Bender, Benedikt; Bruinsma, Bastiaan

Article

Patterns in the Press Releases of Trade Unions: How to Use Structural Topic Models in the Field of Industrial Relations

Industrielle Beziehungen

Provided in Cooperation with:
Verlag Barbara Budrich

Suggested Citation: Bender, Benedikt; Bruinsma, Bastiaan (2022) : Patterns in the Press Releases of Trade Unions: How to Use Structural Topic Models in the Field of Industrial Relations, Industrielle Beziehungen, ISSN 1862-0035, Verlag Barbara Budrich, Leverkusen, Vol. 29, Iss. 2, pp. 91-116, https://doi.org/10.3224/indbez.v29i2.02

This Version is available at:
https://hdl.handle.net/10419/314324

Patterns in the Press Releases of Trade Unions: How to Use Structural Topic Models in the Field of Industrial Relations*

Benedikt Bender, Bastiaan Bruinsma**

Abstract Quantitative text analysis and the use of large data sets have received only limited attention in the field of Industrial Relations. This is unfortunate, given the variety of opportunities and possibilities these methods can address. We demonstrate the use of one promising technique of quantitative text analysis – the Structural Topic Model (STM) – to test the Insider-Outsider theory. This technique allowed us to find underlying topics in a text corpus of nearly 2,000 German trade union press releases (from 2000 to 2014). We provide a step-by-step overview of how to use STM since we see this method as useful to the future of research in the field of Industrial Relations. Until now the methodological publications regarding STM mostly focus on the mathematics of the method and provide only a minimal discussion of their implementation. Instead, we provide a practical application of STM and apply this method to one of the most prominent theories in the field of Industrial Relations. Contrary to the original Insider-Outsider arguments, but in line with the current state of research, we show that unions do in fact use topics within their press releases which are relevant for both Insider and Outsider groups.

Keywords: Insider Outsider Theory, Labour Market, Quantitative Text Analysis, Germany.
JEL: C55, J48, J51, P16

Muster in Pressemitteilungen von Gewerkschaften. Die Anwendung von Structural Topic Models im Feld der industriellen Beziehungen

Zusammenfassung Quantitative Textanalysen und die Verwendung von großen Datensätzen wurden im Bereich der Industriellen Beziehungen bisher nur wenig Aufmerksamkeit geschenkt. Das ist bedauerlich, wenn man bedenkt, welche Vielfalt an Möglichkeiten diese Methoden bieten. Wir zeigen daher wie eine neue und vielversprechende Technik der quantitativen Textanalyse – Structural Topic Model (STM) – eingesetzt werden kann, um große Datenmengen zu reduzieren und die Insider-Outsider Theorie zu testen. Diese Technik ermöglicht es in einem Textkorpus von fast 2000 Pressemitteilungen deutscher Gewerkschaften zwischen 2000 und 2014 die zugrundeliegenden Muster zu finden. Wir geben eine schrittweise Einführung zur Anwendung von STM, weil wir die Zukunft der Forschungsmethodik im Feld der Industriellen Beziehungen auch bei quantitativen Analysetechniken sehen. Bisherige methodologische Literatur zu STM konzentriert sich überwiegend auf die

* Artikel eingegangen: 20.12.2021. Revidierte Fassung akzeptiert nach doppelt-blindem Begutachtungsverfahren: 27.10.2022.

** Benedikt Bender, Institut für Politikwissenschaft, Fachbereich Gesellschaftswissenschaften, Goethe-Universität Frankfurt am Main. Theodor-W.-Adorno-Platz 6, D-60323 Frankfurt am Main. E-Mail: b.bender@soz.uni-frankfurt.de
Bastiaan Bruinsma, Ph.D., Chalmers University of Technology, Department of Computer Science and Engineering, 41296, Göteborg, Schweden. E-Mail: sebastianus.bruinsma@chalmers.se

Mathematik der Methodik und weniger auf deren Umsetzung und Diskussion. Daher geben wir ein praktisches Anwendungsbeispiel der STM und testen eine der bekanntesten Theorien im Bereich der Industriellen Beziehungen. Entgegen den ursprünglichen Annahmen der Insider-Outsider Theorie, aber im Einklang mit dem aktuellen Forschungsstand, zeigen wir, dass Gewerkschaften in ihren Pressemitteilungen Themen ansprechen, die sowohl für die Gruppe der Insider als auch für die Gruppe der Outsider relevant sind.

Schlagwörter: Insider-Outsider Theorie, Arbeitsmarkt, Quantitative Textanalyse, Deutschland

1. Introduction

For the last two decades, the theory of Insider-Outsider representation has been at the centre of a controversial debate. Its arguments as such, presented by Rueda (2005, 2007), Palier and Thelen (2010) and Emmenegger, Hausermann, Palier, and Seeleib-Kaiser (2012), are well known and tested. More recent work rejects the notion of Insider-Outsider representation; trade unions do not to focus only on the Insider, but on the Outsider as well (Oliver & Morelock, 2021). Overall, the field is far from settled. Yet, what is settled are the methods used to study the field. While the theoretical perspectives are many, the methodological perspectives are few. Most, if not all, works rely on either qualitative content analysis, case studies or (multi-level) regression analysis. And while this is not problematic per se, such a lack of variation might lead scholars to miss out on alternative – and perhaps interesting – perspectives.

In this article, we will introduce such an alternative perspective and discuss the new views on the issue it can give us. Structural Topic Modelling (STM) belongs to the field of quantitative text analysis. This field utilises a wide array of statistical techniques to summarise and understand large corpora of textual data. While it has been popular in other fields such as political science (Gross & Jankowski 2020; Lamare & Budd, 2021) or media studies (Hase, Mahl, & Schäfer, 2022), its use is still rare in the field of Industrial Relations (though see Picot & Menéndez, 2017). This is striking, as the wide variety of documents in that field – such as the press releases of trade unions we will use here – are little different from the speeches and manifestos of political parties. As such, the field seems to be missing out on methods that can effectively deal with the large amount of potential data that is available today.

The aim of STM is to find topics: collections of words that underlie a text and that one can use to say what a text is "about". As such, we can use it to test whether trade unions do indeed talk about both "Insider" and "Outsider" related topics. To do so, we will look at the press releases of the four largest trade unions in Germany between 2000 and 2014. We make this choice for three reasons, the first two of which are theoretical and the last of which is methodological. First, choosing Germany allows us to test the traditional assumption that German trade unions focus more on Insiders than Outsiders of the labour market. This assumption relies on the analysis that most of the German industry unions do not show what Dörre (2011) named an "integrative function" for the Outsiders of the labour market. In contrast to their Scandinavian counterparts, e. g. German trade unions do not administer or co-finance unemployment benefits for the whole workforce (the so-called Ghent system),

something which would be reflected in Insider and Outsider topics in their press releases (Durazzi, Fleckenstein, & Lee, 2018). Therefore, it could be expected that most of the topics of the industrial unions in Germany will focus on the Insiders as for the Outsiders. Second, the period between 2000 and 2014 was interesting as it showed a conservative welfare state (Germany) in which the dualistic Insider-Outsider structure was on the rise (Thelen, 2012). During this period, there were the reforms of Agenda 2010 (running from 2003 to 2005) and the introduction of Mini-Jobs, both of which created more Outsiders (Palier & Thelen, 2010), while the introduction of the statutory minimum wage in 2015 again reduced this number (Marx & Starke, 2017; Cantillon, Seeleib-Kaiser, & Veen, 2021). In other words, both Insider and Outsider topics were highly salient during the period between 2000 and 2014. Third, press releases not only represent the trade unions' organisational positions but are also representative of the kind of documents that are often used in quantitative text analysis. As such, they allow us to both see what the trade unions talk about and show how one can use a method such as STM.

Our contribution to the field is thus both methodological and theoretical. On the theoretical side, we show that trade unions in Germany indeed do refer to both Insiders and Outsiders of the labour market during the time we study. On the methodological side, we show that Industrial Relations scholars can use STM to study well-known problems from a different perspective. Yet, we will also make clear that STM – as any other method from the field of quantitative text analysis – is not a *panacea.* As Grimmer and Stewart (2013: 269) note: "there is no globally best method for automated text analysis". Instead, different research goals demand different methods, and while STM is appropriate for our goal here, it will not be so elsewhere where other methods might be more appropriate.

From here, the article will proceed as follows. First, we will review the literature on the Insider-Outsider theory. We do so both to show its continuing relevance and to provide the necessary theoretical background. Then, as the paper is a contribution to a methodological special issue of *Industrielle Beziehungen*, we turn to an in-depth description of STM. Here, we provide step-by-step instructions on how to apply it as well as discuss the various implications of pre-processing, cleaning, and the number of topics one can use. Afterwards, we discuss our findings and interpret our found topics in light of the Insider-Outsider theory. We conclude with some methodological and theoretical suggestions, discuss several of the limitations STM has, and reflect on the further use of STM in the field of Industrial Relations.

