PA

# Selecting More Informative Training Sets with Fewer Observations

**Aaron R. Kaufman**

*Division of Social Science, New York University Abu Dhabi, UAE. E-mail: aaronkaufman@nyu.edu*

## Abstract

A standard text-as-data workflow in the social sciences involves identifying a set of documents to be labeled, selecting a random sample of them to label using research assistants, training a supervised learner to label the remaining documents, and validating that model's performance using standard accuracy metrics. The most resource-intensive component of this is the hand-labeling: carefully reading documents, training research assistants, and paying human coders to label documents in duplicate or more. We show that hand-coding an algorithmically selected rather than a simple-random sample can improve model performance above baseline by as much as 50%, or reduce hand-coding costs by up to two-thirds, in applications predicting (1) U.S. executive-order significance and (2) financial sentiment on social media. We accompany this manuscript with open-source software to implement these tools, which we hope can make supervised learning cheaper and more accessible to researchers.

*Keywords:* supervised learning, text-as-data

## 1 Pain Points in Supervised Learning for Political Science

As sources of text data proliferate in political science, researchers increasingly turn to advanced computational methods to test long-standing theories, develop new hypotheses, and generate novel measures of important concepts across every subfield of the discipline (Grimmer and Stewart 2013). Such tools fall into two categories: unsupervised methods like topic models, factor analysis, or clustering, and supervised methods like LASSO regression, random forests, support vector machines, and neural networks. While unsupervised methods are useful for exploratory and descriptive analysis, supervised methods offer unique promise in political science since they can encode researcher expertise in interpreting documents and extracting useful information from them (Grimmer, Roberts, and Stewart 2021). Supervised methods have quickly gained prominence: nearly 50 such papers have appeared in political science's top three journals since 2018 (Knox, Lucas, and Cho 2022), addressing such literatures as terrorism (Huff and Kertzer 2018), censorship (King, Pan, and Roberts 2013), ideology (Bonica 2018), media bias (Mozer *et al.* 2020), and normative theory (Blaydes, Grimmer, and McQueen 2018).

The standard approach in the social sciences to building a supervised text analysis model follows six key steps: (1) identify a corpus of documents, (2) train human coders to evaluate documents according to the quantity of interest, (3) present those trained coders with a subset of documents drawn at random from the corpus, (4) convert those documents into a low-dimensional representation to train a supervised learning algorithm,[1] (5) generate predictions for the remainder of documents in the corpus, and (6) validate those predictions, typically by examining face validity, calculating accuracy metrics like calibration and area under the receiver-operator characteristic (AUC). For example, a researcher might want to study the partisan leanings of newspaper stories in the United States (Gentzkow and Shapiro 2010; Mozer *et al.* 2020). She might collect a dataset of 10,000 newspaper articles, and then ask research assistants familiar

1 This step may include a number of cleaning, pre-processing, and rebalancing procedures, including stemming, removing stopwords, or correcting class imbalances.

***Political Analysis (2024)***
**vol. 32: 133–139**
**DOI:** 10.1017/pan.2023.19

**Published**
8 June 2023

**Corresponding author**
Aaron R. Kaufman

**Edited by**
Jeff Gill

with the tenor of American politics to read 1,000 articles from that corpus of 10,000 and rate each as either liberal, conservative, or nonpartisan. Then, providing a supervised learning algorithm with both the text of those articles and the ratings her research assistants produced, she would generate predictions for the remaining 9,000 articles, completing her dataset.[2]

The most resource-intensive steps in this workflow are (2) and (3): training human coders to extract relevant information from documents and then paying them to do so, often over the course of weeks or months. The researcher studying news article partisan bias might need to train her students how to identify subtle cues, dog whistles, or framing effects within stories, and even once trained, she might find that her students' ratings diverge, leading her to multiply code each of the 1,000 articles and noisily aggregate those ratings. Even so, she might find that her predictive model performs poorly if her 1,000 articles are too few, if her research assistants' codings disagree or otherwise contain too much measurement error, or for any number of other reasons.

Recognizing this widespread problem, a growing literature aims to improve this practice, reducing labeling costs and improving accuracy. One such approach substitutes fewer expert coders for many nonexperts in a crowdsourcing context (Benoit *et al.* 2016), improving step (2); a complementary approach selects training set documents algorithmically, either in an unsupervised topic modeling (Taddy 2013) or more-costly active-learning (Miller, Linder, and Mebane 2020) or semi-supervised (Gennaro and Ash 2022) framework, reducing effort at step (3).

