Project Title: Machine Learning Algorithms with Application to Political Event Data

Project Description This project will analyze two machine learning algorithms and discuss their application to automated processing of political event data. The first algorithm, Hidden Markov Modeling (HMM), uses observed data to make inferences about an unobserved latent state. The second method, Hierarchical Association Rule Modeling (HARM) selects a subset of decision rules from a set of candidates to classify observations categorically. Applying machine learning to the classification of political event data can greatly reduce the cost in human effort, time, and money.

The motivation for this project is to update the Militarized Interstate Disputes (MID) dataset, which has been widely used in academic research and policy discussions. The MID project relies on humans reading journalistic accounts and manually entering the classification of the event according to a defined schema (Gerner et al., 1994; Grimmer and Stewart, 2013). This dependence on humans is both less accurate, less efficient, and more expensive than automated methods (King and Lowe, 2003; Mikhaylov et al., 2012; Ruggeri et al., 2011). The most recent version of MID data was released in 2004; an update through 2010 is delayed indefinitely and has cost millions of dollars.

This project will explore whether HMM and HARM can help automate the processing of political event indicators to update the MID dataset. The final paper will include an analysis of the algorithms used for HMM and HARM, a comparison in cost and complexity to extant methods for processing event data, and preliminary results for automating the production of MID data.

References

Gerner, D. J., P. A. Schrodt, R. A. Francisco, and J. L. Weddle (1994). The analysis of political events using machine coded data. *International Studies Quarterly 38*(1), 91–119.

Grimmer, J. and B. M. Stewart (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 1–31.

King, G. and W. Lowe (2003). An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International Organization 57*(03), 617–642.

McClelland, C. A. (1961). The acute international crisis. *World Politics 14*(1), 182–204.

Mikhaylov, S., M. Laver, and K. R. Benoit (2012). Coder reliability and misclassification in the human coding of party manifestos. *Political Analysis 20*(1), 78–91.

Obrien, S. P. (2010). Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International Studies Review 12*(1), 87–104.

Ruggeri, A., T.-I. Gizelis, and H. Dorussen (2011). Events data as bismarck's sausages? intercoder reliability, coders' selection, and data quality. *International Interactions 37*(3), 340–361.