Optimal Subset for Text Analysis:

An Approach Using Representative Sampling Strategy

Naijia Liu*, Natalie Ayers†, António Câmara‡, Saki Kuzushima§

Summer Polmeth Meeting, July 2025

Abstract

In text analysis, researchers often rely on random sampling techniques to select training data. However, random sampling can be risky, especially dealing with high-dimensional text data, as it tends to yield high variance in out of sample performance and may overlook important regions of the representation space. To address this challenge, we propose an alternative framework and algorithm that explicitly learns the underlying distribution of the text data and selects the most representative documents for model training. We demonstrate the effectiveness of our approach using both simulated datasets and four different real-world text corpora. We also discuss the scope of our method and its potential integration into more complex contexts, such as active learning and large language models.

Keywords: text analysis, machine learning

*Assistant professor, Department of Government, Harvard University, naijialiu@fas.harvard.edu
†Ph.D. Candidate, Department of Government, Harvard University
‡Ph.D. Candidate, Department of Government, Harvard University
§Independent researcher.

1 Introduction

When dealing with text data, researchers often rely on model training to uncover latent patterns within a corpus and to investigate how textual features relate to an outcome variable of interest—such as in classification tasks. The most widely adopted strategy for initiating such analyses is to construct a training set by randomly selecting a subset of human-labeled documents. This approach is popular in part because it enjoys desirable statistical properties (Hand, 2006): under random sampling, parameter estimates tend to be unbiased. Furthermore, random selection avoids the risk of systematic bias in the training process, making it an attractive default when there is no prior knowledge about the data available.

Despite these advantages, we show in this paper that random sampling can lead to substantial variance in model performance on held-out data, even when the size of the training set is fixed. Figure 1 presents a simple toy exercise conducted using a dataset of BBC news articles (Greene and Cunningham, 2006). The dataset comprises approximately 2,200 articles, each annotated with a binary label indicating whether the article pertains to politics. The X-axis represents the size of the training data, which is randomly sampled for each experiment, while the Y-axis displays the resulting testing accuracy. For each training size, we perform 200 independent simulations to showcase the variability in the sampling process. As shown in the figure, the test accuracy exhibits variability between different random samples. In a one-shot labeling and model training scenario, researchers have no reliable way of knowing whether the resulting model performance is representative or simply the consequence of a favorable (or unfavorable) draw from the data distribution.

This performance variability presents a practical concern. Researchers may expend considerable resources—especially time and annotation costs—on labeling a poorly chosen training subset that does not generalize well. In extreme cases, a "terribly chosen" random training set could significantly degrade the effectiveness of downstream analyses. There are many reasons why a training set might underperform, but in this project, we concentrate specifically on one key dimension: the representativeness of the selected documents. We argue that improving this aspect of training set construction—by selecting examples that better reflect the overall data distribution—can yield large gains in predictive performance, even in the absence of label information.

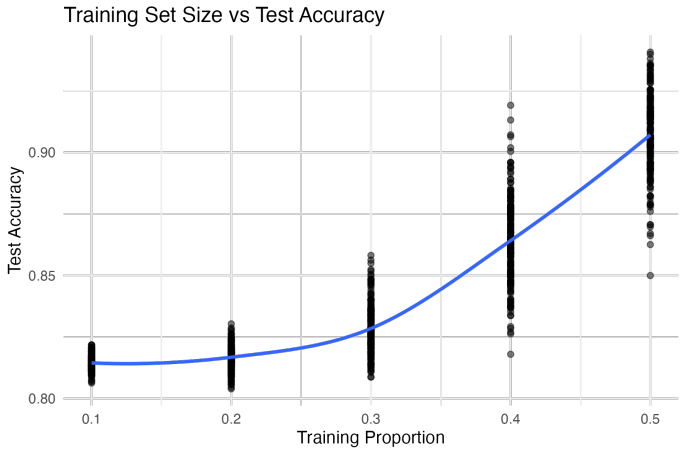

Figure 1: Out-of-sample testing performances vary, based on different training sets

There has been growing attention in the literature to the limitations of random sampling for training set construction, particularly in text classification tasks. For example, Bosley et al. (2022) propose a novel algorithm that combines probabilistic modeling with active learning to more efficiently identify informative documents for labeling.

Similarly, Kaufman (2024) shows that algorithmically selecting documents for hand-coding—using even simple heuristics—can lead to substantial improvements in model accuracy. In their empirical evaluations, performance gains of up to 50% were observed compared to models trained on randomly sampled subsets. These findings underscore the importance of better training data selection, especially in settings where annotation budgets are limited and each labeled example carries significant cost.

Existing studies largely rely on partially known labels to guide the selection of training data. In contrast, this paper will propose algorithms that operate under the assumption that no labels are available from the data, aiming to address the more challenging and realistic scenario of training data selection.

1.1 Related Literature on Sampling

The problem of designing effective sampling strategies for text data remains an active area of research. Many methods have been proposed to improve upon random sampling, each targeting different criteria such as diversity, informativeness, or representativeness.

One widely studied approach is density-based sampling, which aims to identify examples that are most representative of the overall unlabeled data distribution. The core idea is to prioritize instances that lie in densely populated regions of the feature space, under the assumption that labeling such examples provides information about a larger portion of the data. Following Wu et al. (2006), the representativeness of the ith unlabeled observation u_i—out of n total—is defined as:

$$\text{representativeness}(u_i) = \frac{1}{n-1} \sum_{j \neq i} K(u_i, u_j) \tag{1}$$

where $K(x, y)$ denotes a kernel function commonly used in support vector machines (SVM). If a document i has high representativeness under this definition, it implies that, on average, the kernel-based similarity between u_i and the other unlabeled documents is high (i.e., the kernel distance is relatively small). In other words, u_i resides in a densely populated region of the feature space and is therefore likely to reflect common patterns in the data.

Furthermore, Brinker (2003) propose an alternative approach designed for batch-mode active learning, where multiple observations are selected at once rather than sequentially. Their method focuses on maximizing the diversity of the selected batch to avoid redundancy among sampled documents. To quantify diversity, they compute the angles between feature vectors of observations, capturing how distinct each document is from those already selected. Formally, diversity for an unlabeled observation u_i is defined as:

$$\text{diversity}(u_i) = 1 - \max_{s_j \in S} \frac{K(u_i, s_j)}{\sqrt{K(u_i, u_i) K(s_j, s_j)}} \tag{2}$$

where $S = \{s_1, \ldots, s_J\}$ is the current sample. The key idea is to sample observations that exhibit minimal cosine similarity with the current set of selected documents—effectively promoting geometric separation in the feature space. In practice, this helps ensure that the chosen examples span distinct regions of the input space, which can improve the model's generalization ability. It is worth noting that, like the density-based method, this approach is primarily developed in the context of support vector machines (SVM).

Cluster-based sampling (Kang, Ryu and Kwon, 2004; Xu et al., 2003) provides a straightfor-

ward yet effective approach for selecting initial training examples in active learning frameworks. The primary goal of this method is to identify a diverse and representative set of documents at the outset of the labeling process, thereby improving model performance in early rounds of training.

The procedure begins by applying the k-means clustering algorithm—typically with $k = 2$—to partition the unlabeled data into distinct groups. Documents nearest to the resulting cluster centroids are then selected for labeling. Although the centroids themselves are not actual data points, they can either be approximated by their closest real observations or labeled directly and used as modal (i.e., prototypical) examples. Both variants—standard k-means and k-means with modal examples—have been shown to outperform random sampling strategies in empirical studies.

