A Semantic Map of (Migration Discourse in?) the European Parliament

Giorgi Gogelashvili [*] **Samia Haque** [*] **Jakob Kleine** [*]
Dennis Stroh [*] **Quirin Unterguggenberger** [*]

Abstract

Motivated by the rise of populism in Europe since the late 1990s, this study investigates ideological shifts in European Parliament (EP) speeches using natural language processing. Drawing on the novel ParLawSpeech dataset (Schwalbach et al., 2025) which contains 574,199 speeches from 1999 to 2024 alongside metadata on speaker identity, we use sentence embedding models to examine the semantic content and emotional tone of parliamentary debates over time.

We expect that speech embeddings will form clusters reflecting party affiliation and ideological alignment. In step with recent political developments, we further hypothesize an increase in negative sentiment within the immigration debate among centrist and right-wing groups, accompanied by growing semantic similarity between these two factions over the past two decades. Finally, we test whether established migration-related narratives associated with right-wing populism can be identified in parliamentary discourse and how their prevalence has developed over time.

1. Introduction

The continued success of right-wing populist parties in the 21st century is widely regarded as a major threat to European democracy and integration (Fossum, 2023; Rummens, 2017). Populist rhetoric is commonly defined as constructing an antagonism between a 'pure people' and a 'corrupt elite' (Mudde, 2007). Right-wing populism is also closely tied to the issue of immigration. Parties of this ideology have played a central role in the increasing politicisation of immigration (Hutter & Kriesi, 2022), which represents a crucial factor for their political success (Kende & Krekó, 2020). Over the past decade, immigration has become an increasingly salient issue in European election campaigns (Dekeyser & Freedman, 2023) as well as in media coverage (Greussing & Boomgaarden, 2017).

Electoral gains of populist parties have manifested in significant changes of parliamentary discourse (Schwalbach, 2023). A recent quantitative analysis of EP speech embeddings has identified a gradual increase in emotional rhetoric since 1999, with right-wing populist groups leading the trend (Subtil & Verger, 2024). In the German national parliament, an LLM-based study has revealed increasing anti-solidarity messaging around immigration, not only for right-wing, but also christian-conservative and liberal parties (Kostikova et al., 2025). This trend begins around 2015, which marks the onset of the so-called 'refugee crisis' (Brücker et al., 2020).

More fine-grained analyses of the migration discourse have revealed the use of common underlying narratives, defined as 'selective depictions of reality' and 'patterns of interpretation' through which the issue is relayed to the public. Social media posts from populist leaders commonly employ anti-immigrant frames like 'immigrants take our jobs' or anti-establishment narratives such as 'our sovereignty is under threat' (Seiger et al., 2025).

Building on these foundations, we apply computational methods to European Parliament debates to analyze migration discourse. We combine topic modeling to measure issue salience, semantic embeddings to map ideological positioning, and narrative detection to quantify the adoption of populist rhetorical frames across party groups.

This report provides a quantitative assessment of how the growing prominence of right-wing populism and immigration as a salient political issue manifests in debates in the European Parliament, with potential implications for broader societal discourse and legislative outcomes. All parliamentary speeches between 2004 and 2024 as recorded by the ParlLawSpeech dataset inform the analyses.

Speeches are first classified into topics using Latent Dirichlet Allocation (LDA) to identify immigration-related debate and to estimate its prominence over time. Analyses of the distribution of migration-related speeches across predefined

[*]Equal contribution . Correspondence to: AB <first1.last1@student.uni-tuebingen.de>.

Project report for the "Data Literacy" course at the University of Tübingen, Winter 2025/26 (Module ML4201). Style template based on the ICML style files 2025.

debate agendas provide quantitative evidence consistent with agenda-setting strategies employed by right-wing populist groups. With the use of speech embeddings, we examine the semantic dimensions along which party groups can be differentiated and find evidence for an increased use of previously identified anti-immigration narratives by right-wing groups compared to moderate factions.

