CHALMERS
UNIVERSITY OF TECHNOLOGY

Finding the structure of parliamentary motions in the Swedish Riksdag 1971–2015

Quality & Quantity
https://doi.org/10.1007/s11135-023-01802-9

Finding the structure of parliamentary motions in the Swedish Riksdag 1971–2015

Bastiaan Bruinsma[1] · **Moa Johansson**[1]

Accepted: 11 November 2023

Abstract
The current increase in the number of large, open sets of unstructured textual data has created both opportunities and challenges for social scientists. Here, we explore *if* and how we can use such data by looking at a dataset of over 144,000 documents used by parliamentary committees in Sweden. Of these, we aim to understand: (a) the topical content of these motions, (b) how these topics have changed over time, and (c) how these topics differ across political parties. To do so, we use a Structural Topic Model, which allows us to not only find the topics using the textual data itself, but also to include the documents' metadata, such as authorship and date of publication. Doing so, we find 30 topics, which we combine into 9 broader themes. We find that these themes often rise and fall in popularity in line with historical events, and relate to the various political parties as we would expect. Throughout our analysis, we provide a step-by-step overview of how to use structural topic models in practice and also how to handle the type of dataset we use here.

Keywords Structural topic model · Open data · Parliamentary motions

1 Introduction

In recent years, there has been a noticeable increase in the availability of large, open, datasets of textual documents. This growth is partly due to projects like the American Presidents Project[1] and the Comparative Manifesto Project,[2] which enable scholars to analyse a vast number of texts quantitatively. Additionally, many governmental organisations, including the EU[3] and various countries, now provide access to their textual data.

[1] https://www.presidency.ucsb.edu/.

[2] https://manifestoproject.wzb.eu/.

[3] https://data.europa.eu/en.

✉ Bastiaan Bruinsma
sebastianus.bruinsma@chalmers.se

Moa Johansson
moa.johansson@chalmers.se

[1] Department of Computer Science and Engineering, Chalmers University of Technology, Göteborg, Sweden

Published online: 18 December 2023

The question, however, is how useful such collections are in practice. In other words, how easy is it for researchers to use them to answer interesting questions? And if so, what obstacles and problems do they need to overcome before they can? It is well known that large-scale open data presents both theoretical and methodological challenges (Wilkerson and Casas 2017; Brady 2019). On the theoretical side, (e.g. González-Bailón 2013; Tinati et al. 2014; Baden et al. 2022) warn that one needs to have at least some theoretical knowledge about the data, while on the methodological side, there are questions about how to handle large datasets in practice, how to deal with data quality (e.g. Denny and Spirling 2018) and how to deal with bias. Moreover, especially in the case of textual data, the latter and the former are often intertwined (Grimmer et al. 2022).

In this paper, we look at one such dataset and ask ourselves a simple question: *what are the documents about?* Our data in question here are the 144,000 legislative proposals (also known as "motions") submitted to the various committees of the Swedish parliament, the *Riksdag*, since the adoption of the current unicameral legislature in 1971, which we take from the Swedish parliament's open data portal.[4] While we could easily have selected any other dataset, we find the motions intriguing in their own right due to their central role in the Swedish parliamentary system (Strøm 1998), where parliamentary committees are powerful and autonomous (Mattson 2016; Mickler 2022). Therefore, they offer a comprehensive overview of Swedish parliamentary interests over the past few decades and are representative of the documents scholars may wish to examine. Moreover, the large number of texts, the combination of scanned and digital documents, and the wide variety of text lengths make them a good example of the average open dataset one might come across.

To figure out what these documents are about, we turn to *topic models* (Blei et al. 2003). These are exploratory, unsupervised models that use word co-occurrence to find a set of topics that describe a corpus of text. As a result, they are popular amongst scholars who wish to study large, and relatively similar, sets of textual data such as speeches (Curran et al. 2018), or political agendas (Greene and Cross 2017). In particular, as we have access to not only the texts but also the metadata attached to them such as author and date of publication, we chose to use the Structural Topic Model (STM) (Roberts et al. 2019), as this allows us to include it to improve our estimates.

From here on, the article will proceed as follows. First, we will describe the context in which our data were generated, how we collected and corrected them, and which pre-processing steps we took. Then, we will first run a Structural Topic Model without any metadata (also known as a Correlated Topic Model), followed by one with metadata about the date of publication and the political party the author belonged. We will then cluster the topics we obtain from this into 9 overarching themes which we will attempt to label and interpret. After this, we will attempt to validate our topics by looking at how they behaved over time, with which parties they are associated and to which parliamentary committees they were originally sent. We then conclude with an extensive discussion on the usefulness of STM, as well as give certain pointers on how best to deal with large-scale open datasets such as the one we look at here.

2 Data

In total, our data contains 144,337 motions that were submitted between 1971 and 2015. Of the file types offered on the data portal, we choose the XML (Extensible Markup Language) type, as these contain both text and metadata in the same file. As in 606 of them,

[4] https://data.riksdagen.se/.

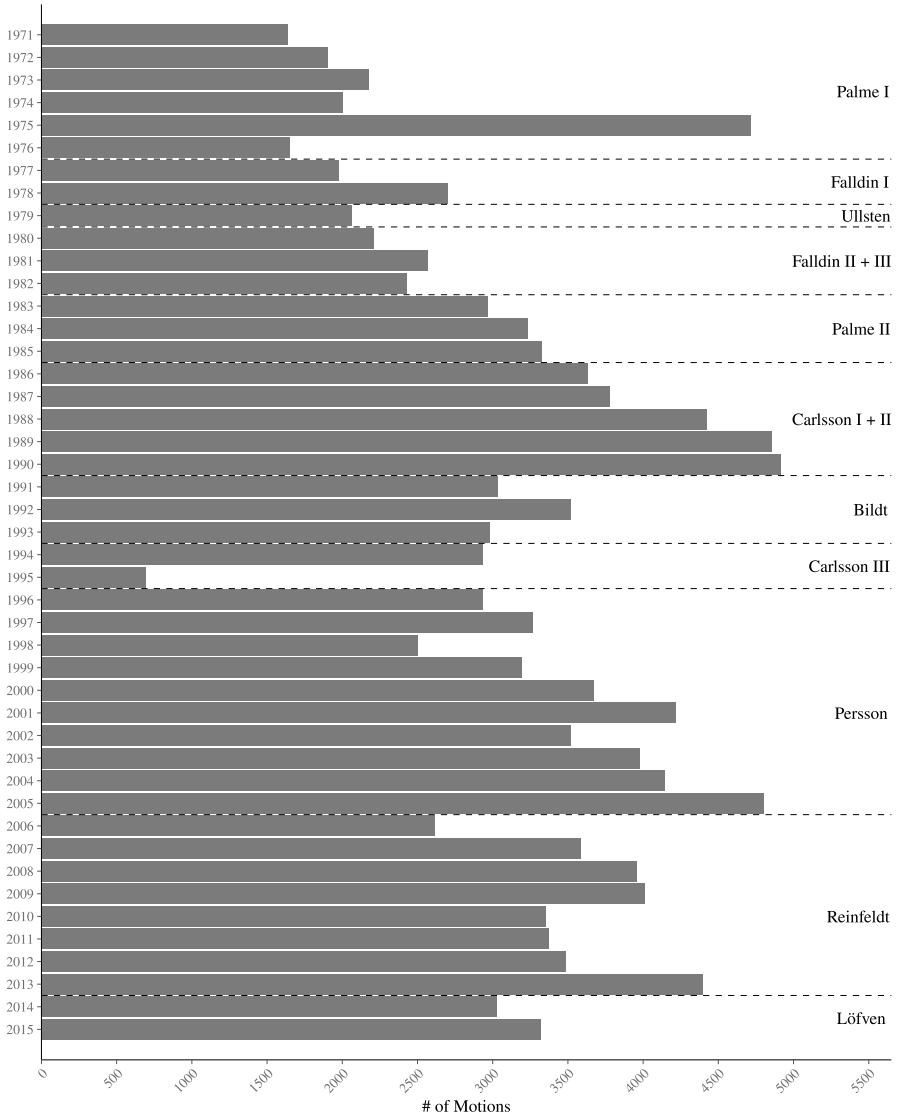

Fig. 1 Number of motions for each year, grouped by the Prime Minister leading the government at that point

this metadata was missing, we drop these, leaving 143,731 motions. On average, this results in 3000 to 3500 motions per year, though this figure is not stable. As Fig. 1 shows, there are some clear and interesting outliers. To begin with, 1990 saw the highest number of motions (4976), while 1995 saw the lowest (735). While interesting, the latter is most likely caused by a change in the budget year that took place that year. More interesting is the consistent pattern where the number of motions increases throughout a single electoral cycle. One reason for this might be that both government and opposition parties need time to get used to a new governmental composition.

As for the committees to which the parliamentarians submitted the motions (see Appendix B for an overview), we find that while some committees are more popular than others, the overall number of motions they receive is roughly similar. The most popular committees are the Health and Welfare (SoU), Transport and Communications (TU) and Education (UbU) committees, while the Defence (FÖU) and Foreign Affairs (UU) committees are the least popular. As for the motions themselves, we can divide them into various types. Of these, the most frequent are the *Enskilda motioner* (individual motions). These have a single author and are like the Private Members' Bills common to the British House of Commons. If they have more than one author, they are either a *Partimotion* when all authors come from the same party or a *Flerpartimotion* if they come from more than one party. If the motion comes from a member of a committee, the motion is also known as a *Kommittemotion* (Committee motion). In addition, if a motion is not tied to a specific topic, it is known as a *Fristående motion* (Free-standing motion), though parliamentarians can only submit these during the General Motions period (Allmänna Motionstiden) at the beginning of the parliamentary year. Finally, a *Följdmotion* (Follow-up motion) is a motion in response to another motion, either in defence or in opposition to it.

3 Pre-processing

Quantitative text analysis means first reducing a text first to words and then to numbers. As such, the input for any text analysis method is the so-called *data-frequency matrix* (DFM). This matrix contains, for each document, a count of how often a certain word appears. The process to arrive at the DFM is known as pre-processing. In our case, this pre-processing consists of three steps. First, we isolate the text we are interested in, then we correct any technical mistakes, and finally, we remove any words we are not interested in.

