An empirical analysis of the training and feature set size in text categorization for e-Discovery

Ali Hadjarian, Ph.D.
Senior Manager
Deloitte Financial Advisory Services LLP
Suite 1200
1001 G Street, NW
Washington, D.C. 20001
ahadjarian@deloitte.com

Jianping Zhang, Ph.D.
Senior Manager
Deloitte Financial Advisory Services LLP
Suite 1200
1001 G Street, NW
Washington, D.C. 20001
jianpzhang@deloitte.com

Shuxing Cheng, Ph.D.
Senior Associate
Deloitte Financial Advisory Services LLP
Suite 1200
1001 G Street, NW
Washington, D.C. 20001
shucheng@deloitte.com

Abstract
While sample size calculations based on confidence levels and confidence intervals are widely used in predictive coding for determining the size of a validation set, their use in determining the size of the training sample, however, is somewhat questionable. In this paper, we argue that the number of documents in the training set has less to do with the total size of the document population and more to do with the complexity of the categorization problem at hand. We further argue that this complexity may be approximated by the number of features (i.e., predictive terms) sufficient for achieving near optimal classification performance. We use empirical results from four real-life legal matters to support the above arguments.

Introduction
Text categorization - often referred to as predictive coding in e-Discovery — is the task of automatically classifying documents into a set of predefined categories[i]. It typically involves two steps: (1) training and (2) prediction - or scoring (Figure 1). In the training step, a supervised learning algorithm is used to build a classification (or predictive) model from a sample of attorney reviewed documents, also referred to as the training set. The classification model is then used to generate class predictions (or scores) for all the documents that have not been yet reviewed. This happens in the prediction step.

To measure the effectiveness of the classification model and as such assess the reliability of the document scores, the performance of the model is generally evaluated on a separate sample of attorney reviewed documents, also referred to as the validation set. The above steps can then be repeated iteratively, with each round adding more attorney reviewed documents to the training set (typically by augmenting it with previous round's validation set), with the aim of improving the classification performance of the model.

Given the high cost of attorney review, it would generally be preferable to use as few training documents as possible when building a classification model. But exactly how many documents should be included in the training set? There seems to be some confusion and misconceptions around this question. Some have suggested statistical sampling as a means to arrive at this number. While statistical measures such as confidence levels and confidence intervals are helpful in calculating the size of the validation set, they don't provide a good means for approximating the required size of the training set, i.e., the number of training documents needed for the classification

model to approach its peak performance.

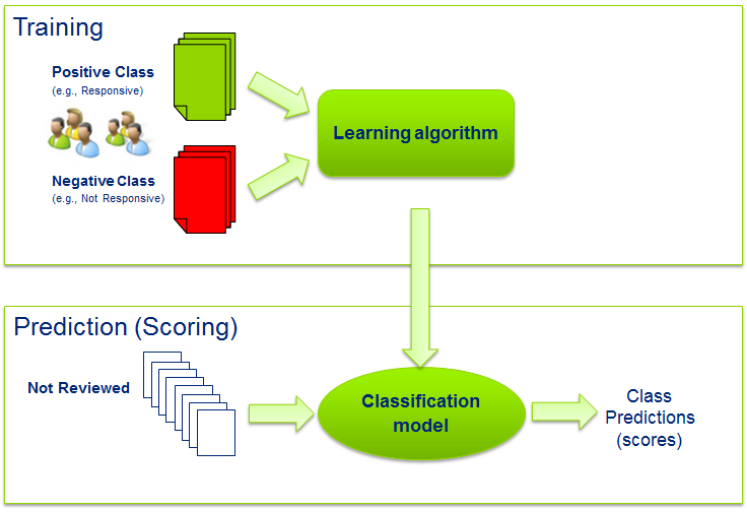

Figure 1. The Text Categorization Process

Here, we suggest that the number of documents needed for the model to approach its peak performance can be quite different from one categorization problem to another. In other words, when it comes to training sets, one size does not fit all. We argue that the size of the training set is less dependent on the total size of the document population and much more dependent on the complexity of the categorization problem at hand. We further argue that the complexity of the categorization problem itself can be approximated by the number of features, or in this case significant terms (e.g., words), required for the classification model to approach its peak performance.

The outline of the paper is as follows. In the next two sections, we will list some of the related work and state the objectives of this study. In the two sections following that, we will discuss our experimental setup and the results obtained on data from four real-life matters. We will then summarize our findings and conclude the study in the last section.

Related work

Questions surrounding the size of training and feature sets have long been of interest to machine learning researchers. Works in computational learning theory attempt to answer these questions using a theoretic framework by which the researchers try to place theoretical bounds on the sample complexity for various learning algorithms[ii]. The theoretical relationship between sample complexity and classification performance has also been examined in the context of text categorization with the support vector machine (SVM) algorithm, which is also the supervised learning algorithm of choice in our study[iii].

