DEUCE: Dual-diversity Enhancement and Uncertainty-awareness for Cold-start Active Learning

Jiaxin Guo[◇†‡] **C. L. Philip Chen**[◇†‡] **Shuzhen Li**[†‡] **Tong Zhang**[◇†‡*]

[◇]Guangdong Provincial Key Laboratory of Computational AI Models and Cognitive Intelligence, School of Computer Science and Engineering, South China University of Technology, Guangzhou, China

[†]Pazhou Lab, Guangzhou, China

[‡]Engineering Research Center of the Ministry of Education on Health Intelligent Perception and Paralleled Digital-Human, Guangzhou, China

cs_guojiaxin@mail.scut.edu.cn philipchen@scut.edu.cn
cslishuzhen@mail.scut.edu.cn tony@scut.edu.cn

Abstract

Cold-start active learning (CSAL) selects valuable instances from an unlabeled dataset for manual annotation. It provides high-quality data at a low annotation cost for label-scarce text classification. However, existing CSAL methods overlook weak classes and hard representative examples, resulting in biased learning. To address these issues, this paper proposes a novel dual-diversity enhancing and uncertainty-aware (DEUCE) framework for CSAL. Specifically, DEUCE leverages a pretrained language model (PLM) to efficiently extract textual representations, class predictions, and predictive uncertainty. Then, it constructs a Dual-Neighbor Graph (DNG) to combine information on both textual diversity and class diversity, ensuring a balanced data distribution. It further propagates uncertainty information via density-based clustering to select hard representative instances. DEUCE performs well in selecting class-balanced and hard representative data by dual-diversity and informativeness. Experiments on six NLP datasets demonstrate the superiority and efficiency of DEUCE.

1 Introduction

Cold-start active learning (CSAL; Yuan et al., 2020a; Zhang et al., 2022b) has gained much attention for efficiently labeling large corpora from zero. Given an unlabeled corpus (i.e., the ‘‘cold-start’’ stage), it aims to acquire a small subset (seed set) for annotation. Such absence of labels can happen due to data privacy concerns (Holzinger, 2016; Li et al., 2023), limited domain experts[1] (Wu et al., 2022), labeling difficulty (Herde et al., 2021), quick expiration of labels (Yuan et al., 2020b; Zhang et al., 2021), etc. In real-world tasks with specialized domains (e.g., medical report classification with rare diseases; De Angeli et al., 2021), the complete absence of labels and lack of *a posteriori* knowledge pose challenges to CSAL.

While active learning (AL) has been studied for a wide range of NLP tasks (Zhang et al., 2022b), the cold-start problem has been hardly addressed. At the cold-start stage, the model is untrained and no labeled data are available for validation. Traditional CSAL applies random sampling (Ash et al., 2020; Margatina et al., 2021), diversity sampling (Yu et al., 2019; Chang et al., 2021), or uncertainty sampling (Schröder et al., 2022). However, random sampling suffers from high variance (Rudolph et al., 2023); diversity sampling is prone to easy examples and vector space noise (Eklund and Forsman, 2022); and uncertainty sampling is prone to redundant examples, outliers, and unreliable metrics (Wójcik et al., 2022). Moreover, existing methods ignore class diversity, where the sampling bias often results in class imbalance (Krishnan et al., 2021). At worst, the *missed cluster effect* (Schütze et al., 2006; Yu et al., 2019) can happen, i.e., clusters of weak classes are neglected. Tomanek et al. (2009) showed that an unrepresentative seed set gives rise to this effect. Learning is misguided, if started unfavorably.

* Tong Zhang is the corresponding author.

[1]Recent studies (Lu et al., 2023; Naeini et al., 2023; Zhang et al., 2023) have shown that state-of-the-art PLMs still underperform human experts in difficult tasks.

Transactions of the Association for Computational Linguistics, vol. 12, pp. 1736–1754, 2024. https://doi.org/10.1162/tacl_a_00731
Action Editor: Sebastian Padó. Submission batch: 5/2024; Revision batch: 9/2024; Published 12/2024.

The key challenge for CSAL lies in how to acquire a diverse and informative seed set. As a general heuristic (Dasgupta, 2011), a proper seed set should strike a balance between exploring the *input space* for instance regions (e.g., diversity sampling) and exploiting the *version space* for decision boundaries (e.g., uncertainty sampling). Such hybrid CSAL strategies have been proposed based on combinations of neighbor-awareness (Hacohen et al., 2022; Su et al., 2023; Yu et al., 2023), clustering (Yuan et al., 2020a; Agarwal et al., 2021; Müller et al., 2022; Brangbour et al., 2022; Shnarch et al., 2022; Yu et al., 2023), and uncertainty estimation (Dligach and Palmer, 2011; Yuan et al., 2020a; Müller et al., 2022; Yu et al., 2023). However, existing methods fail to explore the *label space* to enhance class diversity and mitigate imbalance. Moreover, most methods perform diversity sampling followed by uncertainty sampling, treating both aspects in isolation.

To address these challenges, this paper presents DEUCE, a dual-diversity enhancing and uncertainty-aware framework for CSAL. It adopts a graph-based hybrid strategy to enhance diversity and informativeness. Different from previous works, DEUCE not only emphasizes the diversity in textual contents (textual diversity), but also diversity in class predictions (class diversity). This is termed ***dual-diversity*** in this paper. To achieve this in the cold-start stage, it exploits the rich representational and predictive capabilities of PLMs. For informativeness, the predictive uncertainty is estimated from a one-vs-all (OVA) perspective. This helps mining informative ''hard examples'' for learning. Then, DEUCE further employs manifold learning techniques (McInnes et al., 2020) to derive dual-diversity information. This results in the novel construction of a Dual-Neighbor Graph (DNG). Finally, DEUCE performs density-based uncertainty propagation and Farthest Point Sampling (FPS) on the DNG. While propagation prioritizes ***representatively uncertain*** (RU) instances, FPS enhances the dual-diversity. Overall, DEUCE ensures a more diverse and informative acquisition.

The merits of DEUCE are attributed to the following contributions:

- The dual-diversity enhancing and uncertainty aware (DEUCE) framework adopts a novel hybrid acquisition strategy. It effectively selects class-balanced and hard representative instances, achieving a good balance between exploration and exploitation in CSAL.
- This paper proposes a graph-based dual-diversity enhancement mechanism to select diverse instances with textual diversity and class diversity, tackling class imbalance in CSAL.
- This paper presents an embedding-based uncertainty-aware prediction mechanism to effectively select hard representative instances according to predictive uncertainty.

2 Related Work

2.1 Cold-start Active Learning (CSAL)

According to the taxonomy of Zhang et al. (2022b), CSAL research for NLP can be categorized as informativeness-based, representativeness-based, and hybrid. As most methods are hybrid, the techniques and challenges for informativeness or representativeness are elucidated below.

2.1.1 Informativeness

Uncertainty. The main metric for informativeness in CSAL is uncertainty, as it is more tractable in cold-start stages than others (e.g., gradients). High predictive uncertainty indicates difficulty for the model, thus valuable for annotation. Most existing methods use language models (LMs) for estimation. Common estimators include entropy (Zhu et al., 2008; Yu et al., 2023), LM probability (Dligach and Palmer, 2011), LM loss (Yuan et al., 2020a), and probability margin (Müller et al., 2022). However, several challenges exist in uncertainty estimation: (a) Often, a closed-world assumption is imposed. In other words, predictions are normalized such that they sum to 1. This hinders the expression of uncertainty, as it forces mapping to one of the known classes, ignoring options such as ''none of the above'' (Padhy et al., 2020). (b) PLMs suffer from overconfidence (Park and Caragea, 2022; Wang, 2024). This requires calibration for more robust uncertainty estimation (Yu et al., 2023). (c) Task information is hardly considered. As a result, the uncertainty will not be related to the downstream task (output uncertainty), but rather its intrinsic perplexity (input uncertainty) (Jiang et al., 2021). PATRON (Yu et al., 2023) uses task-related prompts to tackle this issue.

2.1.2 Representativeness

Density. To avoid outliers, density-based CSAL methods prefer ''typical'' instances. The method of Zhu et al. (2008) and TypiClust (Hacohen et al., 2022) prioritize instances with high kNN density. Uncertainty propagation (Yu et al., 2023) is also useful in aggregating density information. A typical group of uncertain examples indicates a region where the model's knowledge is lacking.

Discriminative. Some CSAL methods acquire sequentially or iteratively. They thus discriminate, i.e., prefer an instance if it differs the most from selected ones. Coreset selection (Sener and Savarese, 2018) selects an instance (cover-point) such that its minimum distance to selected instances is maximized. VOTE-k (Su et al., 2023) adopts a greedy approach to select remote instances on a kNN graph.