While this approach does not originally provide a formal mathematical definition of representativeness, it can be expressed as:

$$\text{representativeness}(u_i) = \begin{cases} 1 & \text{if} \quad i = \arg\min \ d(u_i, c_k) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where c_k denotes the centroid of cluster k, and $d(\cdot, \cdot)$ is a distance metric defined over the feature space. Under this binary definition, only the single observation closest to the centroid is considered representative. However, the idea can be naturally extended to a continuous formulation by assigning weights to all observations based on their distance to the centroid. In such a scheme, documents closer to the centroid would receive higher weights, allowing for a soft notion of representativeness that captures the graded structure of the cluster.

1.2 Roadmap

Following the line of literature discussed above, we propose two complementary algorithms in this paper to address the challenge of training data selection in text classification tasks.

The first is a one-shot selection algorithm designed to match or exceed the performance of random sampling from the outset. By identifying a representative and informative subset of documents at the initialization stage, this method enables researchers to achieve strong predictive performance without the need for iterative labeling or additional annotation costs. Our results show that even without multiple rounds of refinement, substantial improvements in model quality can be attained

through principled subset selection.

Beyond the one-shot selection approach, we also consider an iterative labeling procedure tailored for researchers who are willing to engage in multiple rounds of data annotation. This method builds on established principles from active learning, where the training set is updated incrementally based on feedback from the model's performance.

At each iteration, our algorithm evaluates model uncertainty or identifies patterns in prediction error to guide the selection of new examples for labeling. By adaptively targeting the most informative documents, the procedure efficiently improves model performance over time. This strategy is especially valuable in settings with constrained labeling budgets, where incremental gains in accuracy are prioritized and each annotation must be maximally impactful.

We begin by introducing the setup and key components of the proposed algorithm. We then outline the formal framework and underlying assumptions. Next, we demonstrate the algorithm's performance through comparisons on both simulated and real-world datasets. Finally, we offer more discussions on the scope of the method and outline potential next steps.

2 Set up

We consider a dataset X consisting of N documents. The objective is to construct a training subset of size $M \leq N$ that enables accurate prediction of document labels on unseen data.

We assume that the researcher defines a set of k distinct classes: $Y = \{y_1, y_2, \ldots, y_k\}$. Let X_{train} denote the feature matrix derived from the training text data. Given a labeled subset of the data, our goal is to learn the conditional distribution

$$P(Y_{\text{train}} = y \mid X_{\text{train}}), \tag{4}$$

which maps document features to label class probabilities. The trained model is then used to generate predictions for the remaining, unlabeled documents via

$$\widehat{P}(Y_{\text{test}} = y \mid X_{\text{test}}), \tag{5}$$

with the ultimate goal of achieving strong out-of-sample predictive performance.

Let $\mathbf{h}(X)$ denote a representation function that maps each document to a (possibly lower-dimensional) feature space that captures the relevant structure in the data. For example, $\mathbf{h}(X)$ may represent the embedding vectors of a corpus, obtained through techniques such as word2vec (Mikolov et al., 2013). We proceed under the following assumptions:

Assumption 1 *Sufficient information in representation.*

$$Y \perp\!\!\!\perp X \mid \mathbf{h}(X) \tag{6}$$

Assumption 1 also implicates the following equation to be true.

$$P(Y = y) = \int P(Y = y \mid \mathbf{h}(X)) \cdot P\left(\mathbf{h}(X)\right) \; d\mathbf{h}(X) \tag{7}$$

This is to say the outcome labels correspond to the representation of the data distribution. For a given representation $\mathbf{h}(X)$,

Assumption 1 asserts that the representation function $\mathbf{h}(X)$ retains all information necessary for predicting the label Y. In other words, classification performance is unaffected by moving from the original input space to the feature space defined by $\mathbf{h}(X)$.

This mirrors the classical notion of a sufficient statistic, where $\mathbf{h}(X)$ summarizes all information in X relevant to the prediction task. Similar assumptions are common in the representation learning literature. For instance, Dai, Shen and Wang (2022) define a learning-adaptive sufficient embedding, requiring that any two inputs with the same representation have the same conditional label distribution. The idea also parallels the information bottleneck principle, which seeks a transformation $\mathbf{h}(X)$ that retains maximal mutual information with Y while compressing X (Tishby, Pereira and Bialek, 2000; Shwartz-Ziv and Tishby, 2017). These frameworks all implicitly assume that $P(Y \mid X)$ can be expressed in terms of $\mathbf{h}(X)$ alone.

Equation 7 relates the marginal distribution of labels to the distribution over representation space. This follows naturally from the law of total probability under Assumption 1. It is also consistent with the cluster assumption in semi-supervised learning, which posits that high-density regions in the input (or representation) space correspond to coherent class labels (Chapelle, Scholkopf and

Zien, 2005; Zhu and Goldberg, 2009). Under this view, label probabilities reflect the structure of the data distribution. Related ideas also appear in domain adaptation, where models often assume that $P(Y \mid \mathbf{h}(X))$ remains stable across domains, so changes in $P(Y)$ are induced by shifts in $P(\mathbf{h}(X))$ (Ben-David et al., 2007; Ganin et al., 2016).

The assumption on representation sufficienty is frequently embedded—explicitly or implicitly—in the design of modern machine learning models for text data. We will also discuss potential scenarios in which the assumption may be violated, and the implications such violations could have for model performance and subset selection.

Assumption 2 *Representation Coverage: The distribution of representations in the optimal subset better approximates the population distribution, on average:*

$$\mathbb{E}_{opt}\left[D_{\mathrm{KL}}\left(P_{\mathbf{h}}^{opt} \,\|\, P_{\mathbf{h}}^{full}\right)\right] \leq \mathbb{E}_{rand}\left[D_{\mathrm{KL}}\left(P_{\mathbf{h}}^{rand} \,\|\, P_{\mathbf{h}}^{full}\right)\right] \tag{8}$$

where $P_{\mathbf{h}}^{opt}$, $P_{\mathbf{h}}^{rand}$, and $P_{\mathbf{h}}^{full}$ are the distributions of $\mathbf{h}(X)$ under the optimal subset, random subset, and full data, respectively.

Assumption 2 states that , on average, the representation distribution in the optimal subset is closer to the population representation than that of a randomly sampled subset. The expectation is taken over draws of subsets. It holds by design of the optimal subset algorithm, as the algorithm is constructed to approximate the underlying distribution of the text data. Specifically, it aims to select a subset S that minimizes the divergence: $\min_S D_{\mathrm{KL}}\left(P_{\mathbf{h}}^{S} \,\|\, P_{\mathbf{h}}^{\text{full}}\right)$, thereby ensuring closer alignment between the subset and the full data distribution in the representation space. While the random sampling technique can get "lucky" and achieve minimal divergence by chance, such outcomes are rare and not guaranteed.

Theorem 1 *Performance Guarantee: Under Assumption 1 and additional conditions, optimal subset performs no worse than random sampling technique.*

We show that, under Assumption 1 , Assumption 2 and one additional regularity assumption, the optimal subset algorithm can often outperform the random sampling technique. In the proof provided in Section A, we also offer further discussion on the plausibility of these additional assumptions, as well as the conditions under which the two algorithms may yield similar performance.

2.1 Violation of Assumptions

We outline several scenarios in which the assumptions may fail to hold. While not exhaustive, these cases highlight common and practically relevant situations where the decomposition breaks down.