Starting with the first step, the XML files contain much text that we are not interested in. For example, almost all the motions contain a wide variety of headers, footers, and addresses, as well as various tables and figures. In addition, we note that during the digitalisation of the documents, extra text was often added. We remove all this, leaving only the main body of the text. Next, we deal with a wide variety of technical mistakes introduced when the original documents were scanned in. For example, localisation errors seem to have caused non-standard characters like "å" to become "Ã¥". Together with problematic glyphs like "ff" or "ll", we corrected these both manually and by using various regular expressions. Also, we removed "single" letters as we deemed them to be artefacts of the scanning process.

Having isolated and corrected our texts, we then turn to decide which terms to include in our DFM. Of the many techniques we can use to do so, Denny and Spirling (2018) identify seven: removing punctuation, removing numbers, lower casing, stemming, removing stopwords, including n-grams, and removing infrequent terms. As we can run these techniques in any order we like, there are $2^7 = 128$ possible combinations to choose from. As none of these combinations is inherently better than any of the others, it is therefore easy to get lost on one of the many "forked garden paths" (Gelman and Loken 2014). As each of these paths leads to a different dataset, each choice influences the reliability and validity of the result (Maier et al. 2018; Denny and Spirling 2018). As such, the final decision on which steps and which order to choose rests with the researcher and the aim of the

research.[5] As such, we consider it practical to remove symbols and numbers, lowercase our texts, remove the various stop words, and calculate n-grams. We choose not to apply stemming as this procedure has the same goal as the topic modelling itself. These both aim to combine similar words based on their context. As such applying stemming at this point might only make the topic model algorithm's job harder. Also, given that Swedish is more likely to contain compound words, the stemming process is harder (Lucas et al. 2015). As stemming reduces nouns to their root, this could lead to different compound nouns being reduced to a similar root (Proksch and Slapin 2009). As for removing infrequent terms, we can make a similar point. As Greene et al. (2016) note, compound words can lead to many infrequent terms, which might be relevant for our analysis. As such, we skip this step as well.

To carry out this pre-processing, we use the `quanteda` package in R (Benoit et al. 2018) which also allows us to generate the data-frequency matrix. In the end, our matrix counted 47, 330, 840 individual features, of which 544,974 were unique (for a more detailed overview of the number of features, unique features and sentences for each year, see Appendix E). This indicates that many tokens (even after stemming) were unique, leading to a very sparse dataset. When looking at this in the context of the committees, we find that the longest motions (based on the number of sentences) can be found at the Finance Committee (FiU), while the shortest ones occur in the taxation committee (SkU). This is roughly mirrored in the number of unique words (types) which is highest in the Finance committee, but lowest in the taxation committee.

The result of our pre-processing is a data-frequency matrix of 143, 731 long (the number of documents) and 544, 974 wide (the unique number of words). Together with an equally long dataset containing each document its date of publication and its author, this serves as the input for our analysis.

4 Topic models

The method we opt for here to investigate our documents are topic models, whose aim it to find the underlying, or latent, structure of a text. Given that they work without assumptions on what makes up the topics, they are a type of unsupervised method (Grimmer and Stewart 2013). Their underlying idea is that while writing a document a writer first chooses which topics to use, and then to which degree they will do so. Then, from each of the selected topics, they select the words belonging to it to construct the document. As such, topic models can help us say something about what a document is "about".

As with most methods of quantitative text analysis, topic models make three basic assumptions. First, that the word order itself is irrelevant. This approach, also known as the bag-of-words assumption, is common for nearly all such methods and assumes that word

[5] Besides this, Denny and Spirling (2018) suggest comparing all different pre-processing combinations and seeing how sensitive they are towards certain choices. To do so, they suggest comparing the pairwise distances between each of the combinations and then calculating a linear regression with the distances as the dependent variable and the various pre-processing decisions as predictors. Here, however, we run into computational problems as the resulting vectors of the pairwise distances are larger than our computational resources can manage.

order can be discarded without a significant loss of information (Grimmer et al. 2022). Second, documents are similar if they have similar words. Given the lack of word order, two documents that have the same words occurring with the same frequency are seen as fully similar documents. Third, they assume that each document is generated purposefully (that is, not at random) from a certain number of pre-existing topics.

Topic modelling first emerged during the late 1990 s, but only became successful with the introduction by Blei et al. (2003) of Latent Dirichlet Allocation (LDA). This then led to a succession of models that allowed for the inclusion of data other than just the words contained in the texts (e.g. Rosen-Zvi et al. 2004). Here we look at two of the most popular topic models: the Correlated Topic Model (CTM) and the Structural Topic Model (STM). We do so for three reasons. First, CTM is one of the most frequently used implementations of LDA and can serve as a good example of what a standard LDA would look like. Second, STM was designed as an evolution of the CTM, with the difference that STM does allow for *metadata*, while CTM does not. This thus allows us to see to what degree this metadata can help us. Both CTM and STM are available as alternate procedures in the same R package (`stm` Roberts et al. 2019).

4.1 The correlated topic model

CTM was developed by the same authors—Blei and Lafferty (2007)—as the original LDA. Its main aim was to get around one assumption of LDA that held that all topics are independent of each other and thus do not *correlate*. Yet, given that a document that includes the topic of "war", is also more likely to include a topic of "foreign affairs" than it is to include "pensions", this assumption is untenable at best.

As in LDA (Blei et al. 2003), CTM sees documents as a distribution of topics, while the topics themselves are a distribution of words. The idea then is that a word is chosen based on the distribution of topics in that document—θ—and then using that topic's distribution of words—β—to select a word. The distribution used here, a Dirichlet distribution, is used because it is skewed, thus providing just a small set of words with a high probability of occurring (cf. Blei and Lafferty 2007). While this distribution is both practical and functional in LDA, it comes with the downside that it tends to produce independent topic probabilities. To get around this, CTM relies on a logit-normal distribution instead (Blei and Lafferty 2007). This allows it to use the covariance matrix of the normal distribution to calculate the correlations between the topics.

As input, CTM requires textual data, as well as a pre-set number of topics. As there is no "correct" number of topics, this number depends on what we deem to be useful for our analysis (Grimmer et al. 2022). To help here, we follow the suggestions by Roberts et al. (2019) and first run a search function to estimate the range for the number of k topics we should choose. Figure 2 shows the result of this. Here, we find that for most of the indicators, there are large jumps—especially in the number of iterations needed—around 20 topics. This is most clear when we set out the semantic coherence of the topics (the degree of how often words occur together) against their exclusivity (the degree to which words are only associated with a single topic). As we aim for a balance between the two, we decide to focus on those models around 20 topics. Running a model for each of these, we then analyse the topics they generate qualitatively and select the model in which the topics are easiest to interpret.

To choose our constellation of topics, we take two steps. First, by using the words that are associated with each of the topics. These words are chosen based on the FREX

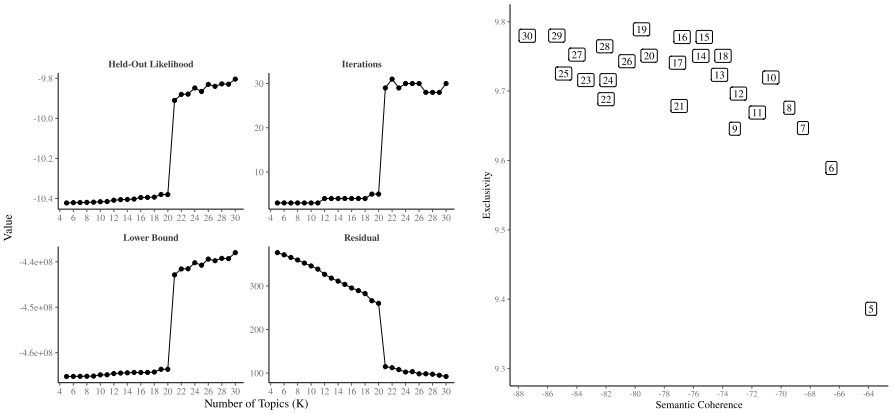

Fig. 2 Model diagnostics for 5–30 topics (left) and their Semantic Coherence versus Exclusivity (right) for the Correlated Topic Model

algorithm suggested by (Roberts et al. 2019). This considers both the frequency and the exclusivity of the words to link them with certain topics. Second, we ask the algorithm to provide us with 50 motions in which the topic is highly present. Doing so allows us to combine both the information from the model parameters with the act of reading the documents, as suggested by Grimmer et al. (2022). Based on this, we decided on a model with 18 topics.

In Appendix D, Fig. A1 shows the prevalence for each of the topics. Here, we find that all topics seem to occur with relatively equal frequency (between 4.9 and 6.2%). While this means that no single topic dominates the others, it also indicates that the topics might be similar in content. This becomes more clear when we look at Table A4, which shows the words most associated with each of the topics, based on their FREX value. To begin with, based on these words it is often difficult to say what the topics are about. For example, the most prevalent topic—Topic 17—contains terms such as cohabitants, transport policy, occupational health, national park, and tuberculosis. Also, when looking at the documents that contain the highest percentage of this topic, we find a similar, wide range of different ideas. This is the same with all other topics, which seem to be mixtures of sub-topics instead of topics of their own. In all, CTM was unable to provide us with a useful overview of our documents.

There are two ways we could address this. First, we could run a more fine-grained model. Yet, doing so (for 30 topics, see Fig. A2 and Table A5 in Appendix D), showed topics with a similar problematic interpretation. Second, we can see if providing more information to the model might help to separate them, which is what we do with the Structural Topic Model.