Batch Diversity. It is more efficient to acquire in batch mode (Settles, 2009), i.e., to select multiple instances at each step. Clustering has been a common technique to enhance batch diversity and avoid redundancy in CSAL. It helps structure the unlabeled dataset by grouping similar instances together. Nguyen and Smeulders (2004) and Kang et al. (2004) first proposed pre-clustering the input space to select representatives from each cluster. Dasgupta and Ng (2009) used spectral clustering on the similarity matrix of documents. Hu et al. (2010) and Yu et al. (2019) used hierarchical clustering to stabilize the process. Zhu et al. (2008) and more recent works (Yuan et al., 2020a; Chang et al., 2021; Agarwal et al., 2021; Müller et al., 2022; Hacohen et al., 2022; Yu et al., 2023) have commonly used k-MEANS for its simplicity and efficiency. However, these clustering methods can be sensitive to outliers. Moreover, clustering in the input space only contributes to textual diversity, regardless of other aspects.

2.2 Missed Cluster Effect

The missed cluster effect (Schütze et al., 2006; Tomanek et al., 2009) is an extreme case of class imbalance. It refers to when an AL strategy neglects certain classes (or clusters within classes). Schütze et al. (2006) first recognized the missed cluster effect in the context of text classification. They suggested more use of domain knowledge. Knowledge extraction from PLMs is in harmony with this suggestion. Dligach and Palmer (2011) proposed an uncertainty-based approach to avoid the missed cluster effect in word sense disambiguation (WSD). However, it is based on task-agnostic LM probability. Marcheggiani and Artières (2014) showed that labeling relevant instances, which reduces the labeling noise, also helps mitigate the missed cluster effect. Label calibration aligns with this finding. While many works are devoted to addressing the missed cluster effect or general class imbalance (e.g., Aggarwal et al., 2020; Fairstein et al., 2024) for general AL, they often rely on a labeled subset. Class diversity enhancement would help mitigate class imbalance issues, but it remains an open question for CSAL.

3 Methodology

In this section, the methodology of the proposed DEUCE is introduced. Section 3.1 first defines CSAL and declares the notations for the rest of this paper. The framework of DEUCE is then elaborated in Section 3.2.

3.1 Problem Formulation

This paper considers CSAL in a pool-based manner. Learning is initiated with a set of N unlabeled documents, $\mathcal{X} := \{x_i\}_{i=1}^N$. A C-way text classification task is defined by a set of classes $\mathcal{Y} := \{y_j\}_{j=1}^C$ taking values in a domain $\mathbb{Y}$.

Given a labeling budget $b \ll N$, a CSAL strategy acquires a subset $\mathcal{X}_s \subset \mathcal{X}$ with a fixed size $|\mathcal{X}_s| = b$, such that the labeled subset $\mathcal{X}_s'$ boosts most performance when used as a training seed set. The performance is evaluated by fine-tuning a PLM $\mathcal{M}_\theta$ with $\mathcal{X}_s'$, and testing for its accuracy.

3.2 The DEUCE Framework

The proposed DEUCE framework is illustrated in Figure 1. Overall, the components of DEUCE serve the same goal—to produce a seed set with high dual-diversity and informativeness.

3.2.1 Embedding Module

In CSAL, data selection starts with only an unlabeled corpus. DEUCE leverages PLM embeddings, which guide the selection process towards more diverse and informative samples.

Specifically, the embedding module implements a prompt-based, verbalizer-free approach

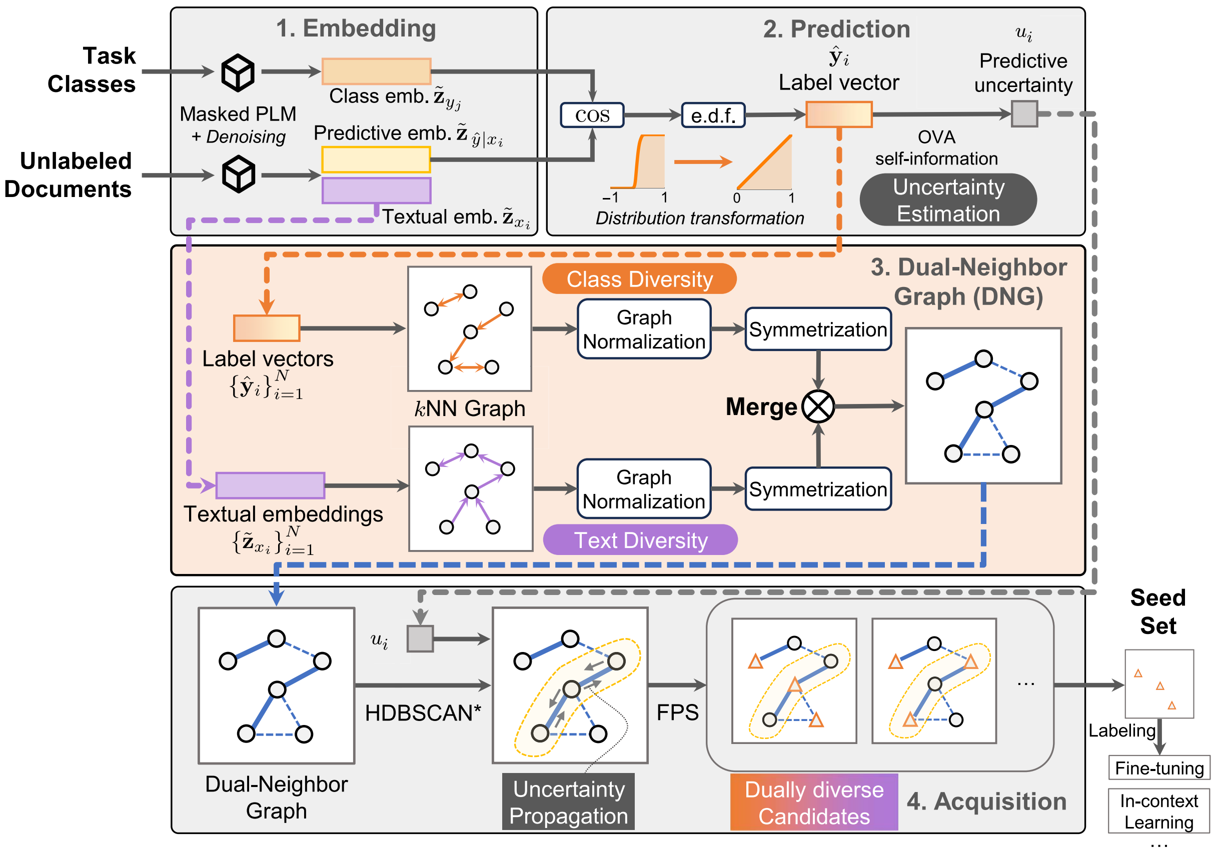


Figure 1: The proposed DEUCE framework.

(Jiang et al., 2022). This requires only a single inference pass per document.

Textual and Predictive Embedding. In a masked PLM, the bidirectional semantics can be condensed into a [MASK] token. In light of this, DEUCE extends Jiang et al. (2022)'s template with double [MASK] tokens:

$$T_x := \boxed{\begin{array}{l}\text{This sentence: ``[X]'' means [MASK].} \\ \text{Its [DOMAIN] is [MASK].}\end{array}},$$

where [DOMAIN] is the target domain $\mathbb{Y}$, such as ''sentiment''. The hidden representations of [MASK] tokens are extracted as the textual $\mathbf{z}_{x_i}$ and predictive embeddings $\mathbf{z}_{\hat{y}|x_i}$. They capture the intrinsic and task-related semantics.

However, raw embeddings suffer from template bias and length bias (Miao et al., 2023). DEUCE further applies *template denoising* (Jiang et al., 2022) to obtain the denoised embeddings $\tilde{\mathbf{z}}$.

Class Embedding. Predictions need to be paired with the known classes. Class embeddings $\tilde{\mathbf{z}}_{y_j}$ are generated from a prompt template T_y, similar to T_x:

$$T_y := \boxed{\text{This [DOMAIN]: ``[Y]'' means [MASK].}},$$

where [Y] is the placeholder for a class y_j.

3.2.2 Prediction Module

This module aims to produce uncertainty-aware labels. With class information, DEUCE gains prior knowledge about potential data distributions. With uncertainty information, DEUCE is informed of potential labeling gain.

