

Márton Rakovics, senior lecturer at Eötvös Loránd Science University, Budapest  
marton.rakovics@tatk.elte.hu

## **Investigating language- and political polarization in Hungary through two decades of parliamentary speeches**

Faculty of Law Osijek, 15<sup>th</sup> September 2023, Room 3/17

Computational methods of social science are becoming increasingly important because of the growing amount of naturally generated data available for research purposes. The aim of this lecture is to show some of the opportunities large-scale computation text analysis provides through a research project of the Research Center for Computational Social Science (RC2S2) at Eötvös Loránd University, which investigates language- and political polarization in the Hungarian parliament.

The first part of the lecture covers the fundamental notions and methods utilized in natural language processing (NLP), like word and document embeddings, topic modeling, transformers, boosting, and a graph perspective on texts, highlighting the differences compared to methods tailored to tabular data. An important, but often overlooked methodological aspect also covered here is the inherent qualitative nature of the results achieved with such computational methods. Since the quantitative analysis process for large-scale texts is much less canonized than the ones for traditional survey data, reflecting on internal and external validity has a more pronounced role in NLP than in some of the other quantitative methods. The second part of the lecture showcases some of the methods in use, applied to the transcripts of the parliamentary speeches held by politicians of the Hungarian Parliament from 1998 to 2022, roughly 400.000 unique texts. Since political communication has undergone significant change over the last two decades, the goal is to draft a picture of language and political polarization through speeches in parliament.

In the research that the lecture covers, after the basic text-preprocessing routines, we first used a time-dependent Structural Topic Model (Roberts et al., 2019) to identify the topics that distinguish groups – e.g., parliamentary factions and parties – of speakers in each year. This model identifies distinct topics of a corpus in an exploratory fashion based-on word usage and assigns a word distribution to each topic and a topic distribution to each text. Then we constructed Word Mover's document Embeddings (Wu et al., 2018) of speeches that allowed us to measure the similarities of the previously identified topics. Tracking the usage and similarities of topics helped reveal the political communication strategy of each party relative to others throughout the period. We also built classification models using XGBoost (Hastie et al., 2019) and BERT (Devlin et al., 2018) to predict the party affiliation of the speakers. Evaluating these models for each party and parliamentary term (or year) enabled us to examine the changes in amplitude and characteristics of language- and political polarization over time.

Based on the results of the analysis, the patterns of change in political communication strategies were mapped both for the governing and the opposition parties, and individual politicians who have been key figures throughout the examined period. Within-party use of topics showed a steady increase in homogeneity for right-wing parties, and a decrease for left-wing parties. Moreover, according to the results of the classification models, increasing language- and political polarization was observed between the governing parties and all other parties of the Hungarian Parliament. The patterns of intra-party homogenization and inter-party polarization evident in the speeches are in line with the characteristics of neopopulism-theory (Csigó 2016), according to which a speculative mediatized political space emerged where politicians observe the strategies employed by other politicians to appear popular and distinguish themselves from others by creating self-referential communication bubbles.

References:

- Csigó, P. (2016). *The Neopopular Bubble. Speculating on "the People" in Late Modern Democracy*. Central European University Press.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.
- Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). STM: An R package for structural topic models. *Journal of Statistical Software*, 91, 1-40.
- Wu, L., Yen, I. E., Xu, K., Xu, F., Balakrishnan, A., Chen, P. Y., ... & Witbrock, M. J. (2018). Word Mover's Embedding: From word2vec to document embedding. arXiv preprint arXiv:1811.01713.