Firstly, when the representation is not well aligned with the labels, $\mathbf{h}(X)$ fails to capture the relevant dimensions of variation that correspond to class labels. In such cases, the assumption 7 breaks down. For instance, consider a situation where $\mathbf{h}(X)$ effectively encodes information about the topic or stylistic features of the input, while the label Y depends primarily on sentiment, which is not represented in $\mathbf{h}(X)$. As a result, integrating over $P(\mathbf{h}(X))$ will not recover the correct marginal distribution of the labels, and the model will likely mischaracterize the label distribution under this mismatch. Similarly, the assumption breaks down if there exist hidden confounders outside the representation, such as metadata, author characteristics, or other contextual signals. In particular, if the label assignment depends on information that is not captured in either the input X or the representation $\mathbf{h}(X)$, the decomposition no longer holds.

We showcase this type of violation using a toy simulation example, as illustrated in Figure 2. In this exercise, we generate a synthetic dataset consisting of 1,000 observations, each with two features and a binary label. The label is assigned according to a Gaussian mixture model based on the values of the two features. To examine how misalignment affects the assumption, we systematically vary the level of noise in the relationship between the label Y and the features $X1$ and $X2$. This allows us to simulate scenarios in which the representation fails to capture the relevant signal for predicting the label.

We compare two distinct procedures in this simulation: (1) random selection of the training sample, and (2) the proposed algorithm, which we will describe in detail in a later section. In both cases, we train a simple support vector machine (SVM) classifier to predict labels on the full dataset and evaluate performance using out-of-sample accuracy.

When there is zero noise in the label-generating process, both algorithms achieve comparable accuracy on the test set. However, as we introduce noise—first increasing it from 0% to 30%, and then to 50%—we observe a clear deterioration in performance for both approaches. The degradation is expected, as the task becomes inherently harder with noisier label assignments. Nevertheless,

due to the uninformed nature of random sampling, it is more likely to overlook informative yet rare patterns in the data, especially under high noise conditions. This often results in higher variance and lower reliability in model performance compared to our proposed method, which is explicitly designed to prioritize more representative and label-informative samples. As shown in Figure 2, the "optimal" algorithm—depicted in the right column— achieves higher classification accuracy across varying levels of noise.

Furthermore, Assumption 2 may be violated in the presence of distribution shift—particularly when the data used to estimate $P(\mathbf{h}(X))$ comes from a different domain than the one generating the labels. In such cases, the representation distribution and the label distribution become misaligned. For example, suppose the model is trained on news articles, while the labels are derived from social media posts. Even if the same labeling schema is applied across both domains, the representation distribution $P(\mathbf{h}(X))$ estimated from news content will not accurately capture the distribution underlying the social media data. This mismatch undermines the connection between the learned representation and the true label-generating process, thereby violating the assumption.

Finally, the decomposition may also fail when the data exhibits multi-label or ambiguous instances, indicating that Y is not deterministic given X. In such settings, the same input may validly correspond to multiple labels, introducing uncertainty that is not fully captured by the representation $\mathbf{h}(X)$. As a result, the marginal label distribution may not align cleanly with $P(\mathbf{h}(X))$.

Following the same simulation exercise presented in Figure 2, we provide additional intuition for why a targeted selection algorithm can improve training set selection. In Figure 3, we plot the distribution of decision boundaries across 1,000 runs of SVM models trained using different selection methods. The results show that the optimal subset algorithm, which we will proposed in next section, produces decision points with noticeably lower variance compared to those obtained from random sampling. This highlights the algorithm's ability to yield more stable and consistent performances.

3 Algorithm

Goodfellow et al. (2014) introduced the Generative Adversarial Network (GAN), a widely

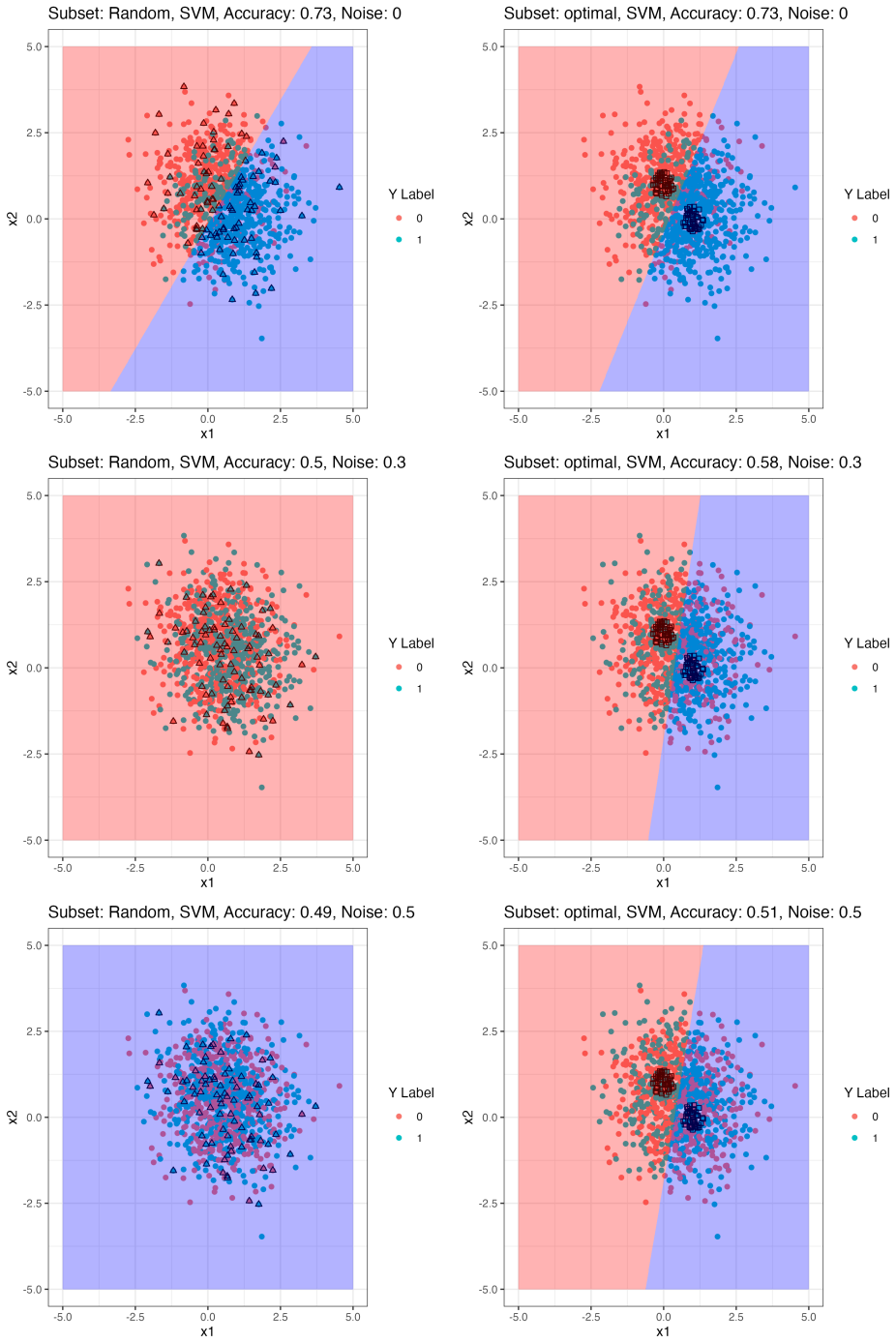

Figure 2: Gaussian mixture model with a binary label, with differing noise level. Top panel: 0% noise; middle panel: 30%*noise*; bottom panel: 50% noise. Triangle or square indicates selected training sample by each algorithm.