The relationship between the training sample size, or feature set size, and model performance in text categorization has also been examined in a number of empirical studies. Examples of these include[iv] [v] [vi] [vii] [viii] [ix] [x]

Objective

The objectives of this study are as follows:

- To conduct an empirical analysis of the relationship between the complexity of a categorization problem and the number of features (i.e., significant terms) required for the associated model to approach its peak performance.
- To conduct an empirical analysis of the relationship between the complexity of a categorization problem and the number of training documents required for the associated model to approach its peak performance.

Since the focus of this study is that of text categorization in e-Discovery, we will rely on results obtained on document populations from real-life legal matters.

Experimental setup

We conducted two groups of experiments, one for each of the objectives as stated above. The data used in all experiments came from four real-life legal matters, each involving millions of documents. Since our empirical analysis makes use of the attorney coding decisions, this analysis was limited to those documents that had been reviewed and coded for responsiveness by the attorneys. These included all of the documents originally used in various iterations of text categorization model training and validation for each of the four matters.

The setup for the feature set size experiments on each of the four datasets was as follows. From the available population of attorney reviewed documents, a statistically significant validation set was held out to perform all the categorization results analysis on. Five classification models were then trained with the remaining documents for varying feature set sizes, namely 50, 100, 250, 500 and all the available features.

Information gain was used as the feature selection criterion in these experiments. Information gain is a statistical measure for calculating the expected reduction in entropy, a common measure in information theory, capturing the impurity of a sample relative to the intended classification[xi]. The information gain value of a given term is generally based on its effectiveness in discriminating between the classes of interest, i.e., the higher the discrimination power, the higher the information gain. The effectiveness of information gain as a feature selection criterion for text categorization tasks has been established in a number of studies, including[xii].

The performance of the classification models with varying feature set sizes was analyzed, using the AUC (Area Under the ROC Curve) as the performance measure. A ROC (Receiver Operating Characteristic) curve, is a two-dimensional plot with the X axis representing the false positive rate and the Y axis representing the true positive rate. The formula for calculating the true positive rate is the same as that of recall. False positive rate, as the name implies, is the ratio of false positive test cases to the total number of negative examples. Generally speaking, an ROC curve captures the trade-off between the true positive and false positive rates for varying classification scores for a classifier with continuous output. This is in contrast to accuracy, for instance, that simply measures performance for a fixed cutoff score value.

In terms of statistics, the AUC captures the probability that a randomly chosen positive data point is assigned a higher score than a randomly chosen negative data point by the classifier[xiii]. The value of AUC generally ranges from 0.5 to 1, with 0.5 representing a random classifier and 1 representing a perfect classifier. AUC is generally considered to be a more effective classification performance measure than the widely used accuracy[xiv].

The setup for the training set size experiments for each dataset followed that of the feature set size experiments described above. Here, however, instead of building different classification models by varying the size of the feature set, we built different models by varying the number of documents in the training set. More specifically, five such models were built for each of the four legal matters, using five different training sample sizes, namely 500, 1500, 3000, 5000, and 6500 documents.

Experimental results

Figure 2 shows the effect of the feature set size on the classification performance for the four legal matters. As it can be seen in the figure, the number of features that it takes for the model to approach its peak performance (in terms of AUC) can vary from one categorization problem to another. Here for example, in the case of Dataset 1, the model's performance improves only slightly when going from 50 to all the available features, whereas the associated performance gain is larger for Datasets 2 and 3 and even larger in the case of Dataset 4.