4.2 Structural topic model

STM (Roberts et al. 2014, 2019) builds on CTM by taking into consideration the metadata to better estimate the *prevalence* of the topics. Thus, it can use information such as the date of publication of a document to see if a certain topic is more likely to occur. It is this feature that makes STM popular, allowing it to be used to study gender differences during the

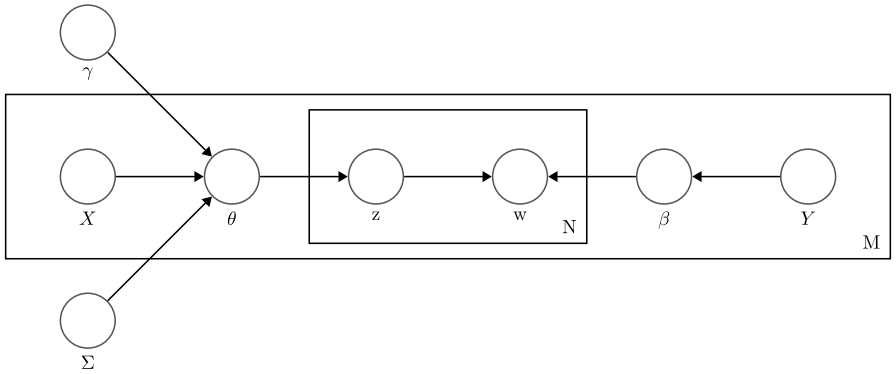

Fig. 3 Plate diagram for STM. Here X, refers to the prevalence metadata; γ, the metadata weights; Σ, the topic covariances; θ, the document prevalence; z, the per-word topic; w, the observed word; Y, the content metadata; β, the topic content; N, the number of words in a document; and M, the number of documents in the corpus

COVID-19 lockdown experience (Czymara et al. 2021) or ideological positions on climate change (Farrell 2015).

To understand the differences with CTM, Fig. 3 shows the plate diagram for STM. Here, as in CTM, an individual word w is part of the number of words in document N, which itself is part of the corpus M. From these word counts, STM then estimates the remaining parameters. The most important of these are θ, which measures to which degree a document belongs to a certain topic, and β, which does the same for each word. To do so, STM uses an expectation-maximisation (EM) algorithm that converges upon reaching a pre-set threshold (Roberts et al. 2019). For both β and θ, the variables X and Y refer to the metadata that governs the likelihood that either a word or a topic occurs in a document. For a complete description of STM and the derivation of the underlying algorithm, see Roberts et al. (2014, 2016).

4.2.1 Choice of prevalence variables

The choice of which metadata to use depends on which data we assume will best predict the prevalence of the topics. In our case, our documents come with (amongst others), metadata on their date of publication, the committee they were submitted to (for documents after 1985), the person(s) who wrote them, and their party affiliation. Of these, we choose *date of publication* and *party affiliation* as our prevalence variables. We do so for three reasons. First, unlike most other metadata, information on both variables is available over the complete period from 1971 onward. Second, we deem it reasonable that both *date of publication* and *party affiliation* will influence which topics will occur.

Besides helping the model, the prevalence variables also serve a second purpose when we later use them to validate our topics. This as studies over the past years have given ample descriptions of how the political system in Sweden has developed (e.g. Lindvall et al. (2019) and Aylott and Bolin (2019)) and which issues are being seen as being "owned" by each political party (e.g. Odmalm (2011) and Martinsson et al. (2013)). Thus, if, for example, a topic becomes more prevalent when it is also historically described to do so, or a party moves to pay more attention to a topic when there is evidence that voters

view this party as closer to owning the issue, this serves as a validation of sorts of our topics.

4.2.1.1 Date of publication As for the date of publication, we reason that political parties are expected to address and actively engage with evolving societal issues (Wagner 2012; Wagner and Meyer 2017). Thus, changing societal preferences over time will eventually be found within the motions as well. As for Sweden, between 1932 and 1976, the Social Democrats dominated the government in a variety of coalitions, until they were replaced by the first Fälldin government (a coalition of the Centre Party, Liberals and Moderates). This government lasted until the early 1980 s when the Social Democrats under Olof Palme again took over to form a new series of governments. It was also this decade that saw Sweden dealing with not only an economic crisis but also the new issues of nuclear power and environmentalism. Both had an impact: the former led to a referendum in 1980, while the latter presaged the rise of the Green Party and its election to the Riksdag in 1988. Even more changes would take place during the 1990 s. First, a coalition led by the Moderates under Carl Bildt rolled back many social policies and reduced the welfare state. Later Social Democratic governments continued this, leading to various pension reforms and a decline of corporatism (Lindvall and Sebring 2005). At the same time, on the international level, after a referendum in 1994, Sweden joined the European Community in 1995, though it chose not to join the Eurozone after another referendum in 2003. Finally, in 2014, a coalition of the Social Democrats and Green Party replaced the centre-right Alliance after the latter lost the elections that year (Berg and Oscarsson 2015). With this, the Social Democrats were back in the position they occupied for much of the previous century.

4.2.1.2 Party As for parties, we expect different parties to have different interests, and thus place different emphasis on certain topics. Note also that these positions are not necessarily stable, but can change over time—for example when new parties enter the scene (Hobolt and Wratil 2015; Meyer and Wagner 2020). In our case, there were nine parties, five of which existed over the whole period and four formed during it (see Appendix A for an overview). Apart from the Green Party in 1988, the Christian Democrats, after a brief period in 1985, would join as a permanent factor in 1991. This election also saw the sudden entry of the populist New Democracy, though they did not last longer than a single session. Later, in 2010, making use of a shift from a focus on socioeconomic to sociocultural issues, the populist Sweden Democrats would join, after taking part in elections since the late 1980 s (Rydgren and van der Meiden 2019). Other parties, such as the Pirate Party and the Feminist Initiative were unsuccessful in gaining seats in the Riksdag, though their policies still had an impact on national-level politics (Cowell-Meyers 2017). Note that, as one of the limitations of STM is that it cannot deal with multiple values in its covariates, it is not possible to run the algorithm with multiple different authors per document (also called a "Flerpartimotion").[6] To circumvent this, the authorship value in the prevalence of the model only included the first author.[7]

[6] There were 6969 of such motions in our dataset. The most common collaboration was between the Liberals, Centre Party and Moderates, with 1331 motions.

[7] Note that there are cases when members switch parties between elections. In those cases, we assigned the document to the party they switched to.

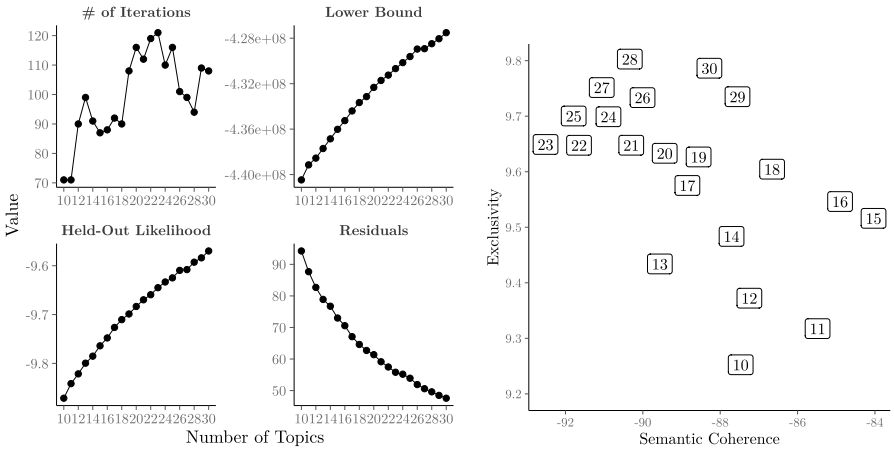

Fig. 4 Model Diagnostics for 10 to 30 topics (left) and their Semantic Coherence versus Exclusivity (right)

4.2.2 Number of topics

As with CTM, we again run a search function to find the optimal number of topics (see also Fig. 4). Here, we find that for most of the indicators, the graphs taper off after 30 topics. This is most clear when we set out the semantic coherence of the topics (the degree of how often words occur together) against their exclusivity (the degree to which words are only associated with a single topic). As we aim for a balance between the two, we decide to focus on those models between 10 and 30 topics. Running a model for each of these, we then analyse the topics they generate qualitatively and select the model in which the topics are easiest to interpret. Following the same approach as with CTM, we eventually decided on a model with 30 topics.

4.2.3 Clustering

To make the interpretation of these topics more manageable, and to acknowledge the fact that the topics are not independent, but are often related to each other, we will further cluster the topics into broader themes. To do so, we will use hierarchical clustering using Ward's method. We do so by using a logarithmic version of the θ matrix (the distribution of topics over each motion) as the input for the distance matrix (following a similar approach in Sánchez-Franco et al. 2021).

Figure 5 shows an overview of the thirty topics we find, as well as how they cluster together. From this, we decided to derive nine larger themes. Note that, as with most dimensionality-reduction methods, there is no optimal solution—that is, there is not one number of clusters. As a result, the choice of clusters is in some ways arbitrary (Theodoridis and Koutroumbas 2008; Müllner 2013). Yet, we can use a combination of various quantitative metrics and qualitative reading to help us. For the first, we draw on the `NbClust` package for R, which provides us with 30 different metrics. Here, we find that the optimal number of clusters is between 6 and 10. We then look at Fig. 5 and consider the implications of each of these cut-off points and the clusters they would lead to. Based on this, we then settle on an overall number of 9 clusters. This is because we feel that a higher

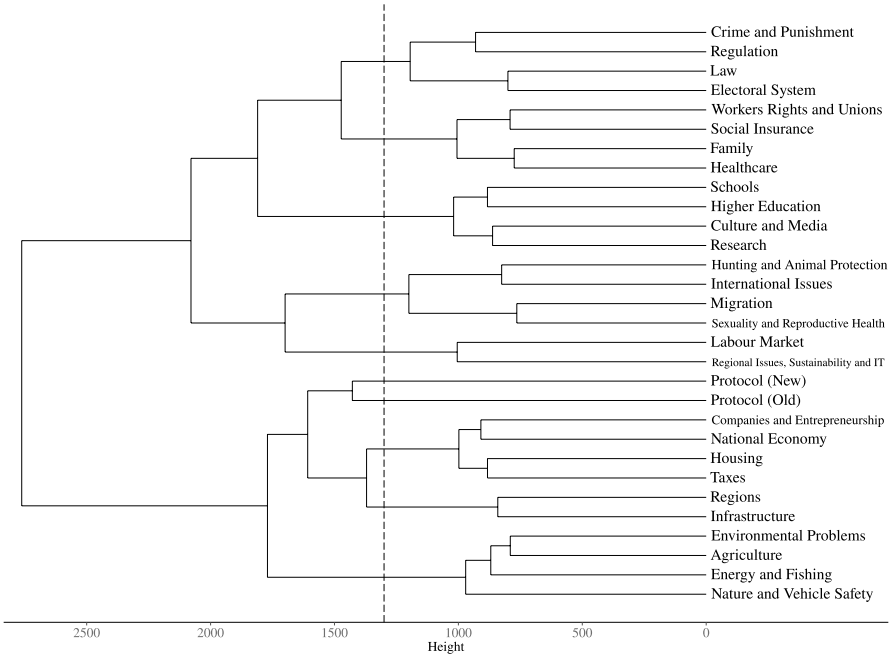

Fig. 5 Hierarchical clustering of the Topics using Ward's method. The dashed line shows the chosen cut-off of 9 clusters (combining both Protocol (Old) and Protocol (New))

number leads to too many clusters, while a lower number would put together topics for which the underlying relationship is less clear. For a further justification of this number of clusters, see Appendix F.

