

# infoVerse: A Universal Framework for Dataset Characterization with Multidimensional Meta-information

Jaehyung Kim<sup>†\*</sup> Yekyung Kim<sup>‡</sup> Karin de Langis<sup>◇</sup> Jinwoo Shin<sup>†</sup> Dongyeop Kang<sup>◇</sup>  
<sup>†</sup>KAIST, <sup>‡</sup>Hyundai Motors Company, <sup>◇</sup>University of Minnesota  
jaehyungkim@kaist.ac.kr dongyeop@umn.edu

## Abstract

The success of NLP systems often relies on the availability of large, high-quality datasets. However, not all samples in these datasets are equally valuable for learning, as some may be redundant or noisy. Several methods for characterizing datasets based on model-driven meta-information (e.g., model’s confidence) have been developed, but the relationship and complementary effects of these methods have received less attention. In this paper, we introduce infoVerse, a universal framework for dataset characterization, which provides a new feature space that effectively captures multidimensional characteristics of datasets by incorporating various model-driven meta-information. infoVerse reveals distinctive regions of the dataset that are not apparent in the original semantic space, hence guiding users (or models) in identifying which samples to focus on for exploration, assessment, or annotation. Additionally, we propose a novel sampling method on infoVerse to select a set of data points that maximizes informativeness. In three real-world applications (data pruning, active learning, and data annotation), the samples chosen on infoVerse space consistently outperform strong baselines in all applications. Our code and demo are publicly available.<sup>1</sup>

## 1 Introduction

The construction of large datasets is one of the essential ingredients for success in various NLP tasks (Wang et al., 2019). However, not all data points are equally important to learn from; many datasets often contain low-quality samples, e.g., incorrect labels (Toneva et al., 2019) or annotation artifacts (Gururangan et al., 2018). Thus, *data characterization* (Roth and Mattis, 1990), a technique for transforming raw data into useful information for

<sup>\*</sup>This work was done while JK was at Minnesota NLP.

<sup>1</sup><https://github.com/minnesotanlp/infoVerse>

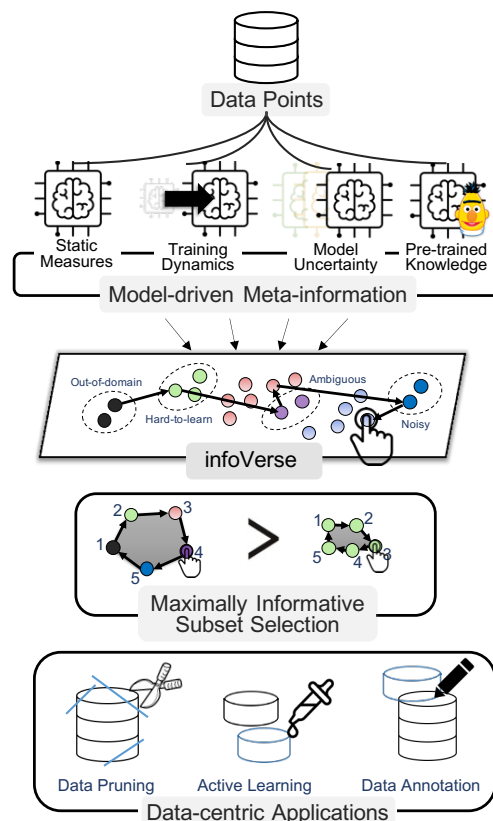


Figure 1: The proposed framework, **infoVerse**. By incorporating multidimensional aspects of data characteristics, infoVerse enables a better dataset characterization. By selecting maximally informative subsets on infoVerse, we improve model performance on a variety of data-centric real-world problems like active learning.

a target task, has a huge potential to improve the model’s performance by trimming the problematic samples (Pleiss et al., 2020) or providing better practices for effective data collection, e.g., active learning (Beluch et al., 2018) and adversarial annotation (Nie et al., 2019).

However, data characterization via human assessment is highly limited due to the huge cost of dealing with a large dataset and the vagueness of the assessment itself. To this end, several model-

driven *meta-information*<sup>2</sup> have been investigated; for example, the model’s confidence is a standard meta-information widely used in active learning (Beluch et al., 2018). Swayamdipta et al. (2020) recently show that the training dynamics of the model’s prediction can indicate the relative importance of training samples. Various types of meta-information are continuously proposed from different intuitions (Salazar et al., 2020; Paul et al., 2021), but their relationship and potential beyond relying on individual one have yet to be explored. Hence, this work answers the following two research questions: (1) *is there a (hidden) complementary effect between various meta-information for better data characterization*, and (2) *is the combined meta-information useful for real-world applications?*

In this paper, we introduce **infoVerse**: a universal framework for better dataset characterization by incorporating multiple aspects of data characteristics. To be specific, infoVerse combines various types of meta-information which offer the different aspects of data characteristics (e.g., how difficult the sample is to learn, how certain multiple models are, and how likely the sample is). Consequently, we can extract richer information about data informativeness from their complementary effect, and infoVerse could guide users (or models) in what samples to focus on for the exploration, assessment, or annotation. To extend the advantages of infoVerse into real-world problems, we further propose a novel sampling method suitable for infoVerse based on determinantal point processes (DPP), which is known to be effective for finding a diverse and high-quality set of samples (Gillenwater et al., 2012; Chen et al., 2018). It enables us to select data points that maximize the information at a set level rather than a sample level on the multidimensional space of infoVerse.

In detail, we first construct infoVerse based on the diverse meta-information, which could be broadly classified into four different categories in Section 3. The complementary effect from the multiple meta-information in infoVerse helps reveal distinct regions in the dataset, such as *hard-to-predict* and *mis-labeled* ones, which are not observable in the original semantic feature space (Section 4). In Section 5, we empirically show that our framework has consistently outperformed the strong baselines in various data-centric applications, like *data prun-*

*ing* (Toneva et al., 2019; Paul et al., 2021), *active learning* (Yuan et al., 2020), and *data annotation* (Xie et al., 2020a), although it is not specifically designed for those problems. This result opens up the potential of infoVerse to other data-centric applications, unlike the application-specified approaches.

Our results show that a dataset could be distinctively characterized when *many different but complementary dimensions are considered together*. We believe that our infoVerse framework could evolve continuously with the development of new meta-information and hence serve as an effective platform for better characterization of datasets and construction of high-quality datasets.





## 2 Related Works

**Quantifying and characterizing dataset.** Although the large quantity of datasets is usually got attention for the success of various NLP tasks, the quality of the dataset is also an important factor. While constructing data with human-in-the-loop is quite reliable like Dynabench (Kiela et al., 2021a), it is expensive and laboring. Hence, some works show the benefits of using a model for quantifying and characterizing datasets; for example, Rodriguez et al. (2021) demonstrates that the model has the ability to annotate, detect annotation errors, and identify informative examples. In this line of work, several model-driven *meta-information* have been proposed (Toneva et al., 2019; Swayamdipta et al., 2020; Beluch et al., 2018), as we provide the detailed explanation in Section 3 and Appendix A. Most of the prior works focuses on finding a new meta-information; however, as they are obtained under different intuition and aspects of data characteristics, one can expect the complementary effect between them to provide richer information about the dataset. Such direction is under-explored from now on, and we try to fill this gap in this work.

**Informative subset selection.** Finding an informative subset is key for various real-world applications; for example, active learning requires selecting the most informative samples among unlabeled samples for labeling (Settles, 2009). The most widely used approaches to finding those samples in active learning are based on *uncertainty*. Although several uncertainty measurements have been successfully applied in various NLP tasks, Dasgupta (2011) pointed out that focusing only on the uncertainty leads to a sampling bias with repetitive patterns. To this end, *diversity*-based

<sup>2</sup>Measurements that quantify implicit information in data points, such as their difficulty to learn.

Table 1: Categorization of used meta-information.

Categories	Meta-information
 <b>Static Measures</b>	Entropy, Confidence, BADGE Task Density, Relative Density score
 <b>Training Dynamics</b>	Average Confidence, Variability Forgetting Number, Area Under Margin
 <b>Model Uncertainty</b>	EL2N score, BALD, Variance Ratio
 <b>Pre-trained Knowledge</b>	Sentence Density, Pseudo-log-likelihood

sampling has been explored (Sener and Savarese, 2018). However, as this approach might select samples that provide little new information, recent works suggest methods combining uncertainty and diversity to take advantage of both methods (Ash et al., 2020; Yuan et al., 2020). Our work provides a better way to select informative samples by effectively incorporating multiple aspects of data characteristics with a single universal framework.

### 3 infoVerse: Universal Framework for Multi-aspect Data Characterization

In this section, we present *infoVerse*, a universal framework for better data characterization. Our high-level idea is extracting the complementary effect between various meta-information, as they are oriented from the different aspects of data characteristics. In Section 3.1, we briefly introduce the used meta-information to construct infoVerse. In Section 3.2, we present a novel sampling method to extend the advantages of infoVerse for solving real-world problems. We remark that our framework can be easily extended with a new meta-information and not limited to specific ones, while the fixed ones are used for the experiments.

#### 3.1 Meta-information for infoVerse

To construct infoVerse, one needs to determine *which meta-information to use*, and it is expected to get better capability for data characterization with diverse meta-information. Hence, we first conduct an extensive investigation of the existing meta-information and find that they could be categorized into four different classes based on how they extract the data characteristics: (1) *Static Measures*, (2) *Training Dynamics*, (3) *Model Uncertainty*, and (4) *Pre-trained Knowledge*. In Table 1, we provide the full list of meta-information for each category, and more details are presented in Appendix A.

**Static Measures.** Meta-information obtained from a single static classifier is arguably the easiest and most natural to use. *Confidence* to true

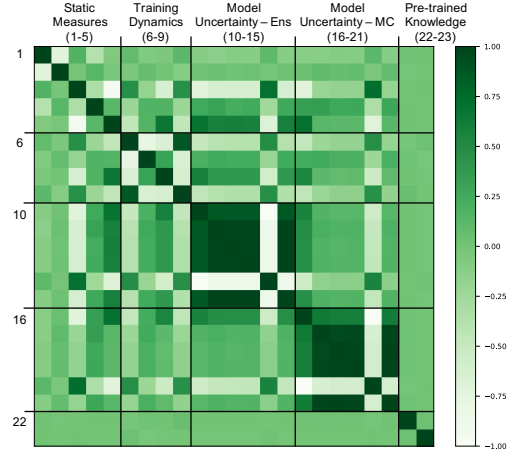


Figure 2: Correlation between meta-information considered in infoVerse on QNLI dataset.

label and *Entropy* (Shannon, 1948) of the predictive probability from the classifier’s output have been popularly used to characterize each data point. *BADGE score* (Ash et al., 2020), a gradient norm of the linear classifier with respect to training loss, is designed to capture the uncertainty of the prediction. Finally, *Task Density* and *Relative Density* defined with kNN distance on the classifier’s feature embeddings (DeVries et al., 2020) effectively estimates the uniqueness of data at a task level.

**Training Dynamics.** Training dynamics of samples largely varies depending on the samples’ characteristics and hence can provide useful information, *e.g.*, the confidence of mislabeled samples slowly increases relative to the normal ones. Swayamdipta et al. (2020) investigate the usefulness of the mean (*Confidence*) and standard deviation (*Variability*) of the model’s prediction to true class across training epochs; they observe that high variable samples are usually useful and low confident ones have a risk of the noisy label. *Forgetting Number* (Toneva et al., 2019), a number of the transitions from being classified correctly to incorrectly during training, is also shown to be an effective measurement for finding redundant samples. Pleiss et al. (2020) reveals that *Area Under Margin* (AUM), a sum of the gap between the logits of true class and most confusing class across training epochs, is different among easy, hard, and mislabeled samples.