2. Data and Methods

2.1. Dataset description

We are using the novel *ParlLawSpeech* (PLS) dataset from Schwalbach et al. 2025 for the investigation of our study. It contains more than 570,000 plenary speeches from legislative periods of the European parliament (EP) between 1999 and 2024. The authors also provide (partially) machine translated text in English for about 40% of the speeches, since the EP stopped providing official translations around the end of 2012. Furthermore, the dataset contains metadata on the speakers and the speeches given, e.g. date and agenda item under which the speech was given, if submission was in written form and/or from multiple *members of parliament* (MEPs), or the speaker's party affiliation (referring to European political parties/groups), among other. We further enriched the dataset with metadata accessible from the public API of the EP's "Open Data Portal", in particular the national party affiliations of each speaker (by using the *EP-ID* of the respective MEPs). This allowed us to link the PLS dataset with the *Chapel Hill Expert Survey* (CHES) from Rovny, Bakker et al. 2025. The CHES dataset estimates party positioning on European integration, ideology (e.g. left/right) and policy issues for national parties in all member states of the European Union (EU). The study surveyed hundreds of experts roughly every four years between 1999 and 2024 and more recently (**TODO**: since when???) also includes ratings of non-EU policy issues such as immigration or anti-elite rhetoric (**TODO:** which are relevant in particular?) Assuming that the ideological orientation of a speaker's affiliated national party roughly reflects his own position, the CHES data set could help us to better control our analyses, as membership of a European party (group) presumably allows for less detailed/granular statements/assumptions.

2.2. Data preprocessing

Handling Duplicates (TODO)

Removing Commentary (TODO) We detect high amount of superfluous commentary in transliterated speeches: markers of the original language, background incidents, and procedural notes. These markers might be source of unwanted bias, which we want to avoid. Fortunately they are predominantly located within parentheses and can be easily removed with rule-based methods. We also observe substantial redundency in the openning and closing sections of the speeches. These sections follow similar rhetorical structures but exhibit substantial lexical variation. To identify low-impact sentences we use TF-IDF algorithm to score the amount of information they contain. We construct separate corpora for opening and closing sentences, and an average TF-IDF score is computed for each sentence. [TODO: Explain how we found cuttoff point]

Translation To keep the speeches most comparable in the embedding space, we use English translations instead of the orignal speeches. Until the year X (TODO), the Parllaw dataset includes a machine translation for each speech. The remaining X (TODO) translations were created using Gemini (2.5-flash, TODO: add citation), a LLM capable of translating texts across various domains (TODO citation). We checked that Gemini *1)* did not re-formulate speeches that were already in English and that *2)* its translations are comparable to Parllaw's in the embedding space. For this, we tested its translations on a random sample of speeches that had already been translated by Parllaw. Gemini *1)* preserved speeches which were already in English [1] and *2)* created translations whose embeddings are very similar to Parllaw's (bootstrapped 0.95 confidence interval of mean cosine similarity: 0.969, n=1001). Thus, we assume that Gemini's and Parllaw's translations are similar enough to conduct our analysis under the assumption that all translations stem from the same source after filling in the missing translations with Gemini's. However, we note that the mixture of two translation approaches might nevertheless introduce a bias to our dataset, that we have to check for. (TODO: did we check for that)

2.3. Methods

2.3.1. Topic Modelling with LDA

To identify how parties talk about migration, we first have to assign speeches to a semantic topic. For this purpose, we use Latent Dirichlet Allocation *(LDA)* [TODO: source]. (TODO why LDA?) LDA is a probabilistic topic model. It assumes that in the analyzed corpus (here: the collection of speeches) there is a set of topics, which are probability distributions over all words in the corpus. It considers each document (here: a speech) as a bag of words that were sampled from these topics. For example, if a topic had high probabilities for the words "fish", "net", "water", then documents covering "fishing" would (under LDA's assumptions)

[1] (TODO add quantifier) Since we created translations before extensively cleaning the dataset, some English speeches included bracketed language flags that led to Gemini re-translating the English speeches. These reformulations are however almost identical to the original speech. Therefore, we accepted those instances where Gemini failed to recognize English texts.

have a high probability of being labelled as that topic.

We tested different parameters (numer of topics, number of iterations over the dataset) and compared the resulting topic coherence (TODO explanation) and the fidelity of the topics through manual inspection. Our final model contains 30 topics (for 10 iterations) — one of which assigns highest probabilities to the words X, Y, Z (TODO words), which we call "migration topic".

For each topic, the model assigns each speech in the corpus a probability of covering that topic. To find speeches covering migration, a minimum topic probability was identified manually: two authors rated a sample of 100 speeches whose migration score (i.e. the assigned probability of the LDA model for the migration topic of that speech) was in the range [X, Y] (TODO range) which is where we suspected the relevance threshold.

