RTextTools

Loren Collingwood,[1,2,*] Tim Jurka[3], Amber Boydstun[3], Emiliano Grossman[4], and Wouter Van Atteveldt[5]

1. Political Science Department, University of Washington
2. Center for Statistics and the Social Sciences, University of Washington
3. Political Science Department, University of California, Davis
4. Centre of European Studies, Sciences Po
5. Communication Science Department, Vrije Universiteit

*Contact author: lorenc2@uw.edu

Keywords: Supervised Learning, Text Analysis, Machine Learning, Social Science

Machine learning has only recently entered the world of social and political science. For years, scholars have used undergraduate research assistants for various classification tasks—such as labeling congressional bill titles, classifying parliamentary speeches, and coding party platforms. Many social scientists are in search of new tools to automate tedious coding tasks, and social scientists have now begun using supervised learning to automate the labeling of documents. *RTextTools* is a recently developed *R* package that provides a uniform interface to several existing *R* algorithms to label text.

From the *tm* package, we include functions to generate document term matrices, stopword removal, and stemming. Currently, the package includes standardized training and classification access to *svm, NaiveBayes, glmnet, randomForest, tree, AdaBoost, Bagging*, and *nnet* algorithms. In addition, we include a C++ maximum entropy train and classification function. We also provide a function for cross validating each algorithm. Finally, several accuracy and ensemble agreement functions are provided to examine how well each algorithm does in terms of predictive accuracy and ensemble agreement. Researchers can quickly identify which text documents are coded with high degrees of accuracy and which documents need to be coded by humans for active learning.