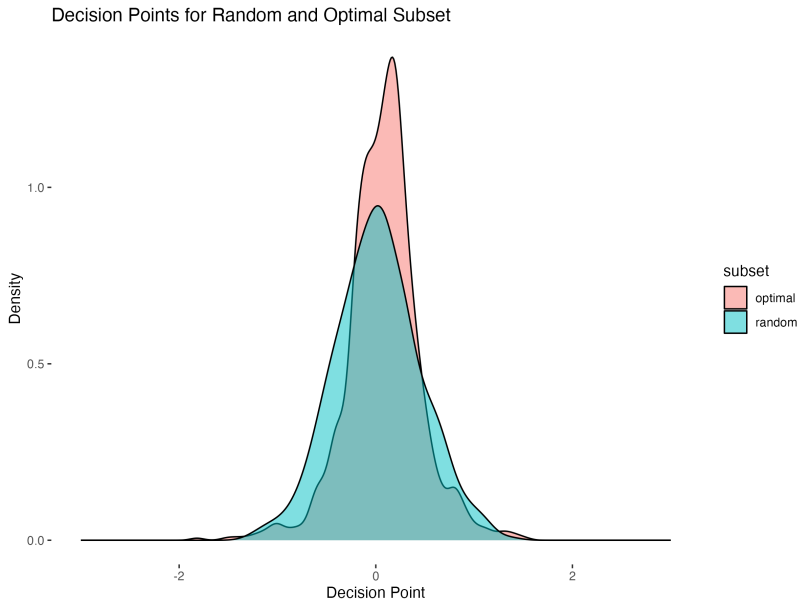

Figure 3: Decision points of SVM for 1000 run.

adopted framework that jointly trains two competing models: a generative model (G) and a discriminative model (D). The role of the generator is to capture and reproduce the underlying data distribution, effectively generating synthetic samples that resemble the real data. In contrast, the discriminator is tasked with distinguishing between genuine data instances and those produced by the generator. The training process is formulated as a two-player minimax game, wherein the generator strives to produce outputs that can successfully fool the discriminator, while the discriminator simultaneously improves its ability to detect synthetic data. The ultimate objective is to reach an equilibrium where the generator produces data that are indistinguishable from the true distribution.

Similarly, the variational autoencoder (VAE) is another generative modeling approach grounded in neural network architectures. It aims to learn meaningful low-dimensional representations—also referred to as latent spaces—that capture the essential structure of the data distribution. VAEs achieve this by encoding inputs into a latent space and then decoding samples from this space back into the data space. This framework allows for both efficient inference and generative sampling from the learned distribution (Kingma, Welling et al., 2019; Pinheiro Cinelli et al., 2021).

For our algorithm, one may adopt either the GAN or VAE framework to effectively model the distribution of all available text data. Both generative approaches are capable of producing or

reconstructing synthetic samples that capture and reflect the underlying structure of the corpus. These generated samples serve as useful tools for approximating the distributional characteristics of the data, particularly when labeled instances are scarce or costly to obtain.

However, a key limitation arises in the interpretability and usability of synthetic text generated by GANs. While these samples function well as distributional proxies, they often lack coherence and semantic clarity, making them difficult to interpret or label reliably—whether by human annotators or automated AI systems.

To address this limitation, we implement a nearest neighbor matching procedure, wherein each synthetic sample is paired with the most similar real document from the original dataset. This strategy enables us to anchor synthetic, distributional representations to actual observed text, thereby identifying real documents that closely resemble representative points in the learned distribution. For this purpose, we adopt the matching approach introduced by Mayer and Timofte (2020), which we formalize in detail in Algorithm 1.

Algorithm 1: Algorithm 1: Optimal Subset with Adversial Learning

Data: Text dataset X: N documents
Result: Optimal subset $\tilde{X}$: $M \leq N$ documents
Compute the feature representation of the dataset:$f(X)$;
while $nrow(\tilde{X}) < M$ **do**
  [1] Select a random subset of data x ;
  [2] Fit GAN or VAE to the subset x ;
  [3] Generate the synthetic sample x' ;
  [4] Compute the feature representation of the synthetic sample:$\tilde{f}(x')$;
  [5] Calculate distance between all documents in X and the synthetic sample $d(f, \tilde{f})$;
  [6] Retrieve top m real samples $\tilde{x}$ with the minimum distance ;
  [7] Add $\tilde{x}$ to the optimal subset $\tilde{X}$
end

In algorithm 1, we outline the proposed adversarial learning procedure for selecting an optimal subset of documents from a larger text corpus. The algorithm begins by computing a lower-dimensional feature representation of the entire dataset using measures chosen by researcher, which serves to capture the most salient variation in the corpus while reducing noise and computational burden. The core of the procedure is an iterative loop, where in each iteration a random subset of the data is sampled and used to train a generative model—either a GAN or a VAE—depending on the implementation. This model learns to approximate the local distribution of the sampled

documents and is then used to generate new synthetic samples that reflect this learned structure.

Once a synthetic sample is generated, its representation is computed and compared to the representations of all real documents in the full dataset. Specifically, the algorithm calculates the distance between the synthetic sample and each document in the corpus, using a chosen distance metric in the feature space. The top m documents with the smallest distances—i.e., those most similar to the synthetic sample—are selected and added to the growing optimal subset $\tilde{X}$. This process repeats until the subset reaches the desired size M. By combining adversarial generation with nearest-neighbor matching, the algorithm is designed to identify a subset of documents that is representative of the underlying data distribution.

3.1 Optimal Sample Selection with Active Learning

In addition to the one-shot selection algorithm, we also explore the integration of active learning into the selection process.

We still consider a dataset consisting of N observations, where each observation is represented by a feature vector $x_i \in \mathbb{R}^P$ for $i \in 1 \dots N$ and is associated with a binary label $y_i \in 0, 1$. A subset of this data is labeled, denoted by $\tilde{X} = (x_1, y_1), \dots, (x_m, y_m)$, and is used to train the initial classifier. In addition to the real data, we generate a set of S synthetic features, denoted by $x'_j \in \mathbb{R}^P$ for $j \in 1 \dots S$, which are produced by a generative model such as a GAN. Predictions are then made on the synthetic features, yielding synthetic labels $\hat{y}'_j \in 0, 1$ for each corresponding x'_j, which are used in the active learning loop to guide the selection of informative real data points for labeling.

Algorithm 2 describes a hybrid approach that integrates generative modeling with active learning to construct an optimal subset $\tilde{X}$ from a larger unlabeled dataset. This procedure, adapted from Mayer and Timofte (2020), begins by fitting a Generative Adversarial Network (GAN) to the full set of real-valued feature representations $x_1, \dots, x_N$, with the aim of generating synthetic data points $x'_1, \dots, x'_S$ that capture the underlying structure of the input space. After the generative model is trained, a small initial set of real samples is labeled to form the initial labeled pool $\tilde{X}_0$. This step seeds the active learning process by providing the classifier with a preliminary basis for training.

