Combining Machine Learning and Qualitative Analysis to Study Public Discourse about Opting out of Testing

Authors: Amy Burkhardt, Terri S. Wilson, Wagma Mommandi and Michele Moses

Background

Through widespread "opt-out" efforts over the past several years, parent and student activists have pressured districts, states, and the federal government to reconsider the extent and limits of state-mandated assessments. One of the primary venues for debates about testing is social media. On Twitter, for instance, as opting out surged across the country in the Spring of 2015, supporters and opponents tweeted about the movement over 45,000 times. These tweets offer intriguing windows into the values and claims at play in discourses about testing. Understanding these patterns of resistance may help researchers and policymakers respond to public concerns about testing, rebuild trust in assessment practices, and ensure the quality of this data. Yet, the sheer volume of data poses challenges for studying these debates.

Objective

We demonstrate the affordances of integrating interpretive analysis into machine learning, showing how these integrated approaches might make the qualitative analysis of "big data" feasible at broad scale. We use statistical classifiers to perform a first pass at grouping large amounts of data into smaller, more selective samples that are then analyzed through qualitative methods.

Research Design

Our study design exploits the basic principles of both qualitative coding and supervised machine learning. One qualitative research technique is applying a coding scheme to a collection of data; each code consists of a word or small phrase and represents a particular characteristic of the data or a phenomenon of interest. Similarly, in supervised machine learning, labels are assigned to text in order to differentiate between two different topics or ideas; a label of 1 indicates the presence of a topic in the text belonging to the positive class, and a label of 0 indicates that the topic is not present and belongs to the negative class. These examples of text are then compiled to create a training set of data, which then trains a classifier to determine if a new piece of text is either positive or negative. We use qualitative codes to label training data; a 1 represents the presence of a code, and 0 is the absence of the code in a tweet. As such, we explore how building training samples, specific to particular qualitative codes, can detect and isolate various perspectives present in a large pool of tweets. We describe a feedback loop in which qualitative coding methods are used to refine the training samples, which in turn refines the coding scheme for a richer and more nuanced exploration of the public's reactions to education policies (Figure 1). Our exploration is guided by the following research questions: First, how does the refinement process of the training data impact model performance? Second, how do different statistical classifiers respond to this refinement process? Lastly, how does this refinement process influence the emerging qualitative analysis of the public discourse on opting out of standardized tests?

Data Collection and Analysis

As a key example of our methodological approach, we describe how we classified tweets in favor of testing. This was a minority position in debates about opting out. Our initial training sample consisted of 205 positive (*Not Critical*) tweets from March 2015, and 700 negative tweets from the same month. We reviewed over 5,000 tweets to identify these positive tweets; a time-intensive process which made it difficult to simply add more positive tweets to improve the model's performance. Instead, we refined our sample through qualitative analysis, identifying two sub-codes: *Pro-testing* tweets (117) and *Encouraging* tweets (88). We then subset the training data and re-trained our classifiers, using these refined samples; that is, instead of building one model per classifier, we built two models, one for each new training sample. Using the normalized counts of unigrams, bigrams, trigrams as features for our models, we compared results from two competing classifiers: naive Bayes (NB) and linear support vector machines (SVM). To evaluate the performance of each model, we computed metrics on three different validation samples. Accuracy, precision, recall and F1 scores were calculated on each of the k-fold training sessions; these metrics were also calculated on a small, held-out portion of the training sample. Additionally, we ran unlabeled data from April 2015 through the trained models (over 13,000 tweets), and then computed precision (i.e., the proportion of tweets that are truly positive). We draw on precision in this initial study as a preferred metric for evaluating the practicality of this novel approach.

Findings

Table 1 and Table 2 present the results detailing the impact of the three training sets (*Not Critical*, *Pro-testing*, and *Encouraging*) on three different validation samples. Figure 2 is a graphical presentation of the results from the classification of April tweets, allowing us to compare the absolute number of tweets classified by the two classifiers for these models. The SVM classifier outperformed the NB classifier. Furthermore, performance improved on the models trained on the refined, smaller samples of tweets: The average precision from the cross-validation samples of the k-fold training sessions with the original training sample is .80, compared to .89 for the model trained on pro-testing tweets, and .94 for the model trained on encouraging tweets. This pattern of improvement across all validation samples -- however, overall degradation in performance is also noted in the April 2015 sample, suggesting that the content of the tweets changes quickly over time.

Conclusion

This approach offers a promising way for education policy researchers to leverage large datasets and capture nuanced information on public discourse. Improving precision by reducing the training sample serves as a possibly counter-intuitive example that less data can actually improve a model's performance. Our experiments with the automatic classification of tweets informed our emerging understanding of the different discourses at play in the opt-out movement. Likewise, qualitative analysis helped to sharpen and refine our training samples, which—in turn—improved the performance of the classifiers, and ultimately, the robustness of the data.

Tables

Table 1.