5 Results

We now turn to the results of our analysis. In each case we will describe the overarching theme as well as the individual topics they are built up of. As with CTM, we use the words that are the highest associated with each theme based on their FREX score, as well as the 50 motions in which the theme was highly present to do so.

5.1 Regulations

The first theme contains the four topics of Crime and Punishment, Regulation, Law, and Electoral System. All share a connection in that they focus on various aspects of the Swedish legal system and deal with multiple regulations. The Crime and Punishment topic deals with regulations focused on crimes. These include issues surrounding the prison system and problems such as drug use that occur in it[8]; issues surrounding the police and

[8] GR02Ju416, GO02Ju511. Note that these abbreviations refer to the unique document ID assigned to each of the documents.

the social services[9]; mentions of violent crimes and the release of offenders[10] and various motions calling for harsher punishments for various crimes[11] (e.g. organised house breakins). The Regulation topic also focuses on regulations but has a prevention focus. As such, it deals with issues such as the government-owned alcohol monopoly Systembolaget, and regulations for various industries, such as taxis, gambling, telesales and consumer credit. The Law and Electoral System topics refer in various specific ways to laws, either in a general context or when focused on the electoral system. Examples of motions here are those focusing on reforms to the electoral system or those with a focus on various individual laws.

5.2 Health and welfare

The second theme deals with various issues surrounding the Swedish social welfare system. One topic here—Social Insurance—deals with regulations and levels of pensions and sick leave. Here, we find references to "rätten_sjukpenning"[12] (entitlement of sickness benefits), "tilläggspensionen" (supplementary pension), but also references to institutions such as the "trafikskadenämnden" (Road Traffic Injuries Commission). Related to this is the topic of Workers' Rights & Unions, which deals with democratic rights in the workplace,[13] rights around strike actions,[14] and laws regulating the order in which an employer can make a worker redundant[15]. While most regulations seem to be on the employee's side, there are exceptions (such as GD02A705, arguing when a blockage is legal). The third topic, Healthcare, deals with various branches of healthcare and medicine. These include geriatrics, psychiatry, care for the chronically ill and care for patients with rare diseases. In addition, there are various motions related to specific diseases and calls for screening programmes. The fourth topic, Family, contains various family issues such as divorce and children,[16] childcare,[17] support for single parents,[18] parental leave[19] and children's rights. There are also some mentions of the reduction of working hours per week[20] and gender equality between parents.

5.3 Education and culture

The third theme covers four topics related to education and culture. The first, Schools, contains motions dealing with lower to middle education, with motions focusing on the school

[9] GT02Ju284.
[10] H002Ju263.
[11] H102Ju278.
[12] Note that this does not say "rätten till sjukpenning" as stopwords were removed before the n-grams were constructed.
[13] G4021439.
[14] G3021003.
[15] GR02A234—references to this show up with the abbreviation LAS, standing for "Lagen om anställningsskydd", or the Employment Protection Act.
[16] GK02L407.
[17] GS02Ub277.
[18] GQ02So377.
[19] GW02Sf218, GT02Sf422.
[20] GV02A406.

system in particular.[21] Also, there are motions dealing with more specific issues such as Swedish as a second language,[22] national tests,[23] actions against bullying,[24] and the grading system.[25] The second topic, Higher Education, deals with similar topics but has a focus on higher education programmes at universities (academic) and colleges (vocational). The Research topic covers the creation of new universities and the funding for research in general. In contrast with the previous two, this topic focuses more on grants and subsidies. This is a feature it shares with the fourth topic—Culture & Media. Here, we find mentions of subsidies and regulations for culture, media, religious associations and leisure. All these issues belong to the same expense area and are part of certain cultural politics (e.g. GP02Kr308). As such, they include calls for the support of handicrafts,[26] grants for public service radio and television,[27] and subsidies for musea.[28]

5.4 International and cultural issues

The fourth theme covers issues either related to international affairs or Swedish culture and nationality. The first, Hunting and Animal Protection, covers motions about hunting,[29] animal protection (e.g. animals at circuses,[30] livestock,[31] and commercial whaling[32]). In addition, this topic also contains motions dealing with the EU and EMU[33] and references the EU constitution. The second, International Issues, deals more with international affairs and foreign policy. As such, here we find references to conflict in the world (such as those in the Middle East or the Horn of Africa), and also various calls for disarmament. The Migration topic covers various aspects of migration policy. Given the controversy surrounding the topic, these are either restrictive[34] or liberal.[35] Some older motions also concern the concept of torture in Swedish law,[36] and temporary work permits for refugees waiting for decisions.[37] Newer motions include references to Christians in Iraq[38] and measures against prostitution and begging.[39] The fourth topic here—Sexuality and Reproductive Health—is one of the more complex topics. Based on the terms associated with it—bisexuella_transpersoner (bisexuel_transgender), lesbiska (lesbians) and abortlagen (abortion laws),

[21] GS02Ub390, GS02Ub390, GQ02Ub417.
[22] GQ02Ub325.
[23] GO02Ub322.
[24] GQ02Ub537.
[25] GR02Ub309, GP02Ub342, GO02Ub231.
[26] GS02Kr300.
[27] GO02Kr9.
[28] GR02Kr207.
[29] GQ02MJ387, GT02MJ478.
[30] GU02MJ320, GT02MJ468.
[31] GU02MJ323, GP02MJ514.
[32] GP02MJ345.
[33] GT02K429, GR02K377.
[34] H102Sf314, H102Ju387, GY02Sf385.
[35] H2021871, H102Sf230, H102Sf223, GS02Sf375.
[36] H302175.
[37] GS02So441.
[38] GZ02U326.
[39] GZ02Sf354.

this topic appears to be about sexuality, reproductive health[40] and the LGBTQ rights.[41] Yet, looking further, we find those motions to be only a small part of a wider mix of various controversial cultural topics. In various cases, motions here were often submitted and then resubmitted many years in a row. These include motions on the introduction of a republic, negative feelings toward the monarchy,[42] or the re-introduction of inheritance between cousins.[43] More recent issues concern the banning of the Islamic call to prayer[44] or the support for secular organisations.[45]

5.5 Labour market and regional development

This theme handles various motions surrounding the labour market as well as various calls for regional investment. The first topic—Labour Market—contains various budget motions with a focus on the labour market. This includes the reduction of employer contributions for young people[46] or calls for universal unemployment insurance.[47] The second topic—Regional Issues, Sustainability and IT—is another mixed topic. Most motions here refer to Expense Area 19 (Regional development), such as those motions referring to regional growth and service.[48] Yet, we also find motions related to IT and computing,[49] as well as motions on sustainability and climate change.[50]

5.6 Economy and taxation

This theme contains motions related to various proposals for different taxes and changes to the national economy. The first—Companies & Entrepreneurship—mentions privatisation,[51] plans to reduce public ownership[52] and technical risk boards.[53] Of interest is that the words associated with this topic contain many misspellings, which is most likely a result of various problems with the OCR procedure used to scan the original documents. The second topic—National Economy—is broader than the first and concerns most often budget questions and national economic policy. The third topic—Taxes—refers to issues such as tax scales,[54] taxes on company cars,[55] and other types of taxes. Given their topic, most of the motions here were submitted to the tax committee (Skatteutskottet). The fourth

[40] H2022605.
[41] GV02A384, GQ02K310.
[42] GY02K359.
[43] GW02C327.
[44] H30266.
[45] GT02L227.
[46] H2022274, H3021947.
[47] GP02A384.
[48] H002N340.
[49] H002N205, GX02T481, GY02Fi242, GV02T448.
[50] H002MJ6, H102MJ22, H002MJ222, H2023064.
[51] GR02N286.
[52] GK02N206.
[53] GS02N432.
[54] G7021494.
[55] GE02Sk335.

topic—Housing—refers to various issues related to the housing market, such as market pricing for rental properties,[56] forms of ownership[57] or the construction of housing.[58]

5.7 Regions

The theme contains motions dealing with the various regions of Sweden. As for the first—Infrastructure—we find motions arguing for funding of railways and roads in various parts of Sweden. Here, we find mentions of various infrastructural projects, such as the rail connection Västkustbanan, railway stations (stockholms_central), and airports (Kastrup). In the second—Regional—we find various issues surrounding regions. Most often this refers to how to solve unemployment there where industries have closed or moved.[59] As such, this topic has some relation to the Regional Issues, Sustainability and IT topic (see *Labour Market and Regional Development*), though here the motions are on the whole from an earlier date. It is also here that we find motions dealing with mining in the north of Sweden[60] the military[61], as well as large-scale plans for regional policy.[62]