In this step, rather than selecting initial documents to be labeled at random, we can apply

Algorithm 2: Optimal Subset + Active (Mayer and Timofte (2020))

Data: Text dataset X: N documents
Result: Optimal subset $\tilde{X}$: $M \leq N$ documents
[Initialization] ;
[1] Fit GAN or VAE to all real features x to generate synthetic features: x';
[2] Label some real data and initialize $\tilde{X}_0$;
[Active + GAN loop] while $|\tilde{X}| < M$ **do**
  [3] Train the classifier using $\tilde{X}_0$;
  [4] Make prediction on the synthetic data to obtain $\{\hat{y}'_1 \dots \hat{y}'_S\}$;
  [5] Compute the uncertainty of the prediction to find the most difficult synthetic data to classify;
  [6] Find real data similar to the most difficult synthetic data to classify found in [5];
  [7] Label the real data found in [6] and update $\tilde{X}$
end

Algorithm 1 to identify the most representative samples from the dataset. This approach leverages the structure of the data distribution to prioritize informative examples, which may result in improved model performance and more efficient use of labeling resources compared to random initialization.

Then, we implement an iterative loop that alternates between classification and subset selection. In each iteration, a classifier is trained using the current labeled set $\tilde{X}$ and then applied to the synthetic samples generated by the GAN or VAE. The classifier's predictions on these synthetic inputs, $\hat{y}'_1, \dots, \hat{y}'_S$, are evaluated for uncertainty—typically using entropy or margin-based measures—to identify the most ambiguous or difficult samples to classify.

In this step, we follow the strategy proposed by Mayer and Timofte (2020) by making predictions on the synthetic data generated by the GAN or VAE. This represents a key difference from classical active learning approaches, such as the one described in Bosley et al. (2022), which typically involve making predictions on all unlabeled real data. When the dataset is large and N is sizable, this can become computationally expensive. Instead, by restricting predictions to a smaller synthetic dataset of size $S < N$, we reduce computational burden while still identifying the most uncertain regions of the feature space. We can then locate real data points that are similar to the synthetic examples with the highest prediction uncertainty. We can compute the uncertainty using the predicted probability of selected samples, $\hat{y}_s = 1$ and $\hat{y}_s = 0$.

These selected real examples are then labeled and added to $\tilde{X}$. This loop continues until the

labeled set reaches the desired size M, ensuring that labels are allocated strategically to the most informative examples.

4 Validation and Application

4.1 Simulation Exercise

We conduct a simulation exercise using a simulated document term matrix to evaluate the performance of our proposed approach. In each run, we apply both random sampling and the proposed algorithm to select the training set. A multinomial naive Bayes classifier is then trained on the resulting dataset, and we assess out-of-sample testing accuracy to compare the two methods.

To further evaluate robustness, we introduce a controlled amount of label noise by deliberately feeding both algorithms a proportion of incorrect labels. This allows us to test their relative resilience to noisy supervision. Compared to the earlier proof-of-concept analysis shown in Figure 2, which relied on a simple Gaussian mixture model, this setup provides a more realistic comparison to both algorithms.

We present the simulation results in Figure 4. From the left to the right panels, we display the comparison under three different levels of label accuracy: 100% correct labels, 90% correct labels, and 70% correct labels, respectively. In each panel, the red dots represent the performance of the random sampling approach, while the blue triangles indicate the results obtained using the optimal subset selection algorithm. The x-axis represents the varying training set size, while the y-axis denotes the corresponding out-of-sample testing accuracy. We show that the proposed algorithm consistently outperforms random sampling, particularly when the training set is small and the labels are correct. This advantage underscores the value of informed subset selection in data-scarce settings: when constrained to a small training set, random sampling tends to introduce higher variance in model performance, whereas a principled selection method can lead to more stable and accurate outcomes.

When Assumption 1 and equation (7) are violated due to the presence of incorrect labels, we observe deteriorating performance for both algorithms. However, the optimal subset algorithm remains competitive and is able to achieve performance comparable to that of random sampling,

even under these challenging conditions.

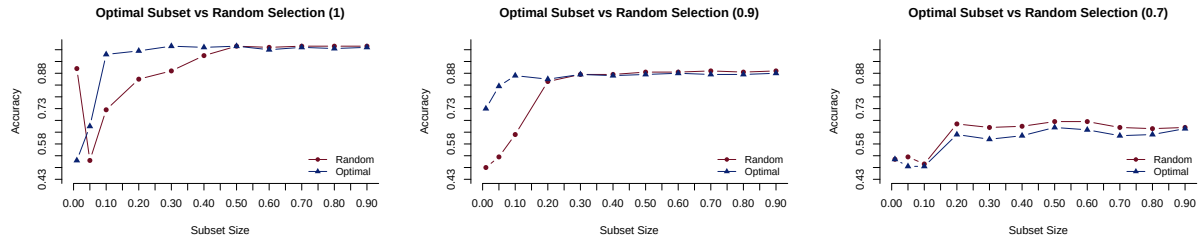

Figure 4: Results from a simulated dataset using a multinomial naive Bayes classifier. From left to right, the panels correspond to datasets with 100% correct labels, 90% correct labels, and 70% correct labels, respectively.

4.2 Human Rights Allegation Dataset

In addition to simulated dataset, we validate our algorithm's performance using the "Human Rights Allegation" dataset (Cordell et al., 2022). This dataset comprises human rights reports collected across 196 countries over a twenty-year span, from 1996 to 2016. Each document in the corpus is annotated with a binary label indicating whether it pertains to allegations of physical integrity rights violations—such as torture, extrajudicial killings, or arbitrary imprisonment.

One important consideration is that the labels in this dataset are highly imbalanced: only 16% of the reports are labeled as involving violations of physical integrity rights. This presents a more stringent test for our algorithm for the following reason: under such imbalance, both the random sampling and optimal subset algorithms face the risk of selecting a training set that underrepresents the minority class. However, the optimal subset algorithm is particularly vulnerable to this issue, as it may favor examples from the majority class that appear more representative of the overall data distribution, thereby increasing the likelihood of being "stuck" with predominantly majority label class. This concern can be addressed by adopting the active learning version of the proposed algorithm (see Algorithm 2), which iteratively refines the training set to better capture difficult examples, including those from underrepresented classes.

We apply our proposed algorithm, alongside the baseline algorithm that selects a random subset of the data, to construct the training sets. Using each selected subset, we then train a predictive model, such as support vector machine (SVM) to classify the documents. Because the

true labels for all texts in the dataset are known, we can objectively evaluate and compare the out-of-sample performance of the models trained under each selection strategy.

For this validation exercise, we vary the training data size from 1% to 40% of the full dataset in order to illustrate performance comparisons across a range of labeling budgets. Figure 5 presents the main results. The x-axis represents the proportion of the dataset used for training, while the y-axis reports the corresponding out-of-sample accuracy achieved by each algorithm. As shown in figure 5,we demonstrate that the optimal subset algorithm outperforms the random sampling technique, yielding higher out-of-sample testing accuracy across various scenarios.

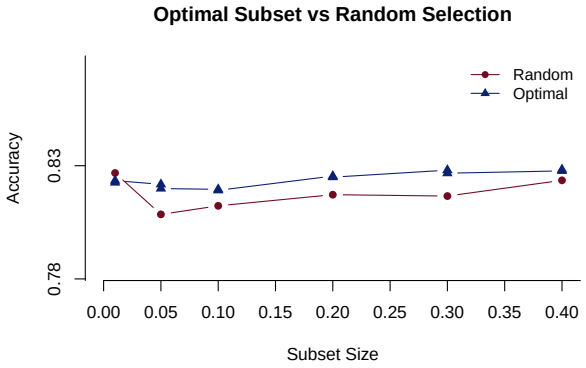

Figure 5: The red dots and line indicate the out-of-sample performance achieved through random sampling, while the blue triangles and line represent the performance obtained using the optimal subset selected by our algorithm.

