Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)

TopicShoal: Scaling Partisanship Using Semantic Search

Sami Diaf , Ulrich Fritsche
Department of Socioeconomics
Universität Hamburg
[sami.diaf,ulrich.fritsche]@uni-hamburg.de

Abstract

Document scaling techniques have been widely used in political science to infer partisanship measures and to rank documents on a scale of ideal points, based on bag-of-word approaches. These approaches typically underestimate the semantic and syntactic patterns contained in the corpus. Recent advances in natural language processing, particularly semantic search models, offer an improved topic coherence due to a semantic space of embedded words and documents, whose structure is able to identify topics without setting their number as a hyperparameter. We propose a scaling technique, namely *TopicShoal*, that extracts meaningful topic vectors using a semantic search technique (*Top2Vec*) and scales partisanship among speakers or parties using a Bayesian factor analysis on the document-topic distances, thereby enabling a semantic explanation of the ideal points' variations. This novelty, suited for both monolingual and multilingual corpora, addresses the bag-of-word constraint by capturing the narrative signals in the corpus and exploiting a coherent and independent topic vector structure. Applied to a corpus of German party manifestos and *Deutsche Bundesbank* executive board members' speeches, *TopicShoal* successfully identifies discourse-level differences among parties and speakers via topic intensities, whose projection on the ideal points' scale reveals common debated themes and other sideline interests that differentiate parties and speakers.

1 Introduction

Text mining in political science comprises distinct families of methods usually applied to monolingual text data. Topic models define probabilistic models used to extract groups of words with a semantic meaning, referred to as topics based on a generative model of texts, while the document scaling family gathers probabilistic as well as non-probabilistic approaches used to infer a unidimensional scale assumed to be a proxy of ideal-points or (ideological) positions prevailing among speakers or parties.

Non-probabilistic scaling techniques are based on pre-established wordlists from *reference* texts (Laver et al., 2003) whose availability outside the English language is limited, while probabilistic techniques are mostly based on the assumption of a Poisson distribution for word frequencies, as for *Wordfish* (Slapin and Proksch, 2008) which infers a unidimensional, normally distributed $\mathcal{N}(0, 1)$ scale for document positions, or the Poisson reduced rank models which permit to endow a time-variability to the learned scale (Jentsch et al., 2020). *Wordshoal* (Lauderdale and Herzog, 2016) uses *Wordfish* estimates over distinct debates to aggregate the results at the level of speakers, where differences in document positions within debates approximate the ideological stance between speakers. Such schemes have been used in political sciences to measure polarization of political parties in the United Kingdom (Goet, 2019), investigate left-right differences (Däubler and Benoit, 2021), in Germany for parties' manifestos (Jentsch et al., 2021) or for economic institutions' forecasting reports (Diaf et al., 2022) and were found to have some drawbacks in applications with small corpora or limited vocabulary (Hjorth et al., 2015) and to text pre-processing choices (Denny and Spirling, 2018). Scaling speakers using topic variations (Vafa et al., 2020) was proposed as a generalization of *Wordshoal* where word contributions are allowed to differ among speakers using a hierarchical Poisson factorization, while Latent Semantic Scaling (Watanabe, 2021) is a semi-supervised approach to scale documents on a specific task, using Latent Semantic Analysis (Deerwester et al., 1990) over sentences or paragraphs, augmented with a wordlist for positive/negative terms. Another hybrid approach learns a *Wordfish* scale that serves as an explanatory variable to a supervised LDA (Diaf and Fritsche, 2021) with the aim of tracking topics' prevalence over time using dynamic word frequencies.

Latent Dirichlet Allocation (Blei et al., 2003) is still the workhorse for topic model applications, despite being a heuristic method yielding relatively unstable results and being highly dependent on the hyperparametrization chosen by practitioners (Airoldi et al., 2014). Further variants were proposed to adapt the algorithm to the corpus specifications' or to add prior information as a semi-supervised approach (Eshima et al., 2020).

The advent of distributional representations helped researchers exploring the field of semantics and overcoming the bag-of-word restrictions by adopting neural architectures able to capture word similarity in context (Mikolov et al., 2013) and facilitate document comparisons (Dieng et al., 2019) even for multilingual documents that require a Zero-shot learning strategy (Bianchi et al., 2021). *Semscale* (Nanni et al., 2019) was proposed as a scaling technique relying on word embedding models, aiming at uncovering party positions from political manifestos and able to capture differences in multilingual manifestos.