Label Vector. For better uncertainty estimation, DEUCE adopts an OVA setup, such that labels $\hat{\mathbf{y}}_i$ do not necessarily sum to 1. First, it computes the inner product ω_{ij} for each pair of predictive and class embeddings:

$$\begin{aligned}\mathbf{\Omega} &= \begin{bmatrix} \tilde{\mathbf{z}}_{\hat{y}|x_1} & \cdots & \tilde{\mathbf{z}}_{\hat{y}|x_N} \end{bmatrix}^\top \begin{bmatrix} \tilde{\mathbf{z}}_{y_1} & \cdots & \tilde{\mathbf{z}}_{y_C} \end{bmatrix} \\ &:= \begin{bmatrix} \omega_{ij} \end{bmatrix}_{i=1,j=1}^{N,C}.\end{aligned}$$

Ideally, similarity ω_{ij} can be linearly transformed to class label $\hat{y}_{ij}$. However, high anisotropy (Gao et al., 2019) was observed in preliminary experiments. As a result, ω_{ij} has a non-uniform distribution over $[-1, 1]$. To tackle this issue, DEUCE uses the empirical distribution function (e.d.f.) of $\mathbf{\Omega}$ to give a calibrated estimate of labels $\hat{\mathbf{Y}}$:

$$\hat{y}_{ij} = \hat{\mathbb{F}}_{\mathbf{\Omega}}(\omega_{ij}) = \frac{1}{NC} \sum_{m=1}^{N} \sum_{n=1}^{C} \mathbb{1}[\omega_{mn} \leq \omega_{ij}],$$

where $\mathbb{1}[\cdot]$ is the indicator function. This gives $\hat{y}_{ij} \sim U(0, 1)$ regardless of the embedding distribution.

Predictive Uncertainty. In CSAL, uncertainty represents the difficulty of an instance. DEUCE adapts entropy, a common measure of uncertainty (§2.1.1).

In information theory, entropy is the expected self-information I of possible events. In an OVA setup, possible events $\{E_i\}$ are "x_i has a high predictive score for *exactly one* class". The probability of event E_i is given by Wójcik et al. (2022):

$$p(E_i) = \max_j \hat{y}_{ij} \prod_{\substack{l=1 \\ l \neq j}}^{C} (1 - \hat{y}_{il}).$$

Therefore, DEUCE adopts the entropy from $\{E_i\}$ as the uncertainty estimate $\mathbf{u}$:

$$u_i = I(E_i) = -\log p(E_i).$$

3.2.3 Dual-Neighbor Graph (DNG) Module

Graphs serve as a powerful tool for data selection by explicitly modeling data interrelationship. This enables the propagation of valuable information (e.g., uncertainty) and the selection of more diverse samples. To integrate textual and class diversity, DEUCE leverages manifold learning techniques (McInnes et al., 2020) on k-Nearest-Neighbor (kNN) graphs of both spaces.[2]

[2] It is worth noting that DEUCE does not utilize or optimize any Graph Neural Network (GNN). With the rich representational capability of PLMs, DEUCE does not require GNNs to learn data representations.

kNN Graph. The use of kNN arises from the neighborhood perspective of diversity. DEUCE aims to avoid selecting neighboring instances. In a kNN graph, an instance x_i is connected with its k nearest neighbors $\{x_{ij}\}$ under some distance function $\Delta(\cdot, \cdot)$. Formally, the two metric spaces of kNN are defined as follows.

- The textual space $(\mathcal{X}, \Delta_{\tilde{z}})$ is defined by textual embeddings under cosine distance, $\Delta_{\tilde{z}}(x_i, x_j) = \frac{1}{\pi} \arccos\left(\tilde{\mathbf{z}}_{x_i}^\top \tilde{\mathbf{z}}_{x_j}\right)$;
- The label space $(\mathcal{X}, \Delta_{\hat{y}})$ is defined by label vectors under ℓ_1 distance, $\Delta_{\hat{y}}(x_i, x_j) = \|\hat{\mathbf{y}}_i - \hat{\mathbf{y}}_j\|_1$.

The kNN graph from each space is denoted by $\mathcal{G}_{\tilde{z}}$ and $\mathcal{G}_{\hat{y}}$, respectively.

Graph Normalization. To unify textual and class diversity, DEUCE merges the two kNN graphs into one for graph-based sampling. However, across two distinct spaces, it is necessary to first normalize the distances (McInnes et al., 2020).

To ease notation, this part omits the subscript as $\mathcal{G} \in \{\mathcal{G}_{\tilde{z}}, \mathcal{G}_{\hat{y}}\}$. For each x_i, DEUCE finds a normalization factor $\tau_i > 0$ that satisfies the equation

$$\sum_{j=1}^{k} \exp\left(-\frac{\Delta(x_i, x_{i_j}) - \rho_i}{\tau_i}\right) = \log_2 k,$$

where ρ_i denotes x_i's distance to its nearest neighbor. The weights $\tilde{w}$ of the normalized (directed) kNN graph $\mathcal{G}$, denoted by $\tilde{\mathcal{G}}$, is defined by

$$\tilde{w}\big(\langle x_i, x_{i_j} \rangle\big) := \exp\left(-\frac{\Delta\big(x_i, x_{i_j}\big) - \rho_i}{\tau_i}\right).$$

After normalization, the original kNN weights $w \in [0, \infty)$ are transformed to $\tilde{w} \in (0, 1]$.

Symmetrization. To identify representative instances, DEUCE performs graph clustering. This requires symmetric kNN graphs.

Let $\tilde{\mathbf{W}}$ denote the sparse weight matrix of $\tilde{\mathcal{G}}$. Since weights $\tilde{w} \in [0, 1]$, they can be interpreted as fuzzy memberships of neighborhood. Hence, symmetrizing $\tilde{\mathbf{W}}$ is equivalent to finding the fuzzy union (Dubois and Prade, 1982) of the neighbors $\tilde{\mathbf{W}}$ and reverse neighbors $\tilde{\mathbf{W}}^\top$:

$$\tilde{\mathbf{W}}_{\text{sym}} = \tilde{\mathbf{W}} + \tilde{\mathbf{W}}^\top - \tilde{\mathbf{W}} \odot \tilde{\mathbf{W}}^\top,$$

where $\odot$ is the Hadamard product. $\tilde{\mathbf{W}}_{\text{sym}}$ defines the weights of the symmetric kNN graph $\tilde{\mathcal{G}}_{\text{sym}}$. Its edges are denoted by $\tilde{\mathcal{E}}_{\text{sym}}$.

Merging. It is now appropriate to merge the two kNN graphs. This unifies textual and class diversity in one graph.

As merged, the DNG is an undirected graph $\mathcal{G}_{\text{dual}} = (\mathcal{V}, \mathcal{E}, w_{\text{dual}})$. The edges $\mathcal{E}$ are the union of edges in $\tilde{\mathcal{G}}_{\tilde{z},\text{sym}}$ and $\tilde{\mathcal{G}}_{\hat{y},\text{sym}}$. Moreover, $\mathcal{E}$ is divided into two types:

- $\mathcal{E}_1$ represents edges which only appear in either kNN graph, called *single-neighbor edges*;
- $\mathcal{E}_2$ represents edges which appear in both kNN graphs, called *dual-neighbor edges*. They connect neighboring documents which are similar in both textual semantics and class predictions.

The weight w_{dual} of an undirected edge $\{x_i, x_j\} \in \mathcal{E}$ is thereby defined as

$$w_{\text{dual}} := \begin{cases} \tilde{w}_{\tilde{z},\text{sym}} \tilde{w}_{\hat{y},\text{sym}} + \gamma \\ \qquad \text{if } \{x_i, x_j\} \in \mathcal{E}_2, \\ \tilde{w}_{\tilde{z},\text{sym}} \\ \qquad \text{if } \{x_i, x_j\} \in \underbrace{\tilde{\mathcal{E}}_{\tilde{z},\text{sym}} \setminus \tilde{\mathcal{E}}_{\hat{y},\text{sym}}}_{\subset \mathcal{E}_1}, \\ \tilde{w}_{\hat{y},\text{sym}} \\ \qquad \text{if } \{x_i, x_j\} \in \underbrace{\tilde{\mathcal{E}}_{\hat{y},\text{sym}} \setminus \tilde{\mathcal{E}}_{\tilde{z},\text{sym}}}_{\subset \mathcal{E}_1}; \end{cases}$$

where γ is a threshold to distinguish dual-neighbor edges $\mathcal{E}_2$ from single-neighbor edges $\mathcal{E}_1$. In essence, DNG assigns greater weights to dual-neighbor edges. As a result, during the subsequent graph clustering and traversal, DEUCE can avoid selecting textual and class neighbors.

3.2.4 Acquisition Module

DEUCE adopts a hybrid acquisition strategy. Overall, the goal is to produce a diverse and informative seed set. To achieve this, the acquisition module performs graph clustering, propagation, and traversal on DNG.

HDBSCAN*. A group of similar documents with high predictive uncertainty indicates an area where the model's knowledge is lacking. By labeling one of the documents, the model predictions can be improved for similar ones in the area. Therefore, it is valuable to identify and prioritize such representatively uncertain (RU) groups for CSAL.

Clustering has been a common technique to group similar instances (§2.1.2). However, traditional clustering methods (e.g., k-MEANS) are ill-suited, as the number of RU groups is unknown. Moreover, they force every instance into a cluster, while some instances may not belong to any RU group. Instead, DEUCE adopts density-based clustering, which identifies RU groups with a sufficient density ($\geq k_r$ similar documents).

Specifically, DEUCE applies HDBSCAN* (Campello et al., 2013, 2015) on the DNG, with minimum cluster size k_r. A document x_i is either (a) clustered in an RU group c_l with membership p_i, or (b) excluded as a non-RU outlier.