2. Insider-Outsider Representation: A Brief Literature Review

The theory of Insider-Outsider representation separates labour market participants into two groups based on their employment status (Lindbeck & Snower, 2002). The first group contains the Insiders – full-time workers with a permanent contract, strong dismissal protection, and high wage earnings (Davidsson & Emmenegger, 2012). The second group contains the Outsiders, who, contrary to the Insiders, are either unemployed or are temporary or part-time workers with weak or no dismissal protections and low wages. As a result of these differences, both groups have different interests in labour market and social policy reforms (Rueda, 2007). Thus, Insiders are more often in favour of cutting unemployment benefits, as their employ-

ment is secure and often have a positive stance towards dismissal protection to keep their employee status safe. By contrast, Outsiders prefer a high standard of unemployment benefits out of self-interest and are often in favour of further flexibilization of the labour market, with the goal of either entering the labour market or obtaining a permanent contract.

In its original inception, the idea of the theory was that trade unions and left-wing parties represented Insider interests (Rueda, 2005, 2007). This idea stemmed from various observations. To begin with, trade unions tended to exist in sectors where the Insiders were dominant (Esping-Andersen, 1999). Also, Insiders tended to join these unions, whereas Outsiders often did not (Rueda, 2006). Furthermore, Insiders were often more active in politics and thus more relevant in an electoral sense, while Outsiders often tended to take the role of observers (Beramendi, Häusermann, Kitschelt, & Kriesi, 2015). Therefore, and because the Insiders were potential new union members (Davidsson, 2009), it was argued to be more efficient for the unions to represent the Insiders than to represent the Outsiders. Based on these ideas, the reasoning was that when there was a left-wing government combined with trade unions that had a strong influence on policy, there would be more reforms that favoured full-time employed workers (i. e. Insiders) (Palier & Thelen, 2010). In line with such arguments, Hassel (2014) found that trade unions were only in favour of more flexibilization for Outsiders if the same reforms ensured that the Insiders were protected against labour market fluctuations (see also Culpepper & Regan, 2014)).

Yet, over time a growing body of evidence challenged this assumption (Marx & Starke, 2017; Benassi & Dorigatti, 2018; Schwander, 2019). For example, Oliver and Morelock (2021) showed that in advanced democracies, trade unions tended to support the whole workforce – both Insiders and Outsiders. Also, Naczyk and Seeleib-Kaiser (2015) found that pension policies – for which a strong sense of solidarity and ideological values still exists within unions – tend to be representative of the whole workforce, instead of only the Insiders. As an example, the authors show that in recent years unions supported occupational pensions for all employees, regardless of their Insider-Outsider status. In line with these arguments, Mosimann and Pontusson (2017) demonstrated similar results whereby unions actively represent the whole workforce, while Benassi and Dorigatti (2018) and Benassi and Vlandas (2016) showed how unions push to create new standards and stricter rules for the labour market, with a special focus on Outsider topics. Finally, Fervers and Schwander (2015) showed that if the unions are well organised and powerful, then the difference in employment conditions between Insiders and Outsiders is reduced.

As such, the current state of research argues that trade unions are representative of the whole workforce, and represent both Insider and Outsiders (Durazzi, Fleckenstein, & Lee, 2018; Visser, 2016). Following this logic, we would thus expect to find evidence of both Insider and Outsider topics in the press releases of the main trade unions. Since the paper is a contribution to a methodological special issue, we discuss the STM we will use to find these topics in detail in the next paragraph.

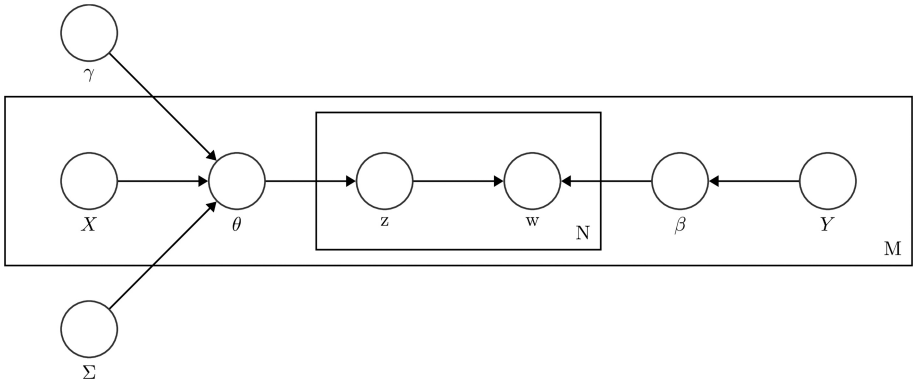
3. Method: The Structural Topic Model

Topic models aim to find underlying "topics" in a corpus of text. The idea is that, when creating the texts, the author draws upon these topics to choose the words they use. Hence, these topics can tell us what texts are about. The process of topic modelling first surfaced in the late 1990's, but only got traction with the introduction of Latent Dirichlet Association (LDA) by Blei, Ng, and Jordan (2003). The success of this model led others to expand upon it by allowing it to work with large data sets (Blei, 2012), including information on authors (Rosen-Zvi, Griffiths, Steyvers, & Smyth, 2004) and dealing with correlations between topics (Blei & Lafferty, 2007). Later, Roberts et al. (2014) (see also Roberts, Stewart, & Tingley 2019) introduced the Structural Topic Model (STM), which combined the functions of all these methods and is furthermore more applicable than others. Given that STM aims to classify words into topics and topics into documents, the technique belongs to the family of classification methods. As it does so without any previous assumptions regarding what the topics are, it is also known as an unsupervised method, or a method using unknown categories (Grimmer & Stewart, 2013). This sets it apart from other text analysis methods such as those that aim to position the documents on an underlying scale (such as Wordscores or Wordfish), or methods that are supervised and have pre-established known categories (such as dictionary analysis or proportions measurement).

STM stands out from other methods of text analysis because of several unique features. To begin with, it allows for mixed membership (that is, a document can contain multiple topics), allowing for correlations between the topics, and including additional information apart from the actual text. Also, one can include a wide variety of metadata (such as the author of the text or the date of publication) to help with the analysis. As a result, STM has become quickly popular in a wide range of social sciences where texts rarely occur in a vacuum. Examples of this wide range of use are measuring the relation between corporate funding and ideological polarization regarding climate change (Farrell, 2015), differences in how both genders experienced the COVID-19 lockdown in Germany (Czymara, Langenkamp & Cano, 2021), or studying global trends for start-up companies (Savin, Chukavina & Pushkarev, 2022). It is also for this reason that we believe STM to be the most suited for our analysis: it seems more than likely that the topics we aim to find will depend on who wrote them and when they were written.

There are several assumptions made by STM that are common for quantitative text analysis methods. To begin with, as with most other text-as-data approaches, topic modelling makes a bag-of-words assumption. That is, a matrix with the individual words in the columns and the documents in the rows replaces the structure of the text. In this matrix, the cells contain the number of times a word occurs in a document. This is the data-term matrix (or the data-frequency matrix), and is, besides the number of topics, the main input of STM. Using the bag-of-words approach means a trade-off between losing the richness of the text (like its word order) and gaining simplicity. Aside from this, all topic models – and thus also STM – make two implicit assumptions. The first is that documents are similar if they have similar words. The second is that the data they use (i. e., the texts) is generated from a certain number of topics. Any topic model algorithm then looks for word co-occurrences and interprets these as topics. Thus, the more often two words are found together, the more likely it is that they belong to the same topic.

Figure 1: Plate diagram of a Structural Topic Model



Source: own compilation

To understand how STM works, Figure 1 shows a plate diagram of the method. Here, we find that an individual word *w* is nested in the number of words in a document *N* and the number of documents in the corpus *M.* In our case, this could refer to a word like "Gesellschaft" being part of a particular IG Metall press release, which itself is part of the complete corpus. The logic of a plate diagram is then that everything that takes place within the box is repeated for as many elements (words or documents) as there are. Seen this way, STM works by taking the words *w* and the document metadata *X* and *Y* (and the number of topics *k*, not shown here). The algorithm then aims to estimate all the remaining variables. An STM most often does this by using an expectation-maximisation (EM) algorithm that first estimates possible parameters and then checks how likely they are, given the data. This project converges as soon as a certain threshold (indicating marginal change) is reached[1].