(TODO finish this)

2.3.2. Semantic Embeddings

Semantic embeddings have been widely used in political text analysis (Miok et al., 2024; Nanni et al., 2021; Rudkowsky et al., 2018). Our aim is to capture patterns in how different political groups address migration. We select candidate embedding models from the MTEB leaderboard (Enevoldsen et al., 2025), based on overall performance and parameter count. Final model selection is based on (i) intra- and interparty cosine similarities, (ii) predictive performance of a logistic regression model with political affiliation as our target variable, and (iii) Kmeans clustering quality measured by homogenity and completeness.

A key concern is that general-purpose semantic embeddings may be primarly capturing stylistic and topical variations and subsequently political group ideologies influence on the embeddings might be neglegible. We test whether intra- and interparty similarity distributions differ substantially with a two-sample Kolmogorov-Smirnov test.

We examine whether party affiliations are encoded in speech embeddings and how these patterns evolve over time. Dimensionality reduction has been used to ascertain parties ideological shift over time and to reveal underlying political dimension with word associations for each reduced axis (Rheault & Cochrane, 2020). Exploratory analysis showed that, although party influence is present, it is not the defining factor of our semantic embeddings. To better understand how party affiliations manifest in the vector space, we aim to identify a subspace of the embedding space in which political and ideological differences become more salient.

To this end, Instead of simply using PCA, we employ Partial Least Squares (PLS). PLS allows us to find directions in the embedding space that are maximally associated with party labels, making it suitable for uncovering latent political dimensions that are not necessarily dominant in the overall variance of the data.

The prevalence of established migration-related rhetoric was assessed using semantic search in a shared embedding space. We used all suitable migration narratives that were identified in a recent report by the European Commission's Joint Research Centre (Seiger et al., 2025, p.130). Each narrative was represented by a short descriptive sentence, which was embedded using the model's built-in 'retrieval-query' prompt. Semantic proximity between narratives and speeches was quantified using cosine similarity.

To validate whether semantic similarity to these narratives captured meaningful political differences, we correlated similarity scores with expert-coded party positions on migration policy and overall ideology from the Chapel Hill Expert Survey (Jolly et al., 2022). Pearson correlation coefficients were evaluated using a Bonferroni-adjusted significance threshold to account for multiple comparisons. Temporal trends and party-block differences in narrative prevalence were analysed as fixed effects of linear mixed-effects models, which incorporated random intercepts and slopes at the party-block level.

3. Results

[Should this be in Discussion???]While a clear interpretation of the underlying political dimensions requires substantial domain knowledge, we believe that combining word associations with extreme examples of speeches along each cardinal direction provides strong clues about their connotations. Based on this analysis, we interpret the first PLS axis as a **conciliatory** ⇔ **oppositional** discourse spectrum, and the second axis as a **moral / human-rights** ⇔ **pragmatic-benefits** debate Figure 2.

Moral outrage and discussion of human rights violations have been consistently key aspects of both green-left blocks and parts of the right block. Along the first axis, we observe little to no movement over the years overall, suggesting that political blocks have largely maintained their characteristic way of conducting discourse. Nevertheless, there is a clear division between centrist and oppositional blocks, with greens often positioned in between. Oppositional blocks exhibit adversarial framing and conflict-driven rhetoric, whereas centrist blocks focus more on consensus-building. On the second axis, we observe a clear shift along the ethical–pragmatic spectrum. Between 2016 and 2020, many parties move from pragmatic policy framing towards more moral debates. Christian conservative and right-wing blocks remain closer to the axis center, while green and left blocks maintain stronger positions on the moral end of the spectrum.