Uncertainty Propagation. To prioritize RU documents, uncertainty information (§3.2.2) is propagated and aggregated in RU groups. This is formulated as a single step of message propagation:

$$\tilde{u}_i = u_i + \sum_{x_j \in c_l \setminus \{x_i\}} w_{\text{dual}}(\{x_i, x_j\})\, p_j u_j.$$

FPS. The final acquisition adopts a combination of diversity sampling and uncertainty sampling. First, DEUCE runs Farthest Point Sampling (FPS; Eldar et al., 1994) on the DNG. As the result only depends on the initial point, FPS is started from documents x_i with top-k degrees. Each produces a candidate seed set $\mathcal{X}_c^{(i)}$, which contains b dually diverse samples. Finally, DEUCE chooses the candidate with the highest propagated uncertainty:

$$\mathcal{X}_s = \underset{\mathcal{X}_c^{(i)}}{\arg\max} \sum_{x_j \in \mathcal{X}_c^{(i)}} \tilde{u}_j.$$

The whole process is described in Algorithm 1.

Algorithm 1 Cold-start acquisition in DEUCE.

Input: unlabeled documents $\mathcal{X}$, classes $\mathcal{Y}$, labeling budget b, number of neighbors k, representativeness threshold k_r, and frozen PLM $\mathcal{M}_\theta$.

$\triangleright$ *Embedding (§3.2.1) and prediction (§3.2.2).*
for all $y_j \in \mathcal{Y}$ **do**
  $\tilde{\mathbf{z}}_{y_j} \leftarrow \textsc{Denoise}\Big(\mathcal{M}_\theta\Big(T_y|_{y_j}\Big)\Big)$
for all $x_i \in \mathcal{X}$ **do**
  $\tilde{\mathbf{z}}_{x_i}, \tilde{\mathbf{z}}_{\hat{y}|x_i} \leftarrow \textsc{Denoise}\Big(\mathcal{M}_\theta\Big(T_x|_{x_i,\mathbb{Y}}\Big)\Big)$
  for all $y_j \in \mathcal{Y}$ **do**
    $\omega_{ij} \leftarrow \tilde{\mathbf{z}}_{\hat{y}|x_i}^\top \tilde{\mathbf{z}}_{y_j}$
    $\hat{y}_{ij} \leftarrow \hat{\mathbb{F}}_{\mathbf{\Omega}}(\omega_{ij})$
  $u_i \leftarrow -\log p(E_i')$ $\triangleright$ *Uncertainty estimation.*
$\triangleright$ *Dual-Neighbor Graph (§3.2.3).*
$\tilde{\mathcal{G}}_{\tilde{z},\text{sym}} \leftarrow \textsc{GraphNorm}(k\text{NN}(\mathcal{X}, \Delta_{\tilde{z}}))$
$\tilde{\mathcal{G}}_{\hat{y},\text{sym}} \leftarrow \textsc{GraphNorm}(k\text{NN}(\mathcal{X}, \Delta_{\hat{y}}))$
$\mathcal{G}_{\text{dual}} \leftarrow \text{DNG}\Big(\tilde{\mathcal{G}}_{\tilde{z},\text{sym}}, \tilde{\mathcal{G}}_{\hat{y},\text{sym}}; \gamma\Big)$
$\triangleright$ *Acquisition (§3.2.4).*
$\mathcal{C} \leftarrow \text{HDBSCAN*}(\mathcal{G}_{\text{dual}}; k_r)$
for all $x_i \in \mathcal{X}$ **do**
  if $\exists c_l \in \mathcal{C} : x_i \in c_l$ **then**
    $\tilde{u}_i \leftarrow \textsc{Propagate}(u_i, c_l)$
for all $x_i \in \arg\text{top-}k \deg(x_i, \mathcal{G}_{\text{dual}})$ **do**
  $\mathcal{X}_c^{(i)} \leftarrow \text{FPS}(\mathcal{G}_{\text{dual}}, x_i; b)$
return $\mathcal{X}_s \leftarrow \arg\max_{\mathcal{X}_c^{(i)}} \sum_{x_j \in \mathcal{X}_c^{(i)}} \tilde{u}_j$

Output: A dually diverse and informative seed set $\mathcal{X}_s \subset \mathcal{X}$.

4 Experiments and Results

4.1 Experimental Setup

Datasets. DEUCE is evaluated on six text classification datasets: IMDb (Maas et al., 2011), Yelp$_{\text{full}}$ (Meng et al., 2019), AG's News (Zhang et al., 2015), Yahoo! Answers (Zhang et al., 2015), DBpedia (Lehmann et al., 2015), and TREC (Li and Roth, 2002). Dataset statistics are shown in Table 1. All the datasets used in the experiments are publicly accessible. The original labels are removed to create a cold-start scenario.

Evaluation Metric. To evaluate the performance of the acquired seed set $\mathcal{X}_s$, it is labeled and used for fine-tuning the PLM. The original labels of the seed set are revealed. The accuracy of the fine-tuned PLM on the test set is then reported. To be consistent with previous methods (Yu et al., 2023), the experiments adopt RoBERTa-base (Liu et al., 2019) as the backbone PLM.

Analysis Metrics. To analyze the effect of dual-diversity enhancement, the class imbalance (IMB) and textual-diversity value of seed sets are reported. Both metrics are computed under budget $b = 128$. IMB (Yu et al., 2023) is defined as:

$$\mathsf{IMB} = \frac{\max_{j=1}^{C} n_j}{\min_{j=1}^{C} n_j},$$

where n_j is the number of instances from class y_j. Textual-diversity value (Ein-Dor et al., 2020; Yu et al., 2023) is defined as:

$$D = \left(\frac{1}{|\mathcal{X} \setminus \mathcal{X}_s|} \sum_{x_i \in \mathcal{X} \setminus \mathcal{X}_s} \min_{x_j \in \mathcal{X}_s} \Delta(x_i, x_j)\right)^{-1},$$

where $\Delta(x_i, x_j)$ is the Euclidean distance of SimCSE embeddings (Gao et al., 2021) of x_i and x_j.

Implementation Details. The fine-tuning setup and hyperparameters are the same as PATRON's (Yu et al., 2023). Notably, the experiment code transplants the original implementation of graph normalization (McInnes et al., 2018) to GPU for acceleration. For DEUCE, $k = 500$, $k_r = 3$, and $\gamma = 1.0$ (since $\tilde{w}_{\text{sym}} \leq 1.0$) are taken. All experiments are run on a machine with a single NVIDIA A800 GPU with 80 GB of VRAM.

Baselines. The following CSAL baseline methods are considered:

- **Random** sampling selects uniformly.
- **Entropy**-based uncertainty sampling (revisited by Schröder et al., 2022) selects data with the highest predictive entropy.
- **Coreset** selection (Sener and Savarese, 2018) iteratively selects data whose minimum distance to the selected data is maximized.
- **ALPS** (Yuan et al., 2020a) computes *surprisal embeddings* from BERT loss as uncertainty. They are then clustered with k-MEANS. Data closest to each centroid are selected.
- **FEW-SELECTOR** (Chang et al., 2021) clusters the text embeddings with k-MEANS.
- **TypiClust** (Hacohen et al., 2022) clusters the text embeddings with k-MEANS, and selects

Dataset	Source domain	Target domain $\mathbb{Y}$	#Class C	#Unlabeled$\lvert\mathcal{X}\rvert$	#Test	Label distribution (bar chart) and names y_j
IMDb	Movie review	Sentiment	2	25,000	25,000	Negative, Positive
$\text{Yelp}_{\text{full}}$	Review	Rating	5	38,352	10,000	1 star, 2 stars, 3 stars, 4 stars, 5 stars
AG's News	News	Category	4	120,000	7,600	World, Sports, Business, Sci/Tech
Yahoo! Answers	Web Q&A	Category	10	300,000†	60,000	Society & Culture, Science & Mathematics, Health, Education & Reference, Computers & Internet, Sports, Business & Finance, Entertainment & Music, Family & Relationships, Politics & Government
DBpedia	Wikipedia lead section	Category	14	420,000†	70,000	Company, Educational institution, Artist, Athlete, Office holder, Mean of transportation, Building, Natural place, Village, Animal, Plant, Album, Film, Written work
TREC	Question	Category	6	5,452	500	‡ Abbreviation, Entity, Description and abstract concept, Human being, Location, Numeric value

Table 1: Statistics of evaluation datasets. †*Yahoo!* and *DBpedia* are the truncated version with 30k samples per class by Yu et al. (2023). ‡*TREC* is an imbalanced dataset.

