Computational Linguistics for Political and Social Sciences

In recent years, an increasing number of studies has been published in the newly emerging text-as-data field. More and more scholars in the areas of political and social science are taking advantage of the ever increasing amount of text available (not only on the internet, but also transcriptions of parliamentary debates, newspaper texts, or party manifestos) to address a heterogeneous set of research questions. While this trend has already brought many promising results, it is not free of risks and challenges. In their seminal paper, Grimmer and Stewart (2013) not only discuss the potential of text-as-data approaches but also highlight the pitfalls that arise when applying NLP methods for the investigation of questions from political and social science.

We therefore argue for a closer collaboration between scholars from the social/political sciences on the one hand, and researchers from the area of computer science, NLP and computational linguistics on the other hand, to overcome those challenges and advance the state of the art for applications in the field of computational social science. As a first step to bridge the gap between the different communities, we organised the 1st Workshop on Computational Linguistics for the Political and Social Sciences (CPSS 2021).[1] The workshop took place in September 2021 as a virtual event, co-located with the Conference on Natural Language Processing (KONVENS 2021) in Düsseldorf, Germany. To meet the diverse needs of our research fields, we not only asked for long and short paper submissions but also for non-archival abstracts, in order to allow researchers to discuss work in progress without committing to a publication. The workshop program included five long and four short paper presentations and six non-archival abstracts that have been presented as posters. The presentations covered a wide range of topics, starting from NLP tools and corpus annotation that support research in the social science (Glaser, Patz, & Stede, 2021; Kahmann, Niekler, & Wiedemann, 2021) to the analysis of framing and formulaic speech (Russo, Comandini, Caselli, & Patti, 2021; Yu & Fliethmann, 2021), work on topic modelling and topic detection for political text analysis, using a variety of supervised and unsupervised techniques (Ahltorp, Dürlich, & Skeppstedt, 2021; Brand, Schünemann, König, & Preböck, 2021; Koh, Boey, & Béchara, 2021; Kreutz & Daelemans, 2021) and methodological studies (De Vos & Verberne, 2021). This JLCL special issue presents four selected long paper contributions from CPSS 2021.[2]

The first paper by De Vos and Verberne addresses a methodological question, namely the problem of data sparsity for the application of machine learning in political research. The authors present a replication study where they investigate the impact of pre-processing when only little data is available, showing the sensitivity of the models to variation regarding training and test splits and pre-processing. Their findings question

[1] https://old.gscl.org/en/arbeitskreise/cpss/cpss-2021

[2] The proceedings of the CPSS 2021 workshop are available online: https://old.gscl.org/media/pages/arbeitskreise/cpss/cpss-2021/workshop-proceedings/352683648-1630596221/cpss2021-proceedings.pdf.

previous results from the literature and highlight the importance of data set size and the validation of model robustness.

The contribution of Yu and Fliethmann studies media framing in German newspaper articles on the European Refugee Crisis (2014–2018). The authors test approaches to frame detection that do not rely on large-scale manual annotations. Their first method is based on LDA topic modelling, the second approach combines static word embeddings with a set of handcrafted keywords based on an expert-curated framing schema. Comparing the two techniques, Yu and Fliethmann show that the embedding-based approach yields better and more interpretable results. This illustrates the benefits to be gained from interdiscipinary work that combines domain knowledge from political science with NLP techniques for exploratory text analyses.

Another approach related to framing is presented in Russo et al. who analyse the use of proto-slogans in political communication before the 2019 European election, based on more than 700,000 comments extracted from the Facebook pages of two Italian leaders of populist parties (Matteo Salvini and Luigi Di Maio). The paper describes how the data has been clustered, followed by a manual annotation step, in order to detect proto-slogans used by the party leaders' supporters. The long-term objective of this work is the identification of stylometric patterns in informal populist social media posts.