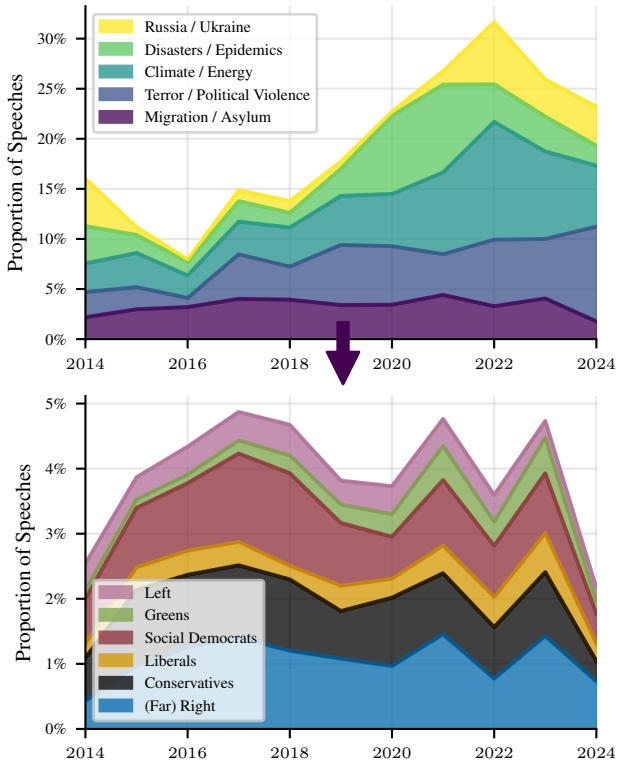

Figure 1. **Top:** Prevalence of selected topics in European Parliament debates over the past decade, as identified by LDA topic modeling (see repository for an interactive version with all topics). **Bottom:** Proportional contributions of political groups to migration topic. In both panels, proportions are computed by dividing by the total number of speeches per year.

4. Discussion & Conclusion

References

Brücker, H., Kosyakova, Y., and Vallizadeh, E. Has there been a "refugee crisis"?: New insights on the recent refugee arrivals in germany and their integration prospects. *Soziale Welt*, 71(1/2):pp. 24–53, 2020. ISSN 00386073. URL https://www.jstor.org/stable/27004992.

Dekeyser, E. and Freedman, M. Elections, Party Rhetoric, and Public Attitudes Toward Immigration in Europe. *Political Behavior*, 45(1):197–209, March 2023. ISSN 0190-9320, 1573-6687. doi: 10.1007/s11109-021-09695-w. URL https://link.springer.com/10.1007/s11109-021-09695-w.

Enevoldsen, K., Chung, I., Kerboua, I., Kardos, M., Mathur, A., Stap, D., Gala, J., Siblini, W., Krzemiński, D., Winata, G. I., Sturua, S., Utpala, S., Ciancone, M., Schaeffer, M., Sequeira, G., Misra, D., Dhakal, S., Rystrøm, J., Solomatin, R., Ömer Çağatan, Kundu, A., Bernstorff, M., Xiao, S., Sukhlecha, A., Pahwa, B., Poświata, R., GV, K. K., Ashraf, S., Auras, D., Plüster, B., Harries, J. P., Magne, L., Mohr, I., Hendriksen, M., Zhu, D., Gisserot-Boukhlef, H., Aarsen, T., Kostkan, J., Wojtasik, K., Lee, T., Šuppa, M., Zhang, C., Rocca, R., Hamdy, M., Michail, A., Yang, J., Faysse, M., Vatolin, A., Thakur, N., Dey, M., Vasani, D., Chitale, P., Tedeschi, S., Tai, N., Snegirev, A., Günther, M., Xia, M., Shi, W., Lù, X. H., Clive, J., Krishnakumar, G., Maksimova, A., Wehrli, S., Tikhonova, M., Panchal, H., Abramov, A., Ostendorff, M., Liu, Z., Clematide, S., Miranda, L. J., Fenogenova, A., Song, G., Safi, R. B., Li, W.-D., Borghini, A., Cassano, F., Su, H., Lin, J., Yen, H., Hansen, L., Hooker, S., Xiao, C., Adlakha, V., Weller, O., Reddy, S., and Muennighoff, N. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*, 2025. doi: 10.48550/arXiv.2502.13595. URL https://arxiv.org/abs/2502.13595.

Fossum, J. E. In What Sense Does Right-Wing Populism Pose a Democratic Challenge for the European Union? *Social & Legal Studies*, 32(6):930–952, December 2023. ISSN 0964-6639. doi: 10.1177/09646639231153306. URL https://doi.org/10.1177/09646639231153306.

Greussing, E. and Boomgaarden, H. G. Shifting the refugee narrative? An automated frame analysis of Europe's 2015 refugee crisis. *Journal of Ethnic and Migration Studies*, 43(11):1749–1774, August 2017. ISSN 1369-183X, 1469-9451. doi: 10.1080/1369183X.2017.1282813. URL https://www.tandfonline.com/doi/full/10.1080/1369183X.2017.1282813.

Hutter, S. and Kriesi, H. Politicising immigration in times of crisis. *Journal of Ethnic and Migration Studies*, 48(2):341–365, January 2022. ISSN 1369-183X, 1469-9451. doi: 10.1080/1369183X.2020.1853902. URL https://www.tandfonline.com/doi/full/10.1080/1369183X.2020.1853902.