Method	Informativeness	Representativeness		
	Uncertainty	Density	Textual diversity	Class diversity
Random	✗	✗	✗	✗
Entropy	✓	✗	✗	✗
Coreset	✗	✗	✓	✗
ALPS	✓	✗	✓	✗
FEW-S.	✗	✗	✓	✗
TypiCl.	✗	✓	✓	✗
PATRON	✓	✓	✓	✗
VOTE-k	✗	✓	✓	✗
DEUCE	✓	✓	✓	✓

Table 2: Comparisons of CSAL methods, which adapt the taxonomy of Zhang et al. (2022b) (§2.1).

data with the highest typicality, i.e., kNN density, from each cluster.

- **PATRON** (Yu et al., 2023) clusters the text embeddings with k-MEANS, and selects from each cluster data with the highest propagated uncertainty. It then iteratively updates the set to refine inter-sample distances.
- **VOTE-k** (Su et al., 2023) iteratively assigns a high score if a data is far from selected data.

Comparisons of the CSAL baselines and DEUCE are presented in Table 2.

4.2 Accuracy Improvement

The main quantitative results of PLM fine-tuning performance with DEUCE and baseline CSAL methods are shown in Table 3. Results for baselines other than VOTE-k are from Yu et al. (2023). To report the standard deviation, each setup is repeated with 10 different random seeds. Figure 2 demonstrates a qualitative visualization of the $b = 128$ seed set from IMDb dataset, acquired by the latest baseline method VOTE-k and the proposed DEUCE. The t-SNE (van der Maaten and Hinton, 2008) method is used for visualization.

From results in Table 3, it can be seen that DEUCE consistently outperforms other baselines, achieving up to a 2.5% gain on balanced datasets and up to 6.2% on the imbalanced dataset, TREC. DEUCE mainly benefits from that it enhances the class diversity as well as textual diversity. This can be concluded from the larger improvements on TREC. In over half of the setups, DEUCE also achieves the lowest standard deviation. In addition, DEUCE improves most when b is small. This aligns with the fundamental goal of AL, which is to maximize performance gains with minimal labeled data. Furthermore, from the visualization in Figure 2, it can be seen that DEUCE's enhancement of dual-diversity leads to a broader and more balanced coverage of both input space and label space. As DEUCE adopts a highest-uncertainty strategy, such coverage also exhibits high predictive uncertainty, thus including more ''hard examples'' which are valuable for annotation.

4.3 Enhancement of Class Diversity

To verify the enhancement of class diversity, the class imbalance value (Yu et al., 2023) under $b = 128$ is reported in Table 4.

From Table 4, it can be seen that DEUCE achieves the lowest average IMB value. This indicates that DEUCE enhances class diversity properly. In contrast, an IMB of ∞ emerges in the pure uncertainty-based (Entropy) and textual-diversity-based (Coreset) method. This indicates the missed cluster effect happens in their acquisition.

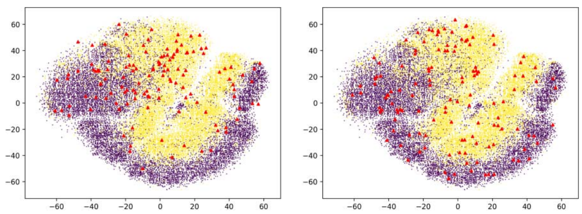
4.4 Enhancement of Textual Diversity

To measure the textual diversity of seed sets, the textual-diversity value (Ein-Dor et al., 2020; Yu et al., 2023) under $b = 128$ is reported in Table 5.

Table 5 shows that DEUCE also achieves the highest average textual-diversity value. This indicates that DEUCE also enhances textual diversity

Dataset	b	Random	Entropy	Coreset	ALPS	FEW-S.	TypiCl.	PATRON	VOTE-k	DEUCE
IMDb	32	80.2±2.5	81.9±2.7	74.5±2.9	82.2±3.0	79.2±1.6	82.8±2.2	85.5±1.5	85.6±1.8	**86.9±0.9**
	64	82.6±1.4	84.7±1.5	82.8±2.5	86.1±0.9	84.9±1.5	84.0±0.9	87.3±1.0	88.0±1.2	**88.5±0.7**
	128	86.6±1.7	87.1±0.7	87.8±0.8	87.5±0.8	88.5±1.6	88.1±1.4	89.6±0.4	89.1±0.7	**90.0±0.3**
Yelp$_{\text{full}}$	32	30.2±4.5	32.7±1.0	32.9±2.8	36.8±1.8	35.2±1.0	32.6±1.5	35.9±1.6	40.1±2.2	**42.6±1.1**
	64	42.5±1.7	36.8±2.1	39.9±3.4	40.3±2.6	39.3±**1.0**	39.7±1.8	44.4±1.1	49.3±1.6	**49.8±1.2**
	128	47.7±2.1	41.3±1.9	49.4±1.6	45.1±1.0	46.4±1.3	46.8±1.6	51.2±0.8	50.8±1.5	**53.4±0.7**
AG's News	32	73.7±4.6	73.7±3.0	78.6±1.6	78.4±2.3	79.1±2.7	80.7±1.8	83.2±0.9	81.8±1.3	**83.7±0.8**
	64	80.0±2.5	80.0±2.2	82.0±1.5	82.6±2.5	82.4±2.0	83.0±2.4	85.3±0.7	84.7±1.3	**86.3±0.6**
	128	84.5±1.7	82.5±0.8	85.2±0.6	84.3±1.7	85.6±0.8	85.7±0.3	87.0±0.6	86.2±1.2	**87.5±0.4**
Yahoo! Answers	32	43.5±4.0	23.0±1.6	22.0±2.3	47.7±2.3	46.8±2.1	36.9±1.8	56.8±**1.0**	54.5±1.6	**58.0**±1.5
	64	53.1±3.1	37.6±2.0	45.7±3.7	55.3±1.8	52.9±1.6	54.0±1.6	61.9±**0.7**	60.8±1.4	**62.8**±1.3
	128	60.2±1.5	41.8±1.9	56.9±2.5	60.8±1.9	61.3±1.0	58.2±1.5	65.1±**0.6**	64.3±0.9	**66.2**±0.9
DBpedia	32	67.1±3.2	18.9±2.4	64.0±2.8	77.5±4.0	83.3±1.0	78.2±1.8	85.3±**0.9**	78.1±2.6	**86.0**±1.7
	64	86.2±2.4	37.5±3.0	85.2±0.8	89.7±1.1	92.7±0.9	88.5±0.7	93.6±**0.4**	92.7±1.3	**94.1**±0.9
	128	95.0±1.5	47.5±2.3	89.4±1.5	95.7±0.4	96.5±0.5	95.7±0.6	97.0±**0.2**	96.4±0.4	**97.3**±0.3
TREC	32	49.0±3.5	46.6±1.4	47.1±3.6	60.5±3.7	60.3±1.5	42.0±4.4	64.0±**1.2**	57.6±2.9	**70.2**±1.7
	64	69.1±3.4	59.8±4.2	75.7±3.0	73.0±2.0	77.3±2.0	72.6±2.1	78.6±1.6	81.8±3.1	**82.2±1.5**
	128	85.6±2.8	75.0±1.8	87.6±3.0	87.3±3.6	87.7±1.5	83.0±3.8	91.1±0.8	89.7±2.6	**92.1±0.8**
Average	32	57.2±3.8	46.1±2.1	53.2±2.7	63.9±3.0	64.0±1.8	58.9±2.5	68.4±**1.2**	66.3±2.1	**71.2**±1.3
	64	68.9±2.5	56.1±2.7	68.5±2.7	71.2±1.9	71.6±1.6	70.3±1.7	75.2±**1.0**	76.2±1.8	**77.3**±1.1
	128	76.6±1.9	62.5±1.7	76.1±1.9	76.8±1.9	77.6±1.2	76.3±1.9	80.2±0.6	79.4±1.4	**81.1±0.6**

Table 3: Evaluation results of DEUCE and CSAL baselines on six datasets and three budgets (denoted by b), each with 10 repetitions. Accuracy (%) of one-round fine-tuned PLM is reported in the format of avg$_{\pm\text{std}}$. The **best** and second best results per setup are emboldened and underlined, respectively.



(a) VOTE-k. (b) The proposed DEUCE.

Figure 2: The t-SNE visualization of the acquired seed set ($b = 128$) on IMDb dataset. Text embeddings are colored by their true labels.

properly. The improvement of textual-diversity value is not significant, compared to IMB value's (Table 4). This signals that DEUCE enhances more of class diversity than textual diversity, compared to other baselines. Such difference can be explained by the highest-uncertainty-candidate strategy, which acquires more information from the label space.

4.5 Quality of Textual Embedding

To analyze the quality of DEUCE's prompt-based, unsupervised text embeddings $\tilde{\mathbf{z}}_{x_i}$ (§3.2.1), they are compared with the supervised Sentence Transformer embeddings (Sentence Transformers, 2024) used in VOTE-k (Su et al., 2023). The correlations are computed across all the possible $\binom{N}{2}$ pairs of their cosine similarity.[3] Results on three datasets are reported in Table 6.