**Model Uncertainty.** As intrinsic randomness within the model affects the samples’ prediction, such uncertainty has been widely used in various fields (Lakshminarayanan et al., 2017; Lee et al., 2021b). There are two popular ways to measure the model uncertainty: Monte-Carlo Dropout (MC-

*Dropout*) (Gal and Ghahramani, 2016) with different Dropout masks for a single model, and *Deep Ensembles* (Lakshminarayanan et al., 2017) with multiple models differently random initialized. Specifically, the following four meta-information are used for uncertainty quantification: 1) *Entropy* of the average predicted class distribution of multiple predictions, 2) *BALD* (Houlsby et al., 2011): mutual information between data samples and classifier, 3) *Variation Ratio* (Beluch et al., 2018): a proportion of predictions different with the majority voted one, and 4) *EL2N score* (Paul et al., 2021): an approximated contribution to the change of training loss. In addition, we also include the average and variability of confidence across different models.

**Pre-trained Knowledge.** As general text representation provides complementary information to task’s one, using the pre-trained language models to extract meta-information is another popular direction. For example, *MLM score* (Salazar et al., 2020), a Pseudo-Log-Likelihood (Wang and Cho, 2019) of Masked Language Model (MLM), gives low values to the sentences with inconsistent context. To reduce the computational cost, we use its approximation following Yuan et al. (2020). Also, *Semantical Density* of each sample based on kNN distance (DeVries et al., 2020) using sentence-BERT (Reimers and Gurevych, 2019) can assess its uniqueness compared to other sentences.

Overall, with 23 meta-information, we construct a new feature space *infoVerse*. Note that some complementary or redundant meta-information is noticed by low or high correlation values in Figure 2, respectively. Remarkably, we observe that similar correlations consistently appear across different datasets and models, meaning that this "meta"-information is quite a dataset- and task-agnostic (see Appendix D).

### 3.2 Maximally-Informative Subset Selection

Consideration of multiple meta-information via *infoVerse* enables better data characterization, but it’s non-trivial to apply this framework to practical problems, such as data pruning and active learning. One of the main challenges is from the multidimensional nature of *infoVerse*, as it requires a new sampling method rather than existing single-score based sample selections; for example, Beluch et al. (2018); Swayamdipta et al. (2020) choose top-ranked samples ordered by a specific meta-information like confidence or uncertainty but such

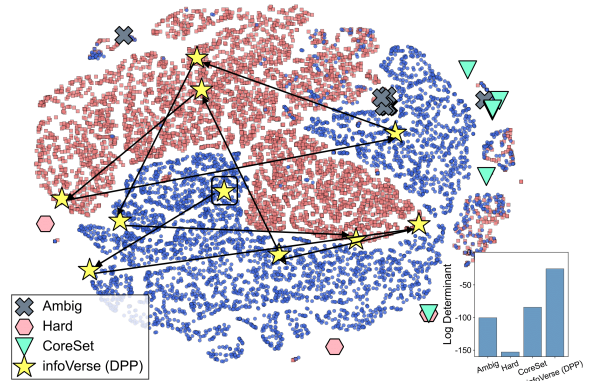


Figure 3: 10 selected samples with different selection methods on QNLI dataset and their log determinant as a proxy measure for set-informativeness.

ordering is hard to be defined in multidimensional space. On the other hand, the single-score strategy cannot capture the relationship between the selected samples in the subset; hence, it suffers from the lack of diversity, especially when the size of the subset is small (see Figure 3). Lastly, the manual verification of the effectiveness of each feature becomes very costly when multiple features are considered. Motivated by this, we provide a new effective subset selection method for *infoVerse* based on *determinantal point process* (DPP). To be specific, we propose to focus on maximizing the informativeness of the subset by leveraging the capability of *infoVerse* for data characterization at a set level; here, DPP provides a way to easily find the effective sampling method by defining the appropriate score and similarity functions.

**Determinantal point processes.** Formally, a DPP on a set of samples  $\mathcal{X} = \{x_1, \dots, x_N\}$  is a probability measure  $\mathcal{P}$  on  $2^{\mathcal{X}}$ , the set of all subsets of  $\mathcal{X}$ . Under the assumption that  $\mathcal{P}$  gives a nonzero probability to the empty set, the probability of each subset  $X \subseteq \mathcal{X}$  is  $\mathcal{P}(X) \propto \det(L_X)$  where  $L \in \mathbb{R}^{N \times N}$  is a real, positive semidefinite (PSD) kernel matrix, and  $L_X$  denotes the sub-matrix of  $L$  which is indexed with the subset  $X$ . Here, we can define the entries of  $L$  as follows:

$$L_{ij} = q(x_i)\phi(x_i)^T\phi(x_j)q(x_j) \quad (1)$$

where  $q(x) \in \mathbb{R}^+$  is a *score* of sample  $x$  which is used to weight the samples with a high quality or desired property such as high confidence or uncertainty. Next,  $S_{ij} = \phi(x_i)^T\phi(x_j) \in [-1, 1]$  represents a *similarity* between samples  $x_i$  and  $x_j$  with a normalized feature vector  $\phi(x) \in \mathbb{R}^d$ ,  $\|\phi(x)\|_2 = 1$ . We note that the determinant  $\det(L_X)$  is pro-

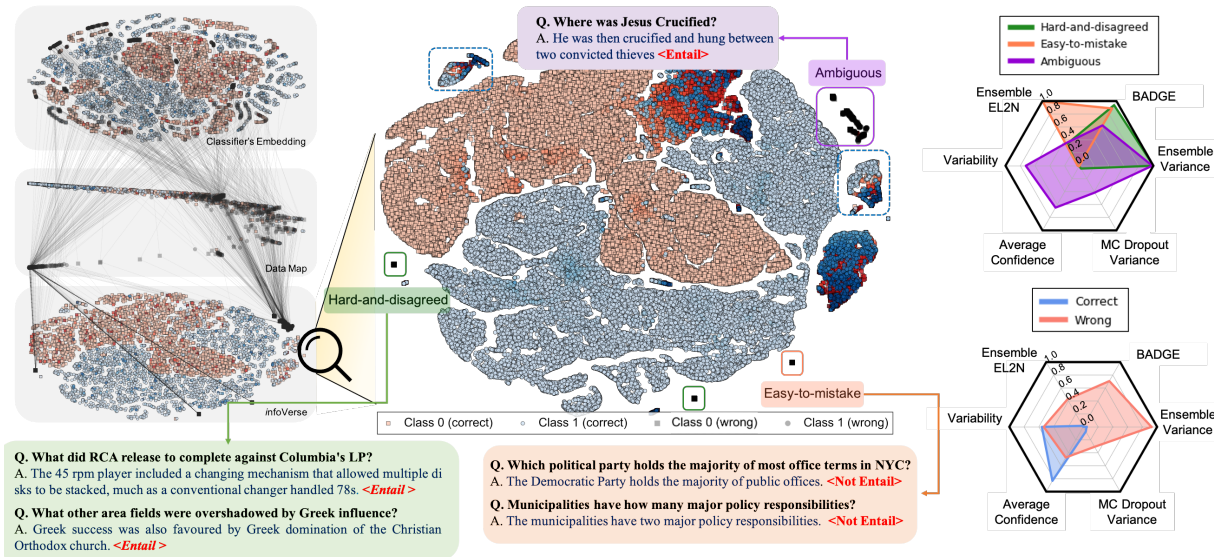


Figure 4: infoVerse (bottom left) on QNLI dataset along with other feature spaces: classifier’s embedding (top left) and *data map* (Swayamdipta et al., 2020) (middle left). (center) We present the zoomed version of infoVerse with some examples of each distinctive region. We set high chroma to high variability samples for better interpretation. (right) Score distribution of each region characterized by infoVerse.

portional to the volume spanned by the vectors  $q(x)\phi(x)$  for  $x \in X$ , and hence the sets with high-quality and diverse samples have the high probability under distribution from DPP. Consequently, DPP provides a natural way to find the maximally-informative subset by selecting the subset with the highest probability among the sets with the same number of samples (Kulesza and Taskar, 2011).

To apply DPP on infoVerse, we consider the following design choices. For  $S_{ij}$ , we use a Gaussian kernel with Euclidean distance (Biyik et al., 2019) on a normalized features  $\tilde{x}$  on infoVerse:

$$S_{ij} = \exp(-\beta \|\tilde{x}_i - \tilde{x}_j\|^2) \quad (2)$$

where we use a fixed value  $\beta = 0.5$ . Regarding score  $q(x)$ , we use a density defined by kNN distance (Carbonera and Abel, 2015)  $D_{\text{KNN}}$  to the same class’ samples on infoVerse for data pruning:

$$D_{\text{KNN}}(x) = -\min_K \{\|\hat{x} - x\|_2 \mid \hat{x} \in \mathcal{X}\} \quad (3)$$

where  $\min_K \{\cdot\}$  is defined as the  $K$ th smallest value in a set. In our experiments, we commonly set  $K = 5$ . As  $D_{\text{KNN}}$  has a negative value, we use its negative inverse to define a positive score function, i.e.,  $q(x) = -1/D_{\text{KNN}}$ . Intuitively, it encourages selecting the samples that preserve the informativeness captured on infoVerse as much as possible. For active learning and data annotation, we use its inverse as  $q(x)$  to select samples that have information hard to be captured by their neighbor

and hence will be beneficial when labeled. Finally, we adopt the efficient greedy method (Chen et al., 2018) for DPP, as finding a set with the highest probability is NP-hard.

Figure 3 shows the 10 selected samples with different selection methods on the QNLI training dataset with a fine-tuned RoBERTa-large classifier. Here, one can observe that single-score based selection methods like *Ambig* and *Hard* (Swayamdipta et al., 2020) actually suffer from the lack of diversity. CoreSet (Sener and Savarese, 2018) or K-means clustering can select diverse samples, but they are known to be vulnerable to the existence of outliers (Georgogiannis, 2016). In contrast, DPP successfully selects informative and diverse samples; as shown in the right below of Figure 3, the log determinant with DPP, i.e., approximate set-informativeness, is much higher than the others.

## 4 Dataset Characterization via infoVerse

In this section, we demonstrate *how infoVerse could help analyze a given dataset via better data characterization*. Specifically, Figure 4 presents infoVerse on QNLI dataset<sup>3</sup> along with other representative feature spaces for data characterization: *classifier’s embedding* (at final layer before linear head) and *data map* (Swayamdipta et al., 2020). As the classifier’s embedding and infoVerse are high dimensional space, we project them to 2D space via t-

<sup>3</sup>Natural language inference dataset derived from SQuAD.

SNE (Van der Maaten and Hinton, 2008) for visualization. First, one can observe that infoVerse maps the samples into distinguishable regions based on their characters; for example, samples with high *variability* are further mapped to some different regions. To be specific, in Figure 4, they have high variability in a common while having a difference in other measures: the regions with dotted boxes have relatively high *Ensemble Entropy* and *MC Entropy*, respectively, which cannot be distinguishable in the data map, showing the advantage of infoVerse in dataset characterization.

This benefit of infoVerse for dataset characterization is more clear when we focus on the incorrectly predicted samples (black squares/circles in Figure 4). As shown in the left of Figure 4, it is hard to find their characteristics on the classifier’s embedding as they are scattered over different regions. Data map (Swayamdipta et al., 2020) maps these samples to regions with low confidence (hard-to-learn) or high variability (ambiguous), but not perfectly as these regions also include correctly predicted samples. In contrast, infoVerse successfully characterizes these incorrectly predicted samples and maps them into three distinct regions with a different distribution of meta-information: 1) *Hard-and-disagreed*, 2) *Easy-to-mistake*, and 3) *Ambiguous*. As shown in the right of Figure 4, both *Hard-and-disagreed* and *Ambiguous* regions have high ensemble uncertainty, but *Hard-and-disagreed* region has relatively low confidence and variability which means that it is also *hard to learn*. It might imply its incorrect prediction due to intrinsic difficulty as one can verify in the examples in Figure 4. In contrast, *Easy-to-mistake* region has much lower uncertainty than other incorrectly predicted regions, which indicates the prediction is certainly wrong. It might indicate that the mistakes happened during annotation even though they are easy ones to annotate correctly. More results of data characterization on other datasets with infoVerse are presented in Appendix D.

## 5 infoVerse for real-world applications

In this section, we demonstrate the advantages of infoVerse on three real-world problems: 1) *Data Pruning* (Swayamdipta et al., 2020), 2) *Active Learning* (Beluch et al., 2018), and 3) *Data Annotation* (Kiela et al., 2021b). These problems commonly *require characterizing and quantifying data to determine which samples to select*, but with dif-

ferent goals; hence, various methods have been explored separately. In contrast, we will demonstrate the potential of infoVerse as a universal framework to deal with such data-centric problems.

