Title: Exploring Applications of Natural Language Processing to Policy Research

Authors: Debbie Kim and Brandon Sepulvado

Affiliation (for both): NORC at the University of Chicago

The digital age and the popularity of social media means that researchers have access to more data than any time in the past. Much of these data are text based, and natural language processing (NLP) approaches—which include topic modeling, vector space modeling, and sentiment analysis—are becoming critical tools as they reveal the structure and meaning of text. We see the benefits of NLP analyses in our everyday life. Google uses NLP to very quickly comb through massive amounts of text data in order to return documents (i.e., websites) most closely related to your search terms; our email servers use NLP to quickly sort incoming emails as Spam or Not Spam; most individuals have experience with text correction on their cellphones.

Policy researchers are beginning to explore ways to apply these extremely powerful, promising techniques to social science data for the public good. Researchers now have the ability to apply statistical models to rapidly analyze newspaper articles, Twitter feeds, government documents, and political propaganda. NLP techniques allow researchers to either surface existing topics (unsupervised models) or measure pre-determined constructs of interest (supervised models). The policy applications of such abilities are vast. For instance, researchers could analyze the difference in meaning of the word “science” between political parties, quickly discern the most salient topics in Congressional hearings, or seek to understand how teachers vs. policymakers define “quality” in teaching and learning. Cutting-edge NLP methods even seek text to identify causal relationships.

At NORC, the Education and Child Development department is working on two ongoing projects leveraging NLP techniques in policy-relevant ways. The first project—through the Institute of Education Sciences-funded National Center for Education Access and Choice—endeavors to create more holistic school quality measures by leveraging text data and NLP. Traditional school quality measures (e.g., achievement data) are limited because they represent only a narrow slice of the overall picture of schooling. For this project, we are partnered with computer scientists, economists, and machine learning experts at Tulane University to analyze text data from the Great Schools website. Parents of K-12 students rate schools on Great Schools, providing a text review for each school. NLP techniques allow us to empirically determine what topics matter most to parents (e.g., school discipline, teacher quality) and how these topics relate to the school choice decisions they make as parents. This analysis would take a tremendous amount of human resources and money to conduct by hand; NLP approaches mitigate these costs by quickly and accurately analyzing this enormous corpus of text data. Ultimately, findings from this text analysis combined with existing quantitative measures of school quality will enable the creation of more holistic quality measures.

The second project seeks to understand how policy ideas diffuse throughout different publics/audiences and the ways education stakeholders engage in consuming and sharing particular forms of information. Social media is a natural source of information to answer these questions because myriad stakeholders are represented and express themselves online, yet the vast amount of text information makes a manual review of these data prohibitive. NORC is piloting a project for the Gates Foundation to test whether NLP techniques (e.g., topic modeling, sentiment analysis) can answer two questions: Can text analysis tell the Gates Foundation the extent to which their education priorities are being represented on social media? And if so, how are they being represented and by whom?

In addition to these two projects, the Center for Excellence in Survey Research at NORC is exploring how natural language processing techniques might complement existing qualitative program analysis workflows. Funding agencies and other stakeholders frequently need actionable insights within a short timeframe, yet qualitative data collection and analysis require much time: from planning interview and focus group questions and identifying participants to transcribing recordings and iteratively reviewing transcripts to identify themes and findings of interest. In particular, we are investigating (1) the quality of automatic transcription programs and (2) how well different topic identification algorithms, such as topic models and co-word networks, produce the same set of topics as human coders when used to analyze the same transcripts. The results from this project will help evaluation experts maximize the efficiency of their work, those in charge of implementing policies (e.g., school officials) more quickly understand the consequences of their actions, and policy makers be able to ascertain in a more agile manner whether their policies achieve the desired outcome(s).

This paper describes and demonstrates the utility of unsupervised natural language processing techniques, notably topic modeling and closely related methods, and their combination with other text analytic tools (i.e., sentiment analysis) for three specific policy challenges: (1) the construction of holistic school quality measures that do not neglect the experiences of families attending the schools; (2) the diffusion of policy ideas among diverse stakeholders, and (3) the reduction of time and cost incurred with qualitative program evaluations. We close with the argument that such methodological advances in natural language processing allow education researchers to scale their current analytic approaches while simultaneously increasing their attention to the empirical nuance in the lived education experience.