Application of Sentiment and Topic Analysis to Teacher Evaluation Policy in the U.S.

Antonio Moretti*
Educator Learning & Effectiveness
Pearson

Kathy McKnight
Educator Learning & Effectiveness
Pearson

Ansaf Salleb-Aouissi
Center for Computational Learning Systems
Columbia University

ABSTRACT

We examine the potential value of Internet text to understand education policy related to teacher evaluation. We discuss the use of sentiment analysis and topic modeling using articles from the New York Times and Time Magazine, to explore media portrayal of these policies. Findings indicate that sentiment analysis and topic modeling are promising methods for analyzing Internet data in ways that can inform policy decision-making, but there are limitations to account for when interpreting patterns over time.

Keywords

Teacher evaluation, topic modeling, sentiment analysis

1. MOTIVATION

In the United States and abroad, teacher evaluation systems are increasingly becoming a common component of school reform efforts. Because teacher effectiveness is central to improving student learning, education policy in the U.S. has targeted teacher evaluation systems, with the rationale that evaluating teachers will lead to improved effectiveness. The result is an often contentious debate among researchers, educators and policy-makers about the utility of these systems in improving teacher effectiveness. Issues include which performance measures to use, how to collect and combine the data, and how it will be used with teachers.

A significant arena for debate about education policy, including teacher evaluations, occurs via the Internet. As the 2013 report "Social Media and Public Policy" notes [Leavy, 2013], use of data produced by Internet users may be useful in understanding policy issues and social problems, and perhaps ultimately, can provide insight to enable governments to develop more informed and better policy. The data may lead to better understanding of policy impact, and could potentially inform the different organizations that deliver public services, such as public education systems.

*Contact author. antonio.moretti@pearson.com

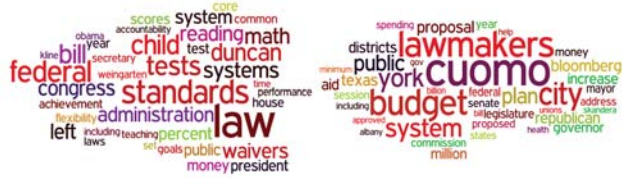
Given the potential value of Internet data to inform policy, our aim for this study is to conduct a preliminary analysis of publicly available Internet data from media outlets reporting on U.S. education policy, to evaluate what might be learned from such data that could inform policy-making regarding teacher evaluation. Therefore, we narrowed the focus to two popular media sources that cover national as well as local education policy – the NY Times, and Time Magazine–to analyze public sentiment and topics of concern regarding education policy focused on teacher evaluation. Given the increased emphasis on teacher evaluations over the past decade, we gathered data from 2004 - 2014. We used two approaches for analyzing data from the online media articles: a topic modeling approach [Blei, 2012] and sentiment analysis [Liu, 2010, pan,]. The research questions we addressed included:

1. What trends, if any, exist in public sentiment regarding teacher evaluation policy over the past decade?
2. What are the recurring topics most associated with media portrayal of teacher evaluation policies?

2. DATA COLLECTION AND ANALYSIS

We used the NY Times API and Time Magazine search query using "teacher evaluation" as the search term. Because there are no tools for collecting the full NY Times and Time Magazine articles,we scraped the websites after retrieving the relevant URLs. We retrieved a total of 348 articles on "teacher evaluation" from the NY Times during the period 2004 to 2014, and 292 articles from Time Magazine during the same period. We examined the articles for their relevance and removed those for which the focus was not primarily on teacher evaluation. The resulting dataset included 171 NY Times articles from 2009 to 2014, and 45 Time Magazine Articles from 2010 to 2014.

For the current study, we used the "topicmodels" package in R [Grün and Hornik, 2011]. We compared two variants of topic modeling: latent dirichlet allocation (LDA) and Correlated Topic Models (CTM). Both approaches are based on Blei [Blei et al., 2003, Blei and Lafferty, 2007]. To determine the number of topics to specify, we used the perplexity score.For our analyses, we specified a ten topic model, i.e. we set $k = 10$ to interpret results. In addition to the entropy measure, we used word clouds to display and make