### 5.1 Data Pruning

The goal of data pruning is selecting the most informative subset of a given training dataset while keeping the performance of the model trained on the subset; hence, measuring the sample’s informativeness becomes a key for data pruning. This problem has been popularly investigated with various meta-information as an important problem for improving the efficiency of various NLP tasks.

**Setups.** For the experiments of data pruning, we first use two datasets, QNLI (Wang et al., 2019) and WinoGrande (Sakaguchi et al., 2020), following the recent work (Swayamdipta et al., 2020). Then, we use three additional datasets, SST-2 (Socher et al., 2013), CoLA (Warstadt et al., 2019), and RTE (Wang et al., 2019), for the comprehensive demonstration. We consider 8 different pruning ratios ( $\{17\%, 34\%, 50\%, 67\%, 75\%, 83\%, 87\%, 91\%\}$ ), which includes more challenging setups compared to the previous works (Swayamdipta et al., 2020; Paul et al., 2021). We run all experiments by fine-tuning RoBERTa-large (Liu et al., 2019), following (Swayamdipta et al., 2020).

To demonstrate the effectiveness of infoVerse-DPP, we compare it with various state-of-the-art approaches to data pruning. We first consider a random-sampling (*Random*); then, we consider three different approaches in (Swayamdipta et al., 2020) (*Easy*, *Hard*, and *Ambig*), which selects the samples by scoring them with a specific meta-information (average confidence and variability). In addition, we introduce two additional data pruning works: *Forget* (Toneva et al., 2019) and *EL2N* (Paul et al., 2021). Finally, we consider density-based approaches as they are arguably the most natural ways to preserve the characteristics of the dataset: *Coreset* (Sener and Savarese, 2018) and *Density* (Yuan et al., 2020). More details of datasets and training can be found in Appendix B.2.

**Results.** Figure 5 shows the performance under varied pruning ratios on WinoGrande (see Appendix C for the other tasks). We first note that the effectiveness of each pruning method significantly varies on the pruning ratio. For example, *Hard* and *Ambig* show good performance at small pruning ratios, but they often fail at large pruning ratios, simi-

Table 2: Average test accuracy of fine-tuned RoBERTa-large over the eight different data pruning ratios.

Dataset	Random	Easy	Hard	Ambig	Forget	EL2N	Coreset	Dense	infoVerse-DPP
WinoGrande	73.1 $\pm$ 0.09	68.1 $\pm$ 0.18	69.6 $\pm$ 0.31	69.3 $\pm$ 0.52	73.8 $\pm$ 0.77	72.2 $\pm$ 1.02	73.9 $\pm$ 0.14	72.9 $\pm$ 0.31	<b>74.6<math>\pm</math>0.24</b>
CoLA	60.7 $\pm$ 0.63	32.2 $\pm$ 0.94	41.1 $\pm$ 0.16	41.0 $\pm$ 0.73	59.1 $\pm$ 0.54	47.6 $\pm$ 1.07	55.5 $\pm$ 0.93	61.2 $\pm$ 0.43	<b>62.5<math>\pm</math>0.14</b>
RTE	76.7 $\pm$ 0.75	73.5 $\pm$ 0.21	56.2 $\pm$ 1.59	56.3 $\pm$ 2.26	71.0 $\pm$ 0.93	61.4 $\pm$ 0.81	56.3 $\pm$ 0.31	78.1 $\pm$ 1.20	<b>78.5<math>\pm</math>0.60</b>
QNLI	92.9 $\pm$ 0.25	70.0 $\pm$ 0.07	79.0 $\pm$ 0.34	80.0 $\pm$ 0.49	92.2 $\pm$ 0.31	82.8 $\pm$ 1.85	84.2 $\pm$ 0.28	92.1 $\pm$ 0.30	<b>93.1<math>\pm</math>0.17</b>
SST-2	95.3 $\pm$ 0.10	65.6 $\pm$ 0.64	88.7 $\pm$ 0.29	93.1 $\pm$ 0.57	95.2 $\pm$ 0.09	90.5 $\pm$ 0.39	92.9 $\pm$ 0.48	94.4 $\pm$ 0.07	<b>95.7<math>\pm</math>0.10</b>

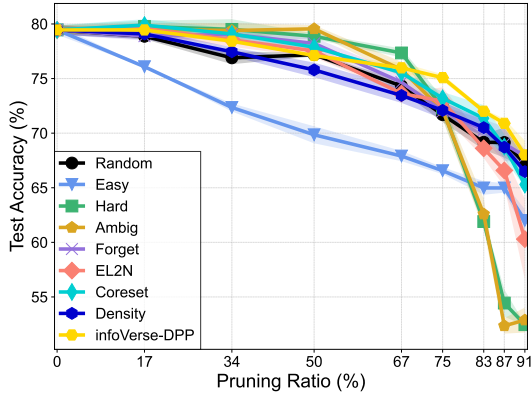


Figure 5: Test accuracy of RoBERTa-large fine-tuned on the pruned training dataset of WinoGrande.

Table 3: Ablation study about the effectiveness of each category of meta-information with infoVerse-DPP.

Categories	WinoGrande	CoLA
Static Measures	72.7 $\pm$ 0.24	60.2 $\pm$ 0.19
Training Dynamics	73.5 $\pm$ 0.35	62.2 $\pm$ 0.41
Model Uncertainty	71.5 $\pm$ 1.47	60.6 $\pm$ 0.36
MC-Model Uncertainty	70.2 $\pm$ 0.80	58.9 $\pm$ 0.82
Pre-trained Knowledge	72.8 $\pm$ 0.84	56.3 $\pm$ 0.51
infoVerse	<b>74.6<math>\pm</math>0.24</b>	<b>62.5<math>\pm</math>0.14</b>

larly observed in (Swayamdipta et al., 2020). On the other hand, density-based methods are robust to the varied pruning ratios, although overall performance is relatively low. Hence, to compare each baseline by considering all the pruning ratios together, we compare a single averaged performance in Table 2 similar to Area Under Curve (Tai, 1994). Here, one can observe that infoVerse with DPP (infoVerse-DPP) consistently outperforms other pruning methods across all datasets.

To demonstrate the advantages of infoVerse-DPP, we present Figure 6 to qualitatively show how the proposed method works for data pruning. Interestingly, types of majority meta-information of selected samples dynamically change as the pruning ratio increases, from *confidence* to *model uncertainty* to *variability*. After the most redundant samples (*i.e.*, high-confidence) are pruned, followed by hard and uncertain samples. In contrast, other baselines do not have any pattern for selection or

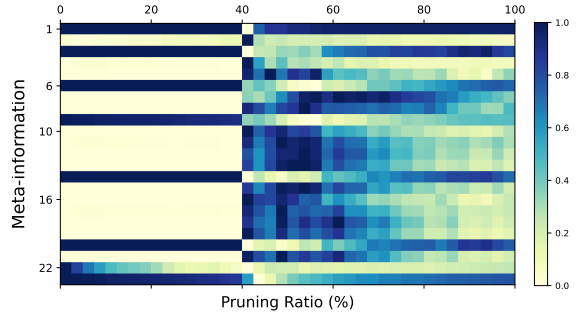


Figure 6: Majority measurement of the selected samples dynamically change as the pruning ratio increases (infoVerse-DPP on SST-2): *confidence* (0-40%)  $\rightarrow$  *model uncertainty* (40-60%)  $\rightarrow$  *variability* (60-85%).

just have static ones, as shown in Figure 15.

It is worth noting that effective pruning strategies would indeed vary with the pruning ratio or given; for instance, (Swayamdipta et al., 2020) disclose that high-confidence samples should be pruned at low pruning ratios due to redundancy, but these samples become essential for training as the ratio increases (e.g., 83%). While (Swayamdipta et al., 2020) could manually check the varied effectiveness of confidence and find the effective pruning strategy based on that, such a manual approach becomes very costly when the number of considered measurements increases. In this aspect, our framework offers an efficient solution as it prunes the samples toward maximizing the informativeness of the remaining samples, rather than focusing on specific measurements. Also, the observed pruning curriculum demonstrates how infoVerse with DPP actually outperforms the other selection methods, by automatically adapting the pruning strategy across varying ratios.

In addition, to verify the complementary effect between different categories of meta-information, we compare the model’s performance by pruning the samples based on each category using DPP. As shown in Table 3, the effectiveness is largely different depending on whether each category can capture the important aspect of the dataset for data pruning. However, when they are combined to construct infoVerse, the performance is significantly

Table 4: Average test accuracy of fine-tuned BERT-base over the nine AL iterations.

Dataset	Random	Entropy	BALD	BERT-KM	FT-BERT-KM	BADGE	ALPS	infoVerse-DPP
AGNEWS	89.9 $\pm$ 0.25	90.6 $\pm$ 0.17	<b>90.8</b> $\pm$ 0.32	89.8 $\pm$ 0.34	90.7 $\pm$ 0.18	90.6 $\pm$ 0.21	89.3 $\pm$ 0.32	90.6 $\pm$ 0.19
SST-2	88.8 $\pm$ 0.62	89.9 $\pm$ 0.81	89.3 $\pm$ 0.63	89.3 $\pm$ 0.57	89.4 $\pm$ 0.64	89.8 $\pm$ 0.63	89.2 $\pm$ 0.66	<b>90.3</b> $\pm$ 0.57
RTE	60.9 $\pm$ 2.80	60.5 $\pm$ 1.92	60.9 $\pm$ 2.20	60.7 $\pm$ 0.06	59.6 $\pm$ 2.39	59.8 $\pm$ 2.29	58.3 $\pm$ 3.00	<b>61.5</b> $\pm$ 2.46

improved which implicitly reveals that they are mutually complementary. More results are presented in Appendix C and E.

## 5.2 Active Learning

Active learning (AL) is a task that finds the most informative subset from unlabeled samples when labeled and used for training the model. AL usually consists of multiple iterations of the following two steps: (1) *select* a subset of unlabeled data under a specific sampling method and expand the labeled data by annotating the subset. (2) Then, *train* the model with the new training dataset. More details and experimental results are in Appendix B.3.

**Setups.** To demonstrate the effectiveness of infoVerse in AL, we compare it with the state-of-the-art AL methods on the various datasets, following the recent works of AL for NLP tasks (Yuan et al., 2020; Margatina et al., 2021). Specifically, we evaluate infoVerse on three datasets: SST-2, RTE, and AGNEWS (Zhang et al., 2015). Also, several AL methods are used as the baselines, which are based on three different strategies: (1) uncertainty-based (*Entropy* and *BALD*, as described in §3.1), (2) diversity-based (*BERT-KM* and *FT-BERT-KM* (Yuan et al., 2020)) which focus to cover data distribution, and (3) hybrid method to consider both aspects jointly (*BADGE* (Ash et al., 2020) and *ALPS* (Yuan et al., 2020)), and *Random* sampling. All experiments are conducted using BERT-base (Devlin et al., 2019). We construct infoVerse of unlabeled datasets with their pseudo-labels (Lee et al., 2013).

**Results.** We first summarize the results in Figure 7 and Table 4. Figure 7 presents the test accuracy of the trained model at each AL iteration on SST-2 dataset (the results of other datasets are presented in Appendix C, due to limited space). In addition, Table 4 shows the average test accuracy across multiple AL iterations and implies the overall performance of each method. Here, one can observe that infoVerse-DPP shows consistent improvements over other baselines; infoVerse-DPP outperforms the baselines in RTE and SST-2, while it shows comparable performance with the highest performing baseline *BALD* in AGNEWS. Conse-

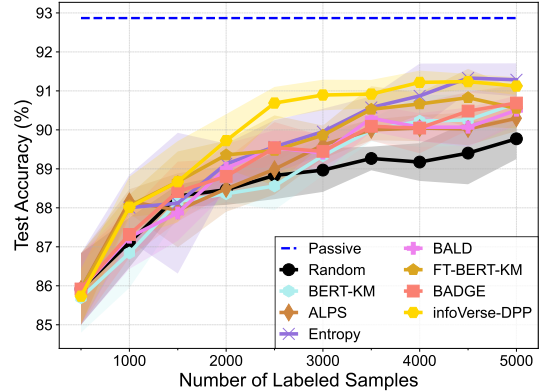


Figure 7: Test accuracy of BERT-base fine-tuned on the labeled samples by each AL method on SST-2.

quently, infoVerse-DPP achieves the lowest average rank (1.3) among the tested AL methods.

