist, but none with the speed and low costs of rJST and JST. For example, one could identify a topic per document or per sentence using human coding or by running a topic model. As a next step one could then use these texts to generate topic-specific sentiment relying on, for example, human crowd coding or dictionary expansion with word embeddings. Human coding is expensive and time-consuming. Word embeddings can be used to populate a sentiment dictionary from a small set of seed words (Rheault et al., 2016). But the method is computationally complex and it requires prohibitively large text corpora (e.g. the whole of Wikipedia or decades of parliamentary speeches) to find meaningful semantic relationships between words. In many applications (e.g. political leader speeches), there are no relevant datasets of this size. Compared to these options, rJST and JST are faster (one-step procedure) and cheaper (no labelled input needed).

## 2.2 The Joint Sentiment-Topic and reverse Joint Sentiment-Topic Models

JST and rJST were originally developed by researchers in computational linguistics. This is particularly relevant, because JST and rJST have been validated on the relatively easy task of predicting sentiment in product or movie reviews (Lin & He, 2009; Lin et al., 2012). Such reviews are commonly written to express a positive or negative opinion about a product, movie, or service (for a discussion, see Maks and Vossen (2013)). Many texts like political speeches and manifestos are more complex and pose additional challenges to these models. The key difference between JST and rJST concerns how they assume a text is generated. rJST assumes that speakers first choose a topic and then a sentiment to go with that topic. rJST thus treats a document as a mixture of topics first and a mixture of sentiments second. In contrast, JST assumes that speakers choose a sentiment first and a topic second (Lin & He, 2009; Lin et al., 2012). For political speech one can easily think of examples in which either sequence makes sense. For instance, a politician may be forced to focus on a topic because of constraints set by the legislative agenda or because of specific audience demands. In such cases, the speaker first chooses the topic, then the sentiment. But a speaker could also be motivated to convey a specific sentiment first. For example, at a rally a politician may want to spread enthusiasm among party activists, regardless of the topic. Both sets of assumptions are equally plausible. Instead of deciding on the most plausible data-generating process, we recommend choosing between JST and rJST based on one’s research question. If you want to track overall sentiment across texts, we recommend using JST. If you want to track sentiment for a specific topic in a text we recommend using rJST.

Table 2.1 shows the generative processes of JST and rJST. We contrast these models with Latent Dirichlet Allocation (LDA) (Blei et al., 2003). LDA is a topic model capable of clustering the content in large text corpora. It treats documents as mixtures of topics and estimates the prevalence of these topics as well as their content. JST and rJST both add a sentiment layer to this. While JST first samples words from a document-level sentiment distribution, rJST first samples them from a topic distribution, just like LDA. Then for each topic rJST chooses a sentiment distribution. In the third step rJST assigns words to topics and sentiment. In contrast, JST first chooses a sentiment distribution, then chooses a topic distribution before assigning words to sentiment and topics.

To estimate JST and reverse JST we have developed the R-package `sentitopics`.<sup>1</sup> The package follows the implementation of Lin and He (2009) using a Gibbs sampler.<sup>2</sup> All `sentitopics` output follows *tidy* principles (Wickham et al., 2019) and is fully compatible with

---

<sup>1</sup><https://github.com/cpibal/sentitopics>

<sup>2</sup><https://github.com/linron84/JST>

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

---

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

**Table 2.1.** Generative processes of LDA, JST and rJST.

widely-used text analysis packages such as `quanteda` (Benoit et al., 2018), `tm` (Feinerer et al., 2008) and `tidytext` (Silge & Robinson, 2016), making it easy for researchers to incorporate JST and rJST into their computational text analysis workflow. In an accompanying online tutorial<sup>3</sup> we show how interested researchers can easily get started with the `sentitopics` package.

## Best practices and recommendations

While JST and rJST require only little manual input, researchers using these model will have to make several decisions about (1) text preprocessing, (2) the supervised input, and (3) model parameter settings.

First, we found that widely used text preprocessing steps had small effects on model results (see appendix 7.1). While performance differences between models estimated on a stemmed (reducing words to their grammatical stems) or trimmed (removing rarely used words) corpus are small, trimming the corpus reduced the average correlation between JST sentiment estimates and human coding by 0.02 while saving only about 10% in computing time. Especially for smaller corpora we therefore recommend estimating JST models on an untrimmed corpus. For all of our subsequent analyses we applied the following preprocessings steps: lowercasing, removing numbers and punctuation, and stemming.

Second, researchers have to select a dictionary to be used as supervised input. In the validation section we show that both general and context-specific dictionaries, as well as auto translated dictionaries, achieve good results, outperforming plain dictionary applications. One concern with dictionaries with high coverage (a large share of dictionary words present in the corpus) is that they leave too little room for the algorithm to find new associations between topics and sentiment. While Lin and He (2009) found the performance of the algorithm to improve when using a filtered dictionary (only keeping dictionary words appearing at least 50 times in the corpus), our models performed best when using the original dictionaries (see appendix 7.1).