(a) Topic 1 (b) Topic 2

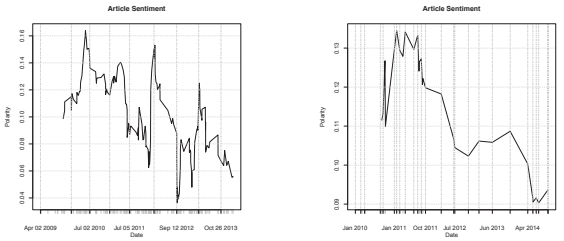
Figure 1: Word clouds for generated from NY Times articles.

sense of the topics generated from topic models. Figure illustrates word clouds for topic 1 and 2 generated from the New York Times articles. The topics that appeared to dominate the NY Times reporting included focus on federal requirements for teacher evaluation systems (e.g., reliance on student test data and relatedly, value-added models, for evaluating teachers); the impact of those requirements on teachers at both a federal and local (NYC) level, e.g., accountability, merit pay, lay offs and budgets; and the reaction of teacher unions to federal and local legislation (e.g., Chicago's teachers strike). In Times Magazine, where coverage of teacher evaluation policy was often combined with coverage of other federal education policies, the focus appeared to be on student achievement testing; changing education policies by the Obama Administration and in Washington DC, led by DC's former Chancellor of Education Michelle Rhee; and policy proposals during the 2012 presidential campaign. Teacher union reactions to teacher evaluation policy were also of focus, including the Chicago teachers strike.

We use the Natural Language Toolkit (NLTK) [Bird, 2006], a leading python platform to harvest textual data. The sentiment analysis tool in NLTK uses naive Bayes classifiers trained on both twitter sentiment as well as movie reviews. In Figure 2, a time series of the sentiment polarity of both the New York Times (left) and Times Magazine (right) articles is presented for 2009 - 2014. We used a simple moving average to plot the sentiment over time. In these graphs, we observe a similar trend in both the NY Times and Times Magazine articles. In both, we see somewhat similar peaks and troughs, as well as a similar trend of decreasing positive sentiment from 2010 to 2014.

3. DISCUSSION & FUTURE WORK

A number of federal and local (to NY) events took place over that period of time, that could be related to the sentiment trends. Nationally, the Obama Administration's Race To the Top (RTTT) legislation was initiated in July 2009, which among other policies, required states to develop and implement teacher evaluation systems that included student achievement as a "significant" component of a teacher's effectiveness rating. In 2010, RTTT was rolled out and the states awarded funding were announced. The state of New York was awarded 700M dollars in August, 2010. A result of this legislation was a contentious battle between lawmakers and the teachers union over the details of the evaluation system, among other policies. In September 2012 in Chicago, teachers took to the streets and went on strike against a range of education policies, including the teacher evaluation system that was to be put in place. NYC and the teachers union settled on an evaluation system in March, 2013.



(a) Topic 1 (b) Topic 2

Figure 2: Article Sentiment over Time

In the case study for this paper, issues regarding the use of student test scores for evaluating teachers; the response of teacher unions to federal and local teacher evaluation system requirements; and the budgets for implementing these systems were just some of the more prominent issues reflected in the results. Our ultimate goal is to advance the understanding of the impact of new policies on the well-being of public schools and teachers. While the methodology is promising, it needs to be harnessed through a useful visualization interface to facilitate the exploration and analysis of the topics produced to make it more useful to leverage in decision making. We acknowledge that there are limitations and potential problems with these approaches. A known challenge is choosing the granularity level of the topics that is related to the number of topics k provided as a parameter. A second challenge is in the interpretation and labeling of the derived topics that require a manual human intervention. In some cases, what is rated as positive or negative analytically might not reflect how human raters would code those words. Moreover, in our example, although sentiment appeared to decline in the negative direction, it still remained on the positive end of the polarity continuum.

4. REFERENCES

[pan,]

[Bird, 2006] Bird, S. (2006). Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics.

[Blei, 2012] Blei, D. M. (2012). Probabilistic topic models. *Commun. ACM*, 55(4):77–84.

[Blei and Lafferty, 2007] Blei, D. M. and Lafferty, J. D. (2007). A correlated topic model of science. *AAS*, 1(1):17–35.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

[Grün and Hornik, 2011] Grün, B. and Hornik, K. (2011). topicmodels: An r package for fitting topic models. *Journal of Statistical Software*, 40(13):1–30.

[Leavy, 2013] Leavy, J. (2013). Social media and public policy: What is the evidence? Technical report, Alliance for Useful Evidence.

[Liu, 2010] Liu, B. (2010). Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, Second Edition. Taylor and Francis Group, Boca*.