Next, we conduct additional experiments to understand in depth how infoVerse-DPP selects the informative unlabeled samples and improves the model’s performance. Specifically, on SST-2, we compare the average of meta-information of the selected samples by infoVerse-DPP and two representative baselines, *Random* and *Entropy*.<sup>4</sup> Figure 8 presents the results; *Entropy* selects the mostly uncertain samples (8(c)), but it relatively suffers to select the unseen samples (8(d)) and also has a risk to select noisy samples (8(a)). In contrast, infoVerse-DPP incorporates the multiple meta-information during the selection; for example, it selects the mostly variable samples with moderately low confidence, which has been demonstrated as a key characteristic for effective training samples (Swayamdipta et al., 2020). Also, the selected samples capture a certain level of uncertainty along with a low sentence-level density (*i.e.*, hence can introduce the new pattern in training samples).

## 5.3 Data Annotation

Finally, we demonstrate the advantage of infoVerse on data annotation (Kiela et al., 2021b), to provide the most effective set of unlabeled samples that are expected to improve the model’s performance after

<sup>4</sup>*Random* is the most simple approach and *Entropy* shows the second best performance.



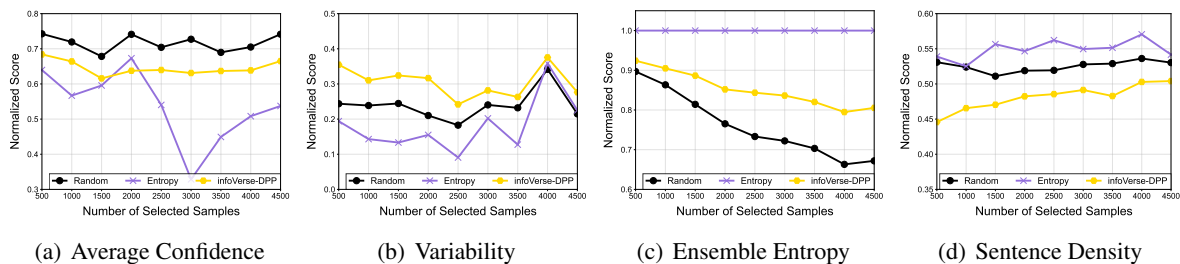


Figure 8: Comparison of the selected samples with different AL methods (*Random*, *Entropy*, and *Ours*) on SST-2.

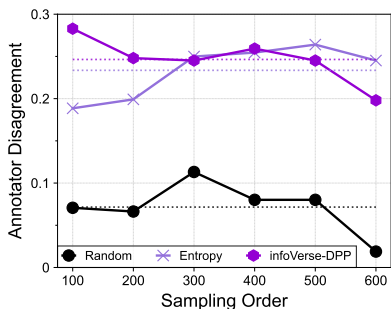


Figure 9: Annotator disagreement on IMP.

Table 5: Test accuracy of RoBERTa-large fine-tuned on the annotated dataset together with the original training dataset. 1,000 samples for SST-5 and 600 samples for IMP are additionally annotated, respectively.

Dataset	Original	Random	Entropy	infoVerse-DPP
SST-5	58.2 $\pm$ 0.66	58.2 $\pm$ 0.72	58.4 $\pm$ 0.83	<b>58.8<math>\pm</math>0.95</b>
IMP	88.6 $\pm$ 0.67	88.9 $\pm$ 0.40	88.8 $\pm$ 0.78	<b>89.0<math>\pm</math>0.50</b>

they are annotated with *human labelers*.

**Setups.** We consider two datasets, SST-5 (Socher et al., 2013) and IMP datasets (Du et al., 2021). Following (Du et al., 2021), we first conduct an unsupervised data retrieval to prepare high-quality 10,000 candidates among 2M unlabeled sentences from Common Crawl (Wenzek et al., 2020) and Reddit corpus.<sup>5</sup> We then apply each selection method to choose final queries for data annotation: 1,000 samples for SST-5 and 600 samples for IMP, respectively. Finally, we ask crowd-workers to annotate the selected samples using Amazon’s Mechanical Turk (Crowston, 2012) with at least three different annotators. We compare the two representative methods, *Random* and *Entropy*, with ours (*infoVerse-DPP*) due to the limited resources. We include more details in Appendix B.4.

**Results.** Table 5 shows the performance with different selection methods on SST-5 and IMP datasets. One can observe that infoVerse with DPP consistently finds more informative sets of

samples leading to extra performance gain than the other sampling methods on both datasets. We further measure disagreement between annotators on the newly-annotated dataset in the IMP task in Figure 9. The order of annotated samples by ours is more linearly aligned with the annotators’ disagreement than other sampling methods, indicating that our method prefers to choose more hard and informative samples first. Consequently, unlike the prior methods relying on single meta-information like confidence (Xie et al., 2020b) or uncertainty (Mukherjee and Awadallah, 2020), our multi-dimensional approach with infoVerse could provide useful contributions for data annotation. Finally, we remark that experimental results and relevant discussions about computational cost and complementary effect of meta-information are presented in Appendix E and F, respectively.

## 6 Conclusion

We propose a new framework, infoVerse to characterize the dataset in various aspects of data informativeness. To be specific, infoVerse utilizes various types of meta-information which offer different aspects of data characteristics. The combination of diverse meta-information helps detect distinct regions of dataset characteristics, which are not observable in the previous feature spaces. In addition, we further propose a novel sampling method to select data points that maximize the information at a set level rather than a sample level on the multidimensional space of infoVerse. We empirically demonstrate the benefit of infoVerse on three applications: data pruning, active learning, and data annotation. infoVerse with the proposed subset selection method shows consistent improvement over the strong baselines of each problem. We believe our framework will emerge with the growth of data-centric approaches and contribute to a better understanding of the dataset and improvement of the dataset’s quality.

<sup>5</sup><https://convokit.cornell.edu/documentation/subreddit.html>

## Limitations

In this paper, we propose a new framework that extracts the various aspect of information about given data, relying on the existing model-driven meta-information from the trained models. Hence, if there are some flaws within the used models, such as biased prediction (Sun et al., 2019) or learning of spurious correlation (Liu et al., 2021), then our framework can be directly affected and may have a risk of inheritance or amplification of such problematic behaviors. However, as our framework is not limited to any specific models and meta-information, one can prevent this problem by using the robustly trained models (Sagawa et al., 2020) or introducing more specialized meta-information (Lee et al., 2021a) for these problems. In addition, despite the empirical gains we find, our subset selection method is not theoretically guaranteed to be (or tightly bound to) the optimal set of max informativeness, which remains an interesting direction. A further study is necessary showing that selected samples from infoVerse could lead to low inter-annotator agreement in manual annotation but provide more accurate information than pseudo-labels. Abnormality detection using infoVerse, like noisy labels, out-of-distribution, or annotation artifacts, could be interesting future directions.

## Broader Impact and Ethical Implications

Our work aims to quantify the data informativeness with multi-perspective for capturing properties that can not be revealed by a single perspective. Especially, infoVerse lends some insight into data by models what we have. Thus, infoVerse has the potential for guiding the construction of high-quality datasets, e.g., removing mis-labeled samples. From these points, it is possible to develop a system or general platform for effectively collecting data like Dynabench<sup>6</sup> and Snorkle<sup>7</sup>. We anticipate that the general platform of infoVerse could be contributing to human-involved machine learning systems.

Although our work empirically demonstrates the improvement over various real-world problems, the current version of infoVerse has a potential risk to be vulnerable to sample undesirable properties (e.g., gender bias (Bordia and Bowman, 2019)) in a dataset, as we construct infoVerse with meta-information measure do not consider such properties. However, it can be easily alleviated by

<sup>6</sup><https://dynabench.org/>

<sup>7</sup><https://www.snorkel.org/>

adding various measurements which represent 'fairness' thanks to the extensibility of our framework. Hence, we believe that our proposed method can be personalized to the purpose of data collection.

## References

- Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep batch active learning by diverse, uncertain gradient lower bounds. *ArXiv*, abs/1906.03671.
- William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. 2018. The power of ensembles for active learning in image classification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Erdem Bıyık, Kenneth Wang, Nima Anari, and Dorsa Sadigh. 2019. Batch active learning using determinantal point processes. *arXiv preprint arXiv:1906.07975*.
- Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. *ArXiv*, abs/1904.03035.
- Joel Luis Carbonera and Mara Abel. 2015. A density-based approach for instance selection. In *2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 768–774. IEEE.
- Laming Chen, Guoxin Zhang, and Hanning Zhou. 2018. Fast greedy map inference for determinantal point process to improve recommendation diversity. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. 2019. Selection via proxy: Efficient data selection for deep learning. *arXiv preprint arXiv:1906.11829*.
- Kevin Crowston. 2012. Amazon mechanical turk: A research tool for organizations and information systems scholars. In *Shaping the future of ict research. methods and approaches*, pages 210–221. Springer.
- Sanjoy Dasgupta. 2011. Two faces of active learning. *Theor. Comput. Sci.*, 412:1767–1781.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.
- Terrance DeVries, Michal Drozdal, and Graham W Taylor. 2020. Instance selection for gans. In *Advances in Neural Information Processing Systems (NeurIPS)*.

- Jingfei Du, Édouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Veselin Stoyanov, and Alexis Conneau. 2021. Self-training improves pre-training for natural language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.
- Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the International Conference on Machine Learning (ICML)*.
- Alexandros Georgogiannis. 2016. Robust k-means: a theoretical revisit. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Jennifer Gillenwater, Alex Kulesza, and Ben Taskar. 2012. Discovering diverse and salient threads in document collections. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Short Papers)*.
- Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*.
- Peiyun Hu, Zachary Chase Lipton, Anima Anandkumar, and Deva Ramanan. 2019. Active learning with partial feedback. *ArXiv*, abs/1802.07427.
- Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. 2017. Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*.
- Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021a. Dynabench: Rethinking benchmarking in nlp. *NAACL*, abs/2104.14337.
- Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. 2021b. Dynabench: Rethinking benchmarking in nlp. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Alex Kulesza and Ben Taskar. 2011. k-dpps: Fixed-size determinantal point processes. In *Proceedings of the International Conference on Machine Learning (ICML)*.
- Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*.
- Jinhee Lee, Haeri Kim, Youngkyu Hong, and Hye Won Chung. 2021a. Self-diagnosing gan: Diagnosing underrepresented samples in generative adversarial networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Kimin Lee, Michael Laskin, Aravind Srinivas, and Pieter Abbeel. 2021b. Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*.
- Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. 2021. Just train twice: Improving group robustness without training group information. In *Proceedings of the International Conference on Machine Learning (ICML)*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.
- Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Subhabrata Mukherjee and Ahmed Hassan Awadallah. 2020. Uncertainty-aware self-training for text classification with few labels. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
- Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. 2021. Deep learning on a data diet: Finding important examples early in training. In *Advances in Neural Information Processing Systems (NeurIPS)*.

- Geoff Pleiss, Tianyi Zhang, Ethan R Elenberg, and Kilian Q Weinberger. 2020. Identifying mislabeled data using the area under the margin ranking. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan L. Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change nlp leaderboards? In *ACL*.
- Steven F Roth and Joe Mattis. 1990. Data characterization for intelligent graphics presentation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 193–200.
- Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *AAAI*.
- Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations (ICLR)*.
- Burr Settles. 2009. Active learning literature survey. *Technical Report from University of Wisconsin-Madison Department of Computer Sciences*.
- Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, A. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Mary M Tai. 1994. A mathematical model for the determination of total area under glucose tolerance and other metabolic curves. *Diabetes care*, 17(2):152–154.
- Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30.
- Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. 2019. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations (ICLR)*.
- Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Alex Wang and Kyunghyun Cho. 2019. Bert has a mouth, and it must speak: Bert as a markov random field language model. *arXiv preprint arXiv:1902.04094*.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations (ICLR)*.
- Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.
- Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Édouard Grave. 2020. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

- Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. 2020a. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020b. Self-training with noisy student improves imagenet classification. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Michelle Yuan, Hsuan-Tien Lin, and Jordan L. Boyd-Graber. 2020. Cold-start active learning through self-supervised language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.
- Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *ArXiv*, abs/1509.01626.

