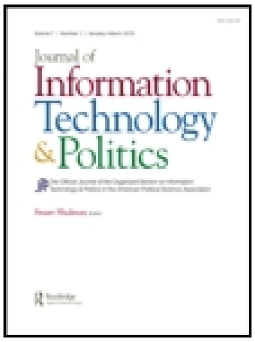


Journal of Information Technology & Politics

ISSN: 1933-1681 (Print) 1933-169X (Online) Journal homepage: https://www.tandfonline.com/loi/witp20

Tradeoffs in Accuracy and Efficiency in Supervised Learning Methods

Loren Collingwood & John Wilkerson

To cite this article: Loren Collingwood & John Wilkerson (2012) Tradeoffs in Accuracy and Efficiency in Supervised Learning Methods, Journal of Information Technology & Politics, 9:3, 298-318, DOI: 10.1080/19331681.2012.669191

To link to this article: https://doi.org/10.1080/19331681.2012.669191

Accepted author version posted online: 28 Feb 2012.
Published online: 05 Jul 2012.

Journal of Information Technology & Politics, 9:298–318, 2012
Copyright © Taylor & Francis Group, LLC
ISSN: 1933-1681 print/1933-169X online
DOI: 10.1080/19331681.2012.669191

Routledge
Taylor & Francis Group

Tradeoffs in Accuracy and Efficiency in Supervised Learning Methods

Loren Collingwood
John Wilkerson

ABSTRACT. Words are an increasingly important source of data for social science research. Automated classification methodologies hold the promise of substantially lowering the costs of analyzing large amounts of text. In this article, we consider a number of questions of interest to prospective users of supervised learning methods, which are used to automatically classify events based on a pre-existing classification system. Although information scientists devote considerable attention to assessing the performance of different supervised learning algorithms and feature representations, the questions asked are often less directly relevant to the more practical concerns of social scientists. The first question prospective social science users are likely to ask is, How well do such methods work? The second is, How much human labeling effort is required? The third is, How do we assess whether virgin cases have been automatically classified with sufficient accuracy? We address these questions in the context of a particular dataset—the Congressional Bills Project—which includes more than 400,000 bill titles that humans have classified into 20 policy topics. This corpus offers an unusual opportunity to assess the performance of different algorithms, the impact of sample size, and the benefits of ensemble learning as a means for estimating classification accuracy.

KEYWORDS. Machine learning, supervised learning, text classification

These days, it seems as though classification algorithms are being applied to almost everything. Some of the most sophisticated and visible are Internet search algorithms that are constantly updated based on user queries and clicks. Classification algorithms are also used to identify geographical features, potential health problems, people, and of course text. With the growth of the Internet and the wealth of new data possibilities, interest in automated text analysis techniques is growing within political science (Cardie & Wilkerson, 2008; Hillard, Purpura, &

Loren Collingwood is a Ph.D. candidate in the Department of Political Science at the University of Washington. His interests include American politics, political behavior, survey methodology, programming in R, and text classification.

John Wilkerson is an associate professor in the Department of Political Science at the University of Washington. His interests include legislative organization, decision-making, and computational methods in political science.

This research was partially funded by the Center for American Politics and Public Policy and the Center for Statistics and the Social Sciences, both at the University of Washington. Earlier versions of this draft were presented at the Southern Political Science Conference, Text as Data Conference, and presentations at Texas A & M University and Rice University. We thank Josh Clinton, Amber Boydstun, Francisco Pedraza, Chris Adolph, Steve Purpura, and three anonymous reviewers for their helpful comments.

Address correspondence to: Loren Collingwood, Department of Political Science, University of Washington, Box 353530, 101 Gowen Hall, Seattle, WA 98195-3530 (E-mail: lorenc2@uw.edu).

Wilkerson, 2008; Hopkins & King, 2010; King & Lowe, 2003; Laver, Benoit, & Garry, 2003; Lazer et al. 2009; Monroe & Schrodt, 2009). Researchers have classified newspaper articles or Internet stories to measure sentiment toward political candidates, and have studied mentions in blog posts and tweets to track public opinion or even happiness (Dodds & Danforth, 2009; O'Connor, Balasubramanyan, Routledge, & Smith, 2010). The possibilities are almost endless.

Many different approaches to automated classification exist, and no single approach is superior to all others. Rather, different approaches have unique advantages and disadvantages. To set the stage for our own work, we will briefly compare dictionary, unsupervised learning, and supervised learning approaches to classifying text into different categories or classes. Suppose that one were interested in categorizing legislation into a limited set of topics. Dictionary- or keyword-based approaches take an axiomatic approach to this classification task. The researcher designates that the presence of a specific keyword or combination of keywords implies that an event belongs to a particular class (Schrodt, Davis, & Weddle, 1994). Thus, there is never any doubt about whether the machine has correctly classified an event. However, the potential drawback of a dictionary approach is that it must include a mapping for every relevant permutation of the data. Depending on the complexity of the data and the classification scheme, constructing this map can be expensive and time consuming.

In contrast to dictionary approaches, machine-learning approaches do not begin with predefined rules. *Unsupervised* learning methods search for hidden structure in unlabeled data. They can therefore be used as a discovery tool (Blei, Ng, & Jordan, 2003; Grimmer & King, 2009). A second, not insignificant advantage is that unsupervised learning methods enable a researcher to categorize a dataset at relatively low cost because they do not require dictionaries or pre-labeled examples. However, a potentially important limitation is that the resulting classes or categories are empirically rather than theoretically derived.

Supervised learning methods begin with an existing classification system. In the initial training phase, the objective is to build a model using pre-labeled examples. In the experiments reported here, the examples are bill titles that humans have assigned to 20 different policy topics (e.g., environment, defense, or health—20 major topics in all). One or more algorithms are used to predict these preassigned classes (topics) using a subset of the data. In essence, the algorithm is a multivariate model where the dependent variable is the class, and the independent variables tested are features (words or characters) contained in the titles. In the next testing phase, the algorithm's success in predicting the topics of a different set of pre-labeled examples is assessed. A particular concern is overfitting. Over-fitting results when the model hews too closely to the training data and as a result does not do a good job predicting new cases (Dietterich, 1995). Out-of-sample testing produces a better test of prediction accuracy. During this training/testing phase, the researcher will experiment with different algorithms, different sampling approaches, and different approaches to specifying the relevant features included in the model. When out-of-sample performance is deemed to be acceptable, in the final classifying stage, the model is used to label virgin cases for topic (i.e., cases that humans have not already labeled).

The performance of supervised learning models is evaluated using the existing labels contained in the training set. Thus the "gold standard" is usually (though not always) a label assigned by humans. It is therefore a less rigorous validation standard than the keywords of a dictionary-based approach. Unsupervised learning approaches require more creative and subjective approaches to validation because they do not begin with pre-existing rules or examples (Quinn, Monroe, Colaresi, Crespin, & Radev, 2010).

RESEARCH OBJECTIVES

The primary objective of this article is to investigate supervised learning methods as a research tool for political scientists. We address several questions related to practical decisions involved in the use of such methods. First, many off-the-shelf algorithms are available. Does the choice of algorithm matter? Second, supervised learning methods required pre-labeled examples. How many pre-labeled examples are required to yield acceptable performance? How much difference does sample size make? Third, do different sampling strategies yield different results, and if so, which is preferred? Finally, once the train/test phase has been completed, how does a user distinguish virgin cases that have been labeled with high accuracy from those that have not?

None of these questions can be answered in definitive fashion because many variables affect performance. Of particular note is overall accuracy. Many factors affect accuracy besides algorithm choice or sample size, including the structure of the data and the classification scheme. If the cases in each category share a unique keyword (feature), then any algorithm is going to perform very well. Or to take another example, if the dataset includes a large number of duplicate cases (e.g., identical bills), then we might see substantially better performance than in the case for a dataset that lacks duplicate cases. Nevertheless, we believe that results presented here are fairly representative of a relatively complex classification task, and we explicitly control for the potential inflationary effects of duplicate records.

In the pages that follow, we first discuss the corpus used in our experiments—the Congressional Bills Project. We then briefly introduce the off-the-shelf machine learning algorithms examined and preprocessing of the data. The remainder of the article then reports on a series of experiments designed to address the questions posed above. In Part I one of the analysis, we assess prediction accuracy for different algorithms and sample sizes while controlling for two potentially confounding effects: we eliminate duplicates cases and ensure the same number of training examples for each of the 20 topics or classes. In Part II of the analysis, we relax these two sampling constraints to allow for a more realistic experiment that includes duplicate cases and is based on a random sample of training examples. Do these differences significantly impact our findings, and if so, how? Finally, Part III addresses the question of how to separate the virgin cases that have been classified with sufficient accuracy in the event that a researcher desires a higher level of overall accuracy than any particular algorithm is able to provide.

CONGRESSIONAL BILLS CORPUS

The Congressional Bills Project (http://www.congressionalbills.org)[1] includes approximately 400,000 public and private bill titles (1947–present).[2] Each bill title in the corpus has been labeled as primarily about one of 20 major topics (19 substantive topics plus "private bills"), and 225 subtopics using the system originally developed for the Policy Agendas Project (http://www.policyagendas.org). For example, a bill "To amend the Clean Air Act of 1970" would be classified as primarily about major topic 7 (Environment) and subtopic 705 (Air Pollution), while a bill "To increase the minimum wage" would be classified as primarily about major topic 5 (Labor and Immigration) and subtopic 505 (Fair Labor Standards).