From Table 6, a weak positive correlation is observed. Moreover, template denoising produces better embeddings, as it removes the biases from raw embeddings. Overall, the quality of textual embeddings is acceptable and adequate for cold-start acquisition.

4.6 Quality of Class Prediction

To analyze the quality of embedding-based class prediction $\hat{\mathbf{y}}_i$ (§3.2.2), they are compared with gold labels. As uncertainty indicates unstable predictions, labels are arranged from the most confident (lowest u_i) to the least. Results are demonstrated in Figure 3.

From Figure 3, a high accuracy of class predictions is consistently observed with high confidence and with denoised embeddings, and vice versa. This demonstrates the good quality of e.d.f. predictions and the derived uncertainty metric.

[3]Semantic similarity benchmarks (e.g., STS) cannot be used here, as the prompt T_x requires a task domain $\mathbb{Y}$.

Dataset	Random	Entropy	Coreset	ALPS	FEW-S.	TypiCl.	PATRON	VOTE-k	DEUCE
IMDb	1.207	6.111	1.000	1.783	1.286	2.765	1.286	1.065	1.169
Yelp$_{\text{full}}$	1.778	3.800	6.000	2.833	2.000	5.200	2.250	1.273	1.450
AG's News	1.462	28.000	2.000	1.667	1.500	1.818	1.500	2.200	1.133
Yahoo! Answers	3.000	12.000	7.000	5.500	2.250	3.333	5.500	3.333	2.125
DBpedia	3.500	∞	9.000	9.000	3.500	9.000	2.333	2.800	3.250
TREC	8.000	16.000	∞	9.500	10.500	21.000	15.000	11.333	6.000
Harmonic avg.	2.128	9.863	3.124	3.138	2.166	3.839	2.338	<u>2.052</u>	**1.779**

Table 4: Label imbalance value (IMB) of acquired seed sets ($b = 128$). Smaller value indicates better class diversity and balance. An IMB of ∞ indicates that the missed cluster effect happens.

Dataset	Random	Entropy	Coreset	ALPS	FEW-S.	TypiCl.	PATRON	VOTE-k	DEUCE
IMDb	0.646	0.647	0.643	0.647	0.687	0.648	0.684	0.669	0.670
Yelp$_{\text{full}}$	0.645	0.626	0.456	0.680	0.685	0.677	0.685	0.657	0.679
AG's News	0.354	0.295	0.340	0.385	0.436	0.376	0.423	0.370	0.448
Yahoo! Answers	0.430	0.375	0.400	0.441	0.470	0.438	0.486	0.451	0.491
DBpedia	0.402	0.316	0.381	0.420	0.461	0.399	0.459	0.434	0.476
TREC	0.301	0.298	0.298	0.339	0.337	0.326	0.338	0.346	0.353
Average	0.463	0.426	0.420	0.485	<u>0.513</u>	0.477	0.512	0.488	**0.520**

Table 5: Textual diversity value D of acquired seed sets ($b = 128$). Larger values indicate better textual diversity.

Dataset	Pearson correlation r	Spearman correlation ρ
IMDb	0.1651	0.1636
w/ denoising	**0.1980**	**0.1889**
Yelp$_{\text{full}}$	0.1424	0.1440
w/ denoising	**0.3072**	**0.2984**
TREC	0.4271	0.4000
w/ denoising	**0.4662**	**0.4368**

Table 6: The quality of textual embeddings, without and with template denoising (Jiang et al., 2022). Both correlation metrics are over $[-1, 1]$; higher values indicate better quality.

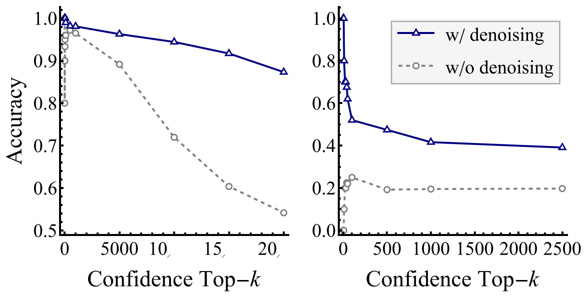


Figure 3: The quality of class predictions with respect to predictive uncertainty u_i. Dataset: IMDb (left) and TREC (right).

5 Discussion

5.1 Comparison with LLM-based Methods

The landscape of NLP is rapidly evolving with generative large language models (LLMs). This section evaluates two potential LLM-based alternatives to DEUCE: serialization for acquisition and zero-shot Chain-of-Thought prompting. The following experiments are conducted with LLAMA 2 7B (Touvron et al., 2023).

5.1.1 Serialization for Acquisition

Inspired by the work of Hegselmann et al. (2023), class and uncertainty information can be serialized into natural language for LLM-based acquisition. The process is designed to involve three passes. In the first pass, each unlabeled text is formalized as a multiple-choice problem for LLM. The prompt template T_1 is used to collect class and uncertainty information:

$T_1 :=$

This sentence: ''[X]'' What is its [DOMAIN]?
Answer Choices: (A) [CLASS A] (B) . . .
Answer: (

In the second pass, LLM decides on whether each text should be selected. Predictive uncertainty is estimated by the entropy of first-pass predictions,

Method	b	IMDb	$\text{Yelp}_{\text{full}}$	AG's News	Yahoo!	DBpedia	TREC	*Average*
	32	81.7	**44.5**	25.2	38.8	62.6	28.4	46.9
Serialization	64	83.8	**51.2**	53.4	55.7	45.9	27.8	53.0
	128	89.6	**56.9**	83.7	63.4	58.6	35.6	64.6
	32	**86.9**	42.6	**83.7**	**58.0**	**86.0**	**70.2**	**71.2**
DEUCE	64	**88.5**	49.8	**86.3**	**62.8**	**94.1**	**82.2**	**77.3**
	128	**90.0**	53.4	**87.5**	**66.2**	**97.3**	**92.1**	**81.1**

Table 7: Fine-tuning results of DEUCE (RoBERTa-base) and LLM serialization (LLAMA 2 7B).

Method	IMDb	$\text{Yelp}_{\text{full}}$	AG's News	Yahoo!	DBpedia	TREC	*Average*
0-shot CoT, w/o choices	63.6	9.2	34.7	23.7	37.1	12.6	32.0
0-shot CoT, w/ choices	72.1	25.4	60.2	43.6	32.3	24.2	43.0
DEUCE, $b = 32$	**86.9**	**42.6**	**83.7**	**58.0**	**86.0**	**70.2**	**71.2**

Table 8: Evaluation results of DEUCE ($b = 32$, RoBERTa-base) and zero-shot Chain-of-Thought prompting (Kojima et al., 2022; LLAMA 2 7B).

bounded by $\log C$. The extended template T_2 is used to combine multiple information:

$T_2 :=$

```
This sentence: ‘‘[X]’’ What is its [DOMAIN]?
Answer Choices: (A) [CLASS A] (B) . . .
Answer: ([ANSWER]) [CLASS]
Uncertainty: [UNCERTAINTY %]
Is it valuable for annotation? Yes or no?
Answer:
```

In the third pass, texts with top-b probabilities of T_2 answered ‘‘yes’’ are selected as the seed set. LLM is then fine-tuned with the seed set under T_1. Finally, T_1 is applied on the fine-tuned LLM to report the test set accuracy.

Due to resource constraints, LoRA (Hu et al., 2022) is used for fine-tuning, with $r = \alpha = 64$. Results are reported in Table 7. Despite utilizing a mid-sized PLM, DEUCE outperforms serialization with LLM in most datasets. The decision process of LLM is also black-box. In contrast, DEUCE adopts graphs to explicitly capture the interplay of information, offering better interpretability.

5.1.2 Zero-shot Chain-of-Thought

Zero-shot Chain-of-Thought (CoT) prompting (Kojima et al., 2022) with LLMs has emerged as a promising method in cold-start scenarios. This paper tests zero-shot CoT without and with explicit choices in prompts. The temperature of generation is set to 0, and a maximum of 256 tokens are generated. Results are shown in Table 8.

Stage	0-shot CoT Energy (kJ)	0-shot CoT Time (sec)	DEUCE Energy (kJ)	DEUCE Time (sec)
Acquisition	–		59.82	81.00
Fine-tuning	–		225.77	208.89
Prediction	2561.58	1967.23	41.99	24.27
Total	2561.58	1967.23	**327.58**	**314.16**

Table 9: Energy consumption and time usage of DEUCE ($b = 32$, RoBERTa-base) and zero-shot Chain-of-Thought prompting (Kojima et al., 2022; LLAMA 2 7B), under the same data amount of 25000.

From the results, fine-tuning PLM with DEUCE still outperforms 0-shot LLM predictions. In class-imbalanced and difficult datasets, performance gaps are greater. Lemon-picking shows that the LLM failed to output a final answer within 256 tokens for many test instances.

In addition, the average total GPU and CPU energy consumption and time usage are measured using Alizadeh and Castor's (2024) method. Results are reported in Table 9. There is a 7.82× difference in energy consumption and 6.26× in time consumption. While increasing the number of output tokens might improve, the added resource consumption cannot be neglected. DEUCE provides an efficient solution for low-resource scenarios.