## A Summary and Formal Definition of Meta-information

In this section, we first present a detailed summarization of considered meta-information in Table 6. Then, we provide a formal definition of each *meta-information* introduced in Section 3.1. Here, we consider a classification task with  $K$  classes for the explanation.  $x$  and  $y$  indicate the input token and the corresponding true label, respectively.  $f_\theta$  indicates the classifier, which is pre-trained Transformer (Vaswani et al., 2017) such as BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019).  $p_\theta = \text{Softmax}(f_\theta)$  is a predictive distribution of classifier and  $z_\theta$  is a contextualized embedding before linear classifier in  $f_\theta = W^T z_\theta$ .

### A.1 Static Measures

Static measures are the meta-information extracted from a single static model, which is the most natural and easy way. In total, 5 different meta-information is used.

#### 1. Task Density (DeVries et al., 2020)

Here, Euclidean distance to  $K$ th nearest sample is used as density following Carbonera and Abel (2015).

$$D_{\text{KNN}}(x) = -\min_K \{ \|\hat{z}_\theta - z_\theta(x)\|_2 \}$$

where  $\hat{z}_\theta \in D_{\text{train}} \setminus \{z_\theta(x)\}$  and  $\min_K \{ \cdot \}$  is defined as the  $K$ th smallest value in a set. In our experiments, we set  $K = 5$ .

#### 2. Relative Density (DeVries et al., 2020)

As the **Task Density** does not utilize the label information, we further consider the relative density which is the difference of kNN density to true class samples and other class samples. Hence, if this value is large, it implies that  $x$  is near to the true class and far from other classes. In our experiments, we set  $K = 5$ .

#### 3. Static Confidence

$$\bar{\mu}(x) = p_\theta(y|x)$$

#### 4. Static Entropy (Shannon, 1948)

$$\bar{H}_{\text{Ent}}(x) = -\sum_{k=1}^K p_\theta(k|x) \cdot \log p_\theta(k|x)$$

#### 5. BADGE (Ash et al., 2020) BADGE is originally

proposed for active learning, to select the diverse and uncertain samples.

$$s_{\text{BADGE}}(x) = \|(p_\theta(x) - y) \cdot z_\theta(x)\|_2$$

### A.2 Training Dynamics

Training dynamics of samples largely varies depending on the samples' characteristic and hence can provide useful information, e.g., the confidence of mislabeled samples slowly increases relative to the normal ones. We totally find 4 corresponding meta-information in this category. Here,  $E$  is the total training epoch.

#### 6. Average Confidence (Swayamdipta et al., 2020)

$$\hat{\mu}(x) = \frac{1}{E} \sum_{e=1}^E p_{\theta^{(e)}}(y|x)$$

#### 7. Variability (Swayamdipta et al., 2020)

$$\hat{\sigma}(x) = \sqrt{\frac{\sum_{e=1}^E (p_{\theta^{(e)}}(y|x) - \hat{\mu}(x))^2}{E}}$$

#### 8. Forgetting Number (Toneva et al., 2019)

$$n_{\text{forget}}(x) = \sum_{e=1}^E \mathbb{1}(\text{acc}_i^{(e)} > \text{acc}_{i+1}^{(e)})$$

where  $\text{acc}^t(x) = \mathbb{1}(\arg \max_k p_{\theta^{(e)}}(k|x) = y)$

#### 9. Area Under Margin (Pleiss et al., 2020)

$$\text{AUM}(x) = \frac{1}{E} \sum_{e=1}^E M^{(e)}(x, y)$$

where  $M^{(e)}(x, y) = \frac{f_{\theta^{(e)}}(y|x) - \max_{k \neq y} f_{\theta^{(e)}}(k|x)}$

### A.3 Model Uncertainty

As intrinsic randomness within the model affects the samples' prediction, such uncertainty has been widely used in various fields. Total 6 different meta-information are considered. As we consider the ensemble from MC-Dropout and the ensemble of multiple random seed models, total 12 measures are considered. Here,  $T$  is the total number of models trained with different random seeds.

#### 10. EL2N score (Paul et al., 2021)

$$s_{\text{EL2N}}(x) = \sum_{t=1}^T \|p_{\theta^{(t)}}(x) - y\|_2$$

Table 6: Categorization of used meta-information to construct infoVerse. The arrow between the parentheses indicates more informative direction for each measure: e.g., less confident data ( $\downarrow$ ) are less likely seen so more informative.

Categories	Meta-information	Description
Static Measures	<i>Confidence</i>	( $\downarrow$ ) • Predictive probability to true label
	<i>Entropy</i>	( $\uparrow$ ) • Entropy of the predictive probability
	<i>BADGE</i>	( $\uparrow$ ) • Norm of the gradient with respect to parameters in the final (linear) layer
	<i>Task Density</i>	( $\downarrow$ ) • Euclidean distance to the $K_{th}$ nearest element on the contextualized embedding from fine-tuned classifier
	<i>Relative Density</i>	( $\downarrow$ ) • Difference of <i>Task density</i> within other class’ samples and true class samples.
Training Dynamics	<i>Confidence</i>	( $\downarrow$ ) • Average predictive probability to true label over the training epochs
	<i>Variability</i>	( $\uparrow$ ) • Variance of predictive probability to true label over the training epochs
	<i>Forgetting number</i>	( $\uparrow$ ) • Summation of <i>forgetting</i> over the training epochs: Sample $i$ undergoes <i>forgetting</i> when accuracy of $i$ decreases between two consecutive epochs
	<i>Area Under Margin</i>	( $\downarrow$ ) • Average <i>margin</i> over the training epochs: <i>Margin</i> captures how much larger the assigned logit is than all other logits
Model Uncertainty (Ens or MC)	<i>EL2N score</i>	( $\uparrow$ ) • Approximation for the gradient norm of expected loss which bounds the expected change in loss for sample $i$ caused by removing it from training
	<i>Entropy</i>	( $\uparrow$ ) • Entropy of predictive probability over multiple models with different random seeds
	<i>BALD</i>	( $\uparrow$ ) • Mutual information between data-points and model’s weights
	<i>Variation Ratio</i>	( $\uparrow$ ) • Proportion of predicted labels that are not coincided with the average prediction
	<i>Confidence</i>	( $\downarrow$ ) • Average predictive probability to true label between the models
Pre-trained Knowledge	<i>Variability</i>	( $\uparrow$ ) • Variance of predictive probability to true label between the models
	<i>PLL</i>	( $\uparrow$ ) • Pseudo-Log-Likelihood (PLL) score from pre-trained Masked Language Models
	<i>Semantical Density</i>	( $\downarrow$ ) • Euclidean distance to the $K_{th}$ nearest element on the contextualized embedding from pre-trained sentence encoder, e.g., sBERT

### 11. Ensemble Entropy (Shannon, 1948)

$$H_{\text{Ent}}(x) = - \sum_{k=1}^K p_{\text{avg}}(k|x) \cdot \log p_{\text{avg}}(k|x)$$

where  $p_{\text{avg}}(k|x) = \frac{1}{T} \sum_{t=1}^T p_{\theta^{(t)}}(k|x)$

### 12. BALD (Houlsby et al., 2011)

$$I_{\text{BALD}}(x) = H_{\text{Ent}}(x) - \frac{1}{T} \sum_{t=1}^T \sum_{k=1}^K -p_{\theta^{(t)}}(k|x) \cdot \log p_{\theta^{(t)}}(k|x)$$

### 13. Variance ratio (Beluch et al., 2018)

$$v(x) = 1 - \frac{f_m(x)}{T}$$

where  $f_m(x) = \sum_{t=1}^T \mathbb{1}(\arg \max_k p_{\theta^{(t)}}(k|x) = \hat{y}_{\text{avg}}(x))$ ,  $\hat{y}_{\text{avg}}(x) = \arg \max_k p_{\text{avg}}(k|x)$

Furthermore, we consider **14. Ensemble Confidence** and **15. Ensemble Variability** which just changes the role of epoch  $E$  to the number of models  $T$ . Also, we further consider the same meta-information by simulating the ensemble with Monte-Carlo dropout (MC-dropout) (Gal and Ghahramani, 2016), *i.e.*, using different Dropout masked models for the ensemble. From this, we

obtain **16. MC EL2N score** to **21. MC Ensemble Variability**.

### A.4 Pre-trained Knowledge

As general text representation provides complementary information to task’s one, using the pre-trained language models to extract meta-information is another popular direction. We use **2** meta-information measures extracted from pre-trained models which are agnostic to the target task.

### 22. Semantical Density (DeVries et al., 2020)

$$D_{\text{KNN}}(x) = - \min_K \{ \|\hat{z}_{\theta} - z_{\theta}(x)\|_2 \}$$

where  $\hat{z}_{\theta} \in D_{\text{train}} \setminus \{z_{\theta}(x)\}$  and  $\min_K \{\cdot\}$  is defined as the  $K$ th smallest value in a set. We set  $K = 5$ . Unlike **Task Density**,  $z$  is extracted from a pre-trained sentence encoder, e.g., sBERT (Reimers and Gurevych, 2019) which is known to capture the relationship between sentences.

### 23. Pseudo-Log-Likelihood (PLL) (Salazar et al., 2020; Yuan et al., 2020)

Originally, Salazar et al. (2020) use the following masked language modeling (MLM) score from pre-trained language model  $\theta^{\text{MLM}}$  as Pseudo-Log-

Likelihood (PLL) (Wang and Cho, 2019).

$$\text{PLL}(x) = \sum_{l=1}^L \log p_{\theta^{\text{MLM}}}(x_l | x_{\setminus l})$$

where  $x_{\setminus l} := (x_1, \dots, x_{l-1}, x_{l+1}, x_L)$ . However, it requires  $L$  times of inference for calculation. Hence, we instead its approximation following Yuan et al. (2020), which just calculates PLL at once without masking tokens.

## B Details of Experiment

### B.1 Dataset

In this section, we provide details of all datasets we used in this work and hyperparameters used for training the models. For all datasets, we used the given standard training and validation sets. We present the details of downstream datasets in Table 7. All of the data we used can be downloaded from HuggingFace dataset <https://huggingface.co/datasets/>. For experiments, we report accuracy using the official test set on WinoGrande and AGNEWS. For those where the label for the official test set is not available (SST-2, RTE, MNLI, and QNLI), we use the given validation set. Also, the maximum sequence length is commonly set to 128.

### B.2 Data Pruning

For data pruning experiments, we commonly fine-tune RoBERTa-large classifier (Liu et al., 2019) which has the 355M parameters, following (Swayamdipta et al., 2020). For fine-tuning, we commonly train it for 10 epochs with learning rate  $1e-5$  and batch-size 16 (except WinoGrande with 64 following (Swayamdipta et al., 2020) due to the optimization difficulty) with Adam optimizer (Kingma and Ba, 2014). For each pruning ratio, selection method, and dataset, we run three times with different random seeds.

### B.3 Active Learning

Active learning (AL) is a task that finds the most informative subset from unlabeled samples when they are labeled and added to the training dataset. AL usually consists of multiple iterations of the following two steps: (1) annotates a subset of unlabeled data chosen by a sampling method, and (2) adds the labeled data to the previous round of the dataset and re-train the model with the new training dataset. For each round, we trained the model

from scratch to avoid overfitting, following Hu et al. (2019).

To be specific, for the experiments in Section 5, we select 100 examples for RTE and 500 examples for CoLA and AGNEWS for each iteration from the training dataset, respectively.<sup>8</sup> Note that the specific selection Then, they are moved to the labeled dataset from the unlabeled pool in each iteration.