Bills are classified primarily by undergraduates as part of a year-long competitive research capstone seminar. Because of the wide variety of bills introduced and the emphasis on primary topic, annotators make subjective decisions based on general coding guidelines rather than the presence or absence of specific keywords. During the first academic quarter, students learn the system by classifying about 100 bills each week and then meeting to compare their results to with the master annotator's (a graduate student or faculty member intimately involved with the project). The group then discusses the cases where inter-annotator reliability is low to arrive at a shared understanding of the correct decision in the opinion of the master annotator. Importantly, discrepancies between the master annotator and student annotators are rarely blatant errors. A bill ending "Don't Ask, Don't Tell" addresses a defense personnel issue (subtopic 1618) but it also addresses a civil rights issue (subtopic 207). When the goal is to partition an entire legislative agenda into just

19 (or 224) categories, even experts will legitimately disagree about what a bill is primarily about.

This process is repeated until average inter-annotator reliability (between the master annotator and individual student annotators) approaches 90 percent major topic and 80 percent subtopic. In the second and third quarters, they are then given independent coding assignments of about 200 bills per week, while the master annotator continues to conduct spot checks to ensure quality results. This system has worked well, but with 10,000 bills introduced every Congress, it is obviously labor intensive. We began experimenting with supervised learning methods several years ago and now rely on them to classify a large proportion of bills at similarly high levels of reliability.

ALGORITHMS AND FEATURE REPRESENTATION

The results reported here rely on the Rtexttools package, an R wrapper for a C program that includes basic preprocessing functionality, four machine learning algorithms, and analytics (Collingwood, 2010). Further work was done via the RTextTools package (Jurka, Collingwood, Boydstun, Grossman, & van Atteveldt, 2011)—a cross-platform extension and more user-friendly version of the original package. Different machine learning algorithms are optimized for different classification domains. For instance, some algorithms perform better at classifying medical records, whereas others may be optimized to classify sentiment. Part of the goal of the present research is to examine how much difference the choice of algorithm makes when congressional bill titles are concerned.

Rtexttools provides four off-the-shelf machine-learning algorithms and basic preprocessing capabilities. These algorithms treat the features of a given title as a nonsequential "bag of words." That is, the ordering of words or characters within the title is not taken into consideration. The four algorithms are considered: Support vector machine (SVM), maximum entropy, naive bayes, and an n-gram tokenizer from the Ling Pipe package for text analysis (Carpenter & Baldwin, 2011). We do not go into detail about these algorithms here, as documentation is plentiful. For instance, support vector machines are described in depth by Boser, Guyon, and Vapnik (1992) and Cortes and Vapnik (1995), and are practically addressed by Hsu, Chang, and Lin (2003). Maximum entropy is developed by Berger, Pietra, and Pietra (1996), with examples from Ratnaparkhi (1997). Naive bayes is covered by Lewis (1998). And readers can peruse chapters six and seven in Carpenter and Baldwin (2011) to review the implementation of character and token n-gram classifiers in the Ling Pipe toolkit.

Each of these algorithms uses similarities and differences among examples (bill titles) preassigned to different classes (topics) to predict the classes of new cases. The more distinguishing the information contained in the examples, the better the performance of the algorithms. For this reason, it is standard practice to remove features that are unlikely to be distinguishing, such as changing all letters to lower case and removing suffixes. We use the Porter Stemmer because it supports alternative forms of words and is known to work well in a variety of situations (Loper & Bird, 2002). In addition, it is also standard practice to remove common "stopwords" such as "the" and "also." The rationale for doing this is that such words have little lexical content, and their inclusion may introduce noise that reduces prediction accuracy (Loper & Bird, 2002). Stopword lists for all languages are commonly available.

Although we do not do so here, a researcher may experiment with assigning additional weight to particular features (e.g., specific keywords, combinations of words, or even external information such as the name of a bill's sponsor) to improve performance. The presence or frequency of particular stopwords in a document may also be informative. The main point is that the researcher has control over which elements or features of a document are deemed to be relevant to the task. Improving feature representation in this way can improve the overall performance of an algorithm by several percentage points (results will vary of course).

ANALYSIS PART I

For the analyses presented in this part, we have removed all duplicate titles from the dataset and have specified an identical number of training examples for each of the 20 topics. Later, we will relax both of these constraints. As discussed, supervised machine learning entails a three-step process. The first two steps entail an iterative process of training an algorithm using prelabeled examples (bill titles coded for topic), testing the performance of the algorithm using a set-aside testing set of prelabeled examples, and tweaking the process to hopefully improve performance (e.g., increasing the sample size or by weighting particular features). The third step is to then use the trained algorithm to classify cases that have not yet been labeled. In this article, we experiment with sample sizes ranging from n = 100 to n = 1000 per topic (yielding total training and test sets from n = 2000 to n = 20,000, respectively).

To address the possibility that the results of any one experiment are dependent on the training sample selected at random from the much larger corpus, we conduct a pseudo-bootstrap where the reported accuracy is based on 1000 experiments using random samples of different sample sizes (e.g., 1000 experiments using a randomly drawn sample of 100 training examples per topic, etc.). The mean of these samples is taken as the point estimate (i.e., average accuracy), and the standard error of this estimate is two times the standard deviation from the mean.

Assessing Predictive Accuracy: Precision, Recall, and the F-Score

Two types of prediction accuracy are relevant to classification tasks. *Precision* refers to how often a case that the algorithm predicts as belonging to a class actually belongs to that class. For example, precision tells us what proportion of bills an algorithm deems to be about defense are actually about defense (based on the gold standard of human-assigned labels). In contrast, *recall* refers to the proportion of bills in a class that the algorithm correctly assigns to that class. In other words, what percentage of actual defense bills did the algorithm correctly classify? Another way of thinking of the difference is that precision seeks to minimize Type I or false positive errors, while recall seeks to minimize Type II or false negative errors. A researcher may be solely concerned about precision or recall. However, the F-Score offers a performance measure that is a weighted average of both precision and recall, where the highest level of performance is equal to 1 and the lowest is equal to 0 (Sokolova, Japkowicz, & Szpakowicz, 2006). Specifically:

$$F - Score = \frac{(\beta^z + 1) * precision * recall}{\beta^z * precision + recall}$$

where the F-Score is evenly balanced between precision and recall (when $\beta = 1$).[3]

Average Algorithm Precision

Due to space considerations, we focus primarily on precision for much of our presentation, as we think that it is the most intuitive measure of accuracy. Table 1 indicates that using 100 examples per topic (for a total sample size of 2000) produces overall precision accuracy (i.e., agreement between the machine and human predictions) of between 54 and 68 percent. For this classification task and sample size, the Ling Pipe n-gram tokenizer performs considerably worse than the other algorithms. Accuracy improves with larger sample sizes, as expected. For a sample of 1000 per category, overall accuracy across the algorithms ranges between 68 and 74 percent. Figure 1 illustrates the improvement associated with increasing the sample size for each algorithm.

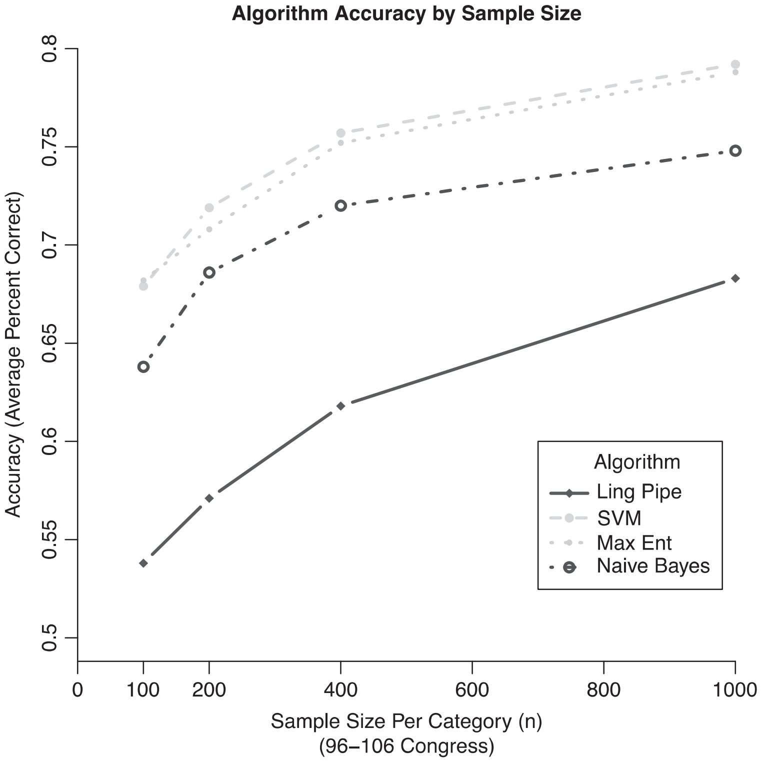
Precision by Topic

Another perspective is to examine the accuracy at the level of the topic (once again by considering mean prediction accuracy across 1000 experiments). Two questions are of interest here. The first is whether there are noteworthy accuracy differences across topics. The second is whether sample size improves accuracy within specific topics and by how much (recall that we have eliminated duplicate bills for this experiment). Table 2 indicates that there

TABLE 1. Algorithm Performance (Precision) Estimates Based on 1000 Experiments

Topic	n = 100	SD 100	n = 200	SD 200	n = 400	SD 400	n = 1000	SD 1000	1000–100
Ling Pipe	0.54	–0.02	0.58	–0.02	0.62	–0.01	0.68	–0.01	0.14
Maximum entropy	0.67	–0.03	0.71	–0.01	0.75	–0.01	0.79	–0.02	0.12
SVM	0.68	–0.03	0.72	–0.01	0.75	–0.01	0.79	–0.01	0.11
Naive bayes	0.64	–0.03	0.68	–0.02	0.71	–0.01	0.74	–0.01	0.1

FIGURE 1. Algorithm performance and training sample size.



are substantial differences in accuracy that can generally be attributed to the challenges associated with different topics. Nearly all private bills (topic 99) include the text "for the relief of," while a large proportion of foreign trade bills (topic 18) propose to "suspend duties." As a result, accuracy is high for these bills. In contrast, the Banking and Finance topics tend to be more diverse. A small training sample does not perform as well in predicting these bills in the test set. However, the marginal benefits of additional examples are also greater. These findings underscore the point that prediction success is dependent on the extent to which the information in the training set is representative of the information in the test set. Sometimes only a small sample is needed; at other times a much larger sample will be required.