4. Analysis

Given its quantitative nature, data are central to any STM. Yet, while this means that the method is grounded in reality, this dependence makes them vulnerable to what Gelman & Loken (2014) call a "garden of forking paths". That is, instead of a generalisable conclusion, one's results are the result of the unique combination of text selection, cleaning, and the method of analysis one employs while analysing the data. To address this, we can employ various techniques, which we will discuss here, after first introducing our data.

a) Data

The time period 2000 and 2014 was chosen because it includes key stages of the most important welfare state reforms in Germany over the last two decades. The large reform package Agenda 2010 was introduced between 2003 and 2005 whilst key developments

1 For a complete overview of STM as well as its mathematical derivation, see Roberts, Stewart, & Tingley (2019).

towards the statutory minimum wage took place from 2005 onwards, a measure which was then introduced in 2015. The time period prior to 2014 was especially interesting for two reasons. Firstly, significant changes of position occurred and stabilised such that all union organisations in our data set were in favour of the minimum wage by the end of 2014. Specifically, the DGB changed their position in the year 2006 and IG Metall as well as IG BCE also moved from a negative to a positive stance towards the minimum wage by the year 2014 (Bender, 2020, pp. 227–239). No more shifts of opinion occurred in 2015. Secondly, the time period from 2013 to 2014 is of particular interest because these two years were the beginning of a new government coalition between CDU/CSU and SPD. Governments tend to present major new projects during their first year in office which are more feasible than the promises made during an election campaign. Therefore, we chose to analyse Insider-Outsider and how unions commented on the new welfare state proposals of this government.

For our data, we draw upon the press releases of the four main trade union organizations in Germany. These are the DGB (*Deutscher Gewerkschaftsbund*), the IG BCE (*Industriegewerkschaft Bergbau, Chemie, Energie*), the IG Metall (*Industriegewerkschaft Metall*) and Ver.di (*Vereinte Dienstleistungsgewerkschaft*). Of these four associations, the DGB is the largest, serving as an umbrella organisation for eight German trade unions, among which are the other three unions. Of the three, IG Metall is the largest, accounting for 38,1 % of its members, followed by Ver.di (32,9 %) and IG BCE (10,4 %) (Deutscher Gewerkschaftsbund, 2021). We chose these organisations for their high political relevance, based on the number of times they were invited as experts to serve on governmental committees (Bender, 2020, pp. 80–99). Furthermore, all these unions have high social relevance based on their numbers of members (Bender, 2020).

We collected the press releases by hand to cover a period from 2000 to 2014: in total, this led us to collect 1990 press releases.[2] The format of the press releases differs between association and over time. For the DGB, IG BCE, Ver.di, and for the IG Metall from 2005 on, we use the online versions of their press releases. Prior to 2005, we use scanned-in versions of the physical press releases which we collected from the IG Metall archive (Bender, 2020, pp. 143–150). While this does not affect the content of the press releases, it does lead to several differences in the pre-processing steps, which we discuss in the next section. In all cases, the documents were available in a selectable PDF format.

The press releases are not equally distributed over all unions. As the main umbrella organisation, the DGB covers nearly half of the press releases, while IG BCE, as the smallest union, covers just above 10 %. Also, in some years, there were more press releases than in others, so that while we collected 212 and 207 press releases in 2002 and 2003, we collected only 96 in 2005 and 2006.

Table 1: Overview of the Number of Press Releases

	DGB	IG BCE	IG Metall	Ver.di
2000	64	19	28	0
2001	83	18	31	60

2 The data collection was part of the ‘Reform Monitor on Political Conflict’ (ReMoPo) based on Bender (2020) and Bender (forthcoming).

	DGB	IG BCE	IG Metall	Ver.di
2002	97	13	46	56
2003	97	24	50	36
2004	53	11	30	20
2005	44	15	21	16
2006	41	7	8	40
2007	64	15	15	36
2008	69	15	23	27
2009	43	9	23	32
2010	52	11	19	45
2011	48	12	25	33
2012	59	7	28	28
2013	48	22	16	28
2014	57	14	11	28
Total	919	212	374	485

Source: own compilation

Press releases were primarily selected because social media posts have developed more recently and were much more limited compared to press releases during the period 2000 to 2014. Secondly, press releases are a carefully prepared summary of the organisational (trade union) position which is officially published at a particular point in time. To create a press release, different opinions within the organisation must be reconciled into one single line of argument. This makes press releases more representative of organisational positions than very short post, tweets, or texts which describe the position of individuals within the organisation (such as interview data). Furthermore, we select only press releases because of pragmatic reasons since these documents are easily found and easily compared with each other and can more reliably be used to demonstrate changes in topics over time.

Yet, as for other empirical material, there are limitations to using these types of data. On the theoretical side, trade unions (and employer associations) often formulate press releases only for the media or the public. As such, press releases do not necessarily reflect the real interests of the organisations or reflect strategic actor positions (Broockman, 2012). Indeed, this can make it difficult to identify topics, as unions may strategically choose to represent both labour market groups in their press releases (Grumbach, 2015). Yet, we note that these limitations also hold for other empirical material, and that the objectivity and accessibility of press releases make them preferable over other sources. Furthermore, on the technical side, press releases are a rather "safe" choice to prove the usability of STM. They have a rather rigid structure compared to other data such as tweets, speeches or interviews. While these risks

having the method appear more useful than it might be, we feel these documents are a useful choice to show how STM could potentially be used.

b) Conversion

Pre-processing, while often seen as the first step, actually does assume some previous work. The text must be present in a (clean) digital format. Webpages might contain many extraneous elements that are not necessary for the later analysis but the text on its own is present. In our case, the original documents were (readable) PDF files, which had to be converted to TXT files (for use in R or Python) in such a way as to represent the original document most faithfully. This involves checking the quality of the scans for smudges, inkblots or other issues; ensuring glyphs like "ff" or "ll" are often not read as such and are separated; dealing with end-of-sentence hyphens; and dealing with single letters resulting from broken-up words. Also, we have to remove the headers, footers, and other objects in the text not part of the "main" body. Finally, given that the lay-out of the documents differs widely, we must perform these steps manually.

During our conversion, we did the following. First, we converted the documents from PDF to TXT using the pdftools package for R (Ooms, 2021) built on the poppler library. Second, we manually removed headers and footers for each of the documents, leaving only the title and the main body of the text. Third, we read all the files into R and removed any unnecessary lines of whitespace. Fourth, we joined together all those words at the end of each sentence, so that the only hyphens left in the text were either those used for lists or used meaningfully within words ("IG Metall-Vorsitzende"). Fifth, we scanned the documents for "single" letters and joined them to the words they belonged to. Sixth, we ran a spell check over each of the documents and corrected any words with obvious spelling mistakes. After having done this, we deemed the TXT files to be fair copies of their PDF files and ready for pre-processing.

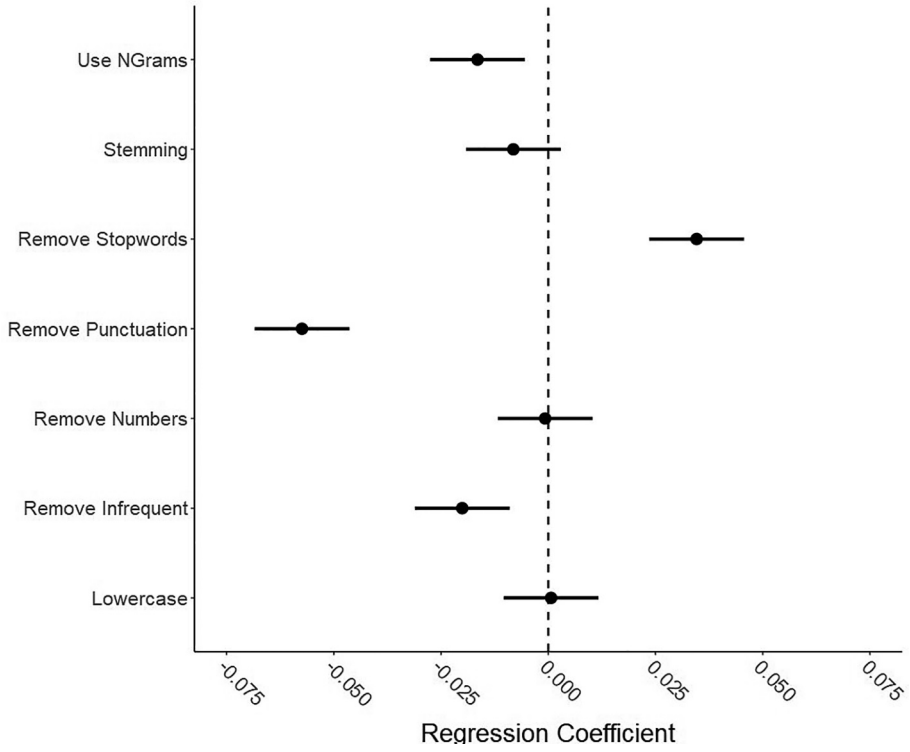
c) Pre-Processing

The second step is the pre-processing. While often overlooked, this step is one of the most important and is where one is most likely to get lost in one of the many "forked garden paths" (Gelman & Loken, 2014). Of the various options for pre-processing, Denny and Spirling (2018) identify seven: removing punctuation, removing numbers, lowercasing, stemming, removing stop words, including n-grams, and removing infrequent terms. For each of these, one has to consider whether they will use the option, and if so, in which order. This means that given seven options, there will be 2^7 = 128 different possible combinations to choose from. As none of these combinations is inherently "better" or more valid than the others, which combination one chooses depends on the goal of the researcher as well as their professional judgment.