Jolly, S., Bakker, R., Hooghe, L., Marks, G., Polk, J., Rovny, J., Steenbergen, M., and Vachudova, M. A. Chapel Hill Expert Survey trend file, 1999–2019. *Electoral Studies*, 75:102420, February 2022. ISSN 02613794. doi: 10.1016/j.electstud.2021.102420. URL https://linkinghub.elsevier.com/retrieve/pii/S0261379421001323.

Kende, A. and Krekó, P. Xenophobia, prejudice, and right-wing populism in East-Central Europe. *Current Opinion in Behavioral Sciences*, 34:29–33, August 2020. ISSN 2352-1546. doi: 10.1016/j.cobeha.2019.11.011. URL https://www.sciencedirect.com/science/article/pii/S2352154619301299.

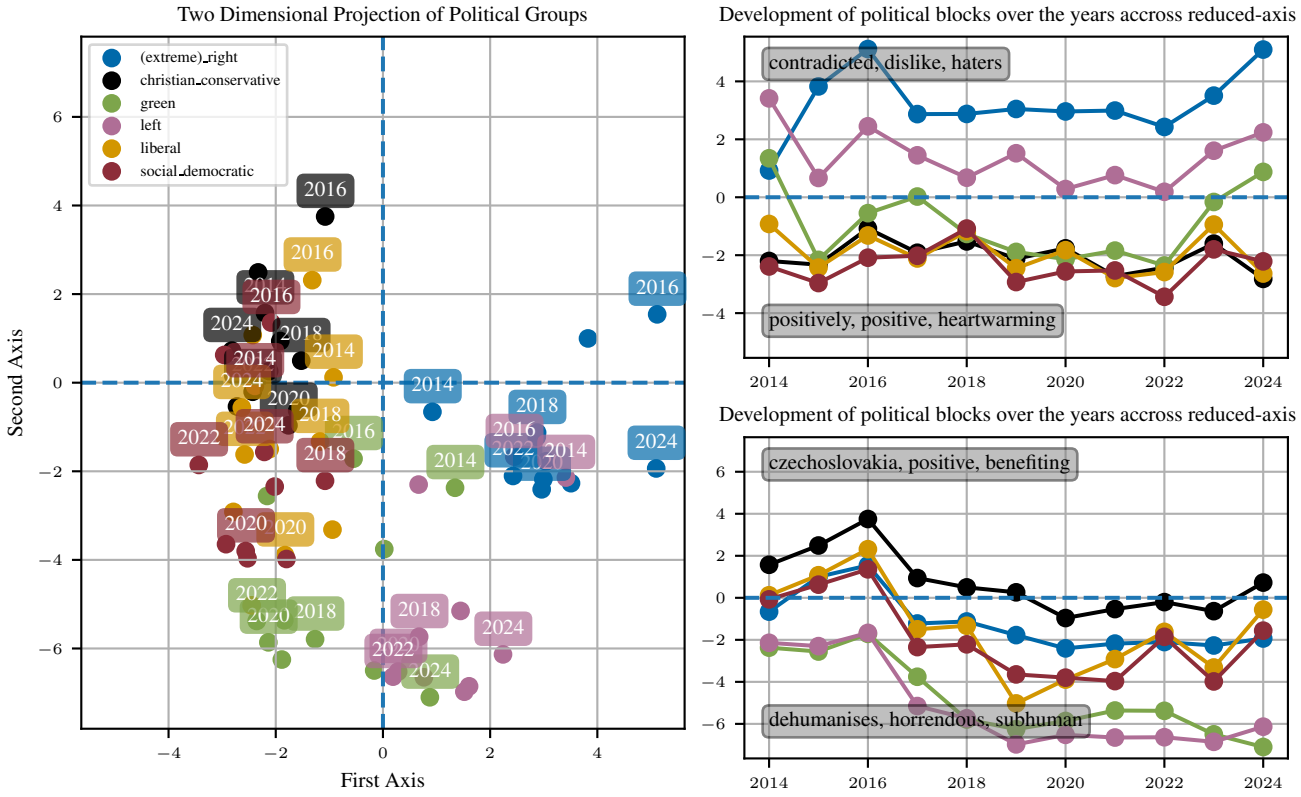

Figure 2. **Left.** Position of each political group **Right.** Movement of political groups over the time displayed separately for each dimension

Kostikova, A., Pütz, O., Eger, S., Sabelfeld, O., and Paassen, B. LLM Analysis of 150+ years of German Parliamentary Debates on Migration Reveals Shift from Post-War Solidarity to Anti-Solidarity in the Last Decade, September 2025. URL http://arxiv.org/abs/2509.07274. arXiv:2509.07274 [cs].