Figure 2 displays precision at the major topic level for each of the algorithms for different sample sizes. As noted earlier, the performance of the different algorithms varies. Support vector machine and maximum entropy perform better than Ling Pipe and naive bayes for this particular project. While this would seem to suggest that there is little benefit to using ling (especially) to

TABLE 2. Support Vector Machine Precision Accuracy (SVM Shows Improvement by Sample Size for a Variety of Categories.)

Topic	n = 100	SD 100	n = 200	SD 200	n = 400	SD 400	n = 1000	SD 1000	1000–100
Banking	0.51	–0.14	0.54	–0.08	0.61	–0.04	0.71	–0.04	0.2
Int'l Affairs	0.58	–0.08	0.62	–0.06	0.69	–0.04	0.76	–0.04	0.18
Law/Crime	0.55	–0.12	0.6	–0.06	0.66	–0.06	0.73	–0.02	0.18
Economics	0.45	–0.06	0.48	–0.06	0.52	–0.04	0.6	–0.02	0.16
Gov't Ops	0.51	–0.12	0.56	–0.1	0.59	–0.06	0.64	–0.02	0.13
Transportation	0.7	–0.1	0.76	–0.04	0.79	–0.04	0.82	–0.02	0.12
Defense	0.67	–0.1	0.72	–0.06	0.75	–0.04	0.79	–0.02	0.12
Civil Rights	0.67	–0.14	0.69	–0.08	0.73	–0.06	0.79	–0.02	0.12
Environment	0.72	–0.12	0.75	–0.08	0.78	–0.04	0.82	–0.02	0.1
Labor	0.7	–0.1	0.74	–0.06	0.76	–0.04	0.79	–0.02	0.09
Agriculture	0.77	–0.1	0.81	–0.06	0.83	–0.04	0.84	–0.02	0.07
Science	0.79	–0.08	0.81	–0.06	0.83	–0.04	0.86	–0.02	0.07
Public Lands	0.74	–0.1	0.73	–0.06	0.77	–0.04	0.8	–0.02	0.07
Education	0.79	–0.08	0.8	–0.06	0.81	–0.04	0.83	–0.02	0.04
Health	0.77	–0.1	0.79	–0.06	0.8	–0.04	0.81	–0.02	0.04
Social Welfare	0.78	–0.1	0.78	–0.04	0.8	–0.04	0.82	–0.02	0.03
Energy	0.84	–0.08	0.86	–0.04	0.86	–0.02	0.87	–0.02	0.03
Housing	0.8	–0.1	0.79	–0.06	0.81	–0.04	0.82	–0.02	0.02
Private Bills	0.92	–0.04	0.93	–0.04	0.94	–0.02	0.95	–0.02	0.02
Foreign Trade	0.85	–0.06	0.87	–0.04	0.86	–0.04	0.87	–0.02	0.01

classify bills, we will show later that using all four algorithms has significant practical benefits.

F-Scores by Topic

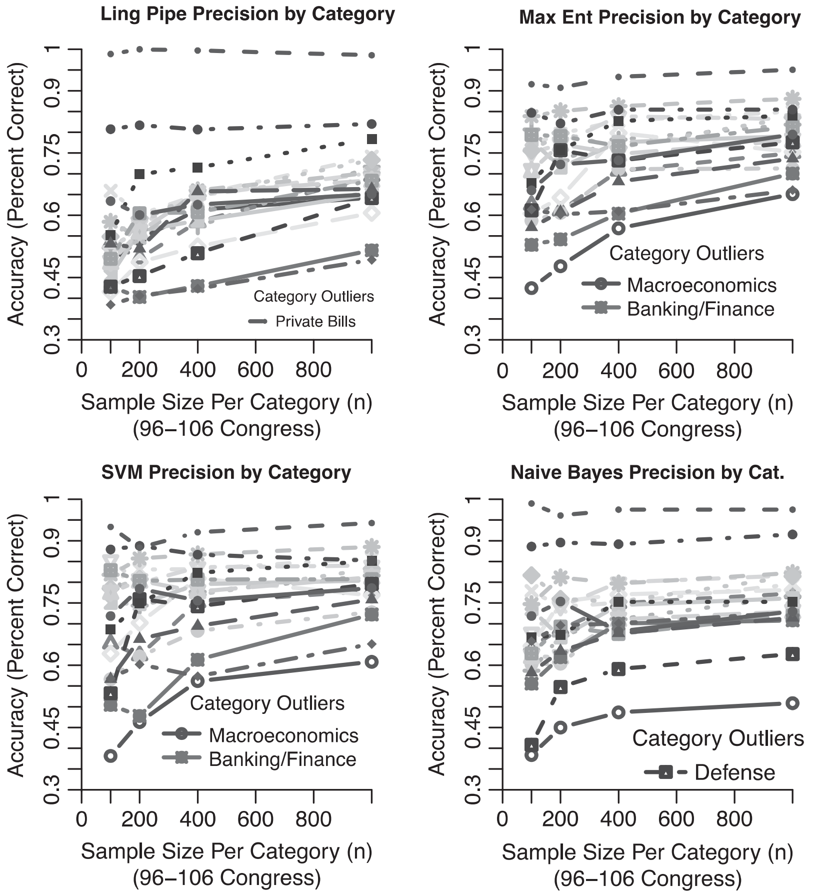
As discussed, an algorithm may perform well in terms of precision (percent of true labels correctly predicted by the algorithm) and less well in terms of recall (percent of predicted labels that correspond to the true labels) or vice versa. Recall is important when the researcher's goal is to avoid missing cases that belong in a class. Precision is important when the goal is to estimate the proportion of cases that belongs in a class. If an algorithm predicted that every case belonged to one out of 20 classes when only 10 percent of the cases truly belonged, then recall for that class would be perfect while precision for that class would be poor.

The F-Score is a weighted average of precision and recall. Table 3 presents the F-scores for topics for the SVM algorithm.[4] Relative performance taking both precision and recall into account only slightly alters the conclusions one would draw from the results for precision alone in Table 2. The same topics are near the top and bottom both in terms of overall accuracy and in terms of improvements due to increasing sample size.

Confusion Matrices

The confusion matrix is an important diagnostic tool that provides an opportunity to simultaneously investigate precision and recall across classes (topics) (Olson & Delen, 2008). When the machine learner makes an error in terms of assigning a bill to a class, do the errors tend to be randomly distributed across the other classes or concentrated in a particular class? In Table 4, the columns indicate the topic the algorithm assigned to bills, while the rows indicate the topic the human annotators assigned to bills. Looking at the column results, SVM classified a total of 1289 bills as about topic 100 (Economics) when the actual number of Economics bills in the test set was 1000 (for a precision percentage of 61 percent). In contrast, SVM classified a total of 966 bills as about Energy (for a much better precision of 89 percent). Looking at the row percentages, the machine learner correctly recalled 784 of 1000 Economics bills (a recall rate of

FIGURE 2. Overall, when the sample size reaches n = 1000 per category, category precision tends to converge between 70–75%. The “Private Bills” category has very high precision, regardless of sample size.



78 percent) and 855 of 1000 Energy bills (a recall rate of 86 percent).

Examining the eighth row in Table 4, we see that the prediction (recall) errors where Energy bills are concerned are broadly distributed across topics. However, for Banking bills, the errors are noticeably biased toward classifying bills about Banking as bills about Economics (13th row). The discovery of this information in the confusion matrix might lead a researcher who was interested in improving performance to take a closer look at these errors. Perhaps the human annotators are systematically assigning bills that are actually about Economics to Banking, or perhaps the supervised machine learning algorithm is keying in on particular nonrelevant features that true Banking bills share with bills about Economics. This process of drawing on the confusion matrix to diagnose prediction errors and adjust the training process in response is a form of active learning (Sammut & Webb, 2011).

ANALYSIS PART II

Earlier we noted that we had removed duplicate bills (representing approximately 34 percent of the cases in the bills corpus) to avoid the potential criticism that their presence would inflate perceived performance. In addition, we stratified our training sample so that each topic had the same number of examples. In this section, we investigate the implications of

TABLE 3. Support Vector Machine F-Scores

Topic	n = 100	SD 100	n = 200	SD 200	n = 400	SD 400	n = 1000	SD 1000	1000–100
Banking	0.489	–0.091	0.549	–0.064	0.605	–0.044	0.671	–0.026	0.182
Law/Crime	0.579	–0.085	0.632	–0.057	0.683	–0.036	0.74	–0.022	0.161
Civil Rights	0.604	–0.087	0.649	–0.058	0.696	–0.039	0.754	–0.022	0.15
Int'l Affairs	0.611	–0.081	0.662	–0.054	0.706	–0.037	0.758	–0.02	0.147
Economics	0.543	–0.07	0.583	–0.047	0.626	–0.035	0.682	–0.021	0.139
Gov't Ops	0.514	–0.086	0.564	–0.061	0.604	–0.041	0.648	–0.026	0.134
Transportation	0.685	–0.08	0.735	–0.053	0.774	–0.032	0.813	–0.02	0.128
Defense	0.656	–0.086	0.703	–0.055	0.743	–0.036	0.78	–0.02	0.124
Environment	0.674	–0.08	0.719	–0.052	0.756	–0.036	0.796	–0.02	0.122
Labor	0.659	–0.084	0.701	–0.052	0.734	–0.035	0.768	–0.02	0.109
Science	0.742	–0.074	0.774	–0.048	0.807	–0.032	0.845	–0.017	0.103
Agriculture	0.732	–0.076	0.766	–0.049	0.796	–0.032	0.831	–0.018	0.099
Energy	0.771	–0.073	0.808	–0.046	0.838	–0.03	0.867	–0.017	0.096
Education	0.74	–0.076	0.775	–0.048	0.805	–0.03	0.836	–0.018	0.096
Public Lands	0.706	–0.076	0.74	–0.05	0.769	–0.032	0.801	–0.021	0.095
Health	0.719	–0.075	0.748	–0.05	0.776	–0.032	0.808	–0.019	0.089
Housing	0.739	–0.075	0.771	–0.047	0.796	–0.031	0.824	–0.018	0.085
Social Welfare	0.736	–0.072	0.758	–0.048	0.779	–0.032	0.803	–0.02	0.067
Foreign Trade	0.803	–0.066	0.828	–0.042	0.844	–0.03	0.86	–0.017	0.057
Private Bills	0.951	–0.031	0.953	–0.022	0.958	–0.015	0.966	–0.011	0.015

relaxing both of these constraints. Does allowing for the presence of duplicate bills dramatically improve performance? Does a random sampling approach lead to worse or better algorithm performance?