To help avoid any potential "cherry-picking" that can result from researchers simply choosing that combination that provides them with the most suitable results, Denny and Spirling (2018) suggest comparing all different pre-processing combinations and assessing how sensitive the documents are towards them. The aim then is to figure out which combi-

nations are unusual (that is, very different from the others) and are most likely to influence the later results of the analysis. To aid with this, they provide a package for R – preText – that compares the differences between the data frequency matrices that result from each pre-processing combination. Using this, the package can calculate a score that shows for each pre-processing option how likely it is to generate a data frequency matrix that is significantly different from the others. These scores, given as regression coefficients, can be either positive (which means that including them increases the likelihood of creating an unusual data set), negative (which decreases it), or zero (which means it does not cause any differences).

Figure 2: Regression Coefficients for each of the seven pre-processing steps



Source: own compilation

Figure 2 shows the output of sensitivity calculation of the data set. Here, if the regression parameter is not significantly different from 0, then either including or excluding that pre-processing step is unlikely to make a difference. When the parameter is negative, it indicates that including this parameter reduces the chance of a data set that leads to unusual conclusions (Denny & Spirling, 2018, p. 183), while a positive parameter increases this chance.

Running this analysis on our data gives us the results shown in Figure 2. From this, we find that while lowercasing, removal of numbers and stemming seem to have little influence on the outcome, including n-grams, removing punctuation and removing infrequent words decreased the likelihood of obtaining an unusual data set while removing stop words increased

it. Especially this latter finding is interesting, given that the removal of stop words is a rather common (and often unconsidered) step in the pre-processing of texts.

While it may be tempting to completely rely on these results, the authors of the package abstain from doing so and consider the overall aim of the research – as well as the theory that it is based on. Here, we should remember that given that we aim to create topics that are substantially different from each other, we want to remove those words that are so common that they are likely to appear in all topics, complicating their analysis. As such, we decide to opt for the removal of the stop words. To study the effect of this decision, Denny and Spirling (2018, p. 183) advise replicating our final analysis with the alternative option (not removing the stop words) as well. Doing so showed no discernible difference in the number of topics and their content, apart from the topic model taking longer to run and requiring more terms to identify the topics.

In addition to showing us information about the effect of the individual cleaning options, preText also informs us in which order they are least likely to generate an abnormal data set. Based on this, and combined with our intuition of what is appropriate, we opt for a process in which we apply the following procedures to our text: lowercasing; removing all symbols, punctuation, separators and unnecessary spaces; removing stop words; stemming the terms; constructing of n-grams; removing of numbers; and removing of infrequent terms. This leads to a data frame with 20,989 unique terms, which form the input for the subsequent analysis.

As for the steps themselves, we take the stop words from the German version of the snowball stop word list and add to this several additional words that cover the names and abbreviations of the organisations. We also carry out the stemming using the snowball stemmer. We note that this is more of a challenge in the German language than in English as the former contains many compound words (Lucas, Nielsen, Roberts, Stewart, Storer, & Tingley, 2015). Given that stemming aims to reduce nouns to their root, this could lead to very different compound nouns being reduced to the same root (Proksch & Slapin, 2009). A similar point can be made for infrequent terms, and while Maier, Niekler, Wiedemann, and Stoltenberg (2020) note that the analysis should not make a difference, Greene, Ceron, Schumacher, and Fazekas (2016) find that removing such words – especially with the compound words – can remove words that might be relevant for the analysis. As for the n-grams, we construct those before we remove the numbers as we deem it likely that certain numbers (such as 3 %-increase) might not be meaningless. Finally, we remove those terms that occurred in only a single document; thus, a word must occur in two or more documents to be included in the final data set. Figure 3 shows the words that were ultimately used in the analysis, visualized by their word count.

Figure 3: Wordcloud of the words in the documents after cleaning

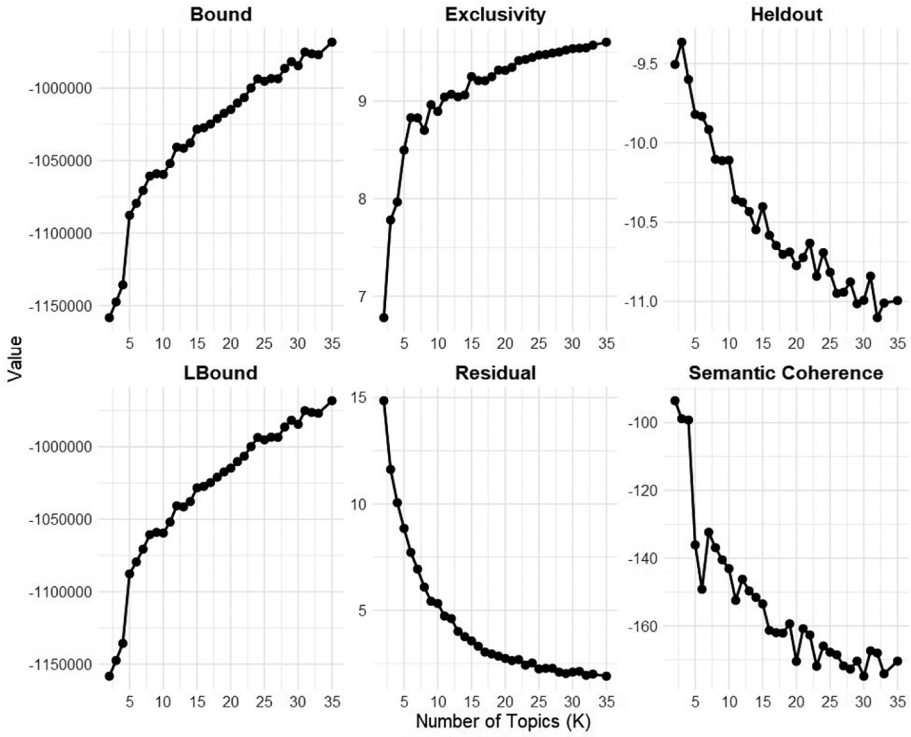
Source: own compilation

d) Number of Topics

As with other methods of unsupervised text analysis, we need to provide STM with the number of topics we wish to extract. To help us with this, we make use of the searchK function implemented into stm, which generates six diagnostics: the held-out likelihood, residuals, bound and lbound, exclusivity and semantic coherence (Roberts, Stewart, & Tingley, 2019). Of these, the "bound" refers to the bound of the maximum likelihood that was used in the stm procedure. The idea is that higher marginal likelihood indicates a better fit of the data. The "lbound" is then simply a correction of the bound by the number of topics, to allow for comparison. "Heldout" is a metric that assesses the ability of the model to predict a part of the data, after having been trained on the other part. The higher the heldout is, the better the model's ability to do so. For all three of these metrics, the relevant data are taken from the dataset used for the subsequent analysis. The idea behind the residuals is that in a correctly specified model, the variance of the residuals equals 1. If this is not the case, this means that we need more topics to deal with the variation that is present. As such, the lower the residuals values, the better specified our model (Taddy, 2012). Semantic coherence then measures how

often the most probable words in a single topic co-occur (Mimno, Wallach, Talley, Leenders, & Mccallum, 2011). Finally, "exclusivity" looks at to what degree words with a high probability under one topic have a low probability under all the other topics (Roberts, Stewart, Tingley, Lucas, Leder-Luis, Gadarian, & Rand, 2014, pp. 1070). The higher the average of these overall topics, the better.

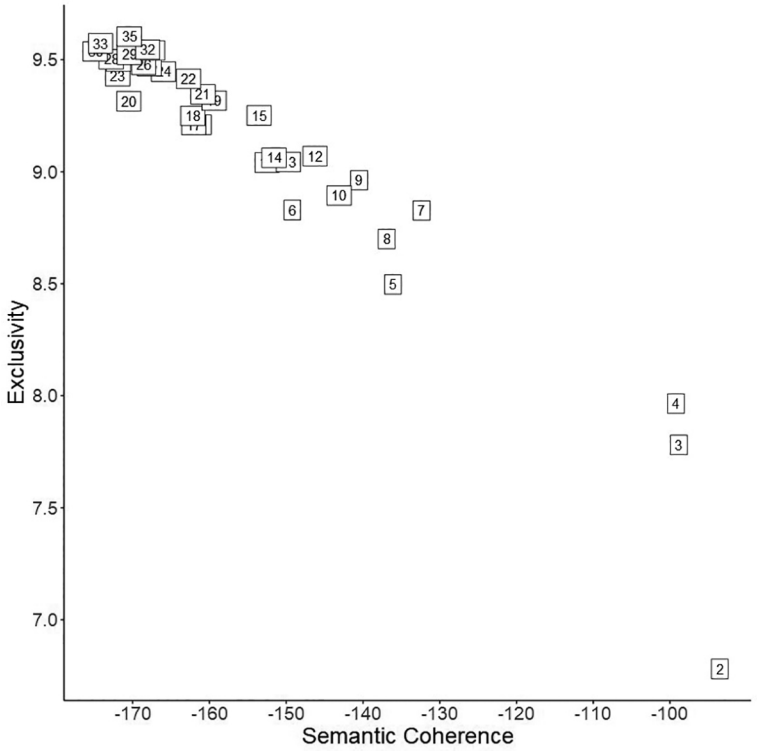
Figure 4: Model Diagnostics for models between 2 and 35 topics



Source: own compilation

First, we run this algorithm with 0 topics. This causes the algorithm to use a proposed measure by Mimno and Lee (2014). While using this algorithm has the advantage of automatically selecting the number of topics, its dependence on randomness means that running it with a different seed will result in different results and a different number of topics. Doing so here leads to a fairly large number of 88 topics (Roberts, Stewart, & Tingley, 2019). We thus turn to another approach and run these metrics for models ranging between 2 and 35 topics. Looking then at Figure 4, we see a sharp increase in the bound just before 10 topics. In the case of the Exclusivity, we find that the sudden increase decreases after this point and find a similar pattern for the Semantic Coherence; also, the Heldout seems to peak around this point. The only diagnostic that is less clear is the Residuals, which keep falling at this point and only seem to slow down later.