Third, the models require researchers to choose between estimating two or three sentiment categories. We found that choosing three categories (including a neutral category) often increased performance (see appendix 7.1). Regarding the number of iterations of the algorithm, especially rJST models benefited from a large number of iterations in our tests. Accordingly, we recommend using at least 1000 iterations when estimating rJST models. The models also allow the tuning of hyperparameter settings which govern the learning behaviour of the algorithm. In our tests, hyperparameter  $\alpha$  (the per-document sentiment-specific topic proportion) had the largest effect on model output. While the default settings worked best for the JST models, setting this parameter to 0.1 produced

---

<sup>3</sup><https://github.com/cpipal/sentitopics-tutorial>

the most coherent topic-sentiment word lists for the rJST models. We therefore recommend trying out multiple values of  $\alpha$  and inspecting model output each time. Similar to LDA models, we also advise researchers to estimate several models with different settings and reading the model output and associated speeches closely to find the best settings.

In addition, we recommend estimating each JST model several times and averaging model predictions over these runs. Like LDA models, JST models do not give the exact same results across different runs (in the literature this is referred to as the issue of multimodality (see Roberts et al. (2016))). As we show in our validations, averaging across model runs results in reliable sentiment estimates. For document-level sentiment, this procedure also allows to compute uncertainty measures around the sentiment prediction using the variation across model runs.

Finally, for topic-specific sentiment estimation using rJST, researchers need to decide what topic prevalence is large enough to provide meaningful results. This is because topics with little presence are expected to have unreliable sentiment estimates. In our examples we found that texts with a minimum estimated topic prevalence of 5% ( $\theta \geq 0.05$ ) indeed covered the respective topic and produced meaningful topic-sentiment. However, we suggest that researchers applying the model read the model output and associated texts carefully to find their optimal solution.

## 2.3 Validating JST and rJST on legislative speech

The key question is whether JST and rJST provide more valid estimates of sentiment than sentiment dictionaries do. In this section we examine this question by validating both models on political texts. Because each model assumes a different data generating process and produces different outputs we need different validity criteria. For JST we evaluate *concurrent validity* of sentiment scores with human coding of the same texts across three languages. Furthermore, we assess *predictive validity* by assessing JST’s ability to capture government/opposition differences. For rJST we evaluate *face validity* and *discriminant validity* by assessing the extent to which rJST places words in different sentiment categories across topics. We also assess *predictive validity* in a way that is similar to JST, but focusing on government-opposition sentiment dynamics within specific topics.<sup>4</sup>

### Validating speech-level sentiment with JST

**Concurrent validity** We start with a multi-language validation of JST with hand-coded speeches from the British House of Commons, the German Bundestag, and the

---

<sup>4</sup>In this paper, we do not systematically assess the identification of topics by JST and rJST because the procedures are so similar to LDA, which has been extensively validated elsewhere (e.g. Blei, 2012)

Dutch Tweede Kamer. From the ParlSpeech v2 dataset (Rauh & Schwalbach, 2020), we sampled 200 speeches from each of these three parliaments, excluding speeches from the chair because those mainly concern parliamentary procedures and announce speaker turns. We recruited 9 native speakers in total (3 per parliament/language) whom we asked to judge if a speech was broadly positive, neutral, or broadly negative. Each speech was coded by 3 coders. Since we were interested in the overall tone of the speech we instructed coders to not focus on particular aspects of the speech, but to rate their overall impression of its sentiment. For this reason we introduced an upper limit of 500 words per speech because for longer speeches, often covering multiple topics, this coding task would have been too hard. We also excluded speeches with fewer than 50 words because they carry little substantive content. Did we expect agreement between coders? High agreement between coders is important when coding policy dimensions or topics. Regarding sentiment we expect coders to disagree to some extent, because differences between coders reflect the ambiguity of language regarding sentiment categories (Andreevskaia & Bergler, 2006; Subasic & Huettner, 2001).

We assigned numeric values to each coding (-1, 0, +1) and took the mean of the three coder ratings for each speech to arrive at a finer-grained continuous sentiment score. We present the distribution of coded sentiment for the three test sets in appendix 7.1. In all three corpora, coders originally judged a majority of speeches as broadly negative (UK: 48%, DE: 50%, NL: 50%) or neutral (UK: 37%, DE: 34%, NL: 28%), with only a much smaller proportion of speeches coded as broadly positive (UK: 15%, DE: 16%, NL: 22%). Such skewed distributions provide a challenge for machine learning algorithms, especially compared to curated and balanced training sets like movie reviews. Parliamentary speeches thus present a challenging test for the JST model.