5.8 Environment

The final theme covers four topics related to various aspects of the environment. The first—Environmental Problems—concerns the regulation of the use of multiple chemicals damaging to human health and the environment. This includes the use of chrome in leather products,[63] chlorine solutions,[64] and flame retardants.[65] Other motions here concern waste management, the introduction of recycling in Sweden,[66] and the protection of the ozone layer.[67] The second-Agriculture—focuses on various crops used in Swedish food production[68] and support to agricultural regions.[69] Also, we find various mentions of different types of animals (sheep, horses, and bees) in this topic. The third topic—Energy—concerns the different types of electricity generation in Sweden. Also, it captures the debate surrounding nuclear power as well as that of alternative sources of fuel for cars. Of interest here is that a few documents related to fishing have been included here[70] most likely as they mention "kW" in the context of fishing boat engines. The fourth topic—Nature and Vehicle Safety—is again somewhat mixed. Here, we find motions about parks, nature reserves and cultural landscapes, but also motions about vehicle safety and different types

[56] GT02Bo12, GO02Bo417.
[57] GZ02C277, GZ02C277, GP02Bo271.
[58] GW02C349.
[59] e.g. G802331, G7021922, G5021505.
[60] G5022253.
[61] GD02Fö3.
[62] GD02A42.
[63] GQ02MJ227.
[64] GC02Jo872.
[65] GO02MJ732.
[66] GA02Jo786, GO02MJ703, GF02So466.
[67] G6021456.
[68] GN02MJ242.
[69] GW02MJ473.
[70] e.g. GU02MJ237.

of terrain vehicles. One example of the latter is the high focus on motions about the reindeer industry, which often focuses on traffic accidents involving reindeer.

5.9 Protocol

This theme covers two topics that cover not so much the actual content of the motions, but as well as how they were written. We refer to this content as the style of protocol that those motions used. There were two versions of this—the protocol used in the 1970 s and those used later, from the 1990 s onwards. In the former, we find mentions of the King ("kungl_maj:ts_proposition") which disappeared after the new constitution in 1974 and references to other persons ("herr") or individual names. In the second, we find various fixed expressions, such as "motionen_anförts" (motion proposed), "budgetåret_anslår" (financial year estimates), and "enlighet_motionen_anförts_beslutar" (in accordance with the motion proposed decides). Overall, all of the motions dominated by this theme are short motions, and as such seems to be "dominated" by the formalia of the protocol text.

6 Validation

As STM, like all other topic models, is a type of unsupervised learning, there is no objective benchmark to validate our findings against. Hence, assessing the outcomes of an STM model using the customary standards of external, internal, and test validity poses challenges (Carmines and Zeller 1979; Zeller and Carmines 1980; Shadish et al. 002a). Instead, their validity is mostly framed in their perceived usefulness and the degree to which one finds the results convincing (Chang et al. 2009). Indeed, the idea of "usability" seems quite ingrained in the text-as-data approach itself (Grimmer et al. 2022). So, what can we do? To begin with, given that validation is establishing whether we are measuring what we aim to measure (King et al. 1994, p.25), our goal here is to see to which degree we are doing so. In this, our goal was to measure what the motions *were about*. Seen this way, our validation would exist of convincing ourselves that the topics we found seem somehow reasonable to occur.

Within our framework, there are three methods to accomplish this, one of which we have already employed during the actual discussion of topics and their subsequent interpretation. This is, in fact, the most commonly used method of validation in papers that use topic models (e.g. Lindstedt 2019). The second is to use the two parameters we included in our model: the date of publication of the motion and the party that submitted it. For example, if the topics we found behave over time as we expect them to based on historical occurrences, this strengthens our conviction that these topics are valid. Third, we can use an outside variable not included in our model—the committees to which the motions have been sent. Here, we would assume that the "Health and Welfare" committee would be mostly associated with the corresponding topic, and less so with, for example, topics related to the environment or the labour market. Taken together, this would then give us reason to believe—or not—that our topics in some way measure what we want them to measure—what the motions were actually about.

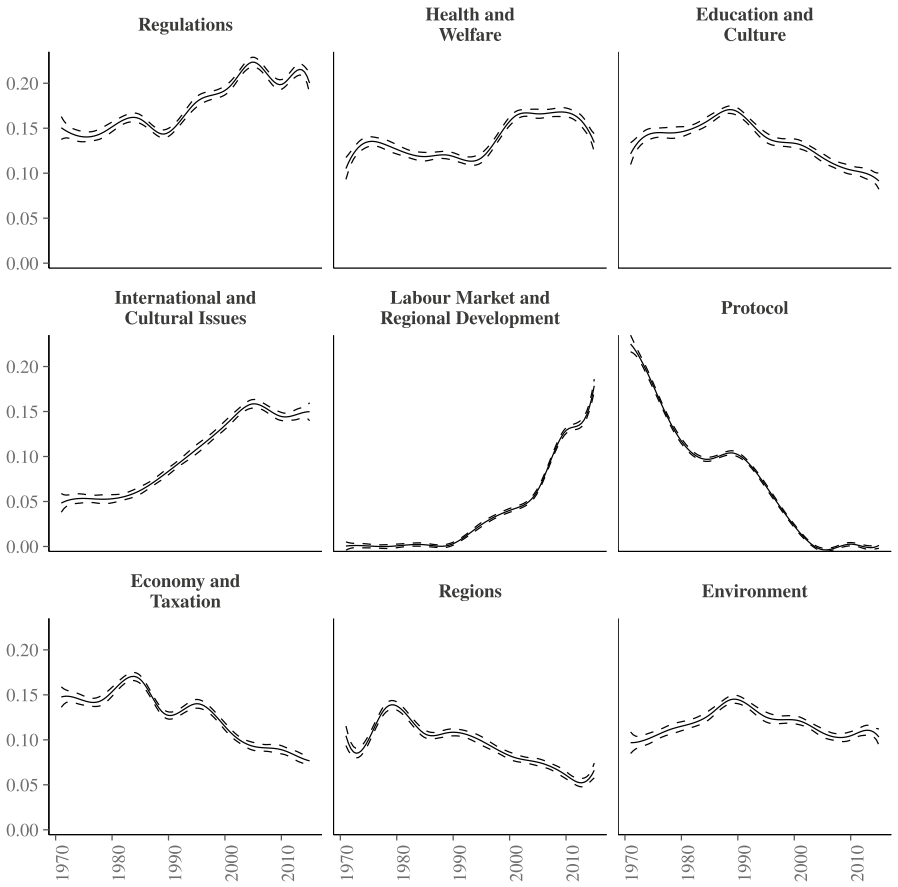

Fig. 6 Prevalence over Date of Publication. The solid line indicates the prevalence of a theme at that point, while the dotted lines indicate the confidence intervals

6.1 Date of publication

We start with the date of publication of the motions. To find the degree to which this variable correlates with the various topics—or, in other words, has an effect on it—we use the `estimateEffect` function from the `stm` package. This function estimates a simple linear regression where the documents are the units, the co-variate the year of publication and the outcome is the topic proportion in each document (given by θ) (Roberts et al. 2019). Figure 6 then shows the results of this, with the topics clustered into their respective themes. For each of these themes, the time scale runs from 1971 to 2015 on the horizontal axis, while the prevalence of the theme for each certain year is shown on the y-axis. Note that for each year, the values of all the graphs add up to one.

Starting with Regulations, we find that, after a rather stable period until the 1990 s, an upward trend led to that theme appearing in about 20% of all the motions around 2005. A similar thing happens with Health and Welfare, where between 1990 and 2000 the prevalence increased from around 12% to 18%. This climb seems to coincide with the

widespread privatisation of the welfare sector around that time (Garpenby 1995; Blomqvist 2004). As for Education and Culture, while there is a sharp increase between 1987 and 1998 (coinciding with the introduction of charter schools), afterwards follows a decrease, which might be caused by a combination of decreased interest of the state in steering the scientific field and the moving of the responsibility for schools from the state to the municipal level.

As for International and Cultural Issues, we see a sharp increase from the mid-1980 s onward. This is most likely a result of the recent growth in the salience of sociocultural politics in Sweden and the growth of importance to the issues (such as citizenship and Swedish culture) that are associated with it (Rydgren and van der Meiden 2019). The same goes for the Labour Market and Regional Development topic, with its prevalence rising sharply between 1989 and 1998, most likely a result of the various new labour market policies to combat the rising unemployment during that time (Carling and Richardson 2004). As for Protocol, we find that its prevalence is as expected, and is as much a part of the language used, as it might be an (unwanted) feature of the motions.

For the Economy and Taxation, we find an overall decline and two periods of strong increase. The overall decline seems to stem from the economy relying less on central governmental steering, while the two increases correspond with periods of strong deregulation. As for Regions, we find a consistent decrease—after an earlier sharp increase at the end of the 1970 s—indicating a decline of interest in regional policies. Finally, for the Environment theme, we find a clear peak during the late 1980 s, coinciding with the rise of the Greens and the increasing awareness of environmental problems during that time (Sundström 2011).

6.2 Parties

Apart from the time the motion was submitted, the second aspect of our model was the inclusion of the party that submitted the motion. Two points are of interest here: how many motions a party submitted, and which themes they were interested in. Recall that the motions only represent a part of the policies of the Riksdag, as governmental parties have a second way of getting their ideas onto the national agenda: the propositions (propositioner). As a result, we expect motions to be more the territory of the opposition than of the government.