Top2Vec (Angelov, 2020) belongs to the semantic search class of topic models where the number of topics, usually set as a hyperparameter, is automatically learned as being equal to the clusters of document representations using UMAP (McInnes et al., 2018) as a non-linear dimensionality reduction technique. As a mixture of three unsupervised models, it uncovers coherent topics and set their hierarchies for a better document-word representation, that could be augmented with pre-trained word embedding models.

This article proposes a novel semantic, topic-based semi-supervised scaling approach that outperforms the existing document scaling techniques in terms of coherence and interpretability, combining topic vectors learned from a semantic space and an aggregation scheme to derive ideal points for an intuitive positional analysis, suited to monolingual and multilingual corpora. It consists, at the first stage, of a semantic search model (*Top2Vec*) that uncovers coherent topics, serving as an input for a Bayesian factor model (Lauderdale and Herzog, 2016) that yields a positional scale with semantic properties through topic intensities. We argue that the usual techniques are constrained by the bag-of-word hypothesis and cannot uncover semantic signals from the corpus, but just similarities in word counts, known as *lexical overlap* (Nanni et al., 2019), that overlook both semantic and syntactic features, in addition of rendering aggregate-level measures sensitive to word frequencies distributions. Moreover, recent applications built upon word embedding models are prone to an information bias transferred from large corpora to small and specific ones for monolingual documents (Papakyriakopoulos et al., 2020) or from one language to another (Bianchi et al., 2021), however, the use of multilingual pre-trained embedding models is mandatory to ensure a language-transferability of topics other than the training set (Bianchi et al., 2021) that requires setting the number of topics.

Two corpora were chosen to test *TopicShoal* at the monolingual and multilingual levels respectively. The corpus of Comparative Manifesto Project (CMP) (Volkens et al., 2021) was used to get the last three legislative elections' manifestos to scale the six main parties forming the current German political landscape, resulting in a scale that identifies partisanship of four parties (CDU/CSU, FDP, Grüne and SPD) in themes related to security, local affairs and economic concerns, in contrast of two parties (AFD, Linke) dominating the two ends of the scale as they have different priorities/focus, hence extending the partisanship spectrum. The corpus of executive members' speeches at the German Central Bank (*Bundesbank*) during the period 2012-2017 (Karim El-Ouaghlidi et al., 2019) is mainly bilingual (German-English) and cannot be analyzed using traditional text mining techniques, however, applying TopicShoal with the help of a multilingual embedding model uncovers a member-specialization strategy from the given addresses with specific interests given to Eurozone, financial stability and digitalization.

2 Methodology

2.1 Top2Vec

Aside from traditional topic models which use variational inference to uncover topics from word counts, *Top2Vec* augments the usual distributional representation methods, as for Word2Vec, by adding a paragraph vector to the neural network (Angelov, 2020) to create a joint word and document representations forming a semantic space able to uncover associations that helps learning coherent topic vectors from dense areas of document using *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN) (McInnes et al., 2017), under the hypothesis that the number of dense areas of documents is equal to the number of topics. Hence, the number of topics is no longer a hyperparameter as for most algorithms.

Top2Vec features a structure of independent, mostly low-correlated, topics because of the HDBSCAN application, ensuring a non-overlapping outcome often found in traditional topic models, hence enabling a robust Bayesian aggregation on independent topics, instead of a debate-structure that might have an intertwined topic prevalence.

2.2 Bayesian factor analysis

We use a modified version of the Bayesian aggregation used in *Wordshoal* (Lauderdale and Herzog, 2016) by setting the document positions as being drawn from a truncated normal distribution, instead of a normal distribution, as the document-topic coefficients are indeed distances mainly on the [0,1] interval.

Let ψ_{ij} defines the score of i^{th} document in the j^{th} topic learned via *Top2Vec*. The Bayesian aggregation used in *Wordshoal* to infer a latent scale, represented by a vector of speakers' positions θ_i is as follows:

TopicShoal

1^{st} Stage: Apply *Top2Vec* and extract the inferred topics:
ψ_{ij} defines the distance between the i^{th} document and the j^{th} topic (based on cosine distance)
2^{st} Stage: Each topic inferred is assumed to form a *debate*:
Inferring ideal points θ_i using the following factor analysis:

$$\psi_{ij} \sim \mathcal{N}(\alpha_j + \beta_j \theta_i, \tau_i)$$
$$\theta_i \sim \mathcal{N}_{trunc}(0, 1)$$
$$\alpha_j, \beta_j \sim \mathcal{N}(0, 0.25)$$
$$\tau_i \sim \mathcal{G}(1, 1)$$

where $\mathcal{N}_{trunc}$ denotes the truncated normal distribution as ψ_{ij} are represent document-topic distances. β_j is a topic polarization parameter.