To get JST estimates for these 200 speeches, we ran JST models with 10,000 randomly sampled speeches from each parliament (using the same criteria with which the 200 coded speeches were initially selected). Since JST is semi-supervised, it requires a sentiment dictionary as a seed dictionary. We compare JST performance when using general purpose and context specific dictionaries as seed dictionaries. This includes the Linguistic Inquiry and Word Count (LIWC) dictionary (Tausczik & Pennebaker, 2010) as a general purpose dictionary, and the Lexicoder Sentiment dictionary (LSD) (Young & Soroka, 2012), originally developed for political news coverage, as a partly-context-specific dictionary. While both dictionaries were originally developed for and validated on English texts, human translated (LIWC) and auto-translated (LSD) (Proksch et al., 2019) versions are available in German and Dutch. We also include two domain-specific dictionaries developed for parliamentary speeches. For the speeches from the UK we use the dictionary from Rheault et al. (2016), and for German speeches the dictionary developed by Rauh (2018). Due to the lack of sentiment dictionaries specifically developed for parliamentary speeches in Dutch we only used the LIWC and LSD dictionaries for

Corpus	Dictionary	r Dictionary-Human Coding	r JST-Human Coding
UK House of Commons	Lexicoder	0.46	<b>0.52</b>
	LIWC	0.28	<b>0.49</b>
	Rheault	<b>0.48</b>	0.47
DE Bundestag	Lexicoder	0.40	<b>0.45</b>
	LIWC	0.31	<b>0.37</b>
	Rauh	0.46	<b>0.48</b>
NL Tweede Kamer	Lexicoder	0.26	<b>0.36</b>
	LIWC	0.33	<b>0.38</b>

**Table 2.2. Mean Pearson correlations between automated measures and human coding.** Dictionary application vs. JST using same dictionary as input. JST results are averaged over multiple JST models with varying topic number  $k$  from 5 to 30 in steps of 5 and 10 runs per model.

estimating JST models using Tweede Kamer speeches. When estimated, JST returns three separate probabilities for the sentiment labels neutral, positive, and negative.

To arrive at a continuous sentiment measure per speech (similar to the dictionary measures) we subtracted the probability for the negative label from the probability for the positive label. Thus, a speech with a JST estimate of 0.11 (Neutral), 0.65 (Positive), 0.24 (Negative) received an overall sentiment score of 0.41. To see if our JST models outperformed dictionary applications, we also calculated sentiment measures using each dictionary alone. To this end we counted the share of words identified by each dictionary category (positive and negative) in each speech, and subtracted the share of negative matches from the share of positive matches. Since choosing an optimal number of topics *a priori* is difficult, we varied  $k$  between 5 and 30 in increments of 5 and estimated a JST model each time. Finally, we estimated each model ten times and averaged their results.

How do JST and dictionary sentiment scores perform relative to each other? Table A5 presents the (Pearson) correlations between (1) sentiment dictionaries and human-coded sentiment and (2) JST estimates and human-coded sentiment. In all but one cases we considered, the sentiment estimates obtained by JST have a higher correlation with human-coded sentiment than dictionary sentiment scores. It is only when we use the highly context-specific Rheault dictionary as seed dictionary that JST does not increase performance. While in this case the JST models perform slightly worse than the dictionary application ( $r = 0.48$  (dictionary),  $r = 0.47$  (JST)), taking the uncertainty introduced by the sample size of  $n = 200$  into account we conclude that these correlations are essentially the same.

Overall, JST is often able to estimate sentiment as well as dictionaries developed specifically for parliamentary rhetoric (Rheault and Rauh). This is exactly one of the goals of JST. All JST models consistently outperform context-free dictionaries (Lexicoder, LIWC). When no context-specific dictionary is available, our results indicate that for political speeches the Lexicoder dictionaries provide the best results when used as the

supervised input for JST models.