4.3 BBC

We conduct a similar comparison using the BBC news dataset, first introduced in figure 1. As with the Human Rights dataset, we apply both the optimal subset selection algorithm and random sampling to the BBC corpus in order to construct training sets. We then train a multinomial naive Bayes classifier to predict binary labels for the news articles—distinguishing between political and apolitical content. To evaluate performance, we vary the size of the training set from 1% to 90% of the full dataset and compare the resulting out-of-sample accuracy. As illustrated in figure 6, the x-axis represents the proportion of the data used for training, while the y-axis reports the out-of-sample accuracy for the binary classification task.

The optimal subsample consistently outperforms the random sampling technique across all

training set sizes. Notably, the proposed algorithm becomes more advantageous when the training size is relatively small, highlighting its effectiveness in low-resource settings where selecting the most informative documents is especially critical.

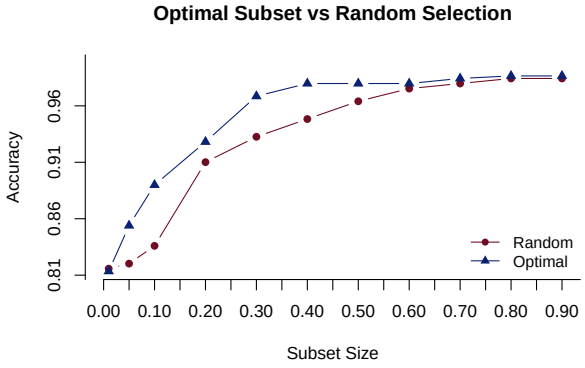

Figure 6: Comparison using the BBC News dataset: Red dots represent results from random sampling, while blue triangles indicate performance using the optimal subset. The classifier used is a multinomial Naive Bayes model for binary labels.

4.4 Multi-class Prediction Task

Finally, we validate our algorithm with two multiclass classification datasets. Multiclass classification is a more challenging task than binary classification, which helps us further evaluate complicated settings

In `SOTU` (Speech of the Union) data, we classify 21,641 quasi-statements from the United States State of the Union presidential addresses into 22 major policy areas (Jones et al., 2023). In `SCOTUS` (Supreme Court of the United States) data, we classify 7,752 case summaries that were argued before the United States Supreme Court into 20 major policy areas (Bird et al., 2009). While both datasets and their expert coding originates from the Comparative Agendas Project, we use the complete and cleaned data from Laurer et al. (2024). Both of the classification tasks involve multi-class outcome variables, which inherently increase the likelihood of violating underlying assumptions. For example, one potential way to violate equation (7) is when certain documents are compatible with multiple label classes, making it unclear which label best captures the intended meaning.

We repeat the same comparison procedure as applied to the Human Rights and BBC datasets.

Specifically, we compare the performance of the optimal subset selection algorithm against random sampling. The results are presented in figure 7, where the left panel displays the comparison on the `SCOTUS` dataset and the right panel on the `SOTU` dataset. As before, we evaluate out-of-sample classification accuracy across varying training set sizes to assess each sampling strategy. As shown in the figure, the optimal subset and random sampling algorithms yield nearly identical classification performance across both datasets. However, the optimal subset algorithm demonstrates a slight advantage when the training set size is relatively small.

One point to highlight is that, due to the complexity of the classification tasks—each involving more than 20 classes—neither algorithm achieves particularly high out-of-sample accuracy. On average, the accuracy remains below 0.5 across both methods. Despite this overall performance limitation, we believe it is still valuable to present these results, as they demonstrate that adopting the optimal subset algorithm remains a safe and reasonable choice even in these settings.

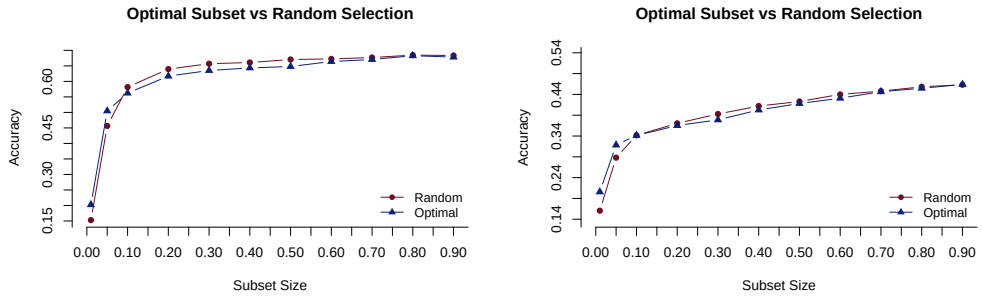

Figure 7: Comparison using the SCOTUS (left panel) and SOTU (right panel) data: Red dots represent results from random sampling, while blue triangles indicate performance using the optimal subset. The classifier used is support vector machine (SVM).

5 Conclusion and Discussion

We propose an algorithm for optimal subset selection by learning the underlying distribution of text data in order to identify the most representative documents for model training. To guide this approach, we introduce a formal framework of assumptions that provides a structured way to think about the problem. The algorithm demonstrates superior performance across a range of settings. The proposed algorithm can be broadly applied to support model training with text data, offering a principled alternative to random sampling and reducing the risk of suboptimal or

unrepresentative training sets.

5.1 Testing the Assumption

We also provide a discussion of potential empirical tests for Assumption 1. One possible way to test this assumption is to train a prediction model for the label Y using $\mathbf{h}(X)$ and compute the corresponding residuals, denoted r_h. We then try to predict r_h using the original input X. If X can predict the residuals with statistical power, this suggests that the residual error from predicting Y with $\mathbf{h}(X)$ is not independent of X, indicating a potential violation of Assumption 1. This approach aligns with the logic of regression-based conditional independence (CI) tests, where both X and Y are regressed on a conditioning variable, and the independence of the resulting residuals is evaluated, under mild assumptions (Zhang, Zhou and Guan, 2018; Zhang et al., 2022; Chernozhukov et al., 2018).

Furthermore, assumption 1 implies that the representation $\mathbf{h}(X)$ retains all label-relevant information from the original input X. We can empirically test this assumption using the concept of mutual information $I(Y; X \mid \mathbf{h}(X))$ (Belghazi et al., 2018; Kraskov, Stögbauer and Grassberger, 2004). This is equivalent to the requirement that the conditional mutual information between Y and X given $\mathbf{h}(X)$ is zero. Intuitively, this means that once we know $\mathbf{h}(X)$, learning X provides no further information about Y.

5.2 Next Steps

Recent advances in large language models (LLMs) have significantly reduced the cost of data labeling by offering cheaper, faster, and often higher-quality annotations. However, rising methodological discussions on the use of LLMs for data annotation urge greater caution (Egami et al., 2024; Fong and Tyler, 2021; Card and Smith, 2018; Chernozhukov et al., 2018; Katsumata and Yamauchi, 2025; Zhao et al., 2021; Ziems et al., 2024). One of the central focuses of the discussion is the prediction error associated with machine-labeled data, and how such error can be calibrated or corrected through small scale human annotations.

The proposed algorithm can potentially be applied in such settings. For example, Egami et al. (2024) introduce a design-based supervised learning (DSL) framework to correct for annotation

errors when using LLM-generated labels. Within the DSL framework, researchers are required to select a small subset of machine-labeled data for validation by human coders or experts. This creates an ideal use case for our subset selection algorithm, which can help identify the most informative or representative examples for expert annotation, thereby improving overall labeling quality and bias-correction efficiency.