Validation results of three models for Linear Support Vector Classifier

Validation Sample	Precision	Accuracy	F1-Score	Recall
3-fold Cross Validation (Average score)				
Not Critical	.80	.79	.56	.43
Pro-testing	.90	.84	.36	.23
Encouraging	.94	.89	.48	.33
Held-out Validation (10% of training set)				
Not Critical	.75	.78	.55	.43
Pro-testing	.80	.86	.50	.36
Encouraging	1.00	.91	.71	.55
Development Test Set (April Tweets)				
Not Critical	.28			
Pro-testing	.36			
Encouraging	.53			

Table 2.

Validation results of three models for Naive Bayes Classifier

Validation Sample	Precision	Accuracy	F1-Score	Recall
3-fold Cross Validation (Average score)				
Not Critical	.89	.81	.59	.45
Pro-testing	.79	.84	.44	.31
Encouraging	.89	.93	.71	.60
Held-out Validation (10% of training set)				
Not Critical	.80	.78	.52	.38
Pro-testing	.67	.85	.47	.36
Encouraging	1.00	.89	.62	.45
Development Test Set (April Tweets)				
Not Critical	.15			
Pro-testing	.09			
Encouraging	.25			

Figures

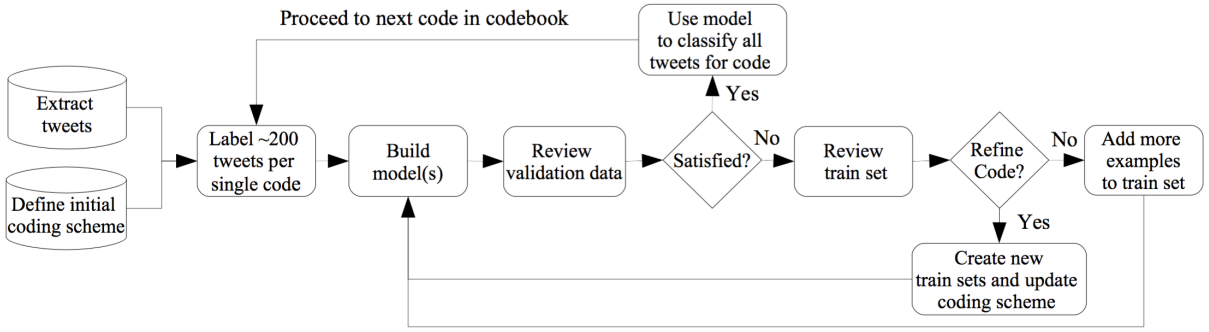

Figure 1 Feedback loop between qualitative research and machine learning

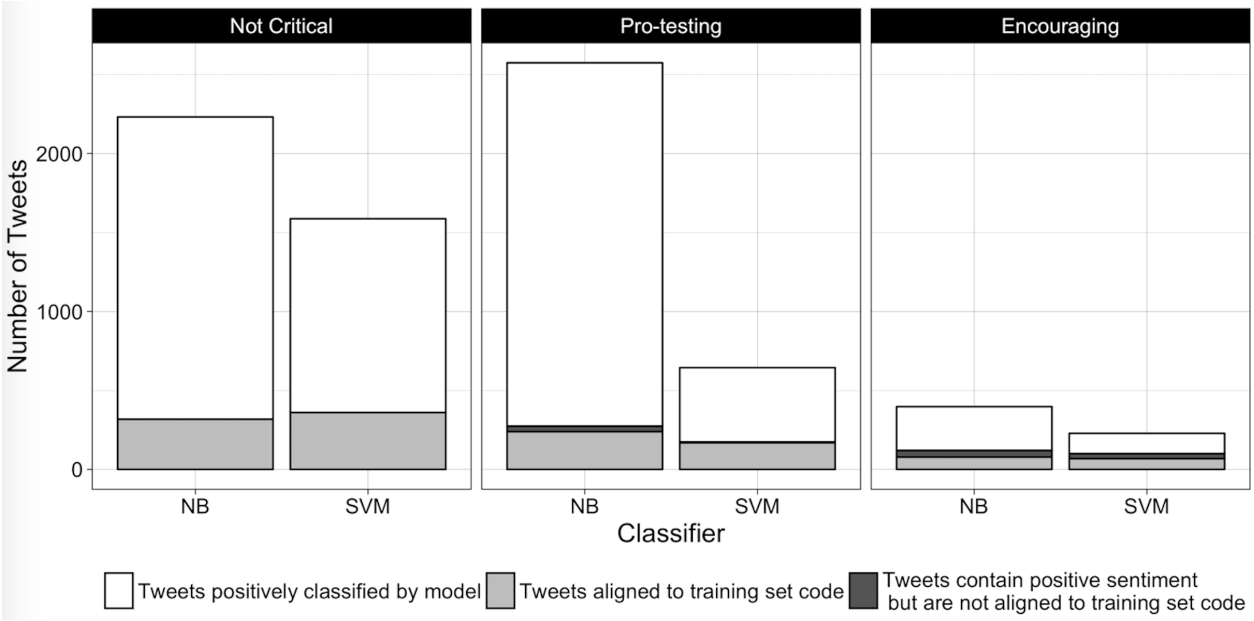

Figure 2 Counts of April Tweets Positively Classified by Three Models for Each Classifier