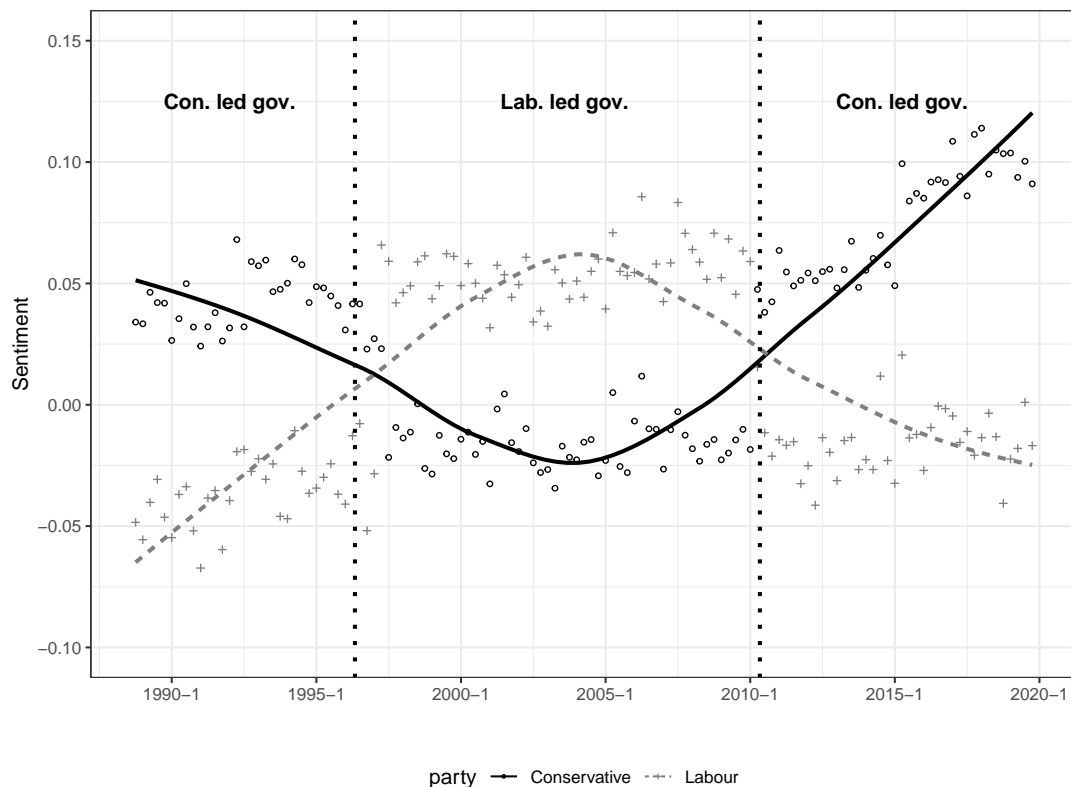
How much does choosing the “right” number of topics influence model performance? The performance does not fluctuate much between JST models with a different number of topics (mean range of  $r$ : 0.06, mean sd of  $r = 0.02$ ). We provide a breakdown of this variation by corpus, dictionary, and topic numbers in appendix 7.1. Therefore, guessing the “right” number of topics does not appear to be of much concern. Similar to our approach, analysts can always estimate several models within a reasonable range of topic numbers (e.g. by using LDA fit measures like perplexity and log-likelihood (Grün & Hornik, 2011)) and average the results.

**Predictive validity** To judge the predictive validity of JST, we probe the model’s ability to replicate a consistent result from previous analyses on sentiment in legislative speech: the finding that government MPs use more positive sentiment than opposition MPs (Crabtree et al., 2020; Kosmidis et al., 2019; Osnabrügge et al., 2021). To this end we again rely on the ParlSpeech v2 corpus examining all speeches from the British House of Commons between 1988 and 2020. In total, our corpus consists of 1,956,223 speeches from 2,175 unique speakers. We aggregated all speeches a speaker delivered in a quarter of a year, and estimated a JST model using these speaker-quarter observations (71,945 documents with 82,690 unique features). We estimated a 180-topic JST model using the Lexicoder sentiment dictionary as input. We arrived at this topic number by using LDA fit measures (see appendix 7.1).

Figure 2.1 displays the estimated JST sentiment scores aggregated by quarter and party in the British House of Commons between 1988 and January 2020 of Conservative and Labour MPs. Overall, JST replicates the previous consistent finding that government politicians speak more positively than opposition politicians: When the Conservative party is in government, its MPs speak more positively than Labour MPs. When they are in opposition, they speak more negatively than Labour MPs.

### **Validating topic-level sentiment with rJST**

To demonstrate the validity of topic-sentiment with rJST we again rely on parliamentary speeches from the UK between 1988 and 2020. We estimated multiple models with a varying number of topics and followed the preprocessing steps and model estimation procedures outlined above, again using the Lexicoder sentiment dictionary as our input dictionary. To increase computational efficiency we used a 10% random sample of Conservative and Labour speeches. Overall, the models were estimated on 135,121 speeches (with 22,262 unique features after preprocessing). We scrutinized the estimated topics qualitatively by inspecting the words highly associated with each topic-sentiment and by reading speeches that scored high on those. Using this procedure, we decided on a model



**Figure 2.1.** JST speech sentiment of conservative and labour MPs in the House of Commons. 1988–2020. Dots (crosses) are quarterly aggregated JST speech sentiment scores of conservative (black) and Labour (grey) members of parliament. Higher scores denote speeches with more positive sentiment. The solid line and dashed lines are fitted regression lines. Conservative and Labour led government periods are indicated with a dotted line.