Furthermore, the proposed algorithm is not limited to text data; it can be readily extended to other high-dimensional contexts, including image data or video data. In these contexts, representations extracted from pretrained models—such as convolutional neural networks for images or spectrogram-based embeddings for audio—can be used as inputs to the subset selection procedure. The optimal subset algorithm can assist in identifying the most informative and representative training examples, thereby improving the performance of downstream classification tasks.

References

Belghazi, Mohamed Ishmael, Arnaud Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville and R Devon Hjelm. 2018. Mutual Information Neural Estimation. In *International Conference on Machine Learning (ICML)*. pp. 530–539.

Ben-David, Shai, John Blitzer, Koby Crammer and Fernando Pereira. 2007. Analysis of representations for domain adaptation. In *NIPS*.

Ben-David, Shai, John Blitzer, Koby Crammer and Fernando Pereira. 2010. "A theory of learning from different domains." *Machine learning* 79(1):151–175.

Bird, Christine, Michelle Whyman, Bryan D. Jones, Frank R. Baumgartner, Sean M. Theriault, Derek A. Epp, Cheyenne Lee and Miranda E. Sullivan. 2009. "Policy Agendas Project: Supreme Court Cases.".

Bosley, Mitchell, Saki Kuzushima, Ted Enamorado and Yuki Shiraito. 2022. "Improving Probabilistic Models in Text Classification via Active Learning." *arXiv preprint arXiv:2202.02629* .

Brinker, Klaus. 2003. Incorporating diversity in active learning with support vector machines. In *Proceedings of the 20th international conference on machine learning (ICML-03)*. pp. 59–66.

Card, Dallas and Noah A Smith. 2018. The importance of calibration for estimating proportions from annotations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. pp. 1636–1646.

Chapelle, Olivier, Bernhard Scholkopf and Alexander Zien. 2005. *Semi-Supervised Learning*. MIT Press.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey and James Robins. 2018. "Double/debiased machine learning for treatment and structural parameters.".

Cordell, Rebecca, K Chad Clay, Christopher J Fariss, Reed M Wood and Thorin M Wright. 2022. "Disaggregating repression: Identifying physical integrity rights allegations in human rights reports." *International Studies Quarterly* 66(2):sqac016.

Dai, Ben, Xiaotong Shen and Junhui Wang. 2022. "Embedding learning." *Journal of the American Statistical Association* 117(537):307–319.

Egami, Naoki, Musashi Hinck, Brandon Stewart and Hanying Wei. 2024. "Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models." *Advances in Neural Information Processing Systems* 36.

Fong, Christian and Matthew Tyler. 2021. "Machine learning predictions as regression covariates." *Political Analysis* 29(4):467–484.

Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand and Victor Lempitsky. 2016. "Domain-adversarial training of neural networks." *Journal of Machine Learning Research* 17(59):1–35.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Yoshua Bengio. 2014. "Generative adversarial nets." *Advances in neural information processing systems* 27.

Greene, Derek and Pádraig Cunningham. 2006. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd international conference on Machine learning*. pp. 377–384.

Hand, David J. 2006. "Classifier technology and the illusion of progress." *Statistical Science* 21(1).

Jones, Bryan D., Frank R. Baumgartner, Sean M. Theriault, Derek A. Epp, Rebecca Eissler, Cheyenne Lee and Miranda E. Sullivan. 2023. "Policy Agendas Project: State of the Union Speeches.".

Kang, Jaeho, Kwang Ryel Ryu and Hyuk-Chul Kwon. 2004. Using cluster-based sampling to select initial training set for active learning in text classification. In *Advances in Knowledge Discovery and Data Mining: 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, May 26-28, 2004. Proceedings 8*. Springer pp. 384–388.

Katsumata, Hiroto and Soichiro Yamauchi. 2025. "Statistical analysis with machine learning predicted variables.".

Kaufman, Aaron R. 2024. "Selecting More Informative Training Sets with Fewer Observations." *Political Analysis* 32(1):133–139.

Kingma, Diederik P, Max Welling et al. 2019. "An introduction to variational autoencoders." *Foundations and Trends® in Machine Learning* 12(4):307–392.

Kraskov, Alexander, Harald Stögbauer and Peter Grassberger. 2004. "Estimating mutual information." *Physical Review E* 69(6):066138.

Laurer, Moritz, Wouter Van Atteveldt, Andreu Casas and Kasper Welbers. 2024. "Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI." *Political Analysis* 32(1):84–100.

Mayer, Christoph and Radu Timofte. 2020. Adversarial sampling for active learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 3071–3079.

Mikolov, Tomas, Kai Chen, Greg Corrado and Jeffrey Dean. 2013. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* .

Mohri, Mehryar, Afshin Rostamizadeh and Ameet Talwalkar. 2012. *Foundations of machine learning*. MIT press.

Pinheiro Cinelli, Lucas, Matheus Araújo Marins, Eduardo Antúnio Barros da Silva and Sérgio Lima Netto. 2021. Variational autoencoder. In *Variational methods for machine learning with applications to deep networks*. Springer pp. 111–149.

Shwartz-Ziv, Ravid and Naftali Tishby. 2017. "Opening the black box of deep neural networks via information." *arXiv preprint arXiv:1703.00810* .

Tishby, Naftali, Fernando Pereira and William Bialek. 2000. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*.

Wu, Yi, Igor Kozintsev, Jean-Yves Bouguet and Carole Dulong. 2006. Sampling strategies for active learning in personal photo retrieval. In *2006 IEEE International Conference on Multimedia and Expo*. IEEE pp. 529–532.

Xu, Zhao, Kai Yu, Volker Tresp, Xiaowei Xu and Jizhi Wang. 2003. Representative sampling for text classification using support vector machines. In *Advances in Information Retrieval: 25th European Conference on IR Research, ECIR 2003, Pisa, Italy, April 14–16, 2003. Proceedings 25*. Springer pp. 393–407.

Zhang, Hao, Shuigeng Zhou and Jihong Guan. 2018. "Measuring Conditional Independence by Independent Residuals: Theoretical Results and Application in Causal Discovery." *Artificial Intelligence* . preprint from 2017.

Zhang, Hao, Shuigeng Zhou, Kun Zhang and Jihong Guan. 2022. "Residual Similarity Based Conditional Independence Test and Its Application in Causal Discovery (SCIT)." *AAAI Conference on Artificial Intelligence* .

Zhao, Zihao, Eric Wallace, Shi Feng, Dan Klein and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*. PMLR pp. 12697–12706.

Zhu, Xiaojin and Andrew B Goldberg. 2009. "Introduction to semi-supervised learning." *Synthesis lectures on artificial intelligence and machine learning* 3(1):1–130.

Ziems, Caleb, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang and Diyi Yang. 2024. "Can large language models transform computational social science?" *Computational Linguistics* 50(1):237–291.

A Proof

Below we show a proof for theorem 1. Suppose we are comparing the performance of random sampling and the proposed optimal subset algorithm for selecting n training samples from the dataset.

Let f_{rand} and f_{opt} denote the classifiers trained on the random and optimal subsets, respectively. Our goal is to demonstrate that, under a set of suitable assumptions, the classifier trained on the optimal subset achieves superior or at least comparable out-of-sample performance relative to the one trained on a randomly selected subset.

$$\mathbb{E}_{X,Y}[\ell(f_{\text{opt}}(\mathbf{h}(X)), Y)] \leq \mathbb{E}_{X,Y}[\ell(f_{\text{rand}}(\mathbf{h}(X)), Y)]$$

where ℓ is the classification loss (e.g., cross-entropy).

Before presenting the proof, we introduce one additional assumption that are necessary for establishing the desired performance guarantee.