We introduce and validate a new method for improving predictive accuracy and reducing the cost of supervised text-as-data models at step (3). Rather than an active-learning or semi-supervised approach, which require researcher time and expertise, our method is fully unsupervised and accompanied by easy-to-use software. We show that selecting the training set using an algorithmic approach rather than either a simple-random sample or a topic model-based method (Taddy 2013) leads to substantial gains in model accuracy,[3] holding constant the number of training examples. Considering a baseline accuracy for a binary classification problem of 50%, this workflow can almost increase the gains in accuracy above baseline by as much as 50%, or produce similar accuracy statistics with as few as one-third of the number of training examples (and therefore cost). This reduction in costs and increase in accuracy can help to improve how accessible machine learning methods are to social science researchers, mitigating structural inequalities in researchers' abilities to conduct empirical work (Sumner, Farris, and Key 2021).

## 2 Selecting Training Observations Intelligently

Considering again our study of news article partisan bias, imagine that our research assistant carefully reads the first two documents, assigning the first a label of "liberal" and the second a label of "conservative." A third document of the category "nonpartisan" adds more information to the training set than another document labeled "liberal" or "conservative". We argue that an intelligent sampling scheme for selecting documents to label must ensure that all "types" of observations are included. For example, one such algorithm divides the covariate space into equal-sized (or equally populated) subspaces and randomly samples from each; another might find the $k$ observations such that the sum of the distances between any two observations is maximized.

We divide this problem into two steps (Mozer *et al.* 2020): (1) representing the text as a numeric vector in reduced dimension and (2) selecting a metric to optimize in that lower-dimensional space. As a foundational approach to this problem, Taddy (2013) represents text with a topic model and selects observations satisfying a "space-filling $d$-optimality" criterion, finding

2 This is an example of a multi-class classification problem—documents may be liberal, conservative, or nonpartisan—though political scientists are often interested in both binary classification problems and regression problems as well.

3 While for tractability AUC is both our accuracy target and the standard target in machine learning, we acknowledge that other features like calibration (coverage) and homoskedasticity are important in political science. See Kaufman (2020) and the Supplementary Material.

encouraging results. But compared with newer neural network-based approaches, the topic model representation is information-poor (Grimmer *et al.* 2021); likewise, simpler representations might better capture variation across documents (Mozer *et al.* 2020). Considering eight low-dimensional representations and seven subset selection methods drawn from natural language processing (Rodriguez and Spirling 2022) plus several baseline methods, we systematically evaluate 57 approaches to select training observations. For a complete accounting of these methods and comparisons between them, see Appendix 1 of the Supplementary Material.

## 3 Empirical Strategy and Results

To validate the usefulness of this approach and to identify the optimal method, we conduct a simulation analysis of two supervised learning problems, one from American politics and one from social media, and show that for any given training set size, most of our approaches produce higher predictive accuracy than either the random sampling or Taddy's (2013) approaches.

Our first application examines presidential unilateral power. Presidents issue a large number of unilateral actions, but most are either ceremonial or inconsequential (Howell 2003). Chiou and Rothenberg (2014) build a measure of unilateral action significance using media mentions, assuming that more heavily reported unilateral actions are more important for policy and governance. Using those scores as training labels, Kaufman and Rogowski (forthcoming) collect the text of every unilateral action issued since 1789 and build a model to estimate unilateral action significance from an action's text. We consider a sample of 10,746 documents labeled as either significant (= 1.8%) or ceremonial (= 0.92%).

Our second example studies StockTwits, a Twitter-like social media platform focused on the stock market. Users optionally tag their tweets as "bullish" (= 1.11%) or "bearish" (= 0.89%). We explore a sample of 6,264 tweets and their binary labels. This application complements the first well: StockTwits documents are much shorter than presidential orders, contain complex tokens like emoji, and use less formal language. Together, these cover two of the most common categories of text analysis in political science.

### 3.1 Simulation Procedure

Our simulation design, detailed in Appendix 2 of the Supplementary Material, consists of four main steps: representation, train/test splitting, document selection, and modeling.

*Representation.* First, we produce low-dimensional representations of each corpus using each of our representation methods.

*Train/Test Splitting.* Second, we partition each corpus into ten 80–20 train/test splits, which we will use to produce uncertainty estimates around each method's accuracy.