Lauderdale and Herzog (2016) assumed debates being independent and serving as a basis to a multiple *Wordfish* scaling within each debate, that renders different word contribution for each scale. While this assumption allows a dynamic word contribution per debate, it ignores a potential topics' prevalence that might differentiate speakers or parties out of the debate dimension. Hence, building an Bayesian factor analysis on semantic topics makes it possible to track their prevalence in the unidimensional scale of positions, using the learned β_j.

In other terms, *TopicShoal* ensures a debate transfer from a time perspective to a topic structure for a better interpretability of the ideal positions. This is motivated by the fact that debates are defined by their occurrence, but usually discuss the same topics or concerns.

3 Application

3.1 German political manifestos

Manifestos of six main German parties (AFD, CDU/CSU, FDP, Grüne, Linke and SPD) for the last three legislative elections (2013, 2017 and 2021) were collected from the CMP (Volkens et al., 2021), consisting of 933 documents coded into 7 manually-annotated different categories (External Relations, Freedom and Democracy, Political System, Economy, Welfare and Quality of Life, Fabric of Society and Social groups).

The prevailing manifestos' interests appear to have a focus on the past and present rhetoric, inline with results found in international manifestos (Müller, 2022), with 20 topics learned, indicating a slight dominance of themes related to society and quality of life, as shown in Table 1.

Topics 14 and 9, respectively criminality and communes/municipalities, polarize the scale to the right-hand side (CDU and AfD) as indicated by positive β_i while most negative topic contributions are related to the left-hand side (Grüne and Linke, negative β_i) of the scale. The 95% confidence intervals offer an idea of parties' interest breadth that are captured by the topic intensities in Table 3. Noticeable are the close ideal points of three parties (Grüne, FDP and SPD), indicating similar interests displayed in their manifestos, and the contrary holds for the AfD, whose position dominates the right-hand scale and appears to be insulated from other parties.

Wordshoal estimation using the same corpus was not convergent [1] in addition of requiring setting an identification constraint[2]. Results do not render a clear partisanship scale, as demonstrated in Figure

[1] Tolerance level set to 10^{-10}

[2] We assumed $\theta_{Linke_{2013}} < \theta_{AFD_{2013}}$

Topic	Top 10 Words
1	fluchtlinge integration asyl bleiberecht gefluchteten asylbewerber antragstellerin optionszwang gefluchtete abschiebungen
2	schulden schuldenbremse eurozone stabilitats europaische eu wachstumspaktes ezb wachstumspakts maastricht
3	russland staaten frieden beziehungen internationale usa internationalen vereinten nationen multilateralen
4	demokratie parteien fußspur nebenverdienste abgeordneten vermengung demokratische transparenz parlamente mandats
5	leistungen versorgung pflege rente medizinische ambulante ambulanten alter medizinischen gesetzlichen
6	arbeitnehmer beschaftigten arbeit arbeitgeber beschaftigte beschaftigung arbeitsplatze tarifvertragen leiharbeit tarifvertrage
7	kultur gedenkkultur kulturelle kunst kulturforderung restitution kulturellen aufarbeitung filmerbe kulturpolitik
8	ehe ehen paare adoptionsrecht adoptionen patchwork fureinander familien verheiratet familie
9	kommunen gemeinden regionen landkreise stadte landlichen lander ort kommunale bund
10	nachhaltige okologische nachhaltigkeit energien nachhaltigen energie okologischen nachhaltiges wachstum erneuerbare
11	bundestagswahl politik merkel koalition steinbruck wahlerinnen marktkonforme koalieren wahlprogramm doch
12	bildung schulen lernen schuler schule schulerinnen lehrer hochschulen unterricht lehr
13	nato bundeswehr militarische abrustung atomwaffen rustung streitkrafte militarischen buchel nuklearen
14	straftaten polizei kriminalitat strafverfolgung tater organisierte fußballstadien aufzuklaren straftater gewalt
15	verbraucher produkte honorarberatung nahrwerte ampel markt wettbewerb finanzprodukten smiley finanzmarkte
16	infrastruktur technologien ausbau deutschlandtakt digitalisierung innovationen digitale anschlussen verkehrswege nutzen
17	wahlt zukunft starken starke grun bekampfen burgernahes schutzen statt stimmt
18	walder natur artenvielfalt tiere klima naturnahe lebensraume umwelt wald klimaschutz
19	engagement zusammenhalt ehrenamtliches feuerwehr ehrenamtlich ehrenamtliche ehrenamt ehrenamtes engagierte feuer
20	landwirtschaft landwirte landwirt ackerbau bauerliche kleinbauerliche landbau agrarbetriebe agrarzahlungen junglandwirte