Figure 5: Semantic Coherence set against Exclusivity



Source: own compilation

Of these metrics, Semantic Coherence and Exclusivity are the most relevant (Roberts et al., 2014; Roberts, Stewart, & Tingley, 2019). Here, we plot both together in Figure 5 and look for those cases where semantic coherence and exclusivity are in balance. In this case, this seems to be for the models containing either 10, 9, 8, 7, or 5 topics. We then inspect the topics of these models for a qualitative interpretation of the meaning of the topics given the words that belong to them. Based on this, we decide to use 9 topics.

e) Running the algorithm

To run our STM, we use the stm package as implemented by Roberts, Stewart, and Tingley (2019) in R. To use it, we have to provide it with our texts in the form of a data-frequency matrix, the number of topics, and several prevalence parameters. These latter distinguish STM from regular topic models and allow us to provide additional information to help the algorithm find the topics. Here, we reason that the topics are most likely dependent on when they were written and who wrote them. As such, we incorporate information on the author (any of the four trade unions) as well as the year in which the press release was published. Given that the

trade unions might not behave the same over time, we also allow for an interaction between the two.

Doing so leads to a model that converges after 139 iterations. The resulting object contains a variety of elements, the most important of which are the topic prevalence (the likelihood of a topic appearing in a text) and the word probabilities (the likelihood of a word appearing in a topic). It is these latter word probabilities that we need to identify "what" the topics are about. Yet, given that many words occur so often that they occur in all topics (and thus have a high probability in all of them), we slightly transform the word probabilities for each word in a metric known as a FREX score (short for FRequency and EXclusivity score). This score balances the frequency with the word occurs in a topic, with how exclusive it is for that topic. To understand the latter, imagine a word that occurs very often in a single topic, but only rarely (or never) in any of the others. These words can then be said to be exclusive to that topic. If the word then also appears quite often in that topic, it can provide a good idea of "what" that topic is about (Roberts, Stewart, & Tingley, 2019). In the next section, we will use these FREX scores to identify those terms which we will use to label our topics.

Results

We will split our results into two sections. First, we will look at each of the 9 topics and the FREX terms that identify them and discuss their meaning. Second, we will look at the prevalence of each of the nine topics between the four associations.

a) Identification of the Topics

Table 2 gives an overview of the 9 topics and the words with the corresponding highest FREX scores. In order to provide a label to the topics and classify them into either an Insider or Outsider perspective, we use both the FREX terms (which are those words that are both highly associated with a topic and are exclusive to it) as well as those documents that are most representative for the given topic. With the latter we mean those documents for which that topic had the highest proportion. Taken together, this allows us to obtain an idea of what a topic is "about". This process is by design subjective and depends on the goals and aims of the researcher. Indeed, the labels are chosen by the researcher based on what they consider the topic is about given their understanding of the words it contains and the documents it refers to, as well as which labels are the most meaningful given the context of the study (Aranda, Sele, Etchanchu, Guyt, & Vaara, 2021).

Given this, we assign the labels to the nine topics and identify two of them (1 and 2) as representing Insider representation, four topics (4, 5, 7, and 9) as doing so for Outsider representation, and two topics (3 and 6) to both. Only Topic 8 has no clear interpretation and thus cannot be assigned to either of Insider or Outsider categories and we leave this topic unlabelled[3]. We will now consider each of the remaining 8 topics in turn.

3 While topic 8 had a topic proportion of 0.12 its terms covered references to the type of publication the press release referred to (*magazin*), words typical in such publications (*kraftig, deutlich*), and references to politicians (*schrod, stoib*) and executives of trade unions (*sagtputzhamm*).

Table 2: Overview of the nine topics, including their label, assigned category, and terms based on the FREX score of each word (high to low)

#	Label	Category	Terms
1	Collective Bargaining	Insider	mindestlohnabfallwirtschaft, allgemeinverbindzuerklar, igz, tarifergebnis, abfallwirtschaft, branchenzuschlag, briefdienstleist, bankgewerb, leiharbeitnehm
2	Strikes	Insider	arbeitnehmerbegehr, tagarbeit, aktionswoch, kundgeb, motto, hamburg, aktion, amazon, arbeiterinnarbeit, hub
3	Political Economy Model	Insider/ Outsider	massiv, sagtmatecki, matecki, kurzarbeit, binnennachfrag, schuldenbrems, monigraan, aufschwung, exitstrategi, eurozon
4	Agenda 2010	Outsider	kinderzuschlag, flachenvertrag, jobcent, vermittlungsausschuss, regelsatz, koalition, hartziv, hartzivempfang, eingliederungsbeitrag, zwangsverrent
5	Minimum Wage	Outsider	telekom, verfug, gestellt, agenda, vw, gmbh, vwgesetz, grundsatzred, tarifein, demografi
6	Workplace Safety	Insider/ Outsider	raffineri, national, umweltschutz, arbeitgesundheitsschutz, effizient, isoldkunkelweb, kraftwerk, arbeitsschutz, akteur, gesundheitswes
7	Poverty	Outsider	armutreichtumsbericht, geringqualifiziert, minijob, sozialhilfeempfang, offentgefordertbeschaft, kombilohn, minijobb, glaubt, praktikum, workingpoor
8	Other	-	magazin, sagtputzhamm, stoib, schrod, wirtschaftswachstum, putzhamm, kraftig, befristetbeschaft, deutlich, finanzpolit
9	Minimum Wage	Outsider	zeitungszustell, mindestlohnausnahm, lufthansa, mindestlohn850euro, ausnahmbeimmindestlohn, allgemeingesetzmindestlohn, hungerlohn, ausnahmgesetzmindestlohn, kenntausnahme, weiterbranch

Source: own compilation

Starting with the Insider topics, we label topic 1 Collective Bargaining. This as most of its words refer to the collective bargaining processes. This topic has a topic proportion of 0.15 (meaning that 15% of the total words belongs to that topic), the largest of the 9. This is not surprising as the negotiation of pay and working conditions between trade unions and employers is quite common in Germany. In fact, this is a core task for the unions, as they aim to achieve and maintain high working standards for their members. As the words related to the topic show, between 2000 and 2014, bargaining processes were highly relevant in different sectors (*abfallwirtschaft, briefdienstleistung, bankgewerbe*). Here, the postal service (*briefdienstleistung*) is especially interesting, given its long and intensive collective bargaining process in 2014 (Stüdemann, 2015). Notably, collective bargaining processes are highly relevant for Insiders, but not so much for Outsiders since these processes mostly affect the employees. Therefore, topic 1 is representative of the Insider category.

Topic 2 (topic proportion 0.11) is also related to collective bargaining processes, but we labelled it Strike, as most of its words relate to industrial disputes. This includes workers' petitions (*Arbeitnehmerbegehren*), action weeks (*aktionswoche*), rallies (*kundgebung*), mottos (*motto*), and strike actions (*aktion*). Not without surprise, the company Amazon also appears in this category as the trade unions – Ver.di in particular – often called for strike actions against the Amazon management in Germany. The Strike topic similarly contains themes relevant to Insider representation, since such strike actions are most relevant for employees in a stable and safe employment relationship. Therefore, we also categorise topic 2 among the Insider topics.

Turning to the Outsider topics, we find topic 7 to be the most prominent topic (topic proportion of 0.13). We label this topic Poverty, as its most frequent term is a reference to a report published by the Federal Ministry of Labour and Social Affairs (Bundesministerium für Arbeit und Soziales), the *armutsundreichtsumsbericht.* This report is the main governmental source of information and data on the increase or decrease of poverty rates in Germany. Other words in the topic relate to poverty risks, such as *sozialhilfeempfang*, working poor and low-skilled workers (*geringqualifiziert*), or people who work in atypical situations often not covered by social insurance systems (*minijob, praktikum*). Also, we find in unions press releases a mention of the introduction of wage subsidies (*kombilohn*). Wage subsidies were part of labour market reforms to help the long-term unemployed to enter the labour market. As most of these terms concern themselves with Outsider topics, we classify this topic in the Outsider category.