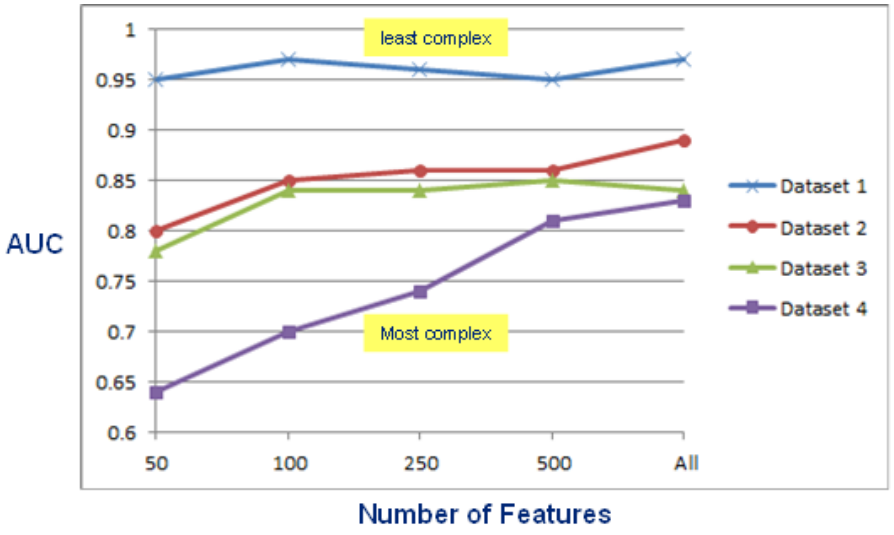

Figure 2. Model performance for varying feature set sizes

Figure 3 captures the performance gain from one feature set size to the next, as a percentage of the model performance at the smallest feature set size (i.e., 50). The maximum gain here was around 30 percentage points for Dataset 4 and the minimum gain was around 2 percentage points for Dataset 1. Incidentally, off the bat, we were able to achieve the highest performance on Dataset 1 and the lowest performance on Dataset 4 (i.e., 95% and 64%, respectively, for 50 features).

Assuming the model performance to be generally linked to the complexity of the categorization problem (i.e., separability of the two classes), we can see a connection between such problem complexity and the feature set size, with more complex problems generally benefitting more from larger feature sets. Again, in the above example, with Dataset 4 at one end of the spectrum, presenting the most challenging problem and as such benefitting the most from an increase in the feature set size, and Dataset 1, on the opposite end of the spectrum, presenting the least challenging problem and as such benefitting the least from an increase in the feature set size, and Datasets 2 and 3 falling somewhere in the middle of the spectrum.

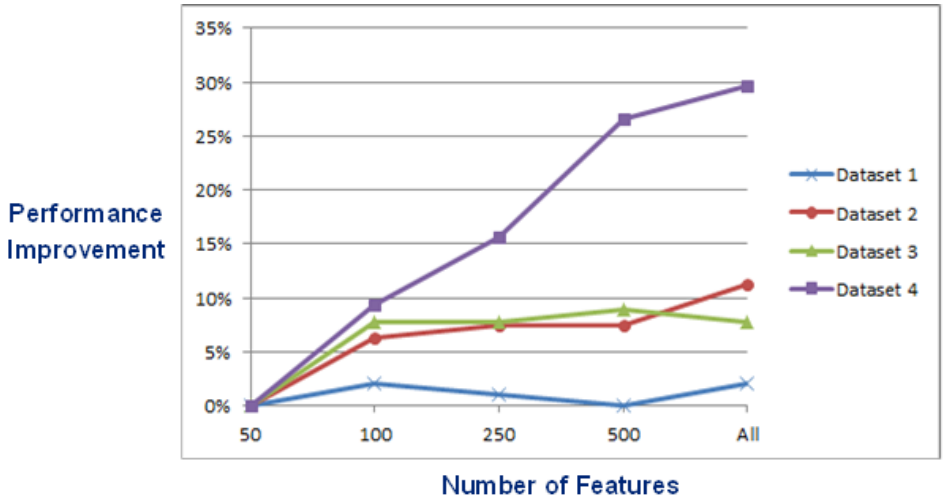

Figure 3. Performance gain for varying feature set sizes

The effect of the training sample size on the classification performance is captured in Figure 4. As it was the case with the feature set size, the gain in performance due to an increase in the sample size seems to be highly correlated with the complexity of the categorization problem. In other words, models with higher AUCs generally seem to benefit less from larger training sets. Figure 5 captures the performance gain from one training sample size to the next, as a percentage of the model performance at the smallest sample size (i.e., 500 documents).

Again, as it was the case for the feature size experiments, Dataset 4, involving the most complex categorization problem seems to benefit the most from an increase in the training sample size, while Dataset 1, involving the least complex categorization problem, seems to benefit the least from such an increase. And once again, Datasets 2 and 3 fall somewhere in the middle.