Table 1: Top 10 words of the topics learned by *Top2Vec* on the German political manifesto corpus.

Topic	Top 10 words
1	eurosystems finanzpolitik eurosystem euroraums finanzkrisen eurozone bankensektors geldpolitik geldpolitischer finanzkrise
2	bargelds geldpolitik geldmarkt bankbilanzen currency monetaren bargeld monetaire wahrung eurosystem
3	bankensektors bankensektor innovationen innovations finanzbranche finanzsektors bankensystems innovation finanzsektor bankensystem
4	eurosystems eurosystem zahlungsverkehr euroraums eurozone zahlungsmittel euroraum kreditvergabe geldmarkt transaktionen
5	repercussions risikoteilung nachhaltig risques risks risiko nachhaltige krisenmaßnahmen risque risk
6	empirical data statistics analyses statistical trends informationen indicators analysen finanzsystems
7	digitalen verbraucher digitale digitalisation consumer digital consumers cyber technologien technologie
8	cyber security sicherheit threat sicherheiten sicherzustellen safeguarding vulnerable secure danger
9	blockchain bitcoins bitcoin geldmarkt currencies zentralbankgeld bankbilanzen bankensystem geldpolitik bankensystems
10	geldpolitik geldpolitischer geldmarkt renminbi geldpolitische geldpolitischen currencies currency zentralbankgeld staatsanleihen

Table 2: Top 10 words of the topics learned by *Top2Vec* on the Bundesbank speeches corpus.

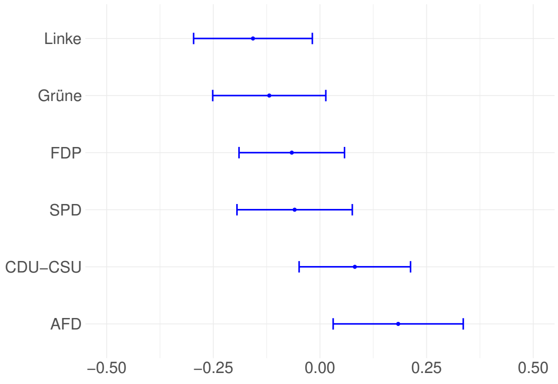

Figure 1: Estimated german parties' ideal points using *TopicShoal*.

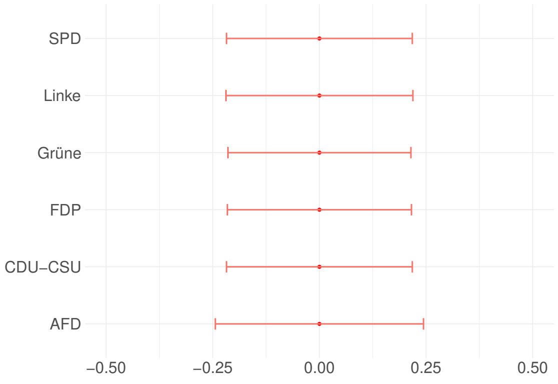

Figure 2: Estimated german parties' ideal points using *Wordshoal* (Lauderdale and Herzog, 2016).

	β_i
Topic 1	0.03
Topic 2	0.02
Topic 3	0.07
Topic 4	-0.16
Topic 5	-0.13
Topic 6	-0.23
Topic 7	-0.25
Topic 8	-0.33
Topic 9	0.15
Topic 10	-0.03
Topic 11	-0.08
Topic 12	-0.04
Topic 13	-0.30
Topic 14	0.30
Topic 15	-0.15
Topic 16	0.02
Topic 17	-0.31
Topic 18	-0.43
Topic 19	0.04
Topic 20	-0.05

Table 3: Estimated topic intensity β_i using *TopicShoal* on the German political manifesto corpus.