Implications of Duplicate Texts for Accuracy

Figure 3 indicates that the performance of the algorithms is not dramatically affected by the large number of duplicate bills in the corpus (based on experiments using a sample of 8000 total bills or 400 examples per topic). Including duplicate bills produces marginally higher accuracy. SVM obtains an accuracy rate of 77.41 percent compared to 75.7 for the deduplicated corpus; ling pipe, 66.18 percent versus 61.8 percent; maximum entropy, 76.84 percent versus 75.20 percent; and naive bayes 73.35 percent versus 72.00 percent. While it is perhaps surprising to see such small differences given the large number of duplicate bills involved, this is a welcome finding from an algorithm-performance perspective.

However, examining the impact of allowing for duplicates at the topic level reveals that the addition of duplicates does not uniformly improve accuracy. As shown in Table 5 (an analysis of SVM accuracy), including duplicates improves prediction accuracy for most topics (e.g., precision for Law and Crime, Science, and Transportation are four to five points higher) but leads to lower precision for others (e.g., precision for the Labor topic is 8.5 points lower when duplicate bills are included). This is an unexpected finding that deserves additional investigation in the future.

Stratified Versus Random Sampling Methods

The pre-existing bills corpus made it possible for us to construct a stratified random sample that included an equal number of training examples per topic. In a more typical project, a researcher would code a simple random sample of cases, which would produce a training sample that was roughly proportional to the distribution of topics in the underlying data. Some topics would have fewer training examples, while others would have substantially more. Does a random sampling approach affect performance, and if so, how? Once again, we utilize n = 8000 training and test samples. Perhaps surprisingly, Figure 4 indicates that the simple

TABLE 4. Support Vector Machine Confusion Matrix

Hand code	Econ	CR	Health	Ag	Labor	Educ	Env	Energy	Tran	LC	SW	Hous	Bank	Def	Sci	FT	Intl	Govt	Lands	PB	n	Pct Right
Economics	784	7	5	6	30	2	2	6	2	6	16	28	36	4	7	14	6	35	3	1	1000	78
Civil Rights	14	730	13	5	18	22	1	2	11	33	20	10	19	11	21	2	18	41	9	0	1000	73
Health	22	3	802	14	14	8	6	0	0	28	34	6	7	32	9	0	0	11	4	0	1000	80
Agriculture	21	1	9	826	7	2	19	3	10	9	5	13	20	2	3	27	8	5	7	3	1000	83
Labor	47	6	19	6	763	18	5	3	7	13	31	8	10	12	5	4	14	24	2	3	1000	76
Education	13	19	11	0	9	830	1	1	5	7	12	6	12	16	13	0	8	26	11	0	1000	83
Environment	17	4	9	20	2	9	757	16	19	8	2	9	18	2	16	9	22	10	50	1	1000	76
Energy	21	2	1	1	1	1	27	855	19	3	2	6	9	5	10	14	4	3	16	0	1000	86
Transportation	20	5	4	1	12	1	21	7	819	14	3	21	14	10	5	8	9	6	18	2	1000	82
Law/Crime	36	29	20	2	21	9	2	1	9	746	19	3	8	11	4	3	21	45	8	3	1000	75
Social Welfare	29	7	37	9	31	16	4	4	4	11	792	12	6	9	5	0	3	14	7	0	1000	79
Housing	26	5	5	13	13	3	8	3	1	9	17	837	17	14	2	0	6	8	11	2	1000	84
Banking	106	17	7	18	14	8	6	17	17	18	5	38	642	4	17	19	14	16	12	5	1000	64
Defense	13	14	16	0	20	16	4	8	11	22	4	8	3	757	2	6	33	33	24	6	1000	76
Science	19	25	4	1	2	16	8	6	8	17	1	2	19	3	825	6	17	14	7	0	1000	82
Foreign Trade	18	5	0	17	4	2	5	7	10	7	1	0	14	2	2	859	32	10	3	2	1000	86
Int'l Affairs	17	17	4	14	7	8	18	6	11	19	2	4	13	17	8	32	768	21	10	4	1000	77
Gov't Ops	53	38	14	3	19	12	7	6	5	44	8	14	12	24	10	4	20	663	17	27	1000	66
Public Lands	13	3	3	7	5	4	40	15	15	11	1	12	9	12	5	1	10	28	805	1	1000	80
Private Bills	0	0	0	0	1	1	0	0	1	0	0	0	1	4	0	0	0	5	0	987	1000	99
n	1289	937	983	963	993	988	941	966	984	1025	975	1037	889	951	969	1008	1013	1018	1024	1047		
Pct. Right	61	78	82	86	77	84	80	89	83	73	81	81	72	80	85	85	76	65	79	94		

FIGURE 3. A comparison of de-duplicated and duplicated datasets yield overall insignificant differences across the methodology. While the duplicated training sets slightly outperform the deduplicated training sets, regardless of algorithm, the differences are minimal.

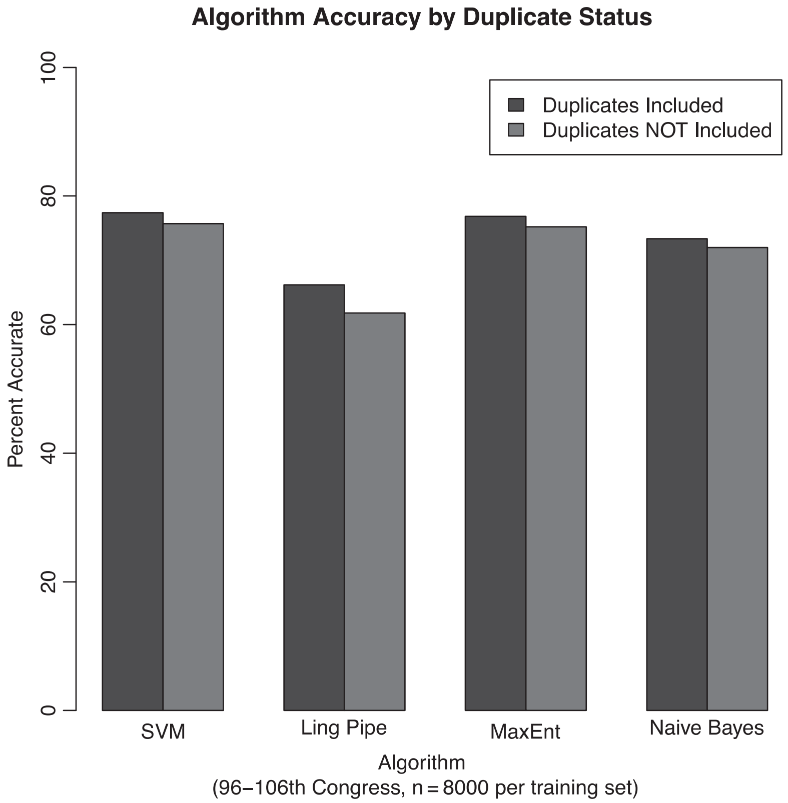
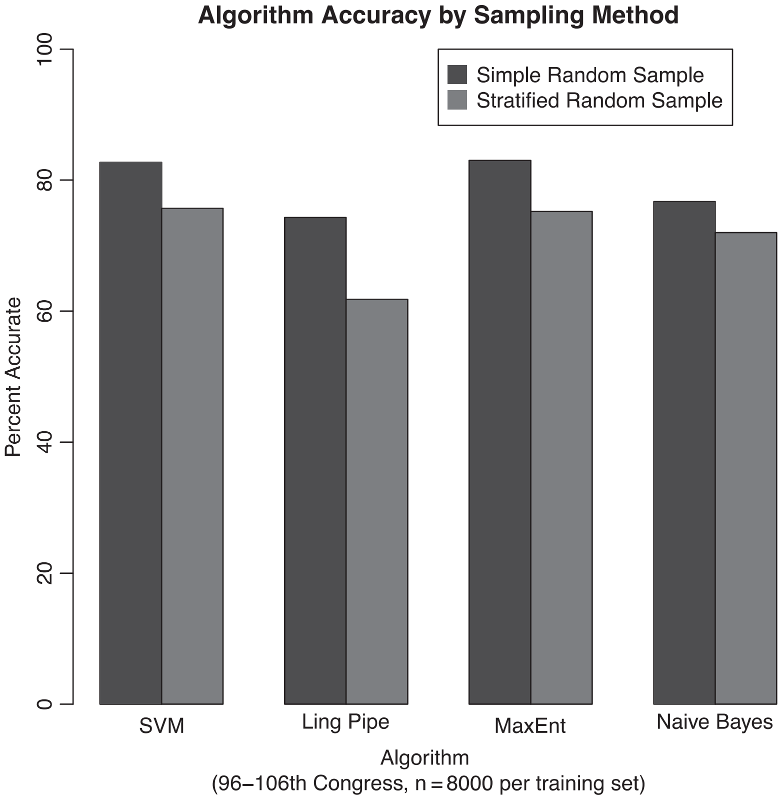


TABLE 5. The Effect of Deduplication

Topic	De-duped	Non-De-duped	Difference
Law/Crime	68.3	74	−5.7
Science	82.28	87.05	−4.77
Transportation	76.96	81.44	−4.48
Energy	86.7	90.03	−3.33
Private Bills	92.04	95.35	−3.31
Health	77.69	80.89	−3.2
Banking	61.32	64.46	−3.14
Housing	80.64	83.73	−3.09
Foreign Trade	86.65	89.18	−2.53
Int’l Affairs	69.47	71.77	−2.3
Education	83.62	85.64	−2.02
Agriculture	83.16	84.87	−1.71
Social Welfare	77.75	79.35	−1.6
Gov’t Ops	57.34	58.82	−1.48
Civil Rights	75.15	76.28	−1.13
Defense	74.19	75.23	−1.04
Environment	79.52	80.21	−0.69
Economics	56.19	56.7	−0.51
Public Lands	75.56	73.53	2.03
Labor	80.93	72.43	8.5

FIGURE 4. A comparison of bills sampled via simple random sampling and stratified random sampling where each category is normalized to a set count shows that simple random sampling is the preferred technique, presumably due to its more accurate representation of the data.