To simulate AL, we sample a batch of  $k$  sentences from the training dataset, query labels for this batch, and Batch size  $k$  is set to 500 for SST-2 and AGNEWS, and 100 for RTE which is a relatively small dataset. For each sampling method and dataset, we run an AL simulation five times with different random seeds. Also, we fine-tune models on five epochs for SST-2 and AGNEWS, and ten epochs for the RTE dataset. We experiment with the BERT-base model which has 110M parameters provided by HuggingFace Transformers (Wolf et al., 2019) with Apache License 2.0. Our implementation is based on existing code repositories<sup>9</sup> with MIT License and used the same hyperparameter (Yuan et al., 2020). We use AdamW (Loshchilov and Hutter, 2019) with a learning rate of  $2e-5$ .

**BERT-KM** (Yuan et al., 2020): As a diversity-based baseline, applying k-means clustering to the 12 normalized BERT output embeddings of the fine-tuned model to select  $k$  data points.

**FT-BERT-KM** (Yuan et al., 2020): Using the same algorithm as BERT-KM except for the BERT embeddings from the previously fine-tuned model are used.

**ALPS** (Yuan et al., 2020): Input sentence is randomly masked, then predict the masked language model (MLM) loss of BERT as a proxy for model uncertainty

### B.4 Data Annotation

Here, we provide the details about the annotation pipeline with crowd workers. During experiments, we annotate the selected unlabeled samples with each selection method for SST-5 and IMP datasets. To this end, we use Amazon’s Mechanical Turk crowd-sourcing platform (Crowston, 2012). Figure 10 and 11 show the interfaces used to collect annotations from crowd workers for each task. The top provides the summary, the middle provides de-

<sup>8</sup>The initial (*i.e.*, the first iteration) labeled samples of all AL methods are commonly selected by random sampling.

<sup>9</sup><https://github.com/forest-snow/alps>



Table 7: Dataset statistics used in experiments

Dataset	Language	Domain	Classes	Train / Dev
QNLI	EN	Natural Language Inference	2	104k / 5.4k
SST-2	EN	Sentiment Analysis	2	67k / 873
WinoGrande	EN	Commonsense Reasoning	2	40k / 1.2k
CoLA	EN	Linguistic Acceptability	2	8.5k / 1.0k
RTE	EN	Natural Language Inference	2	2.5k / 278
AGNEWS	EN	Topic Classification	4	110K / 7.6k

Table 8: Ablation study with each component of infoVerse-DPP. Average test accuracy of finetuned RoBERTa-large classifiers over 8 different pruning ratio on WinoGrande and CoLA are compared, similar to Table 2.

Feature	Sampling	WinoGrande	CoLA
Random	Random	73.1 $\pm$ 0.09	60.7 $\pm$ 0.63
Classifier	Coreset	72.2 $\pm$ 0.91	55.5 $\pm$ 0.93
Classifier	DPP	73.5 $\pm$ 0.25	62.1 $\pm$ 0.19
infoVerse	Coreset	71.9 $\pm$ 0.47	61.0 $\pm$ 0.26
infoVerse	DPP	<b>74.6<math>\pm</math>0.24</b>	<b>62.5<math>\pm</math>0.14</b>

tailed instructions, and then examples are shown. The whole task has 10 items per Human Interface Task (HIT). Workers were paid US\$1.0 per HIT on average, and all workers were paid for their work. To improve the quality of collected preference labels, we only hire the Master workers identified as high-performing workers from Amazon’s Mechanical Turk system. Overall, we gather at least 3 annotations for each sample. For the experiments with annotated samples, we use the same experimental setups with data pruning in Section 5.1. Also, for the annotator disagreement, we report variance within the multiple annotations. We will release the annotated dataset for future research.

## C Additional Results

### C.1 Data Pruning

First, in Figure 13, we plot the test accuracy of fine-tuned RoBERTa-large across different pruning ratios on CoLA, SST-2, RTE, and QNLI datasets; while the baseline methods suffer from inconsistent performance on different pruning ratio (for example, Hard and Ambig show good performance on low pruning ratio, but steeply degraded when the pruning ratio increases), infoVerse-DPP shows the consistently outperforming performance in overall. In addition, we plot the dynamics during data

pruning with infoVerse-DPP in Figure 14, similar to Figure 6. Here, one can observe that infoVerse-DPP automatically finds the effective strategy adaptively. Finally, we present the ablation results of our component (1) infoVerse and (2) DPP-based sampling method. As shown in Table 8, the DPP-based sampling method provides a clear improvement in multidimensional space (vs Coreset). Furthermore, as infoVerse provides a richer feature space than the standard classifier’s embedding, such gain is further enlarged when they are combined.

### C.2 Active Learning

In Figure 12, we present the test accuracy of fine-tuned BERT-base with each AL iteration on RTE and AGNEWS, respectively. Here, one can observe that infoVerse-DPP shows comparable performance with the state-of-the-art baseline of AL.

## D infoVerse on Other Datasets

In this section, we first present the correlation matrices between 23 meta-information on other datasets, similar to Figure 2. As one can see in Figure 16, the correlation matrices between different datasets (and tasks) are quite similar, which implies that the meta-information captures the general characteristic of datasets. In addition, we present infoVerse (bottom left and zoom in right) on other datasets (CoLA, WinoGrande, RTE, and SST-2) along with its classifier’s embedding space (top left) and data map (Carbonera and Abel, 2015) (middle left) in Figures 20, 19, 17, and 18. Here, one can observe that infoVerse successfully reveals distinctive regions again.

## E Experiments to Verify Complementary Effect of Meta-information

To further verify the complementary effect between multiple meta-information, we conduct simple toy experiments in this section. Similar to

Table 9: Detection accuracy (%) of linear classifier trained with multiple meta-information.

Feature Space	Mis-pred	Mis-labeled	OOD	Adv
Classifier Embedding	10.4	89.1	89.0	85.9
*infoVerse	97.5	94.2	90.3	87.1
infoVerse	<b>99.9</b>	<b>94.3</b>	<b>91.4</b>	<b>87.5</b>
- (4)	99.8	94.2	86.3	86.0
- (3)-MC	85.7	94.2	82.9	86.0
- (3)-Ens	77.9	94.2	78.6	86.0
- (2)	49.9	94.1	69.9	86.0

(Swayamdipta et al., 2020), we train a simple linear classifier on each feature space by assuming that gold labels of each task are available for training; for example, given sample is noisy-labeled or not. Here, we consider four different abnormality detection tasks with the QNLI dataset: mis-predicted, mislabeled (or noisy labeled), out-of-distributed (OOD), and adversarial samples (Adv), respectively.

In Table 9, one can verify that the accuracy increases as more meta-information is used; it implies that they are actually complementary and can provide richer information when they are jointly considered. In the end, infoVerse shows a better performance than the original classifier’s semantic embedding in all tested cases. Also, we consider the reduced feature space, \*infoVerse, by applying the PCA-based feature selection method using the correlation matrix, and verify its comparable performance only using the half of meta-information. But, since only small costs are additionally required to use infoVerse compared to \*infoVerse, we use all 23 meta-information for our experiments. It is noteworthy that new meta-information can be easily included in our framework, and contribute to compose more informative feature space. In the remaining part of the section, we further provide the details of this experiment. Here, we use RoBERTa-large classifier (Liu et al., 2019) fine-tuned on QNLI dataset (Wang et al., 2019).

1) Finding mispredicted samples: we train a single linear classifier with SGD optimizer on each feature space to classify whether a given sample is correctly predicted or not. Here, we assume that the binary labels that indicate whether a given sample is correctly predicted or not are available to train the linear classifier. Then, we only measure the performance on the test mispredicted samples.

2) Detecting mis-labeled samples: following the setups in (Swayamdipta et al., 2020), we artificially impose the 10 % label noise into training samples

with high confidence (*i.e.*, easy-to-learn). Then, we train a single linear classifier with SGD optimizer on each feature space to classify whether given samples has corrupted label or not. Here, we assume that the binary labels that indicate whether a given sample is corrupted or not are available to train the linear classifier.

3) Detecting out-of-distribution samples: we consider the samples of QNLI’s development set as inD and samples of MNLI’s development set (both matched and mismatched) as OOD. Then, we train a single linear classifier with SGD optimizer on each feature space to classify whether the given sample is from inD or OOD. Here, we assume that the binary labels that indicate whether a given sample is inD or OOD are available to train the linear classifier.

4) Detecting adversarial sentence: Here, we consider the task to classify whether the normal sentence is given (MNLI) or adversarially gathered sentence is given (ANLI (Nie et al., 2019)). Then, we train a single linear classifier with SGD optimizer on each feature space to classify whether the given sample is normal or adversarial. Here, we assume that the binary labels that indicate whether a given sample is normal or adversarial are available to train the linear classifier.

## F Computational Cost

The computation cost of infoVerse depends on which meta-information is used for constructing it. As introduced in Table 1, we considered meta-information with four categories (static measures, training dynamics, model uncertainty, and pre-trained knowledge). The calculation of these categories requires (1,  $E$ ,  $T$ , and 2) forward passes of the trained model for each sample, where  $E$  denotes the total training epochs, and  $T$  indicates the total number of trained models with different random seeds. In the case of the proposed sampling method, it has  $O(N^2M)$  time complexity when returning  $N$  items from  $M$  total items, but we remark that it can be further reduced with a simple approximation (Chen et al., 2018). Yet, we remark that our method does not require additional training costs since it only utilizes the trained models via standard way (cross entropy and multiple random seeds) and pre-trained models.

For example, we measured the runtime of our approach using the CoLA dataset, with the time consumption in seconds: training/constructing in-

foVerse/DPP sampling consume 3852s/100s/10s, respectively. This demonstrates that our method's overall cost is relatively minor compared to training expenses. Moreover, it is worth noting that the cost of ours is only incurred once at initial construction. It is also worth noting that all meta-information within the same category are obtained with the same forward passes, and there is no additional cost after infoVerse is constructed at once.

In addition, although this work did not focus much on reducing computational overhead, we believe simple practices can further reduce the cost of constructing infoVerse. For example, substituting the additional forward passes by saving the outputs of models during training on the fly (Huang et al., 2017; Tarvainen and Valpola, 2017) or using a proxy of the original model with smaller architectures and fewer training epochs (Coleman et al., 2019).

## Instructions

Summary

Detailed Instructions

Examples

### Sentiment Evaluation for Text

In this study, we want to learn how speakers convey emotional information in language.

For each given text, please rate its sentiment on a scale from Very Negative to Very Positive.

Some of the judgments will be natural. For instance, "This movie was awesome" is a very positive sentence, because it conveys the writer's positive opinion. Another example of a sentence with positive sentiment is "The Rock is going to make a splash bigger than Arnold Schwarzenegger, Jean-Claude Van Damme, or Steven Segal," because it shows the writer's positive sentiment towards The Rock.

It is possible for a piece of text to have a neutral sentiment. For example, "I read the newspaper", "Elections are held next Tuesday", or "Canada is north of the United States" would be considered neutral, because the text does not convey anything about the author's positive or negative sentiment.

Please click on the slide bar **even when your rating for the text is neutral**. The change in color of the slide bar indicates that our system has recorded your answer. **You will not be able to submit unless you click on all the slide bars.**

Summary

Detailed Instructions

Examples

### Sentiment Evaluation for Text

For each given sentence, please rate its sentiment on a scale from Very Negative to Very Positive.

Positive sentiment means the text conveys the author's positive feelings or opinions. Negative sentiment means the text conveys the author's negative feelings or opinions. If the text does not reveal anything about the author's positive or negative feelings, it is neutral.

Some examples of positive texts: "An entertaining, colorful, action-filled crime story with an intimate heart" is positive because it conveys the author's positive opinion on the crime story. "I'm so proud Mils! Keep smiling" is positive because it conveys the author's positive feelings (proud of Mils).

Some examples of negative texts: "@Rainbow6Game felt guilt from playing this game" is negative because it conveys the author's negative feeling (guilt). "Are they the new Kennedy's of American politics.. sadly both mom and daughter are as charming as a toe nail" is negative because it conveys the author's negative opinion of both the mom and daughter.

Please click on the slide bar **even when your rating for the phrase is neutral**. The change in color of the slide bar indicates that our system has recorded your answer. Remember, you will not be able to submit unless you click on all the slide bars.

**This task is highly subjective.** In general, we don't have a right answer in mind, but we will be doing some checking for attentiveness of the examples and to ensure that responses are not made at random.