2, confirming that word counts are not always able to capture parties' partisanship.

3.2 *Bundesbank* speeches

Dataset of *Deutsche Bundesbank* executive board members' speeches (Karim El-Ouaghlidi et al., 2019) is used to test the multilingual version of *TopicShoal* with the help of a multilingual embedding model that ensures a topic-transferability between different languages used in the corpus. The dataset comprises 791 speeches given by nine different executive board members during the period 2012-2017 in four different languages (english, french, german and italian) although english and german share 98% of the corpus. *TopicShoal* is used to extract central bankers positions using multilingual embedding [3] (Reimers and Gurevych, 2019) given to *Top2Vec* that uncovered 10 different topics related to various aspects of monetary policy practices, as for risks and vulnerabilities (topic 5), European concerns (topic 1 and 4), financial innovation (topic 3), security and digitalization (topic 7, 8 and 9) and monetary policy (topic 10) as displayed in Table 2.

The positional analysis, as mentioned in Figures 4 and 5, helps classifying members into small groups of similar interests, given the learned topics, where topics related to classical monetary policy (topics 2 and 10) are polarizing positive members' positions, while risks and crisis-related concerns are mostly linked to negative positions, as reported in Table 4. Positions with wide confidence intervals (Beermann and Böhmler) could be explained by the variety of speeches, members gave during the period, while firm positions with relatively small confidence intervals (Dombret, Weidmann and Thiele) indicate a potential specialization or theme preferences of the members.

	β_i
Topic 1	-0.16
Topic 2	0.44
Topic 3	-0.15
Topic 4	0.22
Topic 5	-0.78
Topic 6	-0.22
Topic 7	0.30
Topic 8	-0.82
Topic 9	-0.01
Topic 10	0.53

Table 4: Estimated topic intensity β_i using TopicShoal on Bundesbank executive board members' corpus.

4 Conclusion

We presented a novel topic-based, scaling technique able to learn ideal points based on the corpus' semantic features and yielding an explanatory positional analysis, for both monolingual and multilingual corpora. It outperforms existing bag-of-word methods, which are not always convergent, and other semantic approaches that directly use bias-prone, pre-trained embedding models. Capturing meaningful topics, in addition to uncovering latent patterns within documents, helps building genuine unidimensional scales to rank speakers or parties without the need of taking the analysis to the multi-dimensional level or requiring further intervention on hyperparameters setting, though such efforts usually add a user-bias and are not time-efficient. *TopicShoal* demonstrated similar interests of four German political parties given to regular debated themes during the last three legislative campaigns, while scaling multilingual speeches at the *Bundesbank* proved to be effective in uncovering preferences and specialization of central bankers related