The final paper by Glaser and collegues argues for using Named Entity Recognition and Named Entity Linking, two well-established NLP tasks, as an alternative source of information for political text analysis that is more transparent, robust and interpretable than topic modelling. The paper presents an add-on to the United Nations Security Council (UNSC) Debates corpus (Schoenfeld, Eckhard, Patz, van Meegdenburg, & Pires, 2019) and compares two approaches for obtaining this information. The pros and cons of each method are discussed, based on an intrinsic evaluation and an exploratory study that asks which entities are mentioned by different political actors in debates on the agenda of Women, Peace and Security.

Due to space limitations, we have only been able to report some of the results from the papers in this volume, and the short summaries given above surely do not do the work justice. Therefore, we invite the reader to form their own opinion and hope that they will find them insightful and intellectually rewarding.

We would like to thank the authors for their fine contributions and the reviewers for their constructive feedback which helped to improve the quality of the manuscripts: Adrien Barbaresi, Julian Bernauer, Chris Biemann, Christian Gawron, Goran Glavas, Annette Hautli-Janisz, Slava Jankin, Jonathan Kobbe, Sebastian Pado, and Esther van den Berg. Finally, we want to thank the editors of the Journal for Language Technology and Computational Linguistics for their support in putting together this special issue. We hope that the reader will enjoy the result!

The guest editors,
Ines Rehbein, Gabriella Lapesa, Goran Glavaš and Simone Paolo Ponzetto.

References

Ahltorp, M., Dürlich, L., & Skeppstedt, M. (2021). Textual contexts for "Democracy": Using topic- and word-models for exploring Swedish government official reports. In *Proceedings of the 1st workshop on computational linguistics for political text analysis* (pp. 45–52).

Brand, A., Schünemann, W. J., König, T., & Preböck, T. (2021). Detecting policy fields in German parliamentary materials with Heterogeneous Information Networks and node embeddings. In *Proceedings of the 1st workshop on computational linguistics for political text analysis* (pp. 53–58).

Glaser, L., Patz, R., & Stede, M. (2021). UNSC-NE: A Named Entity Extension to the UN Security Council Debates Corpus. In *Proceedings of the 1st workshop on computational linguistics for political text analysis* (p. 79-88).

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. doi: 10.1093/pan/mps028

Kahmann, C., Niekler, A., & Wiedemann, G. (2021). Application of the interactive Leipzig Corpus Miner as a generic research platform for the use in the social sciences. In *Proceedings of the 1st workshop on computational linguistics for political text analysis* (p. 39-44).

Koh, A., Boey, D. K. S., & Béchara, H. (2021). Predicting Policy Domains from Party Manifestos with BERT and Convolutional Neural Networks. In *Proceedings of the 1st workshop on computational linguistics for political text analysis* (pp. 67–78).

Kreutz, T., & Daelemans, W. (2021). A semi-supervised approach to classifying political agenda issues. In *Proceedings of the 1st workshop on computational linguistics for political text analysis* (pp. 59–63).

Russo, I., Comandini, G., Caselli, T., & Patti, V. (2021). Share and shout: Discovering proto-slogans in online political communities. In *Proceedings of the 1st workshop on computational linguistics for political text analysis* (p. 25-33).

Schoenfeld, M., Eckhard, S., Patz, R., van Meegdenburg, H., & Pires, A. (2019). *The UN Security Council Debates.* Retrieved from `https://dataverse.harvard.edu/citation?persistentId=doi:10.7910/DVN/KGVSYH`

de Vos, H., & Verberne, S. (2021). Small data problems in political research: a critical replication study. In *Proceedings of the 1st workshop on computational linguistics for political text analysis* (pp. 3–12).

Yu, Q., & Fliethmann, A. (2021). Frame Detection in German Political Discourses: How Far Can We Go Without Large-Scale Manual Corpus Annotation? In *Proceedings of the 1st workshop on computational linguistics for political text analysis* (pp. 13–24).