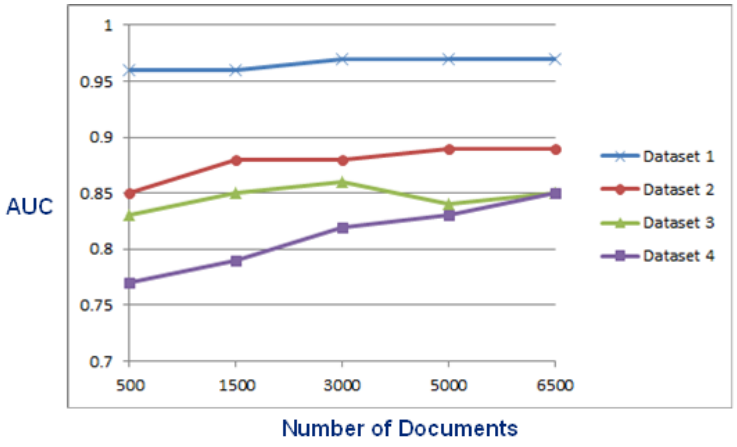

Figure 4. Model performance for varying training sample sizes

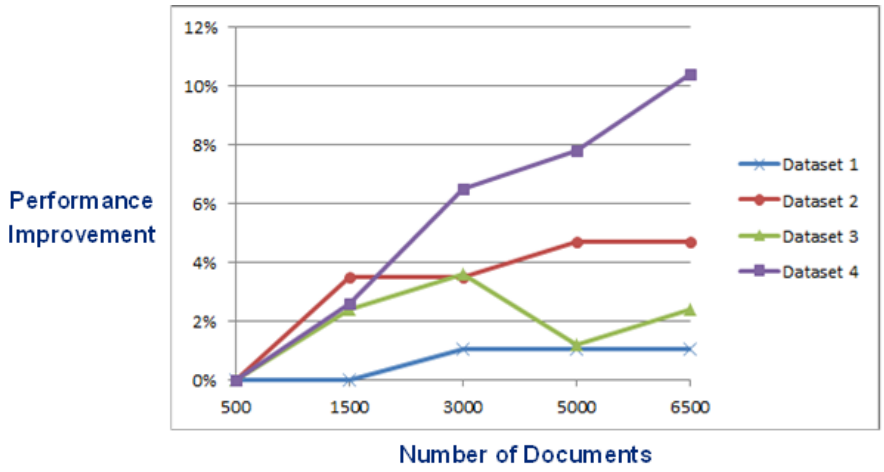

Figure 5. Performance gain for varying training sample sizes

Here, however, as it can be seen in the two figures, Dataset 3 exhibits an atypical characteristic, in that while an increase in the training sample size typically results in an improvement in model performance, this doesn't seem to be the case for this particular dataset. In other words, the model performance for this dataset seems to be at its highest for a training sample with 3000 documents versus 6500 documents for all the other datasets. But even then, as stated previously, this model seems to benefit more from an increase in the training sample size than the one for the least complex categorization problem, namely Dataset 1.

Conclusions

We have conducted an empirical analysis of the size of the feature and training sets in text categorization in e-Discovery. We have demonstrated the relationship between the complexity of the categorization problem at hand and the associated training sample size, with more complex problems generally benefitting more from larger sample sizes. We have further demonstrated how an analysis of the feature set size can help measure such complexity, again with more complex problems generally benefitting more from larger feature set sizes.

Future work will use this empirical analysis as a basis to formulate a mechanism for approximating the size of the training set, one that is hopefully more technically sound than statistical sample size calculations based on confidence levels and intervals and other even more ad hoc standardizations of the training sample size.

References

[i] Sebastiani, F. (2002). Machine learning in automated text categorization. In ACM Computing Surveys, 34(1): 1-47.

[ii] Vapnik, V. N., (1999) An Overview of Statistical Learning Theory. In IEEE Transactions on Neural Networks, Vol. 10, No. 5.

[iii] Joachims, T. (2001). A statistical learning model of text classification for support vector machines. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

[iv] Sahami, M. (1998). Using Machine Learning to Improve Information Access. PhD Thesis, Stanford University, Computer Science Department. STAN-CS-TR-98-1615.

[v] Yang, Y. (1996). Sampling strategies and learning efficiency in text categorization. AAAI Spring Symposium on Machine Learning in Information Access.

[vi] Gabrilovich, E. and Markovitch, S. (2004). Text categorization with many redundant features: using aggressive feature selection to make SVMs competitive with C4.5. In ICML '04: Proceedings of the 21st International Conference on Machine Learning.

[vii] Dumais, S., Piatt, J., Heckerman, D., and Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In Proceedings of the Seventh International Conference on Information and Knowledge Management, CIKM '98.

[viii] Yang, Y. and Pedersen, J.O. (1997). A comparative study on feature selection in text categorization. In Proceedings of International Conference on Machine Learning.

[ix] Rogati, M. and Yang, Y. (2002). High-performing feature selection for text classification. In Proceedings of CIKM 2002.

[x] Taira, H. and Haruno, M. (1999). Feature selection in SVM text categorization. In Proceedings of AAAI - Conference of the American Association for Artificial Intelligence.

[xi] Mitchell, T. M. (1997). Machine Learning. McGraw-Hill.

[xii] Yang, Y. and Pedersen, J.O. (1997). A comparative study on feature selection in text categorization. In Proceedings of International Conference on Machine Learning.

[xiii] Fawcett, T. (2003). ROC graphs: Notes and practical considerations for data mining researchers . Tech report HPL-2003-4. HP Laboratories, Palo Alto, CA, USA.

[xiv] Ling, C.X., Huang, J. and Zhang, H. (2003). AUC: a Better Measure than Accuracy in Comparing Learning Algorithms. Proceedings of 2003 Canadian Artificial Intelligence Conference.