Neutral	Armed Forces/Security		Neutral	European Union	
	Positive	Negative		Positive	Negative
armi	forc	defenc	european	european	eu
afghanistan	secur	royal	treati	europ	european
defenc	arm	ship	union	countri	leav
arm	oper	ministri	europ	britain	union
militari	continu	capabl	eu	british	agreement
personnel	must	navi	constitut	union	negoti
troop	train	aircraft	foreign	germani	uk
soldier	support	forc	articl	franc	deal
regiment	remain	procur	maastricht	french	brexit
veteran	also	air	singl	german	withdraw
afghan	well	equip	referendum	state	vote
royal	reserv	raf	negoti	eastern	trade
command	regular	mod	parliament	nato	remain
deploy	task	base	commiss	world	custom
ministri	commit	strateg	british	foreign	singl
serv	now	carrier	institut	presid	futur
civilian	effort	shipbuild	vote	eu	exit
battalion	serv	helicopt	veto	western	relationship
war	number	arm	sovereignti	join	citizen
british	howev	militari	council	itali	border

**Table 2.3. Highest loading words on senti-topics related to Armed Forces/Security and the European Union.** Both topic-sentiment words lists are drawn from the rJST model with 100 topics. The words are the top 20 words with the strongest association with their category.

with 100 topics. In appendix 7.1 we demonstrate the that rJST works equally well with smaller corpora and lower topic numbers.

**Face validity** To illustrate the rJST output we highlight two topics: armed forces/security and the European Union. Table 2.3 shows word stems that score highest on the three sentiment categories for each of these two topics. Most words appear in the category one would expect based on general sentiment dictionaries. However, stems like “forc” (positive) “royal” (negative) or “war” (neutral) are in a category that make sense in the context of the Afghan war. Importantly, these categories are different from those in the LSD dictionary, where “forc” is scored negative, “royal” positive, and “war” negative. Similarly, for the European Union topic “agreement” and “relationship” are identified as negative by rJST but are likely to be classified as positive by dictionaries (e.g. “agreement” is scored positive in the NRC dictionary). Many positive stems for the European Union topic refer to countries (e.g. “germani”, “french”, “britain”). This is a result of MPs emphasizing the strong, positive relations with these countries.

Placing these context-dependent words in their proper sentiment category is exactly

Speaker	Text	Sentiment (stdz.)
T. May 14 Jan '15	[...] Of course, we have long had detailed plans for dealing with these kinds of attacks. The House will recall the attacks in Mumbai in 2008 when terrorists armed with assault weapons and explosives took the lives of more than 150 people. Since 2010, and learning the lessons of that attack, we have improved our police firearms capability and the speed of our military response, and we have enhanced protective security where possible through a range of other measures. We have improved joint working between the emergency services to deal specifically with marauding gun attacks. Specialist joint police, ambulance and fire teams are now in place in key areas across England, with equivalents in Scotland and Wales, and they are trained and equipped to manage casualties in the event of that kind of an attack. The police and other agencies regularly carry out exercises to test the response to a terrorist attack, and these exercises include scenarios that are similar to the events in Paris. We will ensure that future exercises reflect specific elements of the Paris attacks, so we can learn from them and be ready for them should they ever occur in the United Kingdom. [...] (Topic: Armed Forces/Security)	rJST: 1.73 Dictionary: -1.19 Distance: 2.92
D. Raab 12 Jul '18	The referendum that we saw in 2016 was a brilliant example of a thriving democracy. That vote, whether it had been to leave or to remain, although the majority vote was to leave, was a vote of confidence in our democratic process. It was a vote of confidence in Britain, and it is incumbent on all of us to respect that result and deliver that outcome. The position respects the vote of the people in the referendum of 2016 to leave the EU. It also reflects the will of Parliament. The Government have successfully triggered article 50, pursuant to an Act of Parliament passed last year in the Commons on Second Reading by 498 votes to 114-an overwhelming majority-and they are negotiating for a good outcome that works for both the people and businesses in the UK and those in the EU. As someone who campaigned to leave the European Union, I understand that those who signed the petition are impatient to leave the EU and are asking the Government to leave the negotiations before 2019.[...] (Topic: EU)	rJST: -1.39 Dictionary: 1.75 Distance: 3.14
G. Brown 13 Mar '10	I know that the whole House will wish to join me in paying tribute to the three members of our armed forces from 1st Battalion the Royal Anglian Regiment attached to the Household Cavalry Regiment Battle Group who have lost their lives in Afghanistan this week. Their bravery and the sacrifice they have made for the future of Afghanistan and for the security of the British people will not be forgotten. Our thoughts today are with their families and loved ones as they receive this very sad news. I am sure that the House will also want us to pay respects to Dr. Ashok Kumar, who sadly died this week. He was a tenacious campaigner and a passionate advocate for the people of Teesside, and his expertise and wise counsel will be sorely missed at all times in this House. (Topic: Armed Forces/Security)	rJST: 0.36 Dictionary: 0.22 Distance: 0.14
P. Patel 1 Feb '11	I suggest that timetabling the required number of hours and days for such a debate could be quite challenging, because it would have to cover a vast number of issues. In my view, the British people deserve to know what their Government are planning to do, not only about the powers that the EU seeks to exercise but about those that it currently uses and-dare I say it-abuses, according to some in this House. Like all Conservative Members, I stood on a manifesto that clearly stated:The steady and unaccountable intrusion of the European Union into almost every aspect of our lives has gone too far.Following the ratification of the Lisbon treaty, we made a commitment not to let matters rest, and to negotiate the return to Britain of criminal justice powers and the opt-outs of the charter of fundamental rights and of social and employment legislation. The new clause would give the Government and the Prime Minister an annual opportunity to update the House on the actions being taken to deliver that, and to bring genuine openness and transparency to these proceedings. Forty years ago, when we entered what was then known as the European Economic Community, few could have predicted with any accuracy how deeply integrated and ingrained the EU has now become. Had we known that at the time, I am sure that this Bill would have been even more robust than it is. (Topic: EU)	rJST: 0.00 Dictionary: 0.12 Distance: 0.12