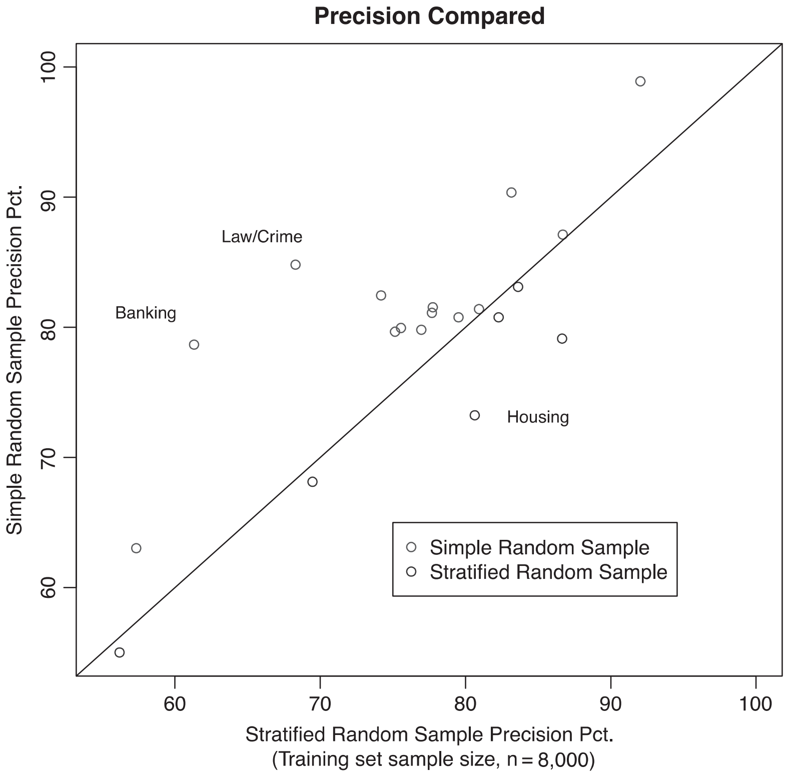
random sampling method outperforms our original stratified random sampling method across all four algorithms. When the sample is drawn via simple random sampling, the SVM algorithm achieves 82.71 percent accuracy versus 75.7 percent when the sample is stratified. Likewise, the ling pipe ratio is 74.29 percent to 61.8 percent, maximum entropy is 83.01 percent to 75.20 percent, and naive bayes is 76.71 percent to 72.00 percent.

A more precise comparison of the benefits of a random sampling approach can be seen in Figure 5, which plots precision percentages at the topic level (SVM) for the stratified random sample (*x* axis) and simple random sample (*y* axis) training approaches. A dot above the diagonal line indicates that the simple random sample is a more accurate predictor, which is the case for almost all of the topics. The likely explanation is that the simple random sampling approach does a better job of predicting the topics with more cases because it has more examples in the training set. A stratified sampling approach trades this benefit for disproportionately high precision in smaller topics and disproportionately low precision in larger topics. To the extent that these findings are generalizable, they are welcome results from an efficiency perspective.

ANALYSIS PART III: ENSEMBLE AGREEMENT AS AN APPROACH TO IMPROVING OVERALL ACCURACY

The results of the previous experiments indicate that we could use the highest performing algorithm (SVM) to classify virgin congressional bills for topic at 78 percent average precision. But what if a researcher desires even higher

FIGURE 5. A comparison of precision rates by category for the SVM algorithm with a training set sample size of n = 8000 shows that simple random sampling outperforms stratified random sampling.

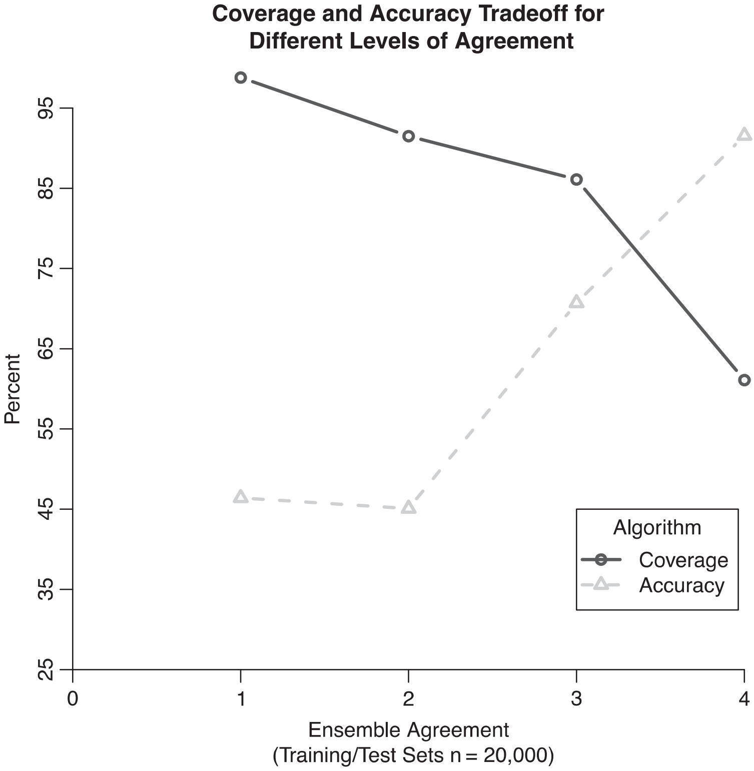


precision? The problem with virgin texts is that we do not know which ones have been correctly coded and which have not. One possibility is to use information from the ensemble of algorithms to identify bills that have been labeled with even higher average precision. If accuracy tends to be higher in the cases where algorithms agree on the class of a bill, then we can use that agreement to infer that the virgin bills have been labeled with higher accuracy. These bills can then be set aside while the researcher focuses on improving the accuracy of the cases where the algorithms disagree. The question, of course, is how many bills can be set aside and how many require follow-up attention using this method?

To answer these questions in the context of the Congressional Bills Project, we examined the relationship between algorithm agreement and accuracy for training and test sets of n = 20,000 randomly drawn bills. The accuracy coverage tradeoffs for different levels of agreements are displayed in Figure 6. The *x* axis corresponds to the number of algorithms in agreement (1 = two algorithms agree, two disagree; 4 = svm, maximum entropy, ling pipe, naive bayes all agree).[5] The *y* axis indicates (dashed line) the percent correctly predicted for different levels of ensemble agreement (accuracy), and (solid line) the percentage of total cases correctly predicted at that level or above (coverage).

Figure 6 indicates that when only one pair of algorithms agree (1), average accuracy (precision) is low, just 45 percent. Coverage is high (99 percent), indicating that at least two algorithms are in agreement for almost all bills. If there is greater agreement than just one pair of algorithms, the average percent of correctly predicted cases is still 45 percent (with 92 percent coverage). When a majority of algorithms agree (three or more), average accuracy improves to 71 percent as coverage declines to about

FIGURE 6. Ensemble agreement demonstrates that supervised learning accuracy varies depending upon the level of algorithm agreement chosen by the researcher.



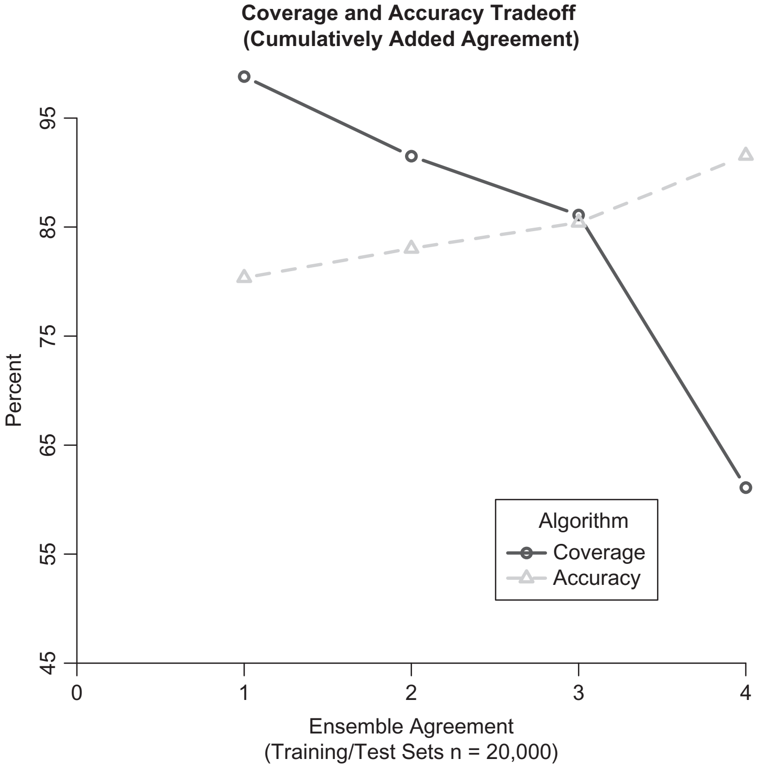
85 percent. Finally, when all of the algorithms assign a bill to the same class, the experiments using the set-aside test set indicate that average accuracy is 92 percent while coverage is just over 60 percent.