Topic 4 (topic proportion 0.12) deals with words related to Agenda 2010 Reform. The Agenda 2010 was a series of social policy and labour market reforms that were introduced between 2003 to 2005 but the debate about the reforms was present until 2014. As with the Poverty topic, most terms refer to the unemployed (*hartziv, hartzivempfang, regelsatz*) or to those who have become Outsiders due to a special form of early retirement (*zwangsverrent*). Also, a term such as *eingliederungsbeitrag* refer to some restrictive forms of labour market activation, because the unemployed have to accept every job offer made by the Employment Agency (*jobcenter*) after a certain period. Other references were those to the mediation committee of the Bundestag (*vermittlungsausschuss*), which unions turned to after failing to stop Agenda 2010, and to extra benefits for unemployed parents (*Kinderzuschlag*). As with the previous topic 7, the words in topic 4 refer to issues related to Outsiders, and we classify the Topic Agenda 2010 in the Outsider category.

Topics 9 and topic 5 then (topic proportions of 0.08 and 0.07) relate to the reform of statutory minimum wage and we label this topic Minimum Wage. In topic 9, we find words such as *mindestlohn850euro* and *allgemeingesetzmindestlohn*, while in topic 5 we find more specific references to the companies involved (*gmbh, telekom, vw, vwgesetz*). During the time covered, the introduction of a statutory minimum wage free of exceptions in all branches, sectors or companies (*mindestlohnausnahmen*, *ausnahmenbeimmindestlohn*, *lufthansa*, *zeitungszustell, weiterbranch*) was an important issue for the unions. As the statutory minimum wage was particularly relevant for Outsiders in the low-wage sector as well as for those with little formal education (Marx & Starke, 2017), we classify both in the Outsider category.

The last two topics, 6 and 3, then belong to both the Insider and Outsider categories. Topic 6 (topic proportion of 0.11) deals with Workplace Safety or Occupational Health (*arbeitsgesundheitsschutz*). The terms it covers all deal with employee wellbeing and accident avoidance in specific companies or sectors such as refineries (*raffineri*) or national power stations (*national, Kraftwerk*). Yet unlike the pure Insider topics, this topic also covers

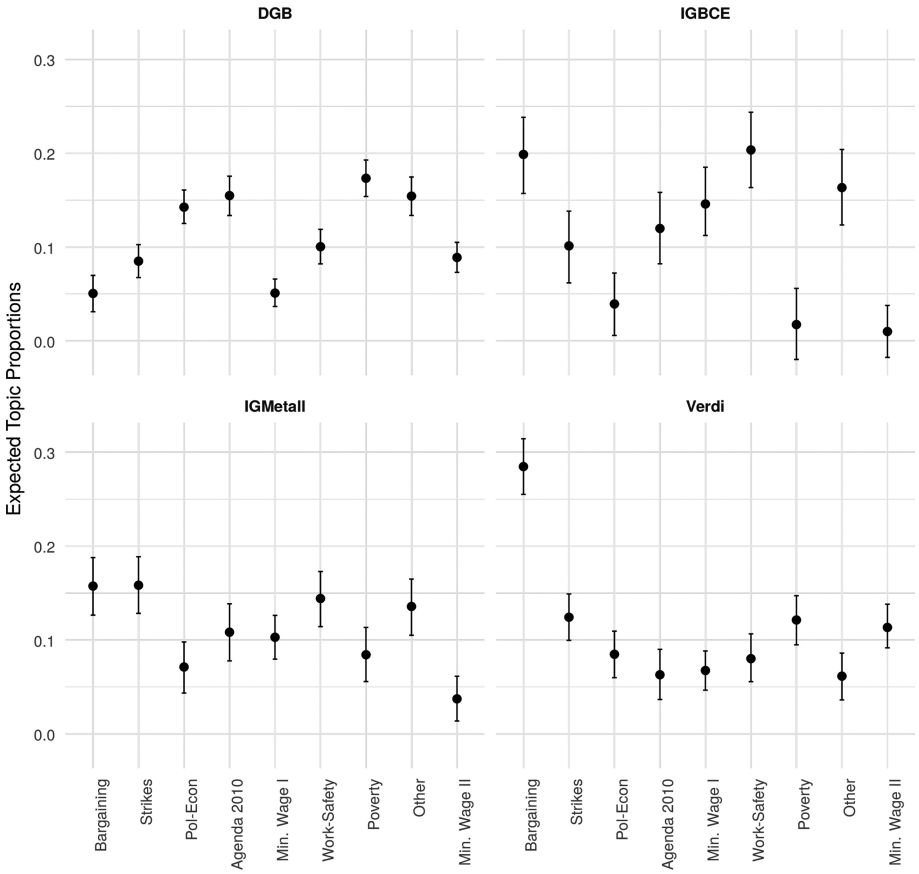
Outsiders, as most unions sought to go beyond the needs of individual employees (*akteur*) and towards improved workplace safety for all employees. Furthermore, any health and workplace safety measures should contribute to society as a whole. This means that they should contribute to better and more efficient environmental protection (*effizient, Umweltschutz*) and should reduce the burden on the healthcare system (*gesundheitswesen*).

The other topic related to the Insider and Outsider categories is topic 3, which refers to the Political Economy (topic proportion of 0.10). Terms here refer to a balanced budget (*Schuldenbremse*), short-time work benefits (*kurzarbeit*) and consumption-led growth as a driving force for economic stimulation (*binnennachfrage, aufschwung*). Also, we find references to the European level (*eurozone*), and these refer to demands by the unions to provide countries with the option to opt out (*exitstrategi*) from the Stability and Growth Pact (active from 2013). This as the unions had criticised the pact for its governmental spending restrictions that could lead to a focus on the supply side of economic growth, instead of stimulating demand. Beside overall concerns on economic growth, fiscal policy and rules to promote economic stability we also find references to one of the leading executive board members of the DGB, Claus Matecki (*matecki and sagtmatecki*). As the DGB focuses on economic growth unrelated to a single sector more than most other trade unions these statements were often made by Matecki, making him prominent in topic 3. As with the previous topic 6, we choose Political Economy topic 3 to refer both categories, to Insiders and Outsiders.

b) Topic Importance for Individual Unions

While the topics are of interest on their own, the strength of STM lies in its ability to provide associations with additional variables. As in our model, we deemed the topics to be dependent on the associations and the year of publication, we now turn to one of these aspects and look at how often the four associations referred to each of the topics. Figure 6 shows this, by giving the topic proportion for each of the topics for the texts of that union. The higher the proportion of a topic, the more references the union made to it.

Figure 6: Prevalence of the Topic Proportions for each of the four associations



Source: own compilation

Our conclusions here are less straightforward but the results confirm existing findings in the current state of the art (Oliver & Morelock, 2021). For example, based on the original Insider-Outsider theory, we would expect that IG BCE and IG Metall would concern themselves more with Insider (1 and 2) and less with Outsider topics (4, 5, 7, and 9). Mainly because both unions represent the Insiders, since the chemical, metal and electrical industries have stable and safe working conditions for most of their employees. Figure 6 shows that only isolated topics displays these union division. The Poverty topic (7) is less relevant for IG BCE and the Minimum Wage topic (9) is indeed less relevant for both industry organizations. Basically, this was because IG BCE and IG Metall feared a weakening of collective bargaining processes in the case of an introduction of the statutory minimum wage (Bender, 2020, pp. 232–239). On the other hand, the Insider topic Collective Bargaining (1) is more relevant for Ver.di and the union represent the service sector which we would expect to focus more on Outsider topics. Furthermore, the pattern of Outsider topics was visible for IG Metall since the Poverty topic (7) is more relevant for them. In sum, we do not find strong evidence for the union

division made by the Insider-Outsider theory. Instead, each of the four unions, independent of the sectors they represented, were concerned with topics relevant for both groups of the labour market.

Limitations

Grimmer & Stewart (2013: 269) noted that "all quantitative models of language are wrong – but some are useful" – and STM is no different. The logic and structure of a text are inherently complex and cannot hope to be captured by simple algorithms making a wide array of assumptions. While the method behind STM has several limitations of its own, most of its limitations are shared with other methods of text analysis.

As with most other methods of quantitative text analysis, there is the assumption that writers (in our case the trade unions) use a limited set of topics when writing their texts and thus that we could be able to "retrieve" these topics. Also, there is the problem of there being no "true" answer to the question of how many topics there are in any corpus of documents or what contents of those topics are. Moreover, as most of the STM algorithms use expectation maximization (EM), different starting values will lead to different topics, and thus different conclusions. Thus, different researchers can reach different conclusions on the number and content of the topics (in our case the press releases). This is an important point because, even when working with the same data, different conclusions on the number of topics might be drawn. While there is no direct way to circumvent this issue, we can minimise it by ensuring transparency, replicability, and by performing a wide array of sensitivity analyses to see how much one's conclusions could or would change given different assumptions. This was also our main consideration when providing our exact procedure above.