Table 1 shows the number of motions per party for nine different periods. From this, we see indeed that parties in the opposition are highly over-represented in the motions. Starting with the Social Democrats under Palme in 1971, while consistently scoring above 40% in the total number of seats, they were responsible for only 15% of the motions. This is while the Moderates, holding around 15% of the seats, are responsible for about 30% of the motions. Later, during the Fälldin years, we see the same effect for the Social Democrats, as their share increases to 33% of the motions, while the share for the Centre, Liberal and Moderates decreases. The Social Democrat figure rose even more during the Bildt era, where they—as the opposition—were responsible for around 37% of the motions. This happened again in the Reinfeldt era, where they again were opposition and again were responsible for 38% of the motions. Of equal interest during this time is the large number

Table 1 Number of motions per party divided into periods per cabinet led by the same prime minister

	V	S	MP	C	L	M	KD	SD
Palme	1206	1977	–	4586	3896	3740	–	–
1971–1975	9.70%	15.89%	–	36.87%	31.32%	30.07%	–	–
Fälldin[a]	2160	4260	–	3201	1807	3301	–	–
1975–1982	16.40%	32.34%	–	24.30%	13.72%	25.06%	–	–
Palme	1461	2464	–	3494	2160	3530	–	–
1982–1985	12.21%	20.60%	–	29.21%	18.06%	29.51%	–	–
Carlsson	2187	4278	1783	5471	4698	5253	–	–
1986–1990	10.12%	19.80%	8.25%	25.32%	21.74%	24.31%	–	–
Bildt	936	3557	–	1448	1589	1499	989	–
1991–1993	9.81%	37.29%	–	15.18%	16.66%	15.72%	10.37%	–
Carlsson	372	932	452	565	542	816	362	–
1994–1995	10.25%	25.67%	12.45%	15.56%	14.93%	22.47%	9.97%	–
Persson	3187	9245	2939	5099	4706	9134	5910	–
1996–2005	8.79%	25.51%	8.11%	14.07%	12.98%	25.20%	16.30%	–
Reinfeldt	2009	10,886	2482	1839	2465	6626	2361	1340
2006–2013	6.98%	37.82%	8.62%	6.39%	8.56%	23.02%	8.20%	4.66%
Löfven	158	1183	200	674	526	2370	587	1058
2014–2015	2.49%	18.63%	3.15%	10.61%	8.28%	37.32%	9.24%	16.66%

[a]Also includes Ullsten between 1978–1979

Some periods span more than one election. Note that representatives from several parties may sign the same motion, which is then counted once for each signing party. The Motions %-row should thus not add up to 100%. The table excludes *Ny Demokrati*, which was active during the Bildt era, taking part in 818 motions, or 8.58% of the total for that era

of motions from the Moderates during this era. While part of the Alliance (a political alliance between the Moderates, Christian Democrats, Liberals and Centre Party), and being the PM's party, they still submitted many motions—23%—roughly the same number as their number of seats.

Turning now to the themes, Fig. 7 shows the interest of each party for each of the nine different themes (estimated in the same manner as the date of publication above). Here, we find that the Centre Party shares a large interest in Regional issues. This is expected, given the agricultural and regional base of their voters and their history as a party focused on the regions (Christensen 1997). For the Moderates, we find Economy and Taxation to be most important, fitting with their market-liberal focus. For the Greens, we find the expected dominance of the Environmental theme, while for the Sweden Democrats, we find a high degree of motions in the Labour Market and Regional Development theme, as well as International and Cultural Issues. Their interest in the first of these can most likely be explained by the success of the party in economically poor regions (Rydgren and Tyrberg 2020). Finally, the Left Party has an expected focus on Health and Welfare, dominating the topic together with the Christian Democrats.

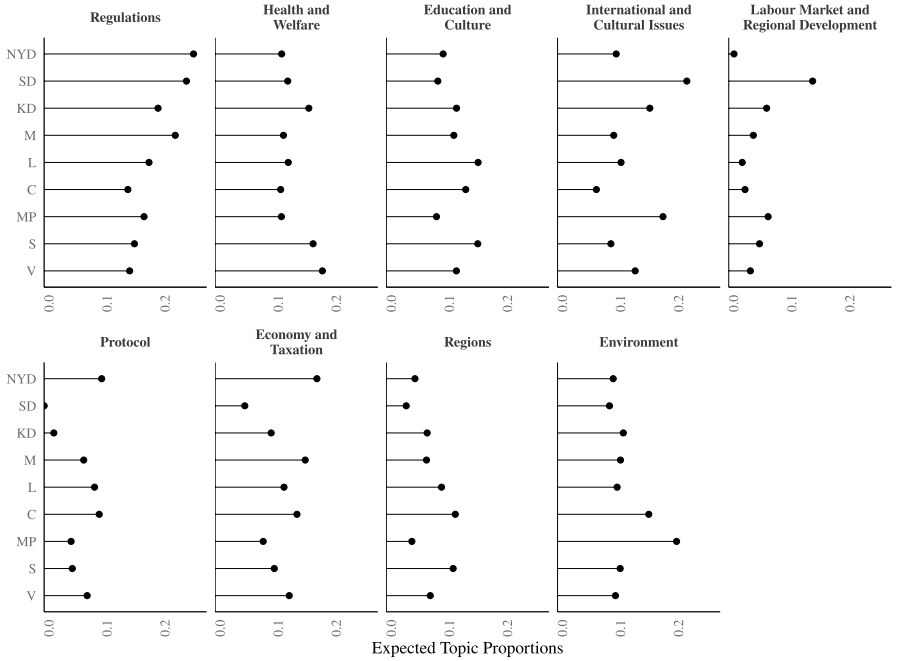

Fig. 7 Topic Prevalences over Parties. As the prevalences have been calculated over party, the prevalence sum to 1 for each party. Party abbreviations are: New Democracy (NYD), Sweden Democrats (SD), Christian Democrats (KD), Moderate Party (M), Liberals (L), Centre Party (C), Green Party (MP), Social Democrats (S) and Left Party (V)

6.3 Committees

While the inclusion of prevalence variables into STM allows us both to reveal a structure in the data that was otherwise hidden and helps us to validate our topics, they also lead to a complication: as we included the date of publication and the party authorship into our STM model as prevalence variables, it should come as no surprise that we find a structure in which the topics develop over time and differ between parties. Indeed, the prevalence variables used by STM are based on how the researcher *believes* the topics are structured. These beliefs then come from the researcher's own experiences and ideas regarding the topics they wish to find. As such, a topic model that includes certain prevalence variables cannot be used to "prove" that this prevalence variable had a certain effect. The strength of the prevalence variables thus lies in the fact that they allow us to reveal patterns that we expect to be there.

We can still object that what we are doing here is simply not more than revealing a parallel structure. Parallel in that there is (in a certain way) already a way in which the motions in the Swedish Riksdag are sorted into various topics: the various committees the motions are sent to. As mentioned, motions in Sweden are sent to either of the 18 committees (see Appendix B). We mentioned earlier that we did not include committee as a prevalence variable given that we lack information on which committee the motion was sent to for the first 15 years. Still, we can ask ourselves to what degree the themes we found reflect the various committees. Drawing on only the estimated prevalences from 1985 onwards, Fig. 8

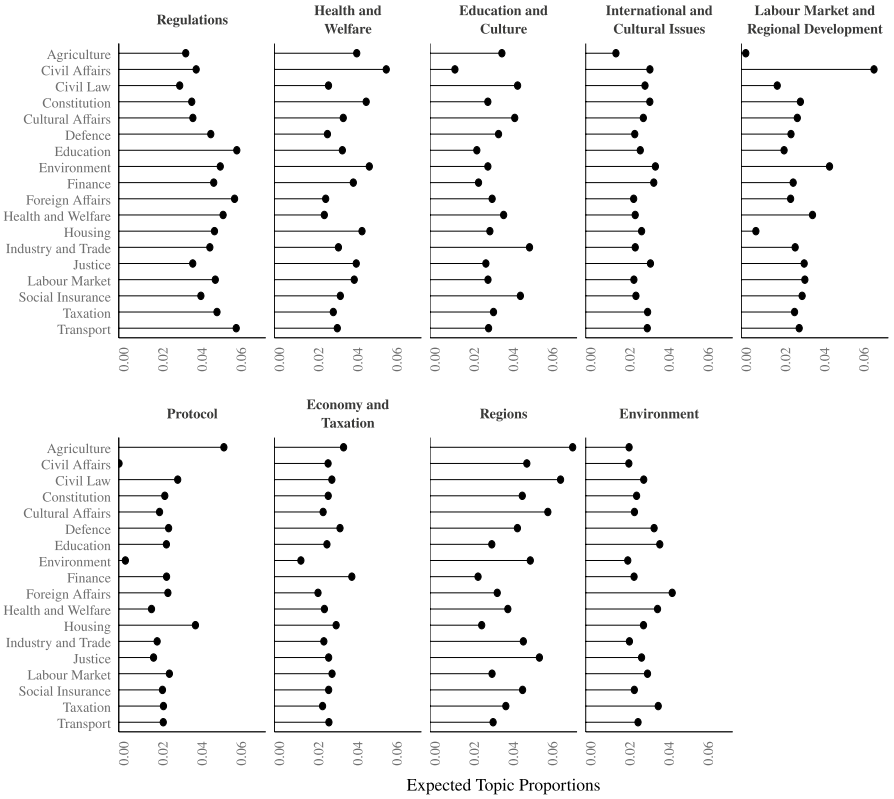

Fig. 8 Topic Prevalences over Committees. As the prevalences have been calculated over the committee, the prevalence sum to 1 for each committee

shows the expected topic proportions for each of the committees. A few points are of interest here. First, for most of the themes, the attention paid to them by the various committees seems quite even. That is, only in the case of "Labour Market and Regional Development" and "Protocol" does one committee (the Civil Affairs and Agriculture committee respectively), dominate. In addition, some of the committees do not dominate those themes we would expect them to. For example, the Education Committee scores significantly higher on Regulations than it does on Education and Culture. Instead, this theme is highly prevalent for the Industry and Trade Committee. Other committees do seem to adhere better to their expected theme though. For example, the Finance Committee scores highest in the Economy and Taxation theme, while Cultural Affairs scores high in Education and Culture.

These mixed findings suggest that simply looking at the Committees to which the motions are sent does not fully reveal the topical content of the motions. Instead, the motions sent to most committees are a "mixed bag" of various topics and themes instead of a domination of a single theme per committee. This holds up even if we consider the original 30 topics separately (see Appendix G).

7 Discussion

Our aim in this paper was to see if and how researchers can make use of the large and open sets of (textual) data that are available nowadays. Using the motions submitted to the committees in the Swedish Riksdag between 1971 and 2015, we run a Structural Topic Model to answer the straightforward question of *what* these documents were about. After selecting our documents and cleaning them, we find that using the metadata attached to them allows us to identify 9 themes. We find that these themes often develop over time as expected and often match the expected profile of the political parties quite well.