5.2 Effect of Labeling Noise

Real-world annotations often involve noise. Northcutt et al. (2021) estimated an average of 2.6% labeling errors across 3 commonly used NLP datasets. To evaluate DEUCE under labeling

DEUCE	b	**IMDb**	**Yelp$_{\text{full}}$**	**AG's News**	**Yahoo!**	**DBpedia**	**TREC**	***Average***
	32	86.9$\pm$0.9	42.6$\pm$1.1	83.7$\pm$0.8	58.0$\pm$1.5	86.0$\pm$1.7	70.2$\pm$1.7	71.2$\pm$1.3
w/o noise	64	88.5$\pm$0.7	49.8$\pm$1.2	86.3$\pm$0.6	62.8$\pm$1.3	94.1$\pm$0.9	82.2$\pm$1.5	77.2$\pm$1.1
	128	90.0$\pm$0.3	53.4$\pm$0.7	87.5$\pm$0.4	66.2$\pm$0.9	97.3$\pm$0.3	92.1$\pm$0.8	81.1$\pm$0.6
	32	67.8$\pm$4.3	38.7$\pm$3.0	72.5$\pm$1.0	49.7$\pm$7.2	61.5$\pm$2.0	69.6$\pm$0.6	60.0$\pm$1.5
w/ noise	64	83.4$\pm$1.3	41.0$\pm$2.7	82.6$\pm$1.4	53.4$\pm$2.7	87.5$\pm$3.3	78.7$\pm$3.3	71.1$\pm$1.1
	128	82.9$\pm$6.3	45.1$\pm$1.7	84.7$\pm$2.4	62.7$\pm$1.3	89.2$\pm$3.7	82.5$\pm$3.8	74.5$\pm$1.5

Table 10: Evaluation results of DEUCE, compared under an expected labeling noise level of 7%.

	b	**IMDb**	**Yelp$_{\text{full}}$**	**AG's News**	**Yahoo!**	**DBpedia**	**TREC**	***Average***
	32	74.5$\pm$**2.9**	32.9$\pm$2.8	78.6$\pm$**1.6**	22.0$\pm$**2.3**	64.0$\pm$2.8	47.1$\pm$**3.6**	53.2$\pm$2.7
Coreset	64	82.8$\pm$**2.5**	39.9$\pm$3.4	82.0$\pm$1.5	45.7$\pm$3.7	**85.2**$\pm$**0.8**	75.7$\pm$3.0	68.5$\pm$2.7
	128	87.8$\pm$**0.8**	49.4$\pm$1.6	85.2$\pm$0.6	56.9$\pm$2.5	89.4$\pm$1.5	**87.6**$\pm$3.0	76.1$\pm$1.9
	32	**83.3**$\pm$4.1	**44.1**$\pm$**0.7**	**83.4**$\pm$2.0	**52.3**$\pm$3.9	**63.2**$\pm$**1.1**	**64.9**$\pm$3.9	**65.2**$\pm$**1.2**
DEUCE w/ rand. pred.	64	**85.9**$\pm$4.5	**48.0**$\pm$**0.3**	**84.6**$\pm$**1.2**	**60.0**$\pm$**0.6**	82.9$\pm$1.7	**78.2**$\pm$**2.0**	**73.3**$\pm$**0.9**
	128	**86.6**$\pm$2.5	**49.5**$\pm$**0.4**	**87.2**$\pm$**0.4**	**63.4**$\pm$**1.3**	**96.8**$\pm$**0.1**	86.8$\pm$**1.3**	**78.4**$\pm$**0.5**

Table 11: Ablation results of DEUCE with random class predictions, compared with Coreset selection (Sener and Savarese, 2018). In this case, the class and uncertainty information are disarranged.

noise, experiments with artificial errors are conducted. As the gold labels may already contain around 3% errors, 7% of seed labels are randomly replaced by wrong labels. The final sets are expected to exhibit an error level of 4–10%. Results are reported in Table 10.

From the results, a decrease in accuracy and an increase in standard deviation occur as expected. However, DEUCE still outperforms 0-shot CoT (Table 8) in nearly all setups, despite the added noise. This shows the robustness of DEUCE for fine-tuning to labeling noise.

5.3 Effect of Class Prediction Failure

For real-world cold-start tasks, the knowledge about classes might not be well exploited by the PLM. In the worst case, the PLM can fail to generate meaningful class predictions. To simulate this scenario, ablation experiments with random class predictions are conducted. In this setup, the predictive embeddings $\mathbf{z}_{\hat{y}|x_i}$ are replaced with random vectors. This ablates class predictions. Results are reported in Table 11.

As class and uncertainty information are disarranged, DEUCE degenerates to single textual diversity and performance degradation occurs as expected. Nonetheless, DEUCE still outperforms Coreset selection (Sener and Savarese, 2018), a CSAL baseline which also purely utilizes textual diversity. This demonstrates DEUCE's effectiveness in real-world cold-start scenarios.

Method	4-shot	8-shot	*Average*
Random	25.1	24.3	24.7
DEUCE	**25.8**	**27.4**	**26.6**

Table 12: Evaluation results of DEUCE (RoBERTa-base) with few-shot Chain-of-Thought prompting (Wei et al., 2022; LLAMA 2 7B) on GSM8K dataset (Cobbe et al., 2021), compared to random sampling.

5.4 Performance of Few-shot Math Reasoning

DEUCE has the potential to generalize on other NLP tasks. To demonstrate this, DEUCE is tested on GSM8K (Cobbe et al., 2021), a dataset of math word problems. However, directly adapting RoBERTa to solving math problems is difficult due to its masked modeling nature. Instead, DEUCE is applied with RoBERTa to produce a seed set.[4] Then, the seeds are taken as examples for few-shot Chain-of-Thought prompting (Wei et al., 2022) with LLAMA 2 7B. From the results, as reported in Table 12, DEUCE is still effective in few-shot math problem solving, compared to random sampling.

[4]For open questions like math problems, there are no concepts of ''classes''. Instead, the predictive embeddings $\tilde{\mathbf{z}}_{\hat{y}|x_i}$ are clustered with HDBSCAN*. The cluster centroids are taken as *meta*-class embeddings $\mathbf{z}_{\hat{y}}$.

6 Conclusion

This paper presents DEUCE, a dual-diversity enhancing and uncertainty-aware CSAL framework via a prompt-based and graph-based approach. Different from previous works, it emphasizes dual-diversity (i.e., textual diversity and class diversity) to ensure a balanced acquisition. This is achieved by the novel construction of Dual-Neighbor Graph (DNG) and Farthest Point Sampling (FPS). DNG leverages the kNN graph structure of textual space and label space from a PLM. In addition, DEUCE prioritizes hard representative examples, so as to ensure an informative acquisition. This leverages density-based clustering and uncertainty propagation on the DNG. Experiments show the effectiveness of DEUCE's dual-diversity enhancement and uncertainty-aware mechanism. It offers an efficient solution for low-resource data acquisition. Overall, DEUCE's hybrid strategy strikes an important balance between exploration and exploitation in CSAL.

Limitations

Backbone LM. DEUCE leverages a discriminative PLM. However, state-of-the-art PLMs are primarily generative. Generative embedding models (e.g., Jiang et al., 2023) or adaptations (Yang et al., 2019; Gong et al., 2019; Zhang et al., 2022a) can be investigated and combined with DEUCE. For such approaches, their quality and efficiency should be carefully minded.

External Knowledge. In DEUCE, the only source of external knowledge is the language model. Incorporation of more domain knowledge, if possible, can improve the performance in the cold-start stage. As DEUCE adopts a prompt-based and graph-based acquisition, prompt engineering and knowledge graphs (Pan et al., 2024) can be investigated.

Acknowledgments

We extend our gratitude to our action editor, Sebastian Padó, and the anonymous reviewers for their constructive comments. We also thank Tianjun Li and Jiangfeng Liu for their helpful feedback on the initial drafts.

This work was funded in part by the National Natural Science Foundation of China grant under number 62222603, in part by the STI2030-Major Projects grant from the Ministry of Science and Technology of the People's Republic of China under number 2021ZD0200700, in part by the Key-Area Research and Development Program of Guangdong Province under number 2023B0303030001, in part by the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (2019ZT08X214), and in part by the Science and Technology Program of Guangzhou under number 2024A04J6310.

References

Deepesh Agarwal, Pravesh Srivastava, Sergio Martin-del-Campo, Balasubramaniam Natarajan, and Babji Srinivasan. 2021. Addressing practical challenges in active learning via a hybrid query strategy. *arXiv preprint arXiv:2110.03785v1*.