Also, the method is only useful for written material – in the eyes of STM, if there is no text an event or occasion simply does not exist. If it does, the user has to be willing to sacrifice the word order and any grammatical structure of that text in favour of a bag of words in which any text is reduced to a data frame of word counts. Besides, they then have to spend a considerable amount of time constructing, cleaning and pruning the data before they can be used, all while ensuring not to add any bias. Finally, they have to be prepared to find no result at all, or results that defy any classification, which was not the case in our application and illustration of STM in "testing" the Insider-Outsider theory

More specific for STM is the limitation that the model stops short of telling us what the optimal number of topics is and what they mean. Therefore, different researchers could come to a different number of topics with different labels, while using the same data. This is because no text corpus has a "true" number of topics with "true" labels. Their number and meaning instead depend on the aim and goal of our research (Aranda et al., 2011; Grimmer & Stewart, 2013; Jacobs & Tschötschel, 2019). Also, even when we can interpret some of the topics, there is always the possibility that we might fall victim to apophenia and identify patterns in what is random noise. Here, we tried to avert this the best we could by basing our selection of the number of topics on a combination between the quantitative information as generated by various tests one can carry out on the data, and the qualitative interpretation of whether the

generated topics carry a meaning that allows for a certain interpretation. Yet, we are well aware that this approach is not a panacea.

While there are no stopgap solutions to any of these limitations, most of them can be (partially) mitigated or assuaged by adopting a responsible research design. This includes being clear about one's intentions from the start, allowing for iteration, describing the cleaning process, and providing extensive arguments for the number of topics and the labels chosen, which we did here in the paper. Also, one should be honest in admitting that no analysis can ever give a "true" representation of the facts and that the truthiness of the research is limited to what can be validated. Indeed, each iteration of the method is merely one of the many possible "forked paths" (Gelman & Loken, 2014) that one could have chosen.

To help deal with these limitations, various techniques are available to the researcher, such as the use of pre-cleaning checks we used here following Denny and Spirling (2018). Or the various statistics and visualisations to provide further insight into the topics such as those provided by the stm and stminsights packages. Furthermore, researchers can implement techniques from other fields, such as discourse analysis to help validate their findings (see for an example Jacobs & Tschötschel, 2019).

Conclusion

The aims of this article were threefold. First, we sought to introduce structural topic modelling (STM) to the field of Industrial Relations. So far, when it comes to the analysis of large textual data, using quantitative methods, scholars in the field of Industrial Relations often lag behind other fields, such as the analysis of party manifestos (Gross & Jankowski 2020; Lamare & Budd, 2021) or media studies (Hase, Mahl, & Schäfer, 2022). This is surprising, since the use of quantitative text analysis, in particular STM, enables us to examine a greater variety and number of potential data sets and use state-of-the-art text analysis research tools.

Second, we set out to provide a step-by-step overview of how one should use STM as most of the methodological published work does not focus on actual practitioners. While others have described the mathematical details of the STM method, little attention has been paid to implementation of STM. Also, as we suspected that the results of an STM analysis would be very sensitive to the decisions made during the analysis, we not only addressed how to use STM in the field of Industrial Relations, but we also explored how the decisions researchers make affect the results. For an overview of these steps, see also Table 3.

Table 3: Overview of Recommended Steps

Step	Aim
Collect documents	Collect all documents needed for analysis
Convert documents	Convert all documents in a common text-type and remove conversion errors
PreText	Check for the risk of performing certain pre-processing steps
Pre-processing	Run chosen pre-processing steps

Step	Aim
Estimate number of Topics	Estimate number of topics in both a quantitative and qualitative sense
Decide on Prevalence	Decide on what is expected to influence the topics in the STM model
Run STM Model	Run the model with the chosen number of topics
Identify Topics	Identify topics by interpreting the words and reading underlying documents
Visualize	Further visualize results

Source: own compilation

Third, we sought evidence against the original predictions made in the Insider-Outsider literature (Rueda, 2005, 2007). In our fist analysis, we found evidence that trade unions represent both Insider and Outsider topics. By extending the period of 2000 to 2014 we had a total dataset of nearly 2,000 press releases published by the unions. This longer period gave a more accurate picture of the topic importance for the most powerful unions in Germany. We found 8 topics that referred to Insiders and Outsiders. In the second analysis, we went on to test theoretical predictions that representation across different sectors would be varied. Our conclusions here were less straightforward. However, we found little empirical evidence for the assumption that Insider topics are more important for the industry unions than for the others. The only differences regarding topic importance were visible for Minimum Wage (9), since it was less relevant for the industry unions (IG Metall, IG BCE) and more relevant for the service sector union (ver.di). The other topics demonstrated that all unions referred to both Insiders and Outsiders, regardless of the sector they mainly represented. As mentioned in the theory section we explain these Insider-Outsider representation patterns by referring to union membership mobilization strategies. In times of declining membership, unions try to mobilize new members regardless of whether they are Insiders or Outsiders. Efforts are made to catch the interest of both groups and to try to balance any conflicts of interest.

While for this case we found that STM is useful, we admit that the case we choose here was a rather safe one. That is, the data we used – press releases – often have a rigid structure and are comparatively clean when compared to other types of data that are often significantly more unstructured – such as tweets, speeches or interviews. The reason we did so was because we aimed to keep the focus on the method itself and not be too limited by any potential problems that unstructured data might cause. Yet, while running an STM might prove to be more difficult on such data, recent research has shown positive results for relatively unstructured data such as those deriving from Facebook posts (Luna, Pérez, Toro, Rosenblatt, Poblete, Valenzuela, Cruz, Bro, Alcatruz, & Escobar, 2022), tweets (Han, Yang & Piterou 2021), or speeches from ECB Executive Board members (Küsters, 2022). Given this, we are confident that STM could also be applied to alternative, more unstructured data, in the field of the Insider-Outsider theory, and plan to return to this theme in further research.

We hope that this article provides researchers with a standard of good practice in using STM. This is because, when done correctly, STM can lead to new and interesting discoveries in the field of Industrial Relations, help researchers to be innovative in their methods and help them to tackle research questions more efficiently.

References

Aranda, A. M., Sele, K., Etchanchu, H., Guyt, J. Y., & Vaara, E. (2021). From big data to rich theory: Integrating critical discourse analysis with structural topic modeling. *European Management Review*, *18*(3), 197–214. https://doi.org/10.1111/emre.12452

Benassi, C., & Dorigatti, L. (2018). The political economy of agency work in Italy and Germany. In V. Doellgast, N. Lillie, & V. Pulignano (Eds.), *Reconstructing solidarity: Labour unions, precarious work, and the politics of institutional change in Europe* (pp. 124–143). Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198791843.003.0006

Benassi, C., & Vlandas, T. (2016). Union inclusiveness and temporary agency workers: The role of power resources and union ideology. *European Journal of Industrial Relations*, *22*(1), 5–22. https://doi.org/10.1177/0959680115589485

Bender, B. (2020). Politisch-Ökonomische Konfliktlinien im sich wandelnden Wohlfahrtsstaat. Positionierung deutscher Interessenverbände von 2000 bis 2014. Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-658-31825-3

Bender, B. (forthcoming): Class Conflict or Consensus? Understanding Social Partner Positions on Social Policy Reforms. Journal of Social Policy (accepted 6 October 2022).

Beramendi, P., Häusermann, S., Kitschelt, H., & Kriesi, H. (Eds.). (2015). *The politics of advanced capitalism.* Cambridge: Cambridge University Press.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, *1*(1), 17–35. https://doi.org/10.1214/07-aoas114

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Broockman, D. E. (2012). The "problem of preferences": Medicare and business support for the welfare state. *Studies in American Political Development*, *26*(2), 83–106. https://doi.org/10.1017/s0898588x12000077

Cantillon, B., Seeleib-Kaiser, M., & Veen, R. (2021). The Covid -19 crisis and policy responses by continental European welfare states. *Social Policy & Administration*, *55*(2), 326–338. https://doi.org/10.1111/spol.12715

Culpepper, P. D., & Regan, A. (2014). Why don't governments need trade unions anymore? The death of social pacts in Ireland and Italy. *Socio-Economic Review*, *12*(4), 723–745. https://doi.org/10.1093/ser/mwt028

Czymara, C. S., Langenkamp, A., & Cano, T. (2021). Cause for concerns: gender inequality in experiencing the COVID-19 lockdown in Germany. *European Societies*, *23*(sup1), 68–81. https://doi.org/10.1080/14616696.2020.1808692

Davidsson, J. B. (2009). The politics of employment policy in Europe: Two patterns of reform. Paper prepared for the ECPR Joint Sessions, Workshop 22 "The Political Economy of Labour Market Reforms in Advanced Capitalist Economies" Lisbon, 14–19 April 2009.