[3] paraphrase-multilingual-MiniLM-L12-v2

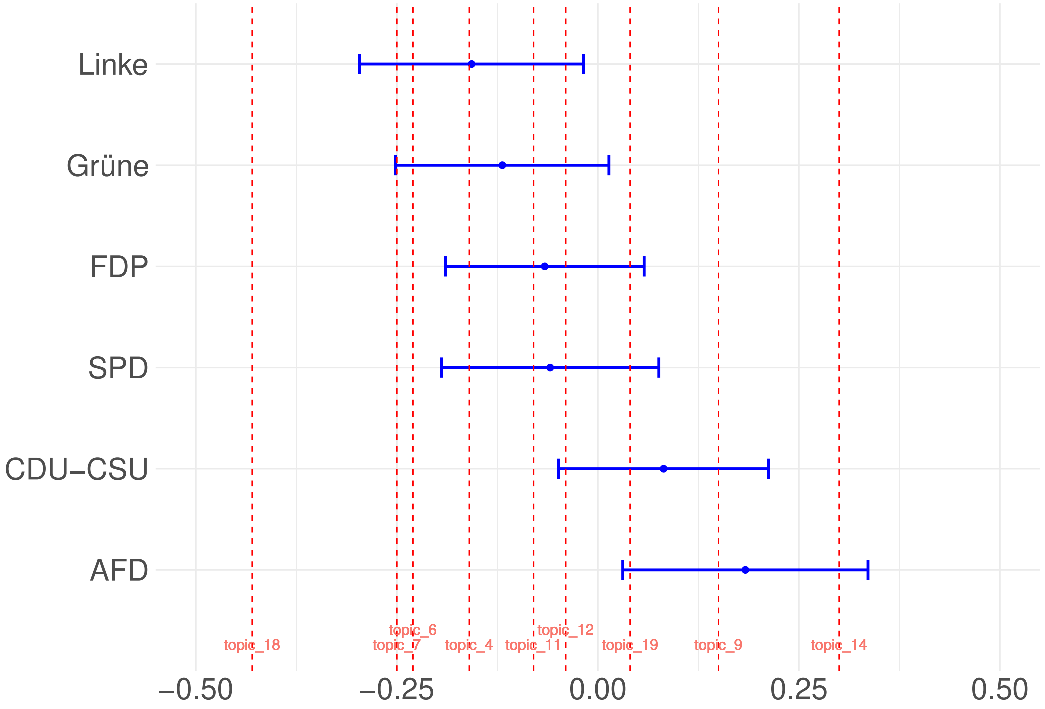

Figure 3: Estimated german parties' ideal points using *TopicShoal* with projected topic contributions.

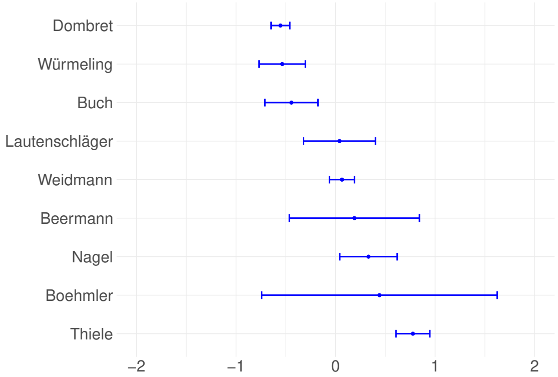

Figure 4: Estimated Bundesbank executive board members' ideal positions using *TopicShoal*.

to modern monetary policy practices and hot topics as for digitalization and financial innovation.

Acknowledgments

Financial support by the *Deutsche Forschungsgemeinschaft* (DFG) trough priority programme 1859, sub-project *Exploring the experience-expectation nexus in macroeconomic forecasting using computational text analysis and machine learning* is gratefully acknowledged.

References

Edoardo M. Airoldi, David Blei, Elena A. Erosheva, and Stephen E. Fienberg, editors. 2014. *Handbook of Mixed Membership Models and Their Applications*. Chapman & Hall / CRC Handbooks of Modern Statistical Methods. Taylor and Francis, Hoboken.

Dimo Angelov. 2020. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021. Cross-lingual contextualized topic models with zero-shot learning. *Association for Computational Linguistics*, pages 1676–1683.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Matthew J. Denny and Arthur Spirling. 2018. Text preprocessing for unsupervised learning: Why it matters when it misleads and what to do about it. *Political Analysis*, 26(2):168–189.

Sami Diaf, Jörg Döpke, Ulrich Fritsche, and Ida Rockenbach. 2022. Sharks and minnows in a shoal of words: Measuring latent ideological positions based