**Table 2.4. Example speeches on the Armed Forces / Security and Europe/EU topics.** rJST topic-sentiment scores and dictionary sentiment scores are standardized to allow the comparison of sentiment scores between methods. The distance between the two scores is their absolute difference. The full text of these example speeches can be found in appendix 7.1.

the goal of rJST.<sup>5</sup> But how context-dependent are these sentiment categories exactly? To answer this question we correlated the rJST topic-sentiment scores to the sentiment scores we would have obtained by just using the Lexicoder sentiment dictionary. For speeches covering the topic of Armed Forces/Security, the Pearson’s correlation is  $r = 0.05$ . For speeches covering the EU topic, the Pearson’s correlation is  $r = -0.04$ . This shows that rJST and sentiment dictionaries can produce very different substantive results.

To further illustrate the value of the rJST model, consider the first two speeches in table 2.4. The first speech was delivered by Theresa May and concerns the terror attacks in Paris and the readiness of the United Kingdom to counter such threats. The

<sup>5</sup>Not all topics are as consistent as the topics above. Comparable to LDA topic models, some topics are just so-called overflow topics.

dictionary scores this passage as negative, as a consequence of words such as “attacks”, “casualties”, and “terrorists”. rJST, however, gives it a positive score because May praises the government’s readiness in the case of terrorist attacks. Within the context of the topic of Armed Forces/security, it makes a lot of sense to consider this speech to be positive. In the second speech, Dominic Raab talks about the UK leaving the European Union. He doesn’t use any negative words, which results in a positive sentiment dictionary score. From our domain knowledge about the European affairs, it is however evident that Raab talks negatively about the EU. The rJST model has picked up on these domain differences and provides a negative score for this speech. Of course, we do not observe such differences for all speeches. For instance, the estimated sentiments using rJST and the plain dictionary are much closer for the speeches by Gordon Brown and Priti Patel in table 2.4. Here both methods give similar results. This again highlights that for short texts that usually only cover one topic, rJST and the dictionary can give similar results.

To sum up, we have given several examples of rJST picking up sentiment specific to a topic, where dictionary approaches fail to account for such context-specificity. An added benefit of rJST is that for longer speeches that contain various topics, the procedure can distinguish between the sentiment on these topics. In contrast, a dictionary procedure can only produce an overall score. Getting topic-specific sentiment scores using a dictionary would require a multi-step procedure.

**Discriminant validity** To what extent is rJST capable of discriminating between words that are always negative or positive, or are sometimes positive and sometimes negative? To answer this question, we selected the 100 most strongly associated words per sentiment-topic combination, relying on the rJST model that we introduced in the previous section. Per unique word we then created a list of the senti-topics with which that word is associated.

Among these 30,000 words (100 topics \* 3 sentiment categories \* 100 most strongly associated words), there are 10,337 unique words. There are fewer unique words than there are total words, since many words carrying sentiment are used across topics. Out of these 10,337 words, 2,343 (22.6%) are strictly associated with neutral sentiment categories. About 40% of the remaining words are only negative and 30% are only positive, while about 30% of words occur in the top 100 most strongly associated words of both negative and positive senti-topics. It is these words we are most interested in, because they show where topics and sentiment have an influence on each other, demonstrating the flexibility of the rJST model. We illustrate this using a selection of example words.

The word ‘difficulti(es)’, for instance, is commonly perceived as negative, but rJST places it in five positive and 4 negative senti-topics. Depending on the topic, it is used to describe difficulties or, contrasting, to solving or dealing with difficult situations. For instance, it is used negatively in the context of drugs/substance abuse, but posi-