Davidsson, J. B., & Emmenegger, P. (2012). Insider-outsider dynamics and the reform of job security legislation. In G. Bonoli & D. Natali (Eds.), *The politics of the new welfare state* (pp. 206–229). Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199645244.003.0010

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

Deutscher Gewerkschaftsbund (2021). *DGB-Mitgliederzahlen ab 2010.* Retrieved on November 10, 2021 from https://www.dgb.de/uber-uns/dgb-heute/mitgliederzahlen/2010?tab=tab_0_0#tabnav

Dörre, K. (2011). Functional changes in the trade unions. From intermediary to fractal organization? *International Journal of Action Research*, *7*(1), 8–48.

Durazzi, N., Fleckenstein, T., & Lee, S. C. (2018). Social solidarity for all? Trade union strategies, labor market dualization, and the welfare state in Italy and South Korea. *Politics & Society*, *46*(2), 205–233. https://doi.org/10.1177/0032329218773712

Emmenegger, P., Hausermann, S., Palier, B., & Seeleib-Kaiser, M. (2012). *The age of dualization: The changing face of inequality in deindustrializing societies.* Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199797899.001.0001

Esping-Andersen, G. (1999). *Social foundations of postindustrial economies.* Oxford: Oxford University Press. https://doi.org/10.1093/0198742002.001.0001

Farrell, J. (2015). Corporate funding and ideological polarization about climate change. *Proceedings of the National Academy of Sciences*, *113*(1), 92–97. https://doi.org/10.1073/pnas.1509433112

Fervers, L., & Schwander, H. (2015). Are outsiders equally out everywhere? The economic disadvantage of outsiders in cross-national perspective. *European Journal of Industrial Relations*, *21*(4), 369–387. https://doi.org/10.1177/0959680115573363

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

Greene, Z., Ceron, A., Schumacher, G., & Fazekas, Z. (2016). The nuts and bolts of automated text analysis. Comparing different document pre-processing techniques in four countries. *OSF Preprints.* https://doi.org/10.31219/osf.Io/ghxj8

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

Gross, M., & Jankowski M. (2020). Dimensions of political conflict and party positions in multi-level democracies: Evidence from the Local Manifesto Project. *West European Politics*, *43*(1), 74–101. https://doi.org/10.1080/01402382.2019.1602816

Grumbach, J. M. (2015). Polluting industries as climate protagonists: Cap and trade and the problem of business preferences. *Business and Politics*, *17*(4), 633–659. https://doi.org/10.1515/bap-2015-0012

Han, C., Yang, M., & Piterou, A. (2021). Do news media and citizens have the same agenda on COVID-19? An empirical comparison of twitter posts. *Technological Forecasting and Social Change*, *169*(S1), 120849, https://doi.org/10.1016/j.techfore.2021.120849

Hase, V., Mahl, D., & Schäfer, M.S. (2022). Der "Computational Turn": ein "interdisziplinärer Turn"? Ein systematischer Überblick zur Nutzung der automatisierten Inhaltsanalyse in der Journalismusforschung. Medien & Kommunikationswissenschaft, *70*(1–2): 60–78. https://doi.org/10.5771/1615-634x-2022-1-2-60

Hassel, A. (2014). The paradox of liberalization – Understanding dualism and the recovery of the German political economy. *British Journal of Industrial Relations*, *52*(1), 57–81. https://doi.org/10.1111/j.1467-8543.2012.00913.x

Jacobs T., & Tschötschel, R. (2019). Topic models meet discourse analysis: a quantitative tool for a qualitative approach. *International Journal of Social Research Methodology*, *22*(5), 469–485, https://doi.org/10.1080/13645579.2019.1576317

Küsters, A. (2022). Applying lessons from the past? Exploring historical analogies in ECB Speeches through Text Mining, 1997–2019. *International Journal of Central Banking*, *18*(1), 277–330. https://doi.org/10.2139/ssrn.3861671

Lamare R. J., & Budd J. W. (2021): The relative importance of industrial relations ideas in politics: A quantitative analysis of political party manifestos across 54 countries. *Industrial Relations*, *61*(1), 22–49. https://doi.org/10.1111/irel.12296

Lindbeck, A., & Snower, D. J. (2002). The insider-outsider theory: A survey (IZA Discussion Paper No. 534). Bonn. Retrieved from: https://docs.iza.org/dp534.pdf

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

Luna, J.L., Pérez, C., Toro, S., Rosenblatt, F., Poblete, B., Valenzuela, S., Cruz, A., Bro, N., Alcatruz, D. & Escobar, A. (2022). Much ado about Facebook? Evidence from 80 congressional campaigns in Chile. *Journal of Information Technology & Politics*, *19*(2), 129–139. https://doi.org/10.1080/19331681.2021.1936334

Maier, D., Niekler, A., Wiedemann, G., & Stoltenberg, D. (2020). How document sampling and vocabulary pruning affect the results of topic models. *Computational Communication Research*, *2*(2), 139–152. https://doi.org/10.5117/ccr2020.2.001.maie

Marx, P., & Starke, P. (2017). Dualization as destiny? The political economy of the German minimum wage reform. *Politics & Society*, *45*(4), 559–584. https://doi.org/10.1177/0032329217726793

Mimno, D., & Lee, M. (2014). Low-dimensional embeddings for interpretable anchor-based topic inference. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1319–1328). Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1138

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & Mccallum, A. (2011). Optimizing semantic coherence in topic models. In R. Barzilay & M. Johnson (Eds.), *EMNLP '11: Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 262–272). Stroudsburg, PA: Association for Computational Linguistics.

Mosimann, N., & Pontusson, J. (2017). Solidaristic unionism and support for redistribution in contemporary Europe. *World Politics*, *69*(3), 448–492. https://doi.org/10.1017/S0043887117000107

Naczyk, M., & Seeleib-Kaiser, M. (2015). Solidarity against all odds: Trade unions and the privatization of pensions in the age of dualization. *Politics & Society*, *43*(3), 361–384. https://doi.org/10.1177/0032329215584789

Oliver, R. J., & Morelock, A. L. (2021). Insider and outsider support for unions across advanced industrial democracies: Paradoxes of solidarity. *European Journal of Industrial Relations*, *27*(2), 167–183. https://doi.org/10.1177/0959680120911221

Ooms, J. (2021). *pdftools: text extraction, rendering and converting of pdf documents.* Retrieved on December 7, 2021 from https://docs.ropensci.org/pdftools/

Palier, B., & Thelen, K. (2010). Institutionalizing dualism: Complementarities and change in France and Germany. *Politics & Society*, *38*(1), 119–148. https://doi.org/10.1177/0032329209357888

Picot, G., & Menéndez, I. (2017). Political parties and non-standard employment: An analysis of France, Germany, Italy and Spain. *Socio-Economic Review*, *17*(4), 899–919. https://doi.org/10.1093/ser/mwx016

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

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). stm: An R package for structural topic models. *Journal of Statistical Software*, *91*(2), 1–40. https://doi.org/10.18637/jss.v091.i02

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

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In M. Chickering & J. Halpern (Eds.) Proceedings of the 20th Annual Conference on Uncertainty in Artificial Intelligence (UAI 2004) (pp. 487–494). Arlington, VA: AUAI Press.

Rueda, D. (2005). Insider–outsider politics in industrialized democracies: The challenge to social democratic parties. *American Political Science Review*, *99*(1), 61–74. https://doi.org/10.1017/S000305540505149x

Rueda, D. (2006). Social democracy and active labour-market policies: Insiders, outsiders and the politics of employment promotion. *British Journal of Political Science*, *36*(3), 385–406. https://doi.org/10.1017/S0007123406000214

Rueda, D. (2007). *Social democracy inside out: Partisanship and labor market policy in advanced industrialized democracies.* Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199216352.001.0001

Savin, I., Chukavina, K., & Pushkarev, A. (2022). Topic-based classification and identification of global trends for startup companies. *Small Business Economics.* https://doi.org/10.1007/s11187-022-00609-6

Schwander, H. (2019). Are social democratic parties insider parties? Electoral strategies of social democratic parties in Western Europe in the age of dualization. *Comparative European Politics*, *17*(5), 714–737. https://doi.org/10.1057/s41295-018-0122-5

Stüdemann, F. (2015). Tarifabschluss mit fadem Beigeschmack. In *Oberpfalznetz.* Accessed June 2015.

Taddy, M. (2012). On estimation and selection for topic models. In N. D. Lawrence & M. Girolami (Eds.), *Proceedings of The Fifteenth International Conference on Artificial Intelligence and Statistics* (pp. 1184–1193). La Palma: Proceedings of Machine Learning Research

Thelen, K. (2012). Varieties of capitalism: Trajectories of liberalization and the new politics of social solidarity. *Annual Review of Political Science*, *15*(1), 137–159. https://doi.org/10.1146/annurev-polisci-070110-122959

Visser, J. (2016). What happened to collective bargaining during the great recession? *IZA Journal of Labor Policy*, *5*(1), 1–35. https://doi.org/10.1186/S40173-016-0061-1