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

Giorgi Gogelashvili [*1] **Samia Haque** [*2] **Jakob Kleine** [*3] **Dennis Stroh** [*4] **Quirin Unterguggenberger** [*5]

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

2. Data and Methods

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.

2.1. Data Cleanup

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]

2.2. 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

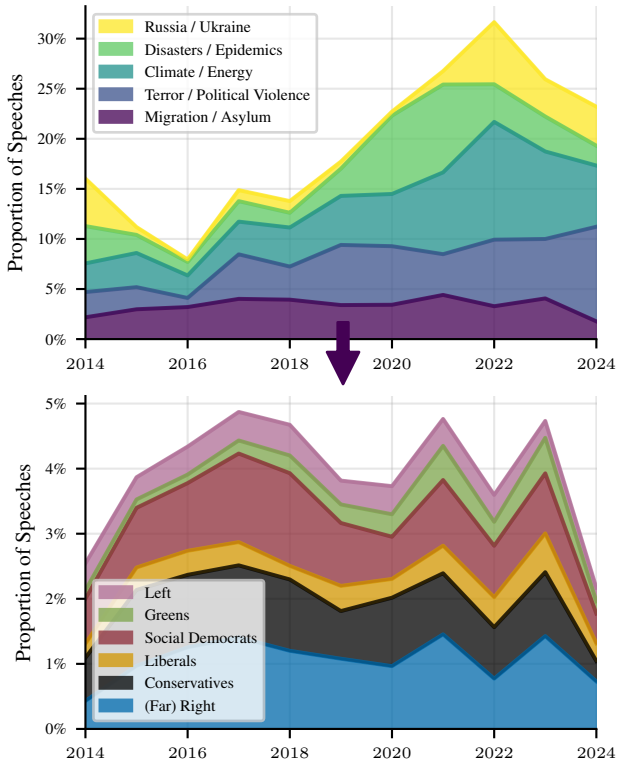

Figure 1. **Top:** Prevalence of selected topics in European Parliament debates over the past decade, as identified by LDA topic modeling (see repository for an interactive version with all topics). **Bottom:** Proportional contributions of political groups to migration topic. In both panels, proportions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

[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.

4. Discussion & Conclusion

References

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.

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.

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

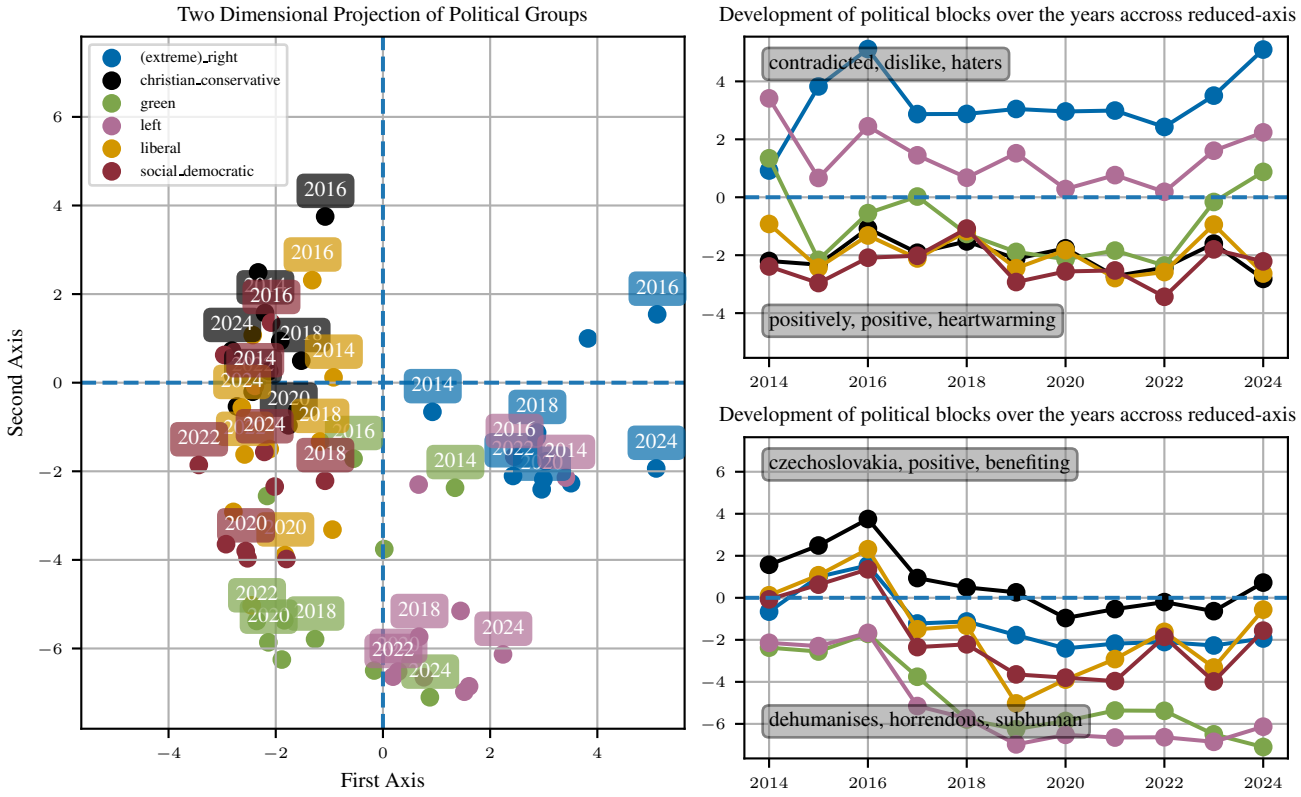

Figure 2. **Left.** Position of each political group **Right.** Movement of political groups over the time displayed separately for each dimension

2024. ISSN 1571-4128. doi: 10.3233/ida-227347. URL http://dx.doi.org/10.3233/IDA-227347.

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.

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.

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.

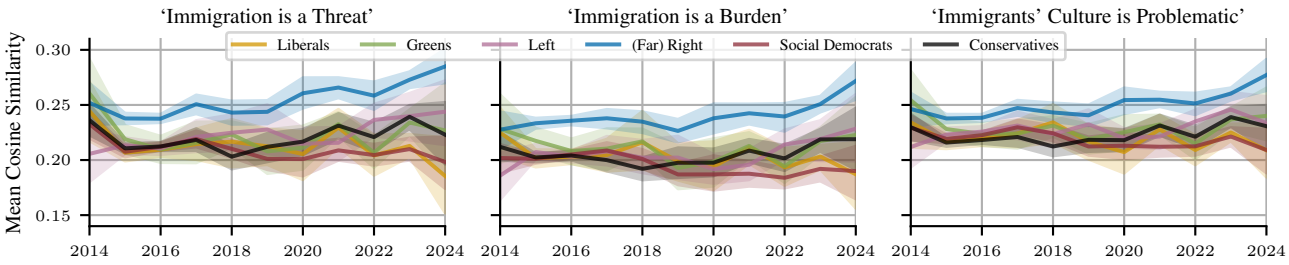

Figure 3. Softer colours represent bootstrapped 95% confidence intervals.