*Document Selection.* Third, for each document selection method, each train/test split, and each corpus, we identify the $k$ optimal documents to include as our training set. We select 50, 100, 250, 500, 750, 1,000, 1,500, 2,000, 2,500, and 3,000 as different values of $k$.

*Modeling.* Finally, we train multinomial Naive Bayes models using each of our sets of selected documents, generate predictions for the remaining observations, and compute the predictive accuracy with AUC.[4]

4 We use this classifier for its computational simplicity as our complete set of simulations is extensive. In practice, many methods select similar documents, though any two rarely contain more than 15% overlap (see the Supplementary Material). For two uses of Naive Bayes in political science, see Nielsen (2017) and Perry and Benoit (2017).

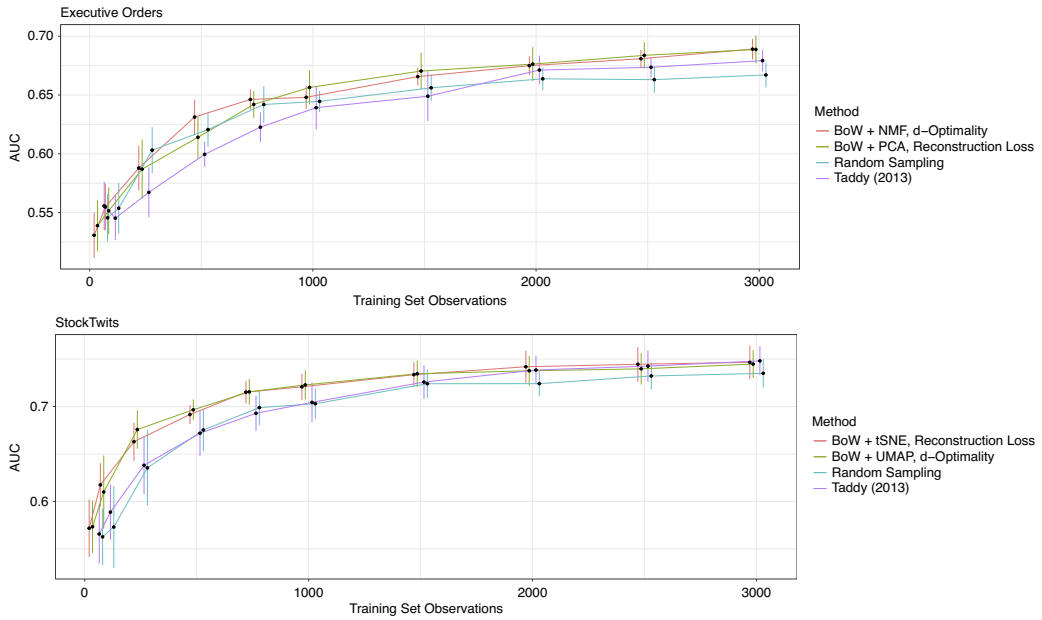


**Figure 1.** The top two methods substantially outperform both the random sampling and the Taddy (2013) approaches, achieving both accuracy gains and cost reductions. The improvement over random sampling is similarly strong for the Executive Orders than for the StockTwits application. Note: the *x*-axis locations of each point are jittered for readability.

## 3.2 Results

Many approaches to training set selection outperform the two baseline methods: random sampling, and that of Taddy (2013). Figure 1 presents the two top-performing methods and the two baselines for each application. In both plots, the *x*-axis indicates the number of training observations, and the *y*-axis indicates model accuracy as measured by AUC. Each point is the average AUC for that method across each of the 10 random train/test splits, with vertical bars indicating two standard errors. Tables 1 and 2 in the Supplementary Material present AUC and standard errors for the complete set of methods.[5]

In the first application, the best-performing method with 3,000 training observations achieves an AUC of 0.69, whereas the random sampling method achieves 0.66 and Taddy achieves 0.67; both are statistically distinct from 0.69. While this change in accuracy may seem small, a completely random classifier would have an AUC of 0.5, so an increase from 0.66 to 0.69 is a 16% improvement above baseline. Another interpretation of this result is that to achieve an AUC of more than 0.66, the random sampling approach requires approximately 3,000 training samples and the Taddy approach requires 2,000; the best-performing word-embedding model only requires fewer than 1,000 observations to achieve the same accuracy, a reduction of 66% and 50%, respectively. And, while trained undergraduates are sufficient for most applications in the social sciences, some applications require coders with specialized knowledge like judges (Kaufman, King, and Komisarchik 2017), legislators (Matsuo and Fukumoto 2021), or physicians (Gulshan *et al.* 2016), rendering these sample size reductions hugely valuable.