As one of the aims of our article was to discuss how to best deal with the type of data we use here, we conclude with several aspects that we wish to stress. These are a) the limits of the model one uses, b) the need for familiarity with the data, c) the pre-processing of the texts, d) the selection of the co-variates, and e) the validation of the topics. Note that while we are not the first to emphasise these points (e.g. Maier et al. 2018; Grimmer and Stewart 2013; Grimmer et al. 2022), we do repeat them here, as we find them crucial to any valid analysis.

Limits of the model No analytical technique is perfect. As Grimmer and Stewart (2013), p.270 note, there is "no globally best method for automated text analysis". The same is the case here with the Structural Topic Model. For example, when a small number of texts share a similar co-variate, STM prefers to group them (Grimmer et al. 2022, p.157–159). Here, we saw this happen with Regional Issues, Sustainability and IT, and Sexuality and Reproductive Health topics. As they all became more prevalent around the same time, the algorithm clustered them together. Yet, such limitations are not so much of a problem, as well as in-baked features the researcher should be aware of.

Familiarity with the data One should be familiar not only with the type of data and its probable content—but also with how these documents are stored. *Riksdagens öppna data* offers its data in various formats: plain text files, XML, HTML, JSON and SQL, each of which comes with its challenges. First, there is incomplete or incorrect metadata for some of the documents. As a result of this, authorship or author party affiliation is often missing, or even incorrect. Also, sometimes information is not recorded. For example, before 1985, there was no information to which committee a motion was sent. Second, there is the text of the documents themselves. As many of the older documents were digitised with OCR software, errors such as glyphs and unwanted dots are common. Also, especially during the early 1990s, a large number of documents were converted with the wrong encoding. This caused several letters (especially those unique to the Swedish alphabet) to appear as unreadable glyphs. As these aspects are not clear on first inspection, investigating the data is the only way to address these issues.

Pre-processing of the texts The choices used for pre-processing should be well considered and argued. For example, here, we decided against stemming our words as this would make it harder for the algorithm to run. We based this on the idea that both procedures aim to do the same thing: grouping similar words. A more difficult choice was the removal of numbers. On the one hand, numbers can provide interesting information, especially when part of n-grams. On the other, they often appear at random places in and outside the main body of the text (such as in page numbers, tables and addresses). As such, we opted to exclude these for our final analysis. Yet, we do agree with Denny and Spirling (2018) and Grimmer et al. (2022) that different decisions would have led to different topics.

Selection of the covariates Which co-variates to use influences both the number and content of the topics. Here, we used the authorship and year of publication to structure our

topics. We did so as we considered that the period between 1971 and 2015 was too long to assume that our topics remained stable. As the political climate changes, so do the topics politicians discuss and how they discuss them. In the same vein, we reason that different parties are bound to have different views on these issues. In addition, we find that including both co-variates helps increase the quality of our topics. As STM co-variates restrain the model, including them makes it easier for the model to find interpretable topics. Indeed, when we run the model without any of the co-variates—in which case it becomes a correlated topic model—the topics are very hard to interpret. We do note the limitation that when including co-variates there can be difficulties when such data is missing, or when the co-variate is different for equal documents (such as in our case when some motions were single-authored and others co-authored).

Validation of the topics The validation of topic models is problematic given the lack of a comparable ground truth. Instead, validation is achieved by establishing whether the topics are in any way likely or useful to the user (Chang et al. 2009). Yet, the process of doing so is precarious and therefore often the main part of the criticism levelled at topic models (e.g. Da 2019; Shadrova 2021). For example, one danger here is the tendency for scholars to submit to confirmation bias and find topics if they expect them to be there. For example, if one expects to find a topic related to traffic, any occurrence of words related to this might incline us to label a topic as such. Here, we aim to reduce the negative impact of these tendencies by looking at the information we do have: that is, the metadata that comes with our texts. By looking at how our topics developed over time and between parties and comparing this with real-world historical events, we aimed to strengthen our case that the topics we found in some way captured what the motions were about. That said, we want to stress that we also agree with Grimmer et al. (2022) in that it is wrong to see methods such as STM as having the aim to retrieve or recreate any *true* ground truth. This is because there are no such things as *real* topics that we could retrieve—topics are a construct in and of itself. Therefore, the validity of a topic model can best be seen in terms of whether it carries an acceptable interpretation of reality according to the expert who uses it.

8 Further work and conclusions

So, where do we go from here? Starting with the method, we could improve our topics further by using metadata to estimate not only the topical prevalence but also the topical content. As such, while the prevalence covariates we included here look at *who* discussed a topic, the content covariates measure *how* they discussed it. For example, using party authorship as a content covariate could show us how word choice differs between parties. This way we could track how different parties write about the same topic and if this changes over time. The main reason we did not do this here was one of practicality. That is, including the nine parties as covariates would lead to a very slow convergence of our model. As Roberts et al. (2019) note, this is due to the model having to replicate the complete dictionary of words for each party. This leads to a very high dimensional space, making the model intractable. As such, when we tried doing so, a single iteration of the EM algorithm took 53,945 s or close to 15 h. As our current model needed 108 iterations to converge, this would make the analysis a matter of days, instead of hours.

As for the data, we could consider models that are less sensitive to mistakes in the text and thus need less cleaning. More interesting are those models for which we do not need to construct a DFM. That is a model for which we can drop the currently dominant *bag*

of words assumption. While the idea that words are independent of their context is unrealistic, it is often taken for granted in most areas of quantitative text analysis (Grimmer et al. 2022). Yet, newer models, such as those using neural networks, do not call for it (e.g. Peinelt et al. 2020; Grootendorst 2022; Zhao et al. 2021). This allows the model to take not only the word into account but also the context in which it appears. As Bianchi et al. (2021) argues, doing so can lead to topics that score well on a wide variety of coherence metrics, though, as Hoyle et al. (2021) notes, this does not reflect in better topics. In most cases, they found that humans prefer topics derived from simple LDA over those using neural networks (Hoyle et al. 2022). Also, such models can exhibit unstable stochastic behaviour, and produce different results even when using the same data. Yet, given that such methods are still new, future work will likely work to address these initial obstacles.

Discussing the challenges of large open datasets in social science, Brady (2019) notes that they allow for all kinds of "new" questions political and social scientists can ask. In the same vein, Grimmer et al. (2022) stress that the social sciences, computer sciences and data sciences, are likely to co-operate even more in the future. This, we see as nothing but a positive development. On the one hand, the computer sciences and data sciences can help the social sciences analyse ever-increasing sizes of datasets, while on the other, social sciences can ensure their validation, quality and usefulness. This way, they can use the large, open datasets of text we discussed here to answer new, interesting, and (perhaps) groundbreaking questions.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11135-023-01802-9.

Funding Open access funding provided by Chalmers University of Technology. The computations in this paper were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Chalmers e-Commons partially funded by the Swedish Research Council through Grant Agreement No. 2018-05973. This work was supported by theWallenberg AI, Autonomous Systems and Software Program - Humanities and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation and the Marcus and Amalia Wallenberg Foundation.

Data availability All data is available from open, public archives and are linked and mentioned in the paper.

Declarations

Conflict of interest Both authors report no conflict of interest.

References

Aylott, N., Bolin, N.: A party system in flux: the Swedish parliamentary election of September 2018. West Eur. Polit. **42**(7), 1504–1515 (2019). https://doi.org/10.1080/01402382.2019.1583885

Baden, C., Pipal, C., Schoonvelde, M., van der Velden, M.A.: Three gaps in computational text analysis methods for social sciences: a research agenda. Commun. Methods Meas. **16**(1), 1–18 (2022). https://doi.org/10.1080/19312458.2021.2015574

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., Matsuo, A.: quanteda: an R package for the quantitative analysis of textual data. J. Open Source Softw. **3**(30), 774 (2018). https://doi.org/10.21105/joss.00774

Berg, L., Oscarsson, H.: The Swedish general election 2014. Elect. Stud. **38**, 91–93 (2015). https://doi.org/10.1016/j.electstud.2014.11.001

Bianchi, F., Terragni, S., Hovy, D.: Pre-training is a hot topic: contextualized document embeddings improve topic coherence. In: Zong, C., Fei Xia, W.L., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp. 759–766. Association for Computational Linguistics (2021)

Blei, D.M., Lafferty, J.D.: A correlated topic model of science. Ann. Appl. Stat. **1**(1), 17–35 (2007). https://doi.org/10.1214/07-AOAS114

Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003). https://doi.org/10.5555/944919.944937

Blomqvist, P.: The choice revolution: privatization of Swedish welfare services in the 1990s. Soc. Policy Adm. **38**(2), 139–155 (2004). https://doi.org/10.1111/j.1467-9515.2004.00382.x

Brady, H.E.: The challenge of big data and data science. Annu. Rev. Polit. Sci. **22**(1), 297–323 (2019). https://doi.org/10.1146/annurev-polisci-090216-023229

Carling, K., Richardson, K.: The relative efficiency of labor market programs: Swedish experience from the 1990s. Labour Econ. **11**(3), 335–354 (2004). https://doi.org/10.1016/j.labeco.2003.09.002

Carmines, E.G., Zeller, R.A.: Reliability and Validity Assessment. Sage, Thousand Oaks (1979)

Chang, J.D., Boyd-Graber, J.L., Gerrish, S., Wang, C., Blei, D.M.: Reading tea leaves: how humans interpret topic models. In: Bengio, Y., Schuurmans, D., Lafferty, J.D., Williams, C.K.I., Culotta, A. (eds.) Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009, vol. 22, pp. 288–296. Curran Associates Inc (2009)

Christensen, D.A.: Adaptation of agrarian parties in Norway and Sweden. Party Polit. **3**(3), 391–406 (1997). https://doi.org/10.1177/1354068897003003007

Cowell-Meyers, K.: The contagion effects of the feminist initiative in Sweden: agenda-setting, niche parties and mainstream parties. Scand. Polit. Stud. **40**(4), 481–493 (2017). https://doi.org/10.1111/1467-9477.12097