Assumption 3 *Loss Continuity: The classification loss $\ell(f(\mathbf{h}(X)), Y)$ is Lipschitz continuous with respect to changes in the distribution of $\mathbf{h}(X)$.*

Again, we emphasize that all three assumptions—Assumptions 1 through 3—are either explicitly stated or implicitly relied upon in many existing text classification methods.

Claim. Under assumptions 1, 2 and 3,

$$\mathbb{E}_{X,Y}[\ell(f_{\text{opt}}(\mathbf{h}(X)), Y)] \leq \mathbb{E}_{X,Y}[\ell(f_{\text{rand}}(\mathbf{h}(X)), Y)]$$

Proof.

Our goal is to compare the population risk of classifiers trained on the optimal and random subsets:

$$\mathcal{L}(f) := \mathbb{E}_{X,Y}[\ell(f(X), Y)] = \mathbb{E}_{X,Y}[\ell(f(\mathbf{h}(X)), Y)].$$

This equality holds by Assumption 1.

Let $\epsilon_{\text{full}}(f)$ denote the expected classification loss of classifier f on the full distribution:

$$\epsilon_{\text{full}}(f) := \mathbb{E}_{X,Y\sim P}\left[\ell(f(\mathbf{h}(X)), Y)\right].$$

We compare the expected performance of f_{rand} and f_{opt}, both trained on subsets of size n. Consider the bound in Theorem 2 of Ben-David et al. (2010):

$$\epsilon_{\text{full}}(f) \leq \epsilon_{\text{train}}(f) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(P_{\mathbf{h}}^{\text{train}}, P_{\mathbf{h}}^{\text{full}}) + \lambda. \tag{9}$$

One point to note is that this result relies on Assumption 3 to be true.

Applying this bound to both classifiers:

$$\mathbb{E}_{\text{opt}}[\epsilon_{\text{full}}(f_{\text{opt}})] \leq \mathbb{E}_{\text{opt}}[\epsilon_{\text{train}}(f_{\text{opt}})] + \frac{1}{2}\mathbb{E}_{\text{opt}}\left[d_{\mathcal{H}\Delta\mathcal{H}}(P_{\mathbf{h}}^{\text{opt}}, P_{\mathbf{h}}^{\text{full}})\right] + \lambda,$$
$$\mathbb{E}_{\text{rand}}[\epsilon_{\text{full}}(f_{\text{rand}})] \leq \mathbb{E}_{\text{rand}}[\epsilon_{\text{train}}(f_{\text{rand}})] + \frac{1}{2}\mathbb{E}_{\text{rand}}\left[d_{\mathcal{H}\Delta\mathcal{H}}(P_{\mathbf{h}}^{\text{rand}}, P_{\mathbf{h}}^{\text{full}})\right] + \lambda.$$

For example, the above inequality states that on average (over many runs algorithm), the expected population error of f_{opt} is no more than: its average training error (on the subset), plus half the average $\mathcal{H}\Delta\mathcal{H}$ divergence between the representation distribution of the optimal subset and the full data, plus the best possible joint error λ (which is a constant lower bound for both classifiers, not depending on the training sample).

Assume the learning algorithm reaches comparable in-sample training loss in both settings:

$$\mathbb{E}_{\text{opt}}[\epsilon_{\text{train}}(f_{\text{opt}})] \approx \mathbb{E}_{\text{rand}}[\epsilon_{\text{train}}(f_{\text{rand}})].$$

Further assume that $d_{\mathcal{H}\Delta\mathcal{H}}$ is upper bounded by a multiple of D_{KL} (as shown in prior work under smooth loss and bounded VC dimension; see Mohri, Rostamizadeh and Talwalkar (2012), Chapter 3):

$$d_{\mathcal{H}\Delta\mathcal{H}}(P, Q) \leq C \cdot D_{\text{KL}}(P \,\|\, Q).$$

Then:

$$\mathbb{E}_{\text{opt}}[\epsilon_{\text{full}}(f_{\text{opt}})] \leq \mathbb{E}_{\text{rand}}[\epsilon_{\text{full}}(f_{\text{rand}})]$$

follows immediately from Assumption 2, since the divergence term is smaller on average for the optimal subset. ■

B More on Simulation

Simulated DTMs were generated to allow us to test the results of our algorithms given differing levels of label accuracy, as this can frequently cause challenges when classifying texts with standard approaches. We generated simulated DTMs containing 1000 words and 1000 documents, with 500 documents assigned to one class and the other 500 assigned to another. We assigned labels to each document with differing levels of accuracy: 100

We also examine the stability of each algorithm by running multiple trials. Figure 8 presents the results using the same simulated data as in the main text. The x-axis represents the subset size, while the y-axis shows the out-of-sample testing accuracy.

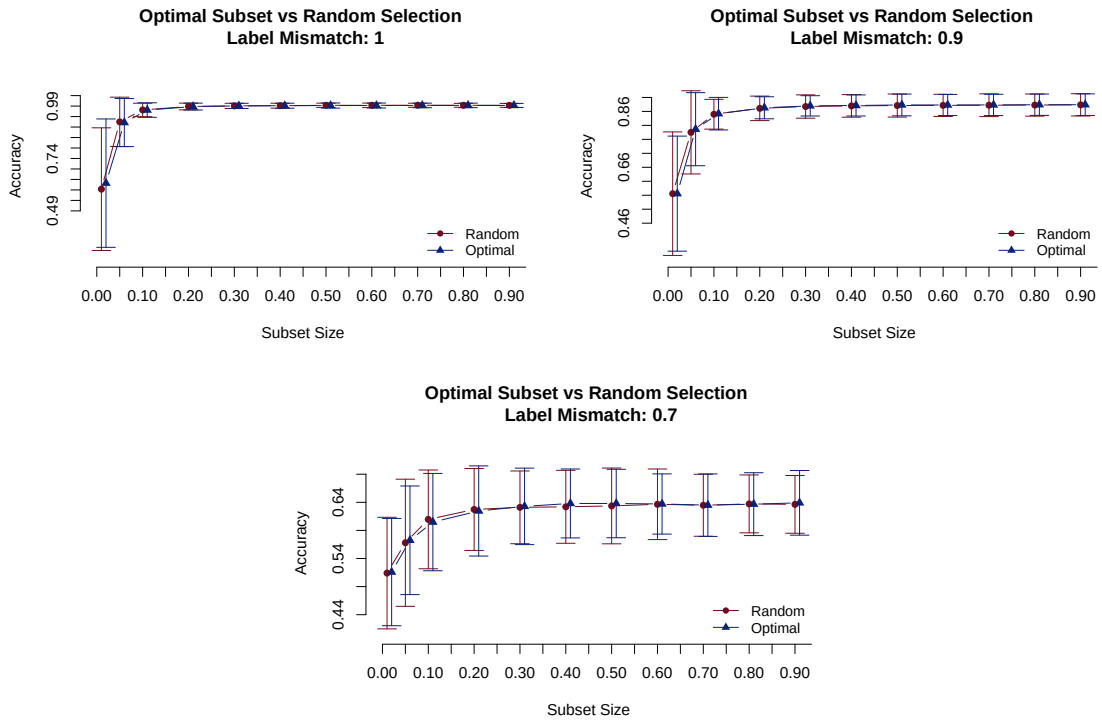

Figure 8: Stability over multiple runs for each algorithm. Top left panel: 100% correct labels; top right panel: 90% correct labels and bottom panel: 70%. Classifier used is support vector machine.