The four-algorithm ensemble agreement standard offers impressive accuracy, but at a cost of limited coverage. However, if we accept bills where at least three algorithms are in agreement (Figure 7)—that is, where either three or four algorithms agree—then we can expect 86 percent average agreement and about 85 percent coverage. This agreement level is substantially higher than what we were able to achieve by relying on the best performing algorithm to classify all bills. It comes at a cost—15 percent of the bills must be reviewed. Although some bills will still need to be manually labeled (the ensemble results tell us which these are), we are able to save substantial time and effort compared to manual coding, while maintaining similarly high levels of accuracy.

DISCUSSION

In the information and computer sciences, supervised machine learning is an established method for automatically classifying large numbers of events according to a pre-existing categorization system. Yet these tools and techniques are only now becoming integrated into the political scientist's methodological toolkit. The timing could not be better, given the growing diversity of digital records of government activity and the promise such methods offer in terms of substantially reducing annotation costs without sacrificing accuracy. This article has investigated the potential benefits of four machine-learning algorithms in the context of one ongoing project (the Congressional Bills Project). Every two years, approximately 10,000 legislative records must be labeled according to a relatively complex and pre-existing topic system. Supervised machine learning offers substantial promise for such a task. Using off-the-shelf

FIGURE 7. Ensemble agreement demonstrates that supervised learning accuracy varies depending upon the level of algorithm agreement chosen by the researcher. The numbers reported here are cumulative, so an ensemble agreement of two is a case where two or more algorithms are in agreement on the predicted label of the document. A score of three is where three or more algorithms are in agreement. Given the tradeoff evinced by agreement and coverage, a three level algorithm agreement is proposed in the current setup.



algorithms and basic preprocessing protocols, we achieved results slightly inferior to what we observe for trained human annotators. However, using information gleaned from the ensemble of algorithms to differentiate bills coded with higher accuracy, we were able to achieve results on par with human annotators, though with less than complete coverage.

Our findings do not provide a definitive answer to the question of how many cases must be manually labeled to achieve good results. The answer depends as much on the coding task as it does on the algorithm. However, for the Congressional Bills Project, we obtained good results for a sample of 100 examples per category and found declining marginal benefits as the number of training examples increased beyond that number. With just 100 examples, three of the four algorithms averaged 65 percent accuracy. Doubling the sample improved performance by about four additional percentage points. With 1000 examples per category, accuracy improved by about 10 additional points.

We also made two reassuring discoveries. The first was that a random sample of training examples performed better than a more labor-intensive approach of creating a stratified random sample of equal numbers of training examples per topic. The better approach to creating a training dataset was to just randomly sample from the data. The second reassuring discovery was that the presence or absence of a large number of duplicate records (32 percent in the case of the Bills Project) did not significantly impact performance. This suggests that duplicate records need not be filtered as part of the training process, and also that results reported for experiments involving large numbers of duplicate records (such as the Bills Projects) are probably representative of performance in contexts where duplicate records are less common.

Finally, ensemble agreement methods can be quite helpful if a researcher desires even higher levels of accuracy than can be achieved by relying on the best performing algorithm (in our experiments this was SVM). Whereas the best we could do with a single algorithm was 78 percent accuracy for 100 percent coverage, we were able to achieve 86 percent accuracy for 85 percent of the cases by focusing on the bills where at least three of the four algorithms "voted" for the same class. Supervised machine-learning methods can make a valuable contribution to larger annotation projects that begin with a preexisting classification system or training data. They are particularly useful in contexts such as the Bills Project where the subjective classification system is sufficiently complex to make keyword approaches prohibitively expensive.

NOTES

1. These bills and accompanying data are freely available.
2. The current analysis is based on bills drawn from the 90th–106th Congresses—a total of 229,037 bills.
3. F-score favors precision when $\beta > 1$, and recall otherwise.
4. F-score tables for the other algorithms can be found in the Appendix.
5. We do not include a point for no agreement, since labeling is entirely arbitrary in that case.

REFERENCES

Berger, A. L., Pietra, V. J. D., & Pietra, S. A. D. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, *22*(1), 39–71.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (pp. 144–152). New York: ACM.

Cardie, C., & Wilkerson, J. (2008). Guest editors' introduction: Text annotation for political science research. *Journal of Information Technology & Politics*, *5*(1), 1–6.

Carpenter, B., & Baldwin, B. (2011). Text analysis with Ling Pipe 4. New York: Ling Pipe Publishing.

Collingwood, L. (2010). *Rtexttools*: Classifies textual documents via automated content analysis. R package version 1.0.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM Computing Survey*, *27*, 326–327.

Dodds, P. S., & Danforth, C. M. (2009). Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies, 11*(4), 441–456.

Grimmer, J., & King, G. (2009). *Quantitative discovery from qualitative information: A general purpose document clustering methodology*. Paper presented at the 2009 American Political Science Association Meeting, Toronto, Ontario, Canada.

Hillard, D., Purpura, S., & Wilkerson, J. (2008). Computer-assisted topic classification for mixed-methods social science research. *Journal of Information Technology & Politics*, *4*(4), 31–46.

Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, *54*(1), 229–247.

Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). *A practical guide to support vector classification*. Unpublished manuscript.

Jurka, T. P., Collingwood, L., Boydstun, A. E., Grossman, E., & van Atteveldt, W. (2011). *RTextTools*: Automatic text classification via supervised learning. R package version 1.3.

King, G., & Lowe, W. (2003). An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International Organization*, *57*(3), 617–642.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, *97*(2), 311–331.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., et al. (2009). Life in the network: The coming age of computational social science. *Science*, *323*(5915), 721–723.

Lewis, D. (1998). Naive (bayes) at forty: The independence assumption in information retrieval. In C. Nédellec & C. Rouveirol (Eds.), *Machine learning: ECML-98* (pp. 4–15). Berlin: Springer.

Loper, E., & Bird, S. (2002). NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural*

Language Processing and Computational Linguistics (Vol. 1; pp. 63–70). Stroudsburg, PA: Association for Computational Linguistics.

Monroe, B. L., & Schrodt, P. A. (2008). Introduction to the special issue: The statistical analysis of political text. *Political Analysis*, *16*(4), 351–355.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). *From tweets to polls: Linking text sentiment to public opinion time series* (Paper 559). Tepper School of Business, Pittsburgh, PA.

Olson, D. L., & Delen, D. (2008). *Advanced data mining techniques*. Berlin: Springer Verlag.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, *54*(1), 209–228.

Ratnaparkhi, A. (1997). A simple introduction to maximum entropy models for natural language processing. *IRCS Technical Reports Series, May*, 97–08.

Sammut, C., & Webb, G. I. (2011). *Encyclopedia of machine learning*. New York: Springer-Verlag.

Schrodt, P. A., Davis, S. G., & Weddle, J. L. (1994). Political science: KEDS A program for the machine coding of event data. *Social Science Computer Review*, *12*(4), 561.

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, f-score and roc: A family of discriminant measures for performance evaluation. In *AI 2006: Advances in artificial intelligence* (pp. 1015–1021). Palo Alto, CA: American Association for Artificial Intelligence.

APPENDIX

TABLE A1. Support Vector Machine Recall Accuracy

Topic	n = 100	SD 100	n = 200	SD 200	n = 400	SD 400	n = 1000	SD 1000	1000–100
Civil Rights	0.58	−0.08	0.62	−0.08	0.66	−0.06	0.72	−0.02	0.15
Energy	0.72	−0.08	0.77	−0.08	0.82	−0.04	0.87	−0.02	0.14
Agriculture	0.69	−0.08	0.73	−0.08	0.78	−0.04	0.82	−0.02	0.14
Environment	0.64	−0.14	0.7	−0.06	0.73	−0.06	0.78	−0.02	0.14
Labor	0.62	−0.14	0.67	−0.08	0.7	−0.06	0.75	−0.04	0.13
Transportation	0.68	−0.12	0.72	−0.1	0.76	−0.06	0.81	−0.04	0.13
Defense	0.64	−0.16	0.7	−0.06	0.74	−0.04	0.77	−0.04	0.12
Law/Crime	0.64	−0.12	0.68	−0.08	0.71	−0.04	0.76	−0.02	0.12
Banking	0.52	−0.1	0.54	−0.06	0.58	−0.04	0.64	−0.04	0.12
Science	0.72	−0.1	0.74	−0.06	0.78	−0.04	0.83	−0.02	0.11
Housing	0.72	−0.08	0.76	−0.06	0.79	−0.06	0.83	−0.02	0.11
Health	0.7	−0.08	0.71	−0.06	0.76	−0.04	0.8	−0.04	0.1
Gov't Ops	0.55	−0.14	0.59	−0.1	0.62	−0.06	0.65	−0.02	0.1
Foreign Trade	0.75	−0.08	0.79	−0.06	0.82	−0.04	0.86	−0.02	0.1
Education	0.74	−0.08	0.76	−0.04	0.79	−0.04	0.84	−0.02	0.1
Int'l Affairs	0.67	−0.1	0.69	−0.08	0.73	−0.04	0.76	−0.04	0.09
Social Welfare	0.7	−0.12	0.72	−0.08	0.77	−0.06	0.79	−0.02	0.09
Public Lands	0.73	−0.08	0.74	−0.06	0.78	−0.04	0.8	−0.02	0.07
Economics	0.74	−0.1	0.76	−0.08	0.77	−0.04	0.79	−0.02	0.05
Private Bills	0.98	−0.02	0.98	−0.02	0.98	−0.02	0.99	−0.02	0