Curran, B., Higham, K., Ortiz, E., Filho, D.V.: Look who's talking: two-mode networks as representations of a topic model of New Zealand parliamentary speeches. PLoS ONE **13**(6), e0199072 (2018). https://doi.org/10.1371/journal.pone.0199072

Czymara, C.S., Langenkamp, A., Cano, T.: Cause for concerns: gender inequality in experiencing the COVID-19 lockdown in Germany. Eur. Soc. **23**(S1), 68–81 (2021). https://doi.org/10.1080/14616696.2020.1808692

Da, N.Z.: The computational case against computational literary studies. Crit. Inq. **45**(3), 601–639 (2019). https://doi.org/10.1086/702594

Denny, M.J., Spirling, A.: Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it. Polit. Anal. **26**(2), 168–189 (2018). https://doi.org/10.1017/pan.2017.44

Farrell, J.: Corporate funding and ideological polarization about climate change. Proc. Natl. Acad. Sci. **113**(1), 92–97 (2015). https://doi.org/10.1073/pnas.1509433112

Garpenby, P.: Health care reform in Sweden in the 1990s: local pluralism versus national coordination. J. Health Polit. Policy Law **20**(3), 695–717 (1995). https://doi.org/10.1215/03616878-20-3-695

Gelman, A., Loken, E.: The statistical crisis in science. Am. Sci. **102**(6), 460–465 (2014). https://doi.org/10.1511/2014.111.460

González-Bailón, S.: Social science in the era of big data. Policy Internet **5**(2), 147–160 (2013). https://doi.org/10.1002/1944-2866.POI328

Greene, D., Cross, J.P.: Exploring the political agenda of the European parliament using a dynamic topic modeling approach. Polit. Anal. **25**(1), 77–94 (2017). https://doi.org/10.1017/pan.2016.7

Greene, Z., Ceron, A., Schumacher, G., Fazekas, Z.: The nuts and bolts of automated text analysis. Comparing different document pre-processing techniques in four countries. OSF Preprints, Center for Open Science (2016)

Grimmer, J., Roberts, M.E., Stewart, B.M.: Text as Data: A New Framework for Machine Learning and the Social Sciences. Princeton University Press, Princeton (2022)

Grimmer, J., Stewart, B.M.: Text as data: the promise and pitfalls of automatic content analysis methods for political texts. Polit. Anal. **21**(3), 267–297 (2013). https://doi.org/10.1093/pan/mps028

Grootendorst, M.: BERTopic: neural topic modeling with a class-based TF-IDF procedure (2022)

Hobolt, S.B., Wratil, C.: Public opinion and the crisis: the dynamics of support for the euro. J. Eur. Publ. Policy **22**(2), 238–256 (2015). https://doi.org/10.1080/13501763.2014.994022

Hoyle, A., Goel, P., Hian-Cheong, A., Peskov, D., Boyd-Graber, J., Resnik, P.: Is automated topic model evaluation broken? The incoherence of coherence. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems, vol. 34, pp. 2018–2033. Curran Associates, New York (2021)

Hoyle, A., Goel, P., Sarkar, R., Resnik, P.: Are neural topic models broken? In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Findings of EMNLP 2022, pp. 1–24. Association for Computational Linguistics (2022)

King, G., Keohane, R.O., Verba, S.: Designing Social Inquiry. Princeton University Press, Princeton (1994)

Lindstedt, N.C.: Structural topic modeling for social scientists: a brief case study with social movement studies literature, 2005–2017. Soc. Curr. **6**(4), 307–318 (2019). https://doi.org/10.1177/2329496519846505

Lindvall, J., Bäck, H., Dahlström, C., Naurin, E., Teorell, J.: Sweden's parliamentary democracy at 100. Parliam. Aff. **73**(3), 477–502 (2019). https://doi.org/10.1093/pa/gsz005

Lindvall, J., Sebring, J.: Policy reform and the decline of corporatism in Sweden. West Eur. Polit. **28**(5), 1057–1074 (2005). https://doi.org/10.1080/01402380500311814

Lucas, C., Nielsen, R.A., Roberts, M.E., Stewart, B.M., Storer, A., Tingley, D.: Computer-assisted text analysis for comparative politics. Polit. Anal. **23**(2), 254–277 (2015). https://doi.org/10.1093/pan/mpu019

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., Adam, S.: Applying LDA topic modeling in communication research: toward a valid and reliable methodology. Commun. Methods Meas. **12**(2–3), 93–118 (2018). https://doi.org/10.1080/19312458.2018.1430754

Martinsson, J., Dahlberg, S., Christensen, L.: Change and stability in issue ownership: the case of Sweden 1979–2010. In: Dahlberg, S., Oscarsson, H., Wängnerud, L. (eds.) Stepping Stones—Research on Political Representation, Voting Behavior, and Quality of Government, pp. 129–144. University of Gothenburg, Göteborg (2013)

Mattson, I.: Parliamentary committees: a ground for compromise and conflict. In: Pierre, J. (ed.) The Oxford Handbook of Swedish Politics, pp. 679–690. Oxford University Press, Oxford (2016)

Meyer, T.M., Wagner, M.: Perceptions of parties' left-right positions: the impact of salience strategies. Party Polit. **26**(5), 664–674 (2020). https://doi.org/10.1177/1354068818806679

Mickler, T.A.: Parliamentary Committees in a Party-Centred Context—Looking Behind the Scenes. Routledge, Abingdon (2022)

Müllner, D.: fastcluster: fast hierarchical, agglomerative clustering routines for R and Python. J. Stat. Softw. **53**(9), 1–18 (2013). https://doi.org/10.18637/jss.v053.i09

Odmalm, P.: Political parties and 'the immigration issue': issue ownership in Swedish parliamentary elections 1991–2010. West Eur. Polit. **34**(5), 1070–1091 (2011). https://doi.org/10.1080/01402382.2011.591098

Peinelt, N., Nguyen, D., Liakata, M.: tBERT: topic models and bert joining forces for semantic similarity detection. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7047–7055. Association for Computational Linguistics (2020)

Proksch, S.O., Slapin, J.B.: How to avoid pitfalls in statistical analysis of political texts: the case of Germany. Polit. Anal. **18**(3), 323–344 (2009). https://doi.org/10.1080/09644000903055799

Roberts, M.E., Stewart, B.M., Airoldi, E.M.: A model of text for experimentation in the social sciences. J. Am. Stat. Assoc. **111**(515), 988–1003 (2016). https://doi.org/10.1080/01621459.2016.1141684

Roberts, M.E., Stewart, B.M., Tingley, D.: stm: an R package for structural topic models. J. Stat. Softw. **91**(2), 1–40 (2019). https://doi.org/10.18637/jss.v091.i02

Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B., Rand, D.G.: Structural topic models for open-ended survey responses. Am. J. Polit. Sci. **58**(4), 1064–1082 (2014). https://doi.org/10.1111/ajps.12103

Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: Chickering, M., Halpern, J. (eds.) Uncertainty in Artificial Intelligence—Proceedings of the Twentieth Conference, UAI '04, Arlington, VA, pp. 487–494. AUAI Press (2004)

Rydgren, J., Tyrberg, M.: Contextual explanations of radical right-wing party support in Sweden: a multilevel analysis. Eur. Soc. **22**(5), 555–580 (2020). https://doi.org/10.1080/14616696.2020.1793213

Rydgren, J., van der Meiden, S.: The radical right and the end of Swedish exceptionalism. Eur. Polit. Sci. **18**(3), 439–455 (2019). https://doi.org/10.1057/s41304-018-0159-6

Sánchez-Franco, M.J., Arenas-Márquez, F.J., Dos-Santos, M.A.: Using structural topic modelling to predict users' sentiment towards intelligent personal agents. An application for Amazon's echo and Google Home. J. Retail. Consum. Serv. **63**, 102658 (2021). https://doi.org/10.1016/j.jretconser.2021.102658

Shadish, W.R., Cook, T.D., Campbell, D.T.: Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin Company, Boston (2002)

Shadrova, A.: Topic models do not model topics: epistemological remarks and steps towards best practices. J. Data Min. Digit. Humanit. **2021**, 1–28 (2021). https://doi.org/10.46298/jdmdh.7595

Strøm, K.: Parliamentary committees in European democracies. J. Legis. Stud. **4**(1), 21–59 (1998). https://doi.org/10.1080/13572339808420538

Sundström, M.R.: The Swedish green party: from alternative movement to third biggest party. Environ. Polit. **20**(6), 938–944 (2011). https://doi.org/10.1080/09644016.2011.623857

Theodoridis, S., Koutroumbas, K.: Pattern Recognition, 4th edn. Academic Press, Burlington (2008)

Tinati, R., Halford, S., Carr, L., Pope, C.: Big data: methodological challenges and approaches for sociological analysis. Sociology **48**(4), 663–681 (2014). https://doi.org/10.1177/0038038513511561

Wagner, M.: When do parties emphasise extreme positions? How strategic incentives for policy differentiation influence issue importance. Eur. J. Polit. Res. **51**(1), 64–88 (2012). https://doi.org/10.1111/j.1475-6765.2011.01989.x

Wagner, M., Meyer, T.M.: The radical right as niche parties? The ideological landscape of party systems in western Europe, 1980–2014. Polit. Stud. **65**, 84–107 (2017). https://doi.org/10.1177/0032321716639065

Wilkerson, J., Casas, A.: Large-scale computerized text analysis in political science: opportunities and challenges. Annu. Rev. Polit. Sci. **20**(1), 529–544 (2017). https://doi.org/10.1146/annurev-polisci-052615-025542

Zeller, R.A., Carmines, E.G.: Measurement in the Social Sciences: The Link between Theory and Data. Cambridge University Press, New York (1980)

Zhao, H., Phung, D., Huynh, V., Jin, Y., Du, L., Buntine, W.: Topic modelling meets deep neural networks: a survey. In: Zhou, Z.H. (eds.) Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, pp. 4713–4720 (2021)