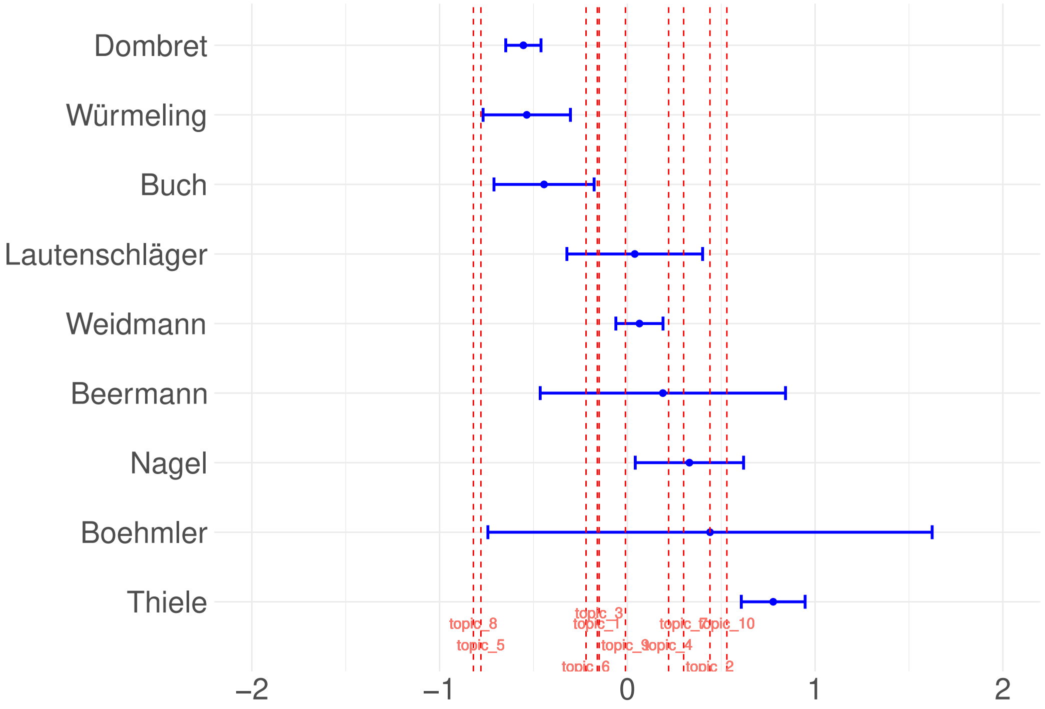

Figure 5: Estimated Bundesbank executive board members' ideal points using *TopicShoal* with projected topic contributions.

on text mining techniques. *European Journal of Political Economy*, page 102179.

Sami Diaf and Ulrich Fritsche. 2021. Topic scaling: A joint document scaling – topic model approach to learn time-specific topics. *arXiv preprint arXiv:2104.01117*, (2104.01117).

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2019. Topic modeling in embedding spaces. *arXiv preprint arXiv:1907.04907*.

Thomas Däubler and Kenneth Benoit. 2021. Scaling hand-coded political texts to learn more about left-right policy content. *Party Politics*, page 135406882110260.

Shusei Eshima, Kosuke Imai, and Tomoya Sasaki. 2020. Keyword assisted topic models. *arXiv preprint arXiv:2004.05964*.

Niels D. Goet. 2019. Measuring polarization with text analysis: Evidence from the uk house of commons, 1811–2015. *Political Analysis*, 27(4):518–539.

Frederik Hjorth, Robert Klemmensen, Sara Hobolt, Martin Ejnar Hansen, and Peter Kurrild-Klitgaard. 2015. Computers, coders, and voters: Comparing automated methods for estimating party positions. *Research & Politics*, 2(2):2053168015580476.

Carsten Jentsch, Eun Ryung Lee, and Enno Mammen. 2020. Time-dependent poisson reduced rank models for political text data analysis. *Computational Statistics & Data Analysis*, 142:106813.

Carsten Jentsch, Enno Mammen, Henrik Müller, Jonas Rieger, and Christof Schötz. 2021. Text mining methods for measuring the coherence of party manifestos for the german federal elections from 1990 to 2021. (8).

Karim El-Ouaghlidi, Matthias Gomolka, and Jens Orben. 2019. Bundesbank speeches: Data report 2019-12. (12).

Benjamin E. Lauderdale and Alexander Herzog. 2016. Measuring political positions from legislative speech. *Political Analysis*, 24(3):374–394.

Michael Laver, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(02).

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11):205.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Stefan Müller. 2022. The temporal focus of campaign communication. *The Journal of Politics*, 84(1):585–590.

Federico Nanni, Goran Glavas, Ines Rehbein, Simone Paolo Ponzetto, and Heiner Stuckenschmidt. 2019. Political text scaling meets computational semantics. *arXiv preprint arXiv:1904.06217*.

Orestis Papakyriakopoulos, Simon Hegelich, Juan Carlos Medina Serrano, and Fabienne Marco. 2020. Bias in word embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ACM Digital Library, pages 446–457, New York,NY,United States. Association for Computing Machinery.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Jonathan B. Slapin and Sven-Oliver Proksch. 2008. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3):705–722.

Keyon Vafa, Suresh Naidu, and David Blei. 2020. Text-based ideal points. *Association for Computational Linguistics*, 2020:5345–5357.

Andrea Volkens, Tobias Burst, Werner Krause, Pola Lehmann, Theres Matthieß, Sven Regel, Bernhard Weßels, Lisa Zehnter, and Wissenschaftszentrum Berlin für Sozialforschung. 2021. Manifesto project dataset.

Kohei Watanabe. 2021. Latent semantic scaling: A semisupervised text analysis technique for new domains and languages. *Communication Methods and Measures*, 15(2):81–102.