In the second application, Taddy performs relatively better with larger training set sizes, but again we observe that the random sampling method's accuracy with 3,000 observations is the same as the best-performing methods' accuracy with 1,000 observations. And, with fewer than 1,000 observations, both Taddy and random sampling are statistically worse-performing than the

5 For definitions and citations for these and the remaining methods, see Appendix 1 of the Supplementary Material.

best methods. However, the absolute gains from the best-performing methods are lower in this application, possibly because the documents are shorter.

Overall, we find that methods perform comparably well across applications, correlating at 0.88. Reconstruction loss is the best-performing distance metric and BERT is the best-performing representation, though in combination the bag-of-words representation yields the greatest accuracy improvements (see Table 3 in the Supplementary Material). Importantly, most of these methods are computationally quick to calculate, most often quicker than the Taddy method. We present full results for all 57 methods, including computing time comparisons, in Appendix 3 of the Supplementary Material.[6]

## 4 Discussion and Conclusion

These analyses have shown that an algorithmic approach to training set selection in text-as-data applications introduces substantial gains by leveraging information that researchers have at their disposal but have not utilized. To applied researchers performing their own supervised learning with text data, applying our methods offers both performance gains and cost reductions: researchers can fix their budget and expect an increase in accuracy up to 50%, or set the accuracy they prefer and save as much as two-third of the cost in achieving it.

We have also developed an approach for evaluating new proposed methods for the same purpose, and in open-sourcing our datasets and replication code, we encourage future researchers to improve upon our method as both algorithms and representations of text data improve. One fruitful path is to combine unsupervised and active learning approaches, using our method to select the initial documents to hand-code before proceeding with active learning as in Miller *et al.* (2020). Even the most computationally intensive method we study is feasible on a laptop without GPUs or accessing APIs, so adopting our method can save researcher resources.[7]

Our approach here may extend to other data domains and problems as well. This paper only considers binary classification problems: we speculate that multi-class classification problems and regression problems will benefit even more from our methods as they have more "types" of observations to capture. Second, text data applications often require little data compared with, for example, audio or video data, and our approach of testing different representations and subset selection methods may greatly reduce the amount of training data needed to classify protest images (Won, Steinert-Threlkeld, and Joo 2017) or measure emotions from audio (Knox and Lucas 2021).

Our analysis falls short, however, in making a singular recommendation. We acknowledge that many of the methods we test perform comparably well to the others, and there is no clear winner from our results. Furthermore, the improvement is most robust for our Executive Orders application, where documents are longer and language is more formalized. As the state-of-the-art of natural language processing (NLP) continues to develop, we speculate that word embedding approaches will improve, but in the meantime, more naive representations like bag-of-words reduced with t-distributed stochastic neighbor embedding (TSNE) or principal components analysis (PCA) are sufficiently better than the current practice to earn our recommendation.

The costs of performing high-quality empirical research in the social sciences is growing, and as publications become more expensive, those who suffer the most are researchers outside R1 institutions and those not on the tenure track, disproportionately women and minority scholars. Much methodological work in causal inference focuses on improving statistical precision, thereby reducing the amount of data necessary to achieve statistical significance. This paper shows

6 We acknowledge that the absolute magnitude of gains in accuracy may seem small. We discuss how these gains are meaningful nonetheless in Appendix 3 of the Supplementary Material.

7 Please see the Replication Materials for the hardware specifications we use in performing our simulations.

that there is much low-hanging fruit in making supervised learning research cheaper and more accessible as well, which we hope broadens adoption of text analysis methods in political science.

## Acknowledgments

The author thanks Matt Blackwell, Sobha Gadi, Andy Harris, Bob Kubinec, Brian Libgober, Luke Miratrix, Jon Rogowski, and Pierre Youssef for helpful comments and feedback during the preparation of this draft; Jeff Gill and three anonymous reviewers provided invaluable feedback as well. Apurva Bhargava and Andrei Kapustin provided excellent research assistance.

## Data Availability Statement

Replication code for this article is available in Kaufman (2023) at https://doi.org/10.7910/DVN/4ROL8S.