TABLE A2. Maximum Entropy Recall Accuracy

Topic	n = 100	SD 100	n = 200	SD 200	n = 400	SD 400	n = 1000	SD 1000	1000–100
Law/Crime	0.57	−0.16	0.65	−0.08	0.71	−0.04	0.76	−0.04	0.19
Banking	0.49	−0.12	0.53	−0.1	0.6	−0.04	0.67	−0.04	0.19
Civil Rights	0.58	−0.1	0.62	−0.08	0.67	−0.06	0.74	−0.04	0.16
Environment	0.63	−0.12	0.69	−0.06	0.73	−0.04	0.78	−0.02	0.15
Gov't Ops	0.51	−0.14	0.55	−0.1	0.6	−0.06	0.65	−0.04	0.15
Transportation	0.66	−0.12	0.71	−0.08	0.76	−0.06	0.8	−0.04	0.14
Defense	0.64	−0.12	0.69	−0.06	0.73	−0.06	0.76	−0.04	0.13
Labor	0.63	−0.14	0.67	−0.08	0.71	−0.06	0.76	−0.04	0.13
Energy	0.73	−0.1	0.78	−0.08	0.82	−0.04	0.86	−0.04	0.12
Int'l Affairs	0.64	−0.12	0.68	−0.08	0.72	−0.06	0.76	−0.04	0.12
Science	0.72	−0.1	0.75	−0.06	0.79	−0.06	0.84	−0.04	0.12
Agriculture	0.7	−0.08	0.74	−0.06	0.77	−0.06	0.81	−0.04	0.12
Housing	0.72	−0.1	0.76	−0.08	0.78	−0.06	0.82	−0.02	0.09
Education	0.73	−0.08	0.74	−0.04	0.78	−0.04	0.82	−0.02	0.09
Health	0.7	−0.08	0.72	−0.06	0.76	−0.04	0.79	−0.02	0.09
Public Lands	0.73	−0.08	0.75	−0.08	0.78	−0.04	0.81	−0.02	0.08
Foreign Trade	0.77	−0.08	0.8	−0.04	0.82	−0.04	0.85	−0.02	0.08
Social Welfare	0.72	−0.12	0.72	−0.08	0.77	−0.06	0.79	−0.04	0.07
Economics	0.68	−0.12	0.71	−0.06	0.72	−0.04	0.74	−0.08	0.06
Private Bills	0.99	−0.02	0.99	−0.02	0.98	−0.02	0.99	0	0

TABLE A3. Naive Bayes Recall Accuracy

Topic	n = 100	SD 100	n = 200	SD 200	n = 400	SD 400	n = 1000	SD 1000	1000–100
Law/Crime	0.52	−0.18	0.6	−0.1	0.65	−0.06	0.69	−0.04	0.17
Int'l Affairs	0.58	−0.12	0.65	−0.08	0.69	−0.04	0.74	−0.02	0.16
Banking	0.4	−0.12	0.42	−0.08	0.49	−0.06	0.55	−0.04	0.15
Gov't Ops	0.41	−0.12	0.45	−0.08	0.49	−0.06	0.54	−0.04	0.14
Transportation	0.62	−0.12	0.66	−0.06	0.69	−0.06	0.74	−0.02	0.13
Civil Rights	0.55	−0.08	0.58	−0.1	0.61	−0.04	0.67	−0.02	0.12
Environment	0.63	−0.14	0.68	−0.04	0.71	−0.04	0.74	−0.04	0.12
Labor	0.56	−0.12	0.6	−0.08	0.64	−0.06	0.67	−0.04	0.11
Energy	0.72	−0.08	0.74	−0.06	0.8	−0.04	0.83	−0.04	0.11
Agriculture	0.7	−0.08	0.72	−0.08	0.75	−0.04	0.78	−0.02	0.08
Public Lands	0.69	−0.1	0.72	−0.06	0.76	−0.02	0.77	−0.02	0.08
Science	0.72	−0.08	0.75	−0.06	0.77	−0.06	0.8	−0.02	0.08
Housing	0.73	−0.1	0.76	−0.06	0.79	−0.06	0.81	−0.02	0.07
Foreign Trade	0.67	−0.1	0.7	−0.06	0.72	−0.06	0.75	−0.04	0.07
Education	0.7	−0.08	0.71	−0.04	0.74	−0.04	0.76	−0.04	0.06
Defense	0.68	−0.1	0.71	−0.06	0.73	−0.04	0.73	−0.02	0.06
Economics	0.73	−0.12	0.76	−0.04	0.76	−0.06	0.78	−0.04	0.05
Health	0.71	−0.08	0.73	−0.06	0.75	−0.04	0.75	−0.04	0.04
Social Welfare	0.71	−0.1	0.71	−0.06	0.74	−0.04	0.74	−0.04	0.03
Private Bills	0.97	−0.02	0.97	−0.02	0.97	−0.02	0.97	−0.02	0

TABLE A4. Ling Pipe Precision Accuracy

Topic	n = 100	SD 100	n = 200	SD 200	n = 400	SD 400	n = 1000	SD 1000	1000–100
Int'l Affairs	0.49	−0.12	0.56	−0.08	0.63	−0.04	0.71	−0.04	0.22
Banking	0.4	−0.12	0.44	−0.12	0.5	−0.08	0.58	−0.04	0.18
Law/Crime	0.47	−0.12	0.53	−0.08	0.59	−0.04	0.65	−0.04	0.17
Gov't Ops	0.36	−0.12	0.41	−0.1	0.47	−0.04	0.53	−0.04	0.17
Civil Rights	0.54	−0.1	0.58	−0.08	0.63	−0.06	0.71	−0.04	0.17
Economics	0.44	−0.12	0.48	−0.08	0.53	−0.06	0.61	−0.04	0.16
Agriculture	0.57	−0.12	0.63	−0.08	0.67	−0.04	0.73	−0.04	0.15
Science	0.59	−0.08	0.62	−0.06	0.66	−0.08	0.73	−0.02	0.15
Education	0.54	−0.12	0.56	−0.08	0.6	−0.06	0.68	−0.04	0.14
Labor	0.48	−0.12	0.5	−0.12	0.53	−0.06	0.61	−0.04	0.13
Foreign Trade	0.58	−0.12	0.61	−0.06	0.64	−0.04	0.7	−0.04	0.13
Health	0.56	−0.12	0.58	−0.08	0.63	−0.08	0.68	−0.02	0.12
Housing	0.58	−0.06	0.59	−0.06	0.62	−0.06	0.7	−0.02	0.12
Public Lands	0.59	−0.12	0.63	−0.06	0.68	−0.06	0.71	−0.02	0.12
Environment	0.58	−0.14	0.6	−0.08	0.64	−0.06	0.69	−0.04	0.12
Energy	0.62	−0.08	0.64	−0.08	0.69	−0.06	0.74	−0.04	0.12
Defense	0.52	−0.1	0.54	−0.08	0.58	−0.06	0.64	−0.04	0.11
Transportation	0.56	−0.1	0.56	−0.08	0.61	−0.04	0.66	−0.04	0.1
Social Welfare	0.59	−0.1	0.59	−0.08	0.62	−0.06	0.67	−0.04	0.08
Private Bills	0.86	−0.08	0.86	−0.06	0.87	−0.06	0.89	−0.02	0.02

TABLE A5. Maximum Entropy Precision Accuracy

Topic	n = 100	SD 100	n = 200	SD 200	n = 400	SD 400	n = 1000	SD 1000	1000–100
Banking	0.51	−0.14	0.54	−0.08	0.61	−0.04	0.71	−0.04	0.2
Int'l Affairs	0.58	−0.08	0.62	−0.06	0.69	−0.04	0.76	−0.04	0.18
Law/Crime	0.55	−0.12	0.6	−0.06	0.66	−0.06	0.73	−0.02	0.18
Economics	0.45	−0.06	0.48	−0.06	0.52	−0.04	0.6	−0.02	0.16
Gov't Ops	0.51	−0.12	0.56	−0.1	0.59	−0.06	0.64	−0.02	0.13
Transportation	0.7	−0.1	0.76	−0.04	0.79	−0.04	0.82	−0.02	0.12
Defense	0.67	−0.1	0.72	−0.06	0.75	−0.04	0.79	−0.02	0.12
Civil Rights	0.67	−0.14	0.69	−0.08	0.73	−0.06	0.79	−0.02	0.12
Environment	0.72	−0.12	0.75	−0.08	0.78	−0.04	0.82	−0.02	0.1
Labor	0.7	−0.1	0.74	−0.06	0.76	−0.04	0.79	−0.02	0.09
Agriculture	0.77	−0.1	0.81	−0.06	0.83	−0.04	0.84	−0.02	0.07
Science	0.79	−0.08	0.81	−0.06	0.83	−0.04	0.86	−0.02	0.07
Public Lands	0.74	−0.1	0.73	−0.06	0.77	−0.04	0.8	−0.02	0.07
Education	0.79	−0.08	0.8	−0.06	0.81	−0.04	0.83	−0.02	0.04
Health	0.77	−0.1	0.79	−0.06	0.8	−0.04	0.81	−0.02	0.04
Social Welfare	0.78	−0.1	0.78	−0.04	0.8	−0.04	0.82	−0.02	0.03
Energy	0.84	−0.08	0.86	−0.04	0.86	−0.02	0.87	−0.02	0.03
Housing	0.8	−0.1	0.79	−0.06	0.81	−0.04	0.82	−0.02	0.02
Private Bills	0.92	−0.04	0.93	−0.04	0.94	−0.02	0.95	−0.02	0.02
Foreign Trade	0.85	−0.06	0.87	−0.04	0.86	−0.04	0.87	−0.02	0.01