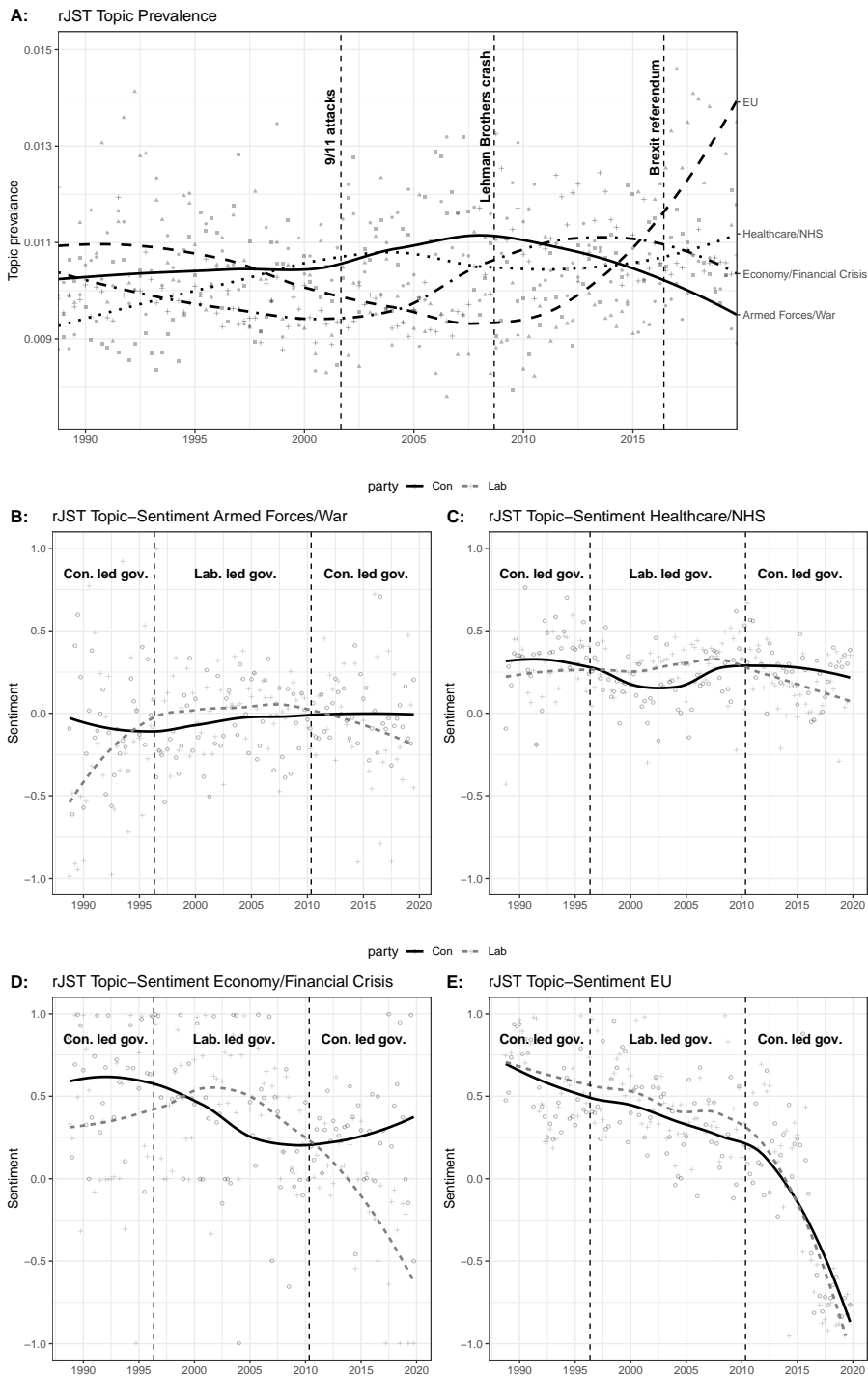
tively in the topic about insurance/finances. ‘Unemploy(ed/ment)’, on the other hand, is used almost exclusively in a negative contexts, for instance when talking about business/privatizations. However, on the issue of jobs/unemployment itself it is also indicative of positive sentiment. This is because in this context, it is often used in connection to “fighting unemployment” or other proposals to reduce unemployment. In contrast, the Lexicoder sentiment dictionary always treats the word as negative. The word ‘love’ is similar in this regards. Except for one topic it is always used in a positive senti-topic. But when talking about crime/suffering, “love” is put in the negative sentiment category. This is because the bi-gram “loved ones”—commonly used to describe the loss of someone—is reduced to the uni-grams “love” and “one” during preprocessing. Nevertheless, rJST is able to identify the contexts the word is used in. The Lexicoder sentiment dictionary, in comparison, always counts the occurrence of the word ‘love’ as positive.

Overall, every dictionary can only treat these words as either positive or negative, or ignore that they carry sentiment in some contexts entirely. From the 2,265 words that appear in both positive and negative rJST senti-topics, 143 appear in the negative word list of the Lexicoder sentiment dictionary, and 179 in the positive word list. Put differently, 6.04% (13.49%) of the words in the negative (positive) Lexicoder word list appear at least once in a rJST senti-topic with opposite polarity. Examples include words such as ‘obstacle’ (Lexicoder: negative, rJST: 66% positive), ‘mean’ (Lexicoder: negative, rJST: 22% positive), and ‘peace’ (Lexicoder: positive, rJST: 66% negative).