## Supplementary Material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2023.19.

## References

Benoit, K., D. Conway, B. E. Lauderdale, M. Laver, and S. Mikhaylov. 2016. "Crowd-Sourced Text Analysis: Reproducible and Agile Production of Political Data." *American Political Science Review* 110 (2): 278–295.

Blaydes, L., J. Grimmer, and A. McQueen. 2018. "Mirrors for Princes and Sultans: Advice on the Art of Governance in the Medieval Christian and Islamic Worlds." *The Journal of Politics* 80 (4): 1150–1167.

Bonica, A. 2018. "Inferring Roll-Call Scores from Campaign Contributions using Supervised Machine Learning." *American Journal of Political Science* 62 (4): 830–848.

Chiou, F.-Y., and L. S. Rothenberg. 2014. "The Elusive Search for Presidential Power." *American Journal of Political Science* 58 (3): 653–668.

Gennaro, G., and E. Ash. 2022. "Emotion and Reason in Political Language." *The Economic Journal* 132 (643): 1037–1059.

Gentzkow, M., and J. M. Shapiro. 2010. "What Drives Media Slant? Evidence from US Daily Newspapers." *Econometrica* 78 (1): 35–71.

Grimmer, J., M. E. Roberts, and B. M. Stewart. 2021. "Machine Learning for Social Science: An Agnostic Approach." *Annual Review of Political Science* 24: 395–419.

Grimmer, J., and B. M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–297.

Gulshan, V., et al. 2016. "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs." *JAMA* 316 (22): 2402–2410.

Howell, W. G. 2003. *Power without Persuasion: The Politics of Direct Presidential Action*. Princeton: Princeton University Press.

Huff, C., and J. D. Kertzer. 2018. "How the Public Defines Terrorism." *American Journal of Political Science* 62 (1): 55–71.

Kaufman, A. R. 2020. "Measuring the Content of Presidential Policy Making: Applying Text Analysis to Executive Branch Directives." *Presidential Studies Quarterly* 50 (1): 90–106.

Kaufman, A., G. King, and M. Komisarchik. 2017. "How to Measure Legislative District Compactness If You Only Know It when You See It." *American Journal of Political Science* 3: 55–58.

Kaufman, A. R., and J. C. Rogowski. forthcoming. " *Divided Government, Strategic Substitution, and Presidential Unilateralism*." *American Journal of Political Science*.

King, G., J. Pan, and M. E. Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107 (2): 326–343.

Knox, D., and C. Lucas. 2021. "A Dynamic Model of Speech for the Social Sciences." *American Political Science Review* 115 (2): 649–666.

Knox, D., C. Lucas, and W. K. T. Cho. 2022. "Testing Causal Theories with Learned Proxies." *Annual Review of Political Science* 25 (1), 419–441.

Matsuo, A., and K. Fukumoto. 2021. "Legislators' Sentiment Analysis Supervised by Legislators."

Miller, B., F. Linder, and W. R. Mebane. 2020. "Active Learning Approaches for Labeling Text: Review and Assessment of the Performance of Active Learning Approaches." *Political Analysis* 28 (4): 532–551.

Mozer, R., L. Miratrix, A. R. Kaufman, and L. J. Anastasopoulos. 2020. "Matching with Text Data: An Experimental Evaluation of Methods for Matching Documents and of Measuring Match Quality." Political Analysis, 28 (4): 445–468.

Nielsen, R. A. 2017. *Deadly Clerics: Blocked Ambition and the Paths to Jihad*. Cambridge: Cambridge University Press.

Perry, P. O., and K. Benoit. 2017. "Scaling Text with the Class Affinity Model." Preprint, arXiv:1710.08963.
Rodriguez, P., and A. Spirling. 2022. "Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research." *The Journal of Politics* 84 (1): 715162.
Sumner, J., E. Farris, and E. M. Key. 2021. "The Costs of Doing Research."
Taddy, M. 2013. "Measuring Political Sentiment on Twitter: Factor Optimal Design for Multinomial Inverse Regression." *Technometrics* 55 (4): 415–425.
Won, D., Z. C. Steinert-Threlkeld, and J. Joo. 2017. "Protest Activity Detection and Perceived Violence Estimation from Social Media Images." In *Proceedings of the 25th ACM International Conference on Multimedia*, edited by Q. Liu, et al., 786–794. New York: Association for Computing Machinery.