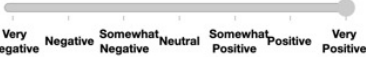
Summary

Detailed Instructions

Examples

#### Good examples

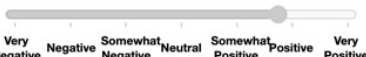
I had a great time at the party.



This sentence clearly conveys the writer's positive opinion of the party. Your opinion of how positive this sentence is might be slightly different, and that's okay. Any answer in the "positive" range on the slider is a good answer.

#### Bad examples

All this makeup is going on sale, but I ain't got the funds. Heartbreaking.



This text should be rated somewhere in the "negative" range, because it conveys the author's negative feelings towards not being able to buy the on sale make up.

Figure 10: Interface to collect annotation from crowd workers for sentiment classification for SST-5

Summary
Detailed Instructions
Examples

**Insult Detection for Text**

In this task, we want to determine when a comment from a conversation would be considered insulting to another participant in the conversation.

Sometimes, this determination is straightforward. For example "You're a moron. Go snuggle yourself in potted meat," is clearly **insulting**.

If a comment disagrees or argues with another participant in the conversation, but does not actually insult them, the comment is **not insulting**. For example, "I don't know why you would say that; your claim is completely ill-founded," is **not insulting**.

Remember that a comment that does not directly insult *another participant* in the conversation is **not insulting**, even if the comment is mean or rude. For example, the comment "LOL Lame old woman, mother of yellow chicken hawks" is considered **not insulting** because although it's mean, there are no insults directed to another participant in the conversation.

This task contains 10 texts. You will not be able to submit until you have responded to all 10.

NOTE: This task contains sentences that may be offensive to readers. They do not represent the views of this task's requesters.

Summary
**Detailed Instructions**
Examples

**Insult Detection for Text**

The important thing to keep in mind during this task is that a comment is only considered insulting if it is insulting toward another participant in the conversation.

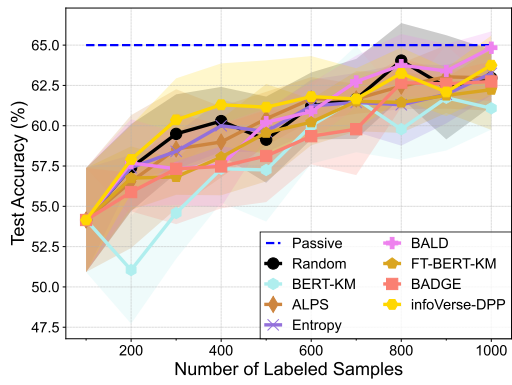
If there is no evidence that the target of the insults is a participant in the conversation, you should assume that he or she is **not** a participant in the conversation. For example, in the comment "Brian's far too stupid to be in 'the game'," we can assume that Brian is not a participant in the conversation. Since Brian is not a participant in the conversation, this comment is **not insulting**.

Some comments may be racist or otherwise offensive, but if they do not directly insult another participant in the conversation, these comments are not insulting. For example, "'It's the old white men and women who saved our nation in World War 11 and many are still alive who will help save our nation again. This time from President Obama and his administration when they vote in November," has strong white supremacist elements, but as it does not insult someone else in the conversation, it should be labeled **not insulting**.

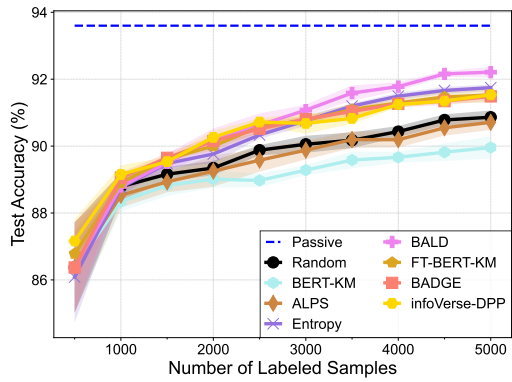
Summary
Detailed Instructions
**Examples**

Good examples	Bad examples
<p>If you mean awesome at killing jobs and reducing America to Greece, then you are absolutely right.</p> <p><input type="radio"/> Insulting <input checked="" type="radio"/> Not insulting</p> <p><b>Explanation:</b> Although the commenter is clearly upset, this is not insulting because the comment does not insult another person in the conversation.</p>	<p>A muslim Grandma says to her daughter, Oh, they blow up so fast.</p> <p><input checked="" type="radio"/> Insulting <input type="radio"/> Not insulting</p> <p><b>Explanation:</b> While this comment invokes offensive stereotypes, since it is not directly insulting another person in the conversation, it should be labeled <i>not insulting</i>.</p>
<p>you seem to be one of those turd minions I see when I've been playing Diablo</p> <p><input checked="" type="radio"/> Insulting <input type="radio"/> Not insulting</p> <p><b>Explanation:</b> This is insulting because the commenter uses the phrase "turd minions" to insult the person he or she is talking to.</p>	<p>You are just plain wrong.</p> <p><input checked="" type="radio"/> Insulting <input type="radio"/> Not insulting</p> <p><b>Explanation:</b> While the commenter strongly disagrees with someone, the comment does not actually insult that person, so this should be labeled <i>not insulting</i></p>

Figure 11: Interface to collect annotation from crowd workers for insult detection for IMP dataset.

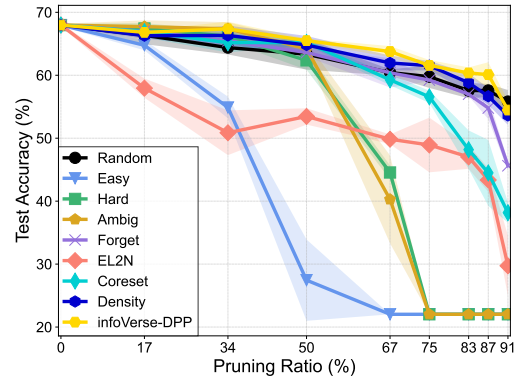


(a) RTE

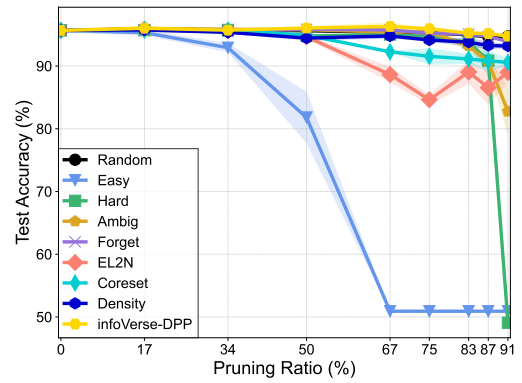


(b) AGNEWS

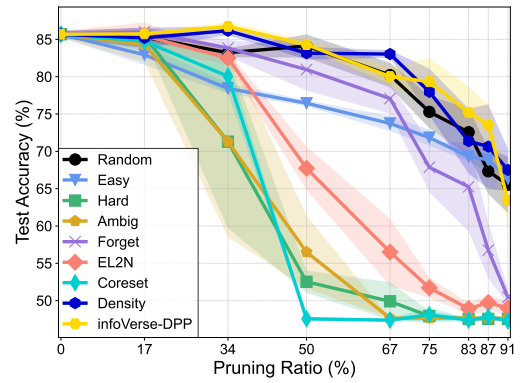
Figure 12: Test accuracy of BERT-base classifier fine-tuned using the labeled samples by each AL method on (a) RTE and (b) AGNEWS.



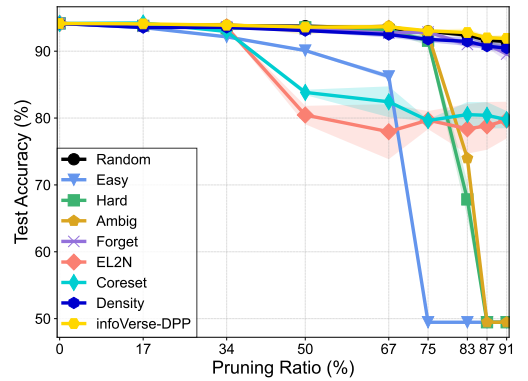
(a) CoLA



(b) SST-2

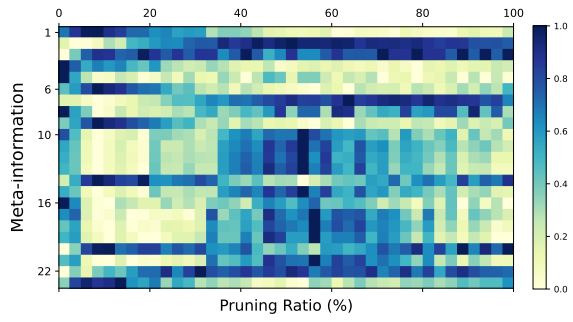


(c) RTE

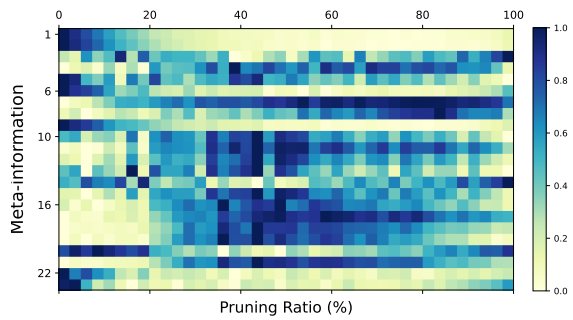


(d) QNLI

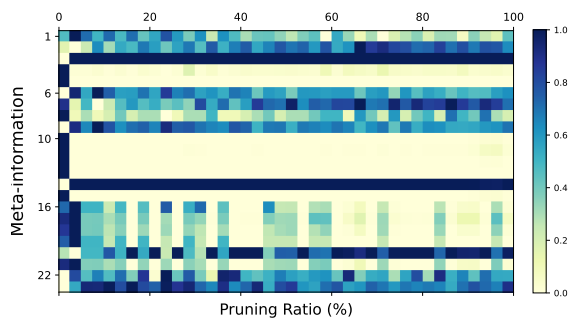
Figure 13: Data pruning performance on different datasets: (a) CoLA, (b) SST-2, (c) RTE, and (d) QNLI.



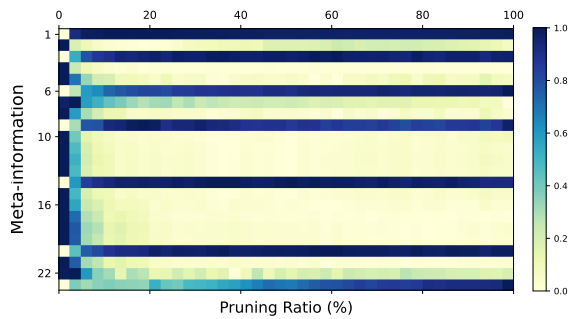
(a) CoLA



(b) WinoGrande

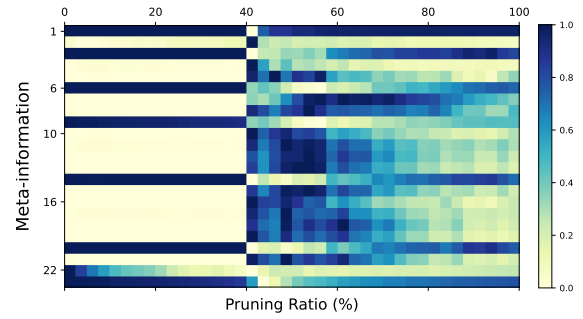


(c) RTE

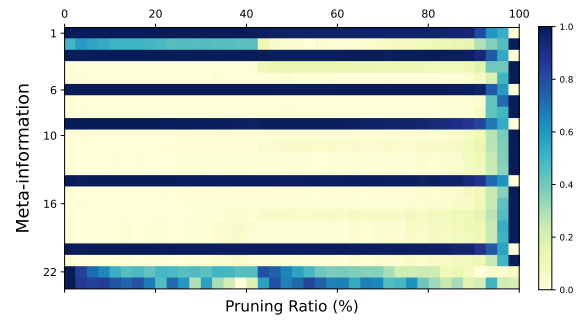


(d) QNLI

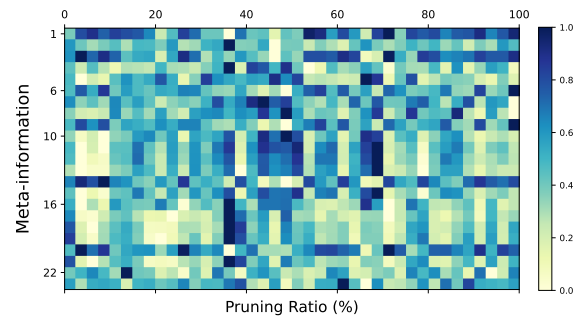
Figure 14: Dynamics of data pruning with infoVerse-DPP on different datasets: (a) CoLA, (b) WinoGrande, (C) RTE, and (d) QNLI.