Miok, K., Hidalgo Tenorio, E., Osenova, P., Benítez-Castro, M.-A., and Robnik-Sikonja, M. Multi-aspect multilingual and cross-lingual parliamentary speech analysis. *Intelligent Data Analysis*, 28(1):239–260, February 2024. ISSN 1571-4128. doi: 10.3233/ida-227347. URL http://dx.doi.org/10.3233/IDA-227347.

Mudde, C. *Populist Radical Right Parties in Europe*. Cambridge University Press, Cambridge, 2007. ISBN 978-0-521-85081-0. doi: 10.1017/CBO9780511492037. URL https://www.cambridge.org/core/books/populist-radical-right-parties-in-europe/244D86C50E6D1DC44C86C4D1D313F16D.

Nanni, F., Glavas, G., Rehbein, I., Ponzetto, S. P., and Stuckenschmidt, H. Political text scaling meets computational semantics. *ACM/IMS Transactions on Data Science*, 2(4):1–27, November 2021. ISSN 2691-1922. doi: 10.1145/3485666. URL http://dx.doi.org/10.1145/3485666.

Rheault, L. and Cochrane, C. Word embeddings for the analysis of ideological placement in parliamentary corpora. *Polit. Anal.*, 28(1):112–133, January 2020.

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, S., and Sedlmair, M. More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures*, 12(2–3):140–157, April 2018. ISSN 1931-2466. doi: 10.1080/19312458.2018.1455817. URL http://dx.doi.org/10.1080/19312458.2018.1455817.

Rummens, S. Populism as a Threat to Liberal Democracy. In Kaltwasser, C. R., Taggart, P., Espejo, P. O., and Ostiguy, P. (eds.), *The Oxford Handbook of Populism*, pp. 0. Oxford University Press, October 2017. ISBN 978-0-19-880356-0. doi: 10.1093/oxfordhb/9780198803560.013.27. URL https://doi.org/10.1093/oxfordhb/9780198803560.013.27.

Schwalbach, J. Talking to the populist radical right: A comparative analysis of parliamentary debates. *Legislative Studies Quarterly*, 48(2):371–397, 2023.

Schwalbach, J., Hetzer, L., Proksch, S.-O., Rauh, C., and Sebők, M. Parllawspeech. (Version 1.0.0) [Data set]. GESIS, Cologne. https://doi.org/10.7802/2824, 2025.

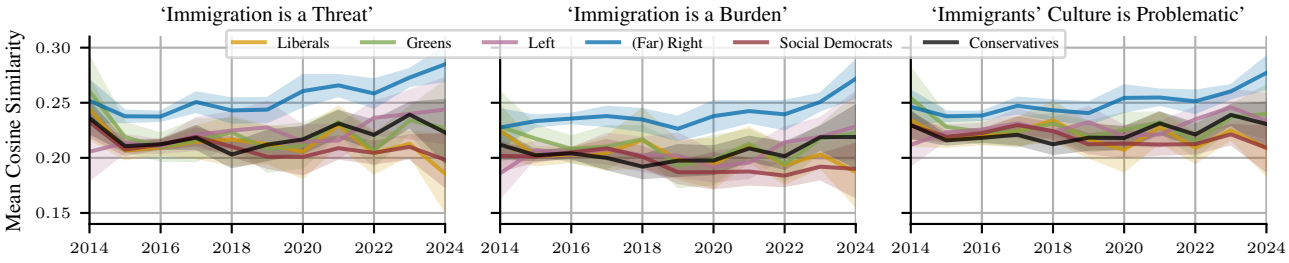

Figure 3.

Seiger, F., Kajander, N., Neidhardt, A.-H., Scharfbillig, M., Dražanová, L., Deuster, C., Krawczyk, M., Blasco, A., Icardi, R., Tzvetkova, M., Bakker, L., Olivo Rumpf, K., and European Commission (eds.). *Navigating migration narratives: research insights and strategies for effective communication*. Publications Office, Luxembourg, 2025. ISBN 978-92-68-28629-6. doi: 10.2760/2350572.

Subtil, H. and Verger, V. Emotional rhetoric and the rise of populism in the european parliament. Policy Brief 108, IPP, June 2024.