TABLE A6. Naive Bayes Precision Accuracy

Topic	n = 100	SD 100	n = 200	SD 200	n = 400	SD 400	n = 1000	SD 1000	1000–100
Defense	0.46	−0.12	0.53	−0.04	0.57	−0.06	0.62	−0.04	0.16
Health	0.62	−0.1	0.66	−0.04	0.71	−0.04	0.75	−0.04	0.13
Transportation	0.68	−0.12	0.74	−0.08	0.78	−0.06	0.79	−0.02	0.12
Science	0.66	−0.08	0.7	−0.08	0.73	−0.04	0.77	−0.04	0.11
Social Welfare	0.62	−0.1	0.65	−0.06	0.69	−0.04	0.73	−0.02	0.11
Law/Crime	0.61	−0.1	0.65	−0.06	0.67	−0.04	0.71	−0.02	0.11
Environment	0.67	−0.06	0.7	−0.06	0.74	−0.06	0.77	−0.04	0.09
Int’l Affairs	0.6	−0.1	0.63	−0.08	0.67	−0.04	0.69	−0.02	0.09
Gov’t Ops	0.61	−0.14	0.66	−0.08	0.69	−0.04	0.7	−0.02	0.09
Labor	0.68	−0.1	0.73	−0.08	0.74	−0.06	0.77	−0.04	0.09
Civil Rights	0.67	−0.12	0.69	−0.06	0.73	−0.04	0.76	−0.02	0.09
Agriculture	0.71	−0.1	0.76	−0.06	0.77	−0.04	0.79	−0.02	0.08
Banking	0.63	−0.14	0.66	−0.08	0.68	−0.04	0.71	−0.04	0.08
Economics	0.42	−0.06	0.44	−0.06	0.47	−0.04	0.5	−0.02	0.07
Housing	0.65	−0.1	0.65	−0.06	0.7	−0.04	0.72	−0.02	0.07
Education	0.74	−0.08	0.75	−0.06	0.78	−0.04	0.8	−0.02	0.06
Energy	0.77	−0.08	0.78	−0.06	0.8	−0.04	0.82	−0.02	0.05
Public Lands	0.7	−0.08	0.7	−0.04	0.72	−0.04	0.73	−0.02	0.03
Foreign Trade	0.89	−0.08	0.89	−0.04	0.9	−0.02	0.91	−0.02	0.02
Private Bills	0.98	−0.02	0.98	−0.02	0.98	−0.02	0.98	0	0

TABLE A7. Support Vector Machine F-Scores

Topic	n = 100	SD 100	n = 200	SD 200	n = 400	SD 400	n = 1000	SD 1000	1000–100
Banking	0.489	−0.091	0.549	−0.064	0.605	−0.044	0.671	−0.026	0.182
Law/Crime	0.579	−0.085	0.632	−0.057	0.683	−0.036	0.74	−0.022	0.161
Civil Rights	0.604	−0.087	0.649	−0.058	0.696	−0.039	0.754	−0.022	0.15
Int’l Affairs	0.611	−0.081	0.662	−0.054	0.706	−0.037	0.758	−0.02	0.147
Economics	0.543	−0.07	0.583	−0.047	0.626	−0.035	0.682	−0.021	0.139
Gov’t Ops	0.514	−0.086	0.564	−0.061	0.604	−0.041	0.648	−0.026	0.134
Transportation	0.685	−0.08	0.735	−0.053	0.774	−0.032	0.813	−0.02	0.128
Defense	0.656	−0.086	0.703	−0.055	0.743	−0.036	0.78	−0.02	0.124
Environment	0.674	−0.08	0.719	−0.052	0.756	−0.036	0.796	−0.02	0.122
Labor	0.659	−0.084	0.701	−0.052	0.734	−0.035	0.768	−0.02	0.109
Science	0.742	−0.074	0.774	−0.048	0.807	−0.032	0.845	−0.017	0.103
Agriculture	0.732	−0.076	0.766	−0.049	0.796	−0.032	0.831	−0.018	0.099
Energy	0.771	−0.073	0.808	−0.046	0.838	−0.03	0.867	−0.017	0.096
Education	0.74	−0.076	0.775	−0.048	0.805	−0.03	0.836	−0.018	0.096
Public Lands	0.706	−0.076	0.74	−0.05	0.769	−0.032	0.801	−0.021	0.095
Health	0.719	−0.075	0.748	−0.05	0.776	−0.032	0.808	−0.019	0.089
Housing	0.739	−0.075	0.771	−0.047	0.796	−0.031	0.824	−0.018	0.085
Social Welfare	0.736	−0.072	0.758	−0.048	0.779	−0.032	0.803	−0.02	0.067
Foreign Trade	0.803	−0.066	0.828	−0.042	0.844	−0.03	0.86	−0.017	0.057
Private Bills	0.951	−0.031	0.953	−0.022	0.958	−0.015	0.966	−0.011	0.015

TABLE A8. Maximum Entropy F-Scores

Topic	n = 100	SD 100	n = 200	SD 200	n = 400	SD 400	n = 1000	SD 1000	1000–100
Banking	0.493	−0.096	0.558	−0.063	0.615	−0.041	0.674	−0.034	0.181
Law/Crime	0.564	−0.093	0.629	−0.058	0.687	−0.037	0.744	−0.036	0.18
Civil Rights	0.578	−0.087	0.635	−0.055	0.689	−0.041	0.753	−0.027	0.175
Int'l Affairs	0.588	−0.083	0.642	−0.055	0.694	−0.038	0.751	−0.026	0.163
Gov't Ops	0.5	−0.09	0.558	−0.062	0.604	−0.043	0.653	−0.031	0.153
Transportation	0.66	−0.086	0.721	−0.054	0.767	−0.034	0.807	−0.029	0.147
Environment	0.649	−0.086	0.704	−0.052	0.747	−0.037	0.79	−0.025	0.141
Economics	0.561	−0.075	0.605	−0.051	0.646	−0.037	0.69	−0.046	0.129
Public Lands	0.672	−0.078	0.72	−0.051	0.759	−0.034	0.797	−0.024	0.125
Defense	0.644	−0.084	0.691	−0.055	0.73	−0.037	0.767	−0.051	0.123
Labor	0.644	−0.084	0.69	−0.054	0.725	−0.036	0.762	−0.03	0.118
Science	0.726	−0.075	0.767	−0.047	0.804	−0.031	0.842	−0.02	0.116
Agriculture	0.718	−0.081	0.759	−0.049	0.792	−0.033	0.825	−0.022	0.107
Education	0.728	−0.078	0.762	−0.049	0.795	−0.031	0.827	−0.02	0.099
Health	0.705	−0.076	0.74	−0.05	0.771	−0.033	0.803	−0.023	0.098
Energy	0.77	−0.072	0.805	−0.045	0.834	−0.029	0.86	−0.019	0.09
Housing	0.73	−0.073	0.76	−0.046	0.785	−0.031	0.815	−0.021	0.085
Social Welfare	0.716	−0.073	0.745	−0.048	0.769	−0.033	0.794	−0.021	0.078
Foreign Trade	0.785	−0.065	0.812	−0.042	0.833	−0.03	0.852	−0.019	0.067
Private Bills	0.953	−0.03	0.959	−0.021	0.964	−0.015	0.972	−0.01	0.019

TABLE A9. Naive Bayes F-Scores

Topic	n = 100	SD 100	n = 200	SD 200	n = 400	SD 400	n = 1000	SD 1000	1000–100
Banking	0.465	−0.103	0.523	−0.07	0.572	−0.045	0.621	−0.027	0.156
Law/Crime	0.56	−0.09	0.617	−0.06	0.662	−0.038	0.702	−0.023	0.142
Gov't Ops	0.477	−0.099	0.532	−0.072	0.574	−0.047	0.613	−0.028	0.136
Transportation	0.637	−0.088	0.692	−0.057	0.732	−0.035	0.765	−0.023	0.128
Civil Rights	0.586	−0.088	0.632	−0.059	0.67	−0.039	0.71	−0.022	0.124
Int'l Affairs	0.596	−0.086	0.643	−0.057	0.682	−0.038	0.718	−0.022	0.122
Defense	0.554	−0.083	0.599	−0.059	0.637	−0.04	0.672	−0.023	0.118
Environment	0.641	−0.084	0.688	−0.053	0.723	−0.037	0.754	−0.021	0.113
Labor	0.611	−0.09	0.658	−0.058	0.689	−0.039	0.719	−0.024	0.108
Health	0.649	−0.075	0.69	−0.052	0.721	−0.035	0.748	−0.022	0.099
Energy	0.723	−0.076	0.767	−0.047	0.796	−0.032	0.82	−0.019	0.097
Science	0.689	−0.075	0.723	−0.05	0.755	−0.033	0.784	−0.019	0.095
Education	0.692	−0.082	0.73	−0.052	0.76	−0.035	0.783	−0.021	0.091
Social Welfare	0.649	−0.073	0.685	−0.049	0.712	−0.034	0.738	−0.022	0.089
Housing	0.675	−0.08	0.71	−0.05	0.736	−0.035	0.761	−0.02	0.086
Economics	0.525	−0.07	0.554	−0.049	0.58	−0.034	0.605	−0.021	0.08
Agriculture	0.701	−0.075	0.732	−0.05	0.756	−0.033	0.779	−0.02	0.078
Public Lands	0.687	−0.075	0.714	−0.049	0.736	−0.034	0.752	−0.021	0.065
Foreign Trade	0.77	−0.074	0.789	−0.049	0.804	−0.034	0.817	−0.021	0.047
Private Bills	0.974	−0.021	0.974	−0.016	0.974	−0.013	0.975	−0.01	0.001