**Predictive validity** To probe the predictive validity of rJST and demonstrate its full output, we investigate topic prevalence and topic sentiment. Similar to our discussion of JST, we expect to see (1) an opposition/government dynamic, with MPs from the latter being generally more positive, and (2) responsiveness to high-profile events.

Figure 2.2 shows quarterly average topic prevalence and topic-sentiment for four selected topics over time. We use the same rJST model as in previous sections. The upper panel A shows the quarterly topic prevalence of four selected topics: Armed Forces/Security, Health care/NHS, Economy/Financial Crisis, and Europe/EU. In line with what we would expect, we see a increased salience of military and security issues as a topic after the 9/11 terror attacks. Similarly, economy/financial crisis as a topic was most prominent during the great financial crisis following the Lehman Brothers crash in 2008. Responsiveness is most evident when considering the European Union as a topic: its salience increased already before the 2016 Brexit referendum, and increased even more sharply after the vote to leave the EU. The lower panels B (armed forces/security), C (healthcare/NHS), D(economy/financial crisis), and E (Europe/EU) display the respective quarterly aggregated topic-sentiment.

Turning to topic-sentiment, in three of four topics we replicate the government-opposition dynamics we observe when plotting JST sentiment scores: politicians from



**Figure 2.2.** rJST topic-prevalence and topic-sentiment of four selected topics over time. The top panel A displays quarterly aggregated estimated topic-prevalence of four selected topics in UK parliamentary speeches using a 100 topic rJST model. Panels B - E display the quarterly aggregated topic-specific sentiment of these topics by party using the same rJST model. Sentiment of conservative (labour) MPs is indicated by a solid (dashed) line. For the rJST topic-sentiment scores only speeches with a topic probability of at least 0.05 for each respective topic were used.

the government party speak more positively than opposition candidates, also about specific topics. This is, however, not the case for the issue of the European Union, where for the most time Labour MPs speak more positively than Conservative MPs. In addition, following the Brexit referendum of 2016, politicians of both parties spoke increasingly negative about the EU with an, as expected, disappearing government-opposition effect. These results show that rJST can help us examine government-opposition dynamics in more depth: opposition speakers are comparatively more negative than government speakers on some topics but not on other topics.

## 2.4 Discussion

In this paper, we set out to demonstrate the validity and utility of estimating sentiment and topics simultaneously in social text using JST/rJST. Our findings underscore the added value that both models have for communication- and social scientists who want to conduct sentiment analysis on textual data. First, we have shown that sentiment scores estimated from JST are, in almost all cases we explored, better predictive of human-coded sentiment of legislative speeches than sentiment dictionaries. JST outperformed both generalized and context-specific sentiment dictionaries on this metric in English, Dutch and German speeches. Additionally, rJST allows for fast and efficient estimation of topic-specific sentiment. Not only is this a useful quantity of interest for many communication science applications (how positive or negative is a politician about one topic versus another?), it could also serve as a tool for constructing topic-specific sentiment dictionaries. Substantively, we demonstrated that rJST senti-topics track widely documented government-opposition dynamics and respond in expected ways to real-world events.

Plain dictionary applications are the workhorse tool for automated text analysis in the social sciences (Baden et al., 2021). JST and rJST provide a substantive improvement over these tools while being easy and fast to implement. They are also easier, faster and cheaper than methods relying on more sophisticated language models (e.g. word embeddings, transformer models) and local training data. JST and rJST thus account for context in sentiment in a way that is feasible for most researchers.

JST and rJST contrast with dictionary methods in two important ways. First, whereas dictionaries assume that a word has a fixed sentiment, JST and rJST learn how certain sentiment words can have positive meaning in some topics and negative or neutral meaning in others. This reduces the risk of measurement error and mitigates measurement bias. Secondly, both JST and rJST allow for generating uncertainty estimates around sentiment scores, something which is not feasible with dictionary approaches, regardless of whether they are general or domain-specific. Assigning a sentiment score to a word is a notoriously difficult thing to do, and JST and rJST offer applied researchers the oppor-

tunity to model this complexity as a function of the topic in which that word appears in a text. Importantly, in this paper we demonstrated that this more flexible approach of JST and rJST improves on dictionary methods in various important ways.

JST and rJST do require careful model specification. Especially with a smaller number of documents, sentiment scores and sentiment topics fluctuate across model runs, a problem that plagues topic models more generally and that has been referred to as the issue of multimodality. To account for this, we recommend estimating JST and rJST multiple times and averaging over model runs. As we have shown in this paper, this generates sentiment scores that consistently outperform dictionary methods and it produces meaningful sentiment-topics. Future research will need to investigate in more detail how data hungry these models are. More generally, JST and rJST are semi-supervised tools and so they require extensive validation by the analyst (Grimmer & Stewart, 2013). To this end, researchers cannot escape from close reading of key texts to ensure that the model results make good sense.