Machine learning on legislative text in EU and EU countries to provide a quantitative backbone

Master of Science thesis project

Dec 25, 2018

Machine learning in Political Science

Survey data, the engine of the behavioral revolution of the social sciences is about to run its course, with low response rate and poorly representative samples being the norm rather than the exception. Fortunately, vast amount of new information from social media, via digitalized governmental archives, to population registries are opening up new exiting avenues for innovative social science research, such as paternity leave and children's performance in school, extent of censorship in Chinese online new reporting, or conditions for receptiveness to fake news. Moreover, the new data availability in combination with tools from machine-learning has spurred an interest in prediction and sophisticated policy-recommendations, ranging from optimize relocation of immigrants given their skill-set and local labor market needs, via probabilistic detection of election fraud, to forecasting of popular unrest and civil war. The undertaking of such research questions was, until recently, outside the realm of social science. There are however limits to the amount of new insights that can be obtained purely from richer data and "black-box" import of machine-learning tools. More robust, new insights require similar steps to be taken in the development of applied, testable, theoretical models to facilitate direct empirical evaluations of the model dynamics and the consistency of the model with the data. Such a step requires a solid grounding in computing.

The aim of the project is to develop an algorithm based on Machine Learning methods to help determine where EU directives and regulations originate from. The download/upload model states that part of the legislative process in the EU is member countries *uploading* parts of their existing national body of legislation to be incorporated to the EU legislation. If this is the case one should be able to identify parts of the native legislative texts from their country of origin in the EU-legislation. Based on this a machine learning algorithm will be implemented to compare and examine legislative texts.

By taking as input the national legislation of the countries before an EU regulation is made, and the finished EU-regulation, and search for similarities, for instance how much of the respective countries national legislation is to be found in the EU regulation, one can use this as a proxy for assessing which country or countries has the most influence on the resulting regulation. As all EU-laws are translated to all languages in the EU, language differences does not need to be taken into consideration.

It is expected to find various levels of what Padgett calls synthesis and emulation [1] (meaning mixes of several states, synthesis, or more or less copying from one state, emulation), depending of what field one has to consider, but what is said about this earlier is qualitative and inconsistent, so this quantitative approach might serve as a backbone for further research.

To evaluate this model one can follow the similarities into the download-phase, where the countries of the EU have to implement the laws following the guidelines set by the directive and see to what degree the member countries follow it. That is, how compliant the country is, which is a subject that has been more studied and thus have more material to compare with. So one can see if this method can replicate the findings of various articles written on the compliance of EU countries to EU law [2-5].

When this is done, I will look at how these factors change over time, to see if the dominant countries become more dominant or less, or if they change.

The milestones are as follows

1. Spring 2019: Develop, based on recurrent neural networks, reiforcement learning and autoencoders [6], code and theory for analyzing text. In particular, develop code with the aim to extract specific phrases and sentences.

2. Fall 2019: Start including selected texts from the EU and apply the above Machine Learning techniques to these. Start analyzing the data.

3. Spring 2020: Final analysis of data and wrap up of thesis.

The thesis is expected to be handed in May/June 2020.

References.

1. Stephen Padgett, *Between synthesis and emulation: EU policy transfer in the power sector*, Journal of European Public Policy **10**, 227 (2003).

2. Dimiter Toshkov, *Embracing European Law: Compliance with EU Directives in Central and Eastern Europe*, European Union Politics **9**, 379 (2008).

3. Tanja A. Borzel, *Why there is no southern problem. On environmental leaders and laggards in the European Union*, Journal of European Public Policy **7**, 141 (2000).

4. Dimiter Toshkov, *In search of the worlds of compliance: culture and transposition performance in the European Union*, Journal of European Public Policy **14**, 933 (2007).

5. Eva Thomann and Asya Zhelyazkova, *Moving beyond (non-)compliance: the customization of European Union policies in 27 countries*, Journal of European Public Policy **24**, 1269 (2017).

6. Aurelien Geron, *Hands-on Machine Learning with Scikit-Learn and TensorFlow*, O'Reilly, 2017.