(a) infoVerse-DPP

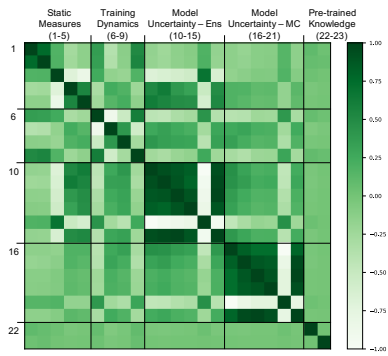


(b) Variability

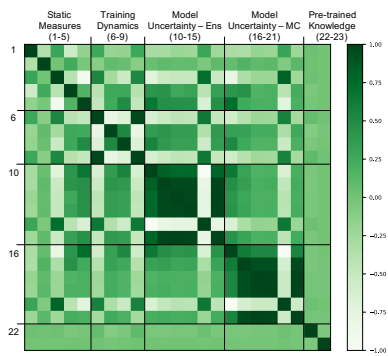


(c) Random

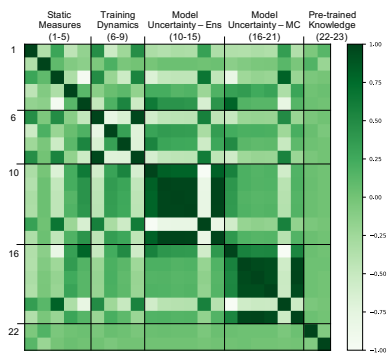
Figure 15: Dynamics of data pruning with different selection method on SST-2 dataset: (a) infoVerse-DPP, (b) Variability, and (c) Random.



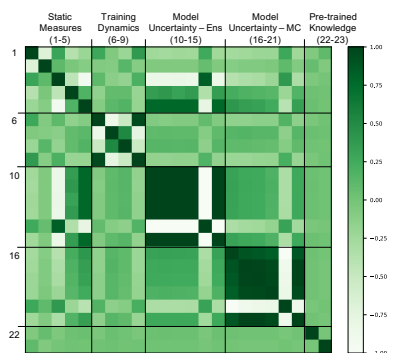
(a) WinoGrande



(b) SST-2



(c) CoLA



(d) RTE

Figure 16: Confusion matrices between meta-information on different datasets: (a) WinoGrande, (b) SST-2, (C) CoLA, and (d) RTE.



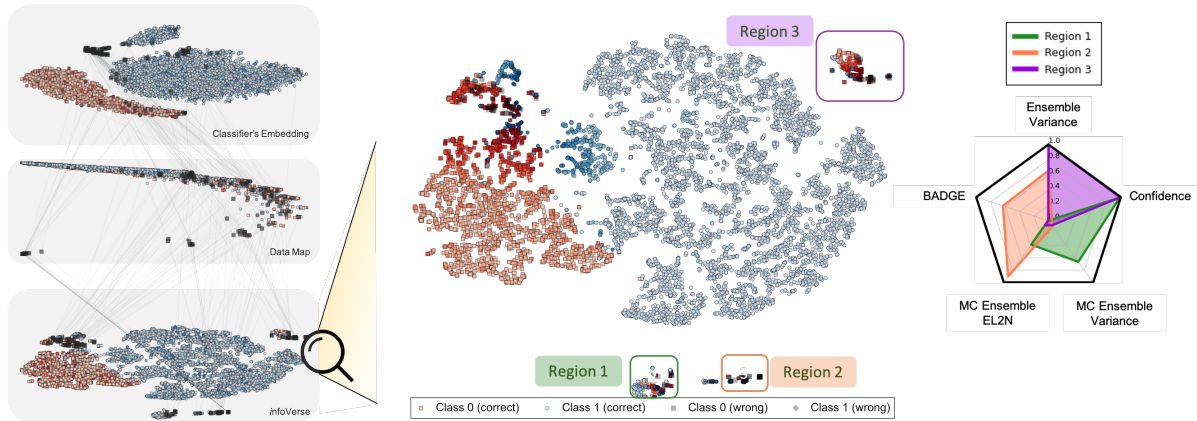


Figure 17: infoVerse (bottom left) on CoLA along with other feature spaces: classifier embedding (top left) and *data map* (Swayamdipta et al., 2020) (middle left). (middle) Zoomed version of infoVerse is presented. (right) Score distribution of each wrong region characterized by infoVerse.

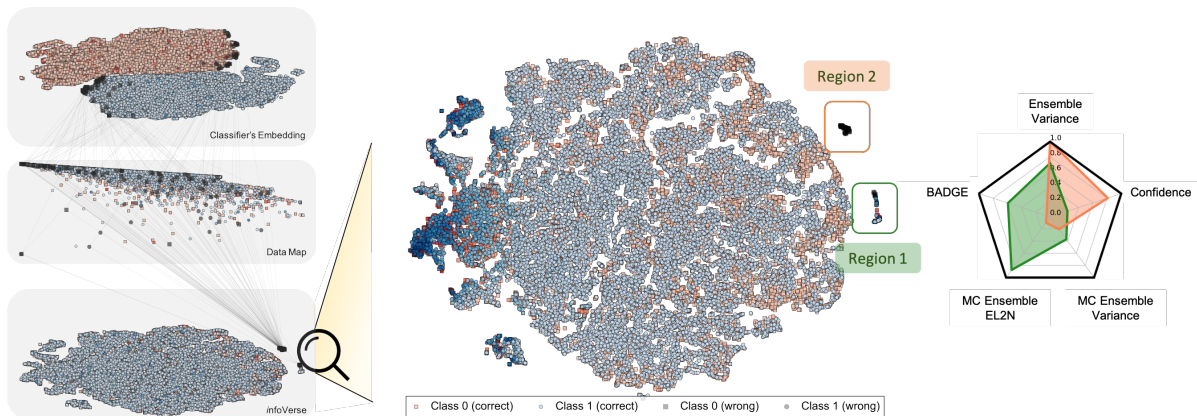


Figure 18: infoVerse (bottom left) on WinoGrande along with other feature spaces: classifier embedding (top left) and *data map* (Swayamdipta et al., 2020) (middle left). (middle) Zoomed version of infoVerse is presented. (right) Score distribution of each wrong region characterized by infoVerse.

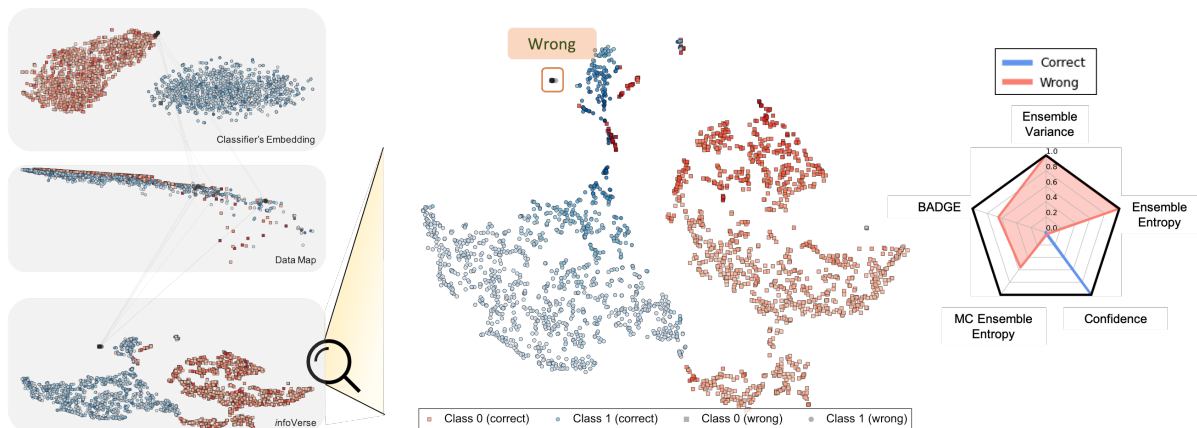


Figure 19: infoVerse (bottom left) on RTE along with other feature spaces: classifier embedding (top left) and *data map* (Swayamdipta et al., 2020) (middle left). (middle) Zoomed version of infoVerse is presented. (right) Score distribution of each wrong region characterized by infoVerse.

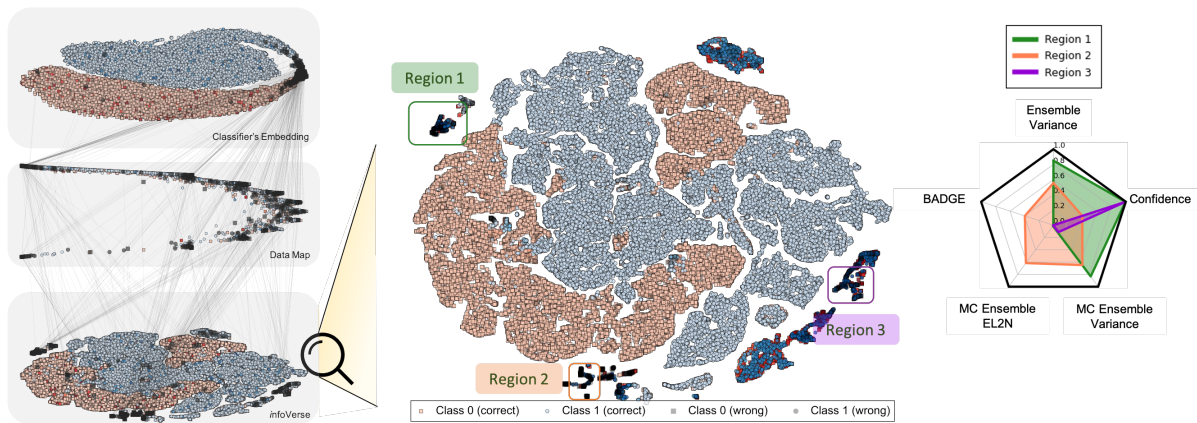


Figure 20: infoVerse (bottom left) on SST-2 along with other feature spaces: classifier embedding (top left) and *data map* (Swayamdipta et al., 2020) (middle left). (middle) Zoomed version of infoVerse is presented. (right) Score distribution of each wrong region characterized by infoVerse.

## ACL 2023 Responsible NLP Checklist

---

### A For every submission:

- A1. Did you describe the limitations of your work?  
*we discussed the limitations in Limitation section.*
- A2. Did you discuss any potential risks of your work?  
*we discussed the potential risks in Broader Impacts and Ethical Implications.*
- A3. Do the abstract and introduction summarize the paper’s main claims?  
*we summarized our main claims in Abstract and Section 1.*
- A4. Have you used AI writing assistants when working on this paper?  
*Left blank.*

### B Did you use or create scientific artifacts?

*we used scientific artifacts in Section 5.*

- B1. Did you cite the creators of artifacts you used?  
*Section 5 and Appendix B*
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?  
*Appendix B*
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?  
*Not applicable. Left blank.*
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?  
*Not applicable. Left blank.*
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?  
*Table 6 in Appendix B*
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.  
*Table 6 in Appendix B*

### C Did you run computational experiments?

*Section 5*

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?  
*Appendix B,E*

---

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

*Appendix B,E*

- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

*Section 5*

- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

*Not applicable. Left blank.*

**D  Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*Section 5.3*

- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

*Figure 11, 12 in Appendix*

- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

*Appendix B.4*

- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

*We specified the task in instruction (Figure 11, 12 in Appendix)*

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

*Not applicable. Left blank.*

- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

*Not applicable. Left blank.*