Comment: Evaluating Model Performance in Fictitious Prediction Problems

Justin Grimmer *

December 31, 2012

I congratulate the author on an impressive and important contribution. Multinomial Inverse Regression (MIR) has wide applicability to a diverse array of problems. Further, the model is technologically impressive and the software is easy to apply to real problems. I hope that MIR will gain widespread use in the machine learning literature and across the social sciences.

This comment will focus on a basic question that arises when applying MIR: how do we know that it is performing well? The daunting challenge is that when using MIR to *discover* characteristics of words, the usual evaluation methods are inappropriate. The standard evaluation for methods like MIR is to assess its performance in out of sample prediction tasks. When this is the goal—as in supervised learning tasks—then this is the ideal evaluation. And as the author shows, MIR excels at this task, with superior classification accuracy for the set of restaurant reviews.

Text classification is certainly important, but perhaps the most promising application of MIR is to measuring how words convey sentiment. For example, the author uses MIR to identify differences in language between Democrats and Republicans (and differences in language based on constituency characteristics). Social scientists use other methods to make similar inferences. For example, Gentzkow and Shapiro (2010) use Congressional speeches to identify partisan words, then use those words to evaluate the partisan content of newspapers. Diermeier et al. (2011) use text to predict legislator ideology to better understand its content and Monroe, Colaresi and Quinn (2008) perform a series of evaluations, using text to predict differences in partisanship, gender, and geographic location.

These are *fictitious* prediction problems. They are fictitious because the goal is never to perform out of sample predictions. Gentzkow and Shapiro (2010), for example, had no interest in predicting whether future text was from a Democrat or Republican. Rather, Gentzkow and Shapiro (2010) want to identify words that convey a *Republican* and *Democratic* ideology. Likewise, Diermeier et al. (2011) interest in predicting the ideology of legislators is only

*Assistant Professor, Department of Political Science, Stanford University; Encina Hall West 616 Serra St., Stanford, CA, 94305

incidental. Rather, they use the prediction problem to identify the words and phrases that distinguish legislators' ideologies. To accomplish this task, the models merely entertain the *fiction* that prediction is a useful way to measure how words relate to sentiment. The real objective is more vague: to capture the *sentiment* each word conveys.

If accurately measuring sentiment is the goal, then evaluation is much more difficult to perform. Evaluation is harder because accurate out of sample prediction does not imply that a model is identifying words that capture particular sentiment categories. For example, Chang et al. (2009) shows that models that have accurate out of sample predictions receive low evaluations from human coders. Other model based measures of statistical fit—such as AIC, BIC, or DIC—tend to have a loose relationship with evaluations from human coders.

Rather than measures based on fit, the reliable use of methods like MIR requires the development of task specific evaluations. The development of the evaluations is one of the biggest open questions in the application of machine learning techniques to social scientific problems. What is immediately clear when developing the evaluations is that there needs to be clarity in the goals of the measurement. For example, Monroe, Colaresi and Quinn (2008) compare a set of different models, using their detailed knowledge about the Senate and political communication to offer impressionistic assessments of different facets of validity. Gentzkow and Shapiro (2010) present evidence that their word list is face valid. Certainly both validations are valuable, but formalizing these evaluations will allow clear comparisons across different methods for measuring how words relate to sentiment.

To perform the task specific evaluations authors will be forced to clarify *what* they are attempting to measure when they use methods like MIR. This is important, because what one measures when predicting the partisanship of a speaker can be a mixture of the sentiment of interest—say Republican or Democratic ideologies—and other features (Grimmer and Stewart, 2013). For example, when identifying words that distinguish Republican and Democratic representatives, the goal appears to be to measure words that are indicative of partisan conflict. But because Republicans and Democrats tend to come from different parts of the country, the party labels are conflated with language that is particular to the geographic region. The author recommends including covariates to mitigate this particular problem, but without clarifying the goal of measurement, it is hard to know what covariates should be included or excluded from the model. The obscured goals also complicate the evaluation of any covariate's value to the model.

Thanks to impressive articles like the author's, the application of machine learning techniques in the social sciences is growing. With the application of the methods there has also been an impressive number of new methods introduced. My hope is that social scientists will develop a parallel literature on the evaluation of the models. Together, this will lead to the accurate measurement of sentiment across large data sets and facilitate the discovery of new facts about the world.

References

Chang, Jonathan, Jordan Boyd-Graber, Chong Wang, Sean Gerrish and David M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *Neural Information Procesing Systems*.

Diermeier, Daniel, Jean-Francois Godbout, Bei Yu and Stefan Kaufmann. 2011. "Language and Ideology in Congress." *British Journal of Political Science* . Forthcoming.

Gentzkow, Matthew and Jesse M. Shapiro. 2010. "What Drives Media Slant?" *Econometrica* 78(1).

Grimmer, Justin and Brandon Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* . Forthcoming.

Monroe, Burt, Michael Colaresi and Kevin Quinn. 2008. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." *Political Analysis* 16(4):372.