Umang Aggarwal, Adrian Popescu, and Céline Hudelot. 2020. Active learning for imbalanced datasets. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1417–1426. https://doi.org/10.1109/WACV45572.2020.9093475

Negar Alizadeh and Fernando Castor. 2024. Green AI: A preliminary empirical study on energy consumption in DL models across different runtime infrastructures. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI*, CAIN '24, pages 134–139, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3644815.3644967

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep batch active learning by diverse, uncertain gradient lower bounds. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020*. OpenReview.net.

Etienne Brangbour, Pierrick Bruneau, Thomas Tamisier, and Stéphane Marchand-Maillet. 2022. Cold start active learning strategies in the context of imbalanced classification. *arXiv preprint arXiv:2201.10227v1*.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-37456-2_14

Ricardo J. G. B. Campello, Davoud Moulavi, Arthur Zimek, and Jörg Sander. 2015. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data*, 10(1). https://doi.org/10.1145/2733381

Ernie Chang, Xiaoyu Shen, Hui-Syuan Yeh, and Vera Demberg. 2021. On training instance selection for few-shot neural text generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 8–13, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-short.2

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168v2*.

Sajib Dasgupta and Vincent Ng. 2009. Mine the easy, classify the hard: A semi-supervised approach to automatic sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 701–709, Suntec, Singapore. Association for Computational Linguistics. https://doi.org/10.3115/1690219.1690244

Sanjoy Dasgupta. 2011. Two faces of active learning. *Theoretical Computer Science*, 412(19):1767–1781. Algorithmic Learning Theory (ALT 2009). https://doi.org/10.1016/j.tcs.2010.12.054

Kevin De Angeli, Shang Gao, Mohammed Alawad, Hong-Jun Yoon, Noah Schaefferkoetter, Xiao-Cheng Wu, Eric B. Durbin, Jennifer Doherty, Antoinette Stroup, Linda Coyle, Lynne Penberthy, and Georgia Tourassi. 2021. Deep active learning for classifying cancer pathology reports. *BMC Bioinformatics*, 22(1). https://doi.org/10.1186/s12859-021-04047-1, PubMed: 33750288

Dmitriy Dligach and Martha Palmer. 2011. Good seed makes a good crop: Accelerating active learning using language modeling. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 6–10, Portland, Oregon, USA. Association for Computational Linguistics.

Didier Dubois and Henri Prade. 1982. A class of fuzzy measures based on triangular norms: A general framework for the combination of uncertain information. *International Journal of General Systems*, 8(1):43–61. https://doi.org/10.1080/03081078208934833

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active learning for BERT: An empirical study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.638

Anton Eklund and Mona Forsman. 2022. Topic modeling by clustering language model embeddings: Human validation on an industry dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 635–643, Abu Dhabi, UAE. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-industry.65

Yuval Eldar, Micahel Lindenbaum, Moshe Porat, and Yehoshua Y. Zeevi. 1994. The farthest point strategy for progressive image sampling. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 2 - Conference B: Computer Vision & Image*

Processing. (Cat. No.94CH3440-5), volume 3, pages 93–97. https://doi.org/10.1109/ICPR.1994.577129

Yaron Fairstein, Oren Kalinsky, Zohar Karnin, Guy Kushilevitz, Alexander Libov, and Sofia Tolmach. 2024. Class balancing for efficient active learning in imbalanced datasets. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 77–86, St. Julians, Malta. Association for Computational Linguistics.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. 2019. Representation degeneration problem in training natural language generation models. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019*. OpenReview.net.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.552

Xin-Rong Gong, Jian-Xiu Jin, and Tong Zhang. 2019. Sentiment analysis using autoregressive language modeling and broad learning system. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1130–1134. https://doi.org/10.1109/BIBM47256.2019.8983025

Guy Hacohen, Avihu Dekel, and Daphna Weinshall. 2022. Active learning on a budget: Opposite strategies suit high and low budgets. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8175–8195. PMLR.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. 2023. TabLLM: Few-shot classification of tabular data with large language models. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 5549–5581. PMLR.

Marek Herde, Denis Huseljic, Bernhard Sick, and Adrian Calma. 2021. A survey on cost types, interaction schemes, and annotator performance models in selection algorithms for active learning in classification. *IEEE Access*, 9:166970–166989. https://doi.org/10.1109/ACCESS.2021.3135514

Andreas Holzinger. 2016. Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Informatics*, 3(2):119–131. https://doi.org/10.1007/s40708-016-0042-6, PubMed: 27747607

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25–29, 2022*. OpenReview.net.

Rong Hu, Brian Mac Namee, and Sarah Jane Delany. 2010. Off to a good start: Using clustering to select the initial training set in active learning. In *Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference, May 19–21, 2010, Daytona Beach, Florida, USA*. AAAI Press.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. 2023. Scaling sentence embeddings with large language models. *arXiv preprint arXiv:2307.16645v1*.

Ting Jiang, Jian Jiao, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Denvy Deng, and Qi Zhang. 2022. PromptBERT: Improving BERT sentence embeddings with prompts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8826–8837, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.603

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? On the calibration of language models for question answering. *Transactions of the Association for Computational*

Linguistics, 9:962–977. https://doi.org/10.1162/tacl_a_00407

Jaeho Kang, Kwang Ryel Ryu, and Hyuk-Chul Kwon. 2004. Using cluster-based sampling to select initial training set for active learning in text classification. In *Advances in Knowledge Discovery and Data Mining*, pages 384–388, Berlin, Heidelberg. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-24775-3_46

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Ranganath Krishnan, Alok Sinha, Nilesh A. Ahuja, Mahesh Subedar, Omesh Tickoo, and Ravi R. Iyer. 2021. Mitigating sampling bias and improving robustness in active learning. In *Proceedings of Workshop on Human in the Loop Learning (HILL) in International Conference on Machine Learning (ICML 2021)*.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. DBpedia – a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195. https://doi.org/10.3233/SW-140134

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*. https://doi.org/10.3115/1072228.1072378

Yansong Li, Zhixing Tan, and Yang Liu. 2023. Privacy-preserving prompt tuning for large language model services. *arXiv preprint arXiv:2305.06212v1*.

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:1907.11692v1*.

Yuxuan Lu, Bingsheng Yao, Shao Zhang, Yun Wang, Peng Zhang, Tun Lu, Toby Jia-Jun Li, and Dakuo Wang. 2023. Human still wins over LLM: An empirical study of active learning on domain-specific annotation tasks. *arXiv preprint arXiv:2311.09825v1*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605.

Diego Marcheggiani and Thierry Artières. 2014. An experimental comparison of active learning strategies for partially labeled sequences. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 898–906, Doha, Qatar. Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1097

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.51

Leland McInnes, John Healy, and James Melville. 2020. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426v3*.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861. https://doi.org/10.21105/joss.00861

Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2019. Weakly-supervised hierarchical text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6826–6833. https://doi.org/10.1609/aaai.v33i01.33016826

Pu Miao, Zeyao Du, and Junlin Zhang. 2023. DebCSE: Rethinking unsupervised contrastive sentence embedding learning in the debiasing perspective. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, pages 1847–1856, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3583780.3614833

Thomas Müller, Guillermo Pérez-Torró, Angelo Basile, and Marc Franco-Salvador. 2022. Active few-shot learning with FASL. In *Natural Language Processing and Information Systems; 27th International Conference on Applications of Natural Language to Information Systems, NLDB 2022, Valencia, Spain, June 15–17, 2022, Proceedings*, pages 98–110, Cham. Springer International Publishing. https://doi.org/10.1007/978-3-031-08473-7_9

Saeid Alavi Naeini, Raeid Saqur, Mozhgan Saeidi, John Michael Giorgi, and Babak Taati. 2023. Large language models are fixated by red herrings: Exploring creative problem solving and Einstellung effect using the Only Connect Wall dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Hieu T. Nguyen and Arnold Smeulders. 2004. Active learning using pre-clustering. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 79, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/1015330.1015349

Curtis Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Shreyas Padhy, Zachary Nado, Jie Ren, Jeremiah Liu, Jasper Snoek, and Balaji Lakshminarayanan. 2020. Revisiting one-vs-all classifiers for predictive uncertainty and out-of-distribution detection in neural networks. In *ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning*.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–20. https://doi.org/10.1109/TKDE.2024.3352100

Seo Yeon Park and Cornelia Caragea. 2022. On the calibration of pre-trained language models using mixup guided by area under the margin and saliency. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5364–5374, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.368

Kara E. Rudolph, Nicholas T. Williams, Caleb H. Miles, Joseph Antonelli, and Ivan Diaz. 2023. All models are wrong, but which are useful? Comparing parametric and nonparametric estimation of causal effects in finite samples. *Journal of Causal Inference*, 11(1):20230022. https://doi.org/10.1515/jci-2023-0022

Christopher Schröder, Andreas Niekler, and Martin Potthast. 2022. Revisiting uncertainty-based query strategies for active learning with transformers. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2194–2203, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-acl.172

Hinrich Schütze, Emre Velipasaoglu, and Jan O. Pedersen. 2006. Performance thresholding in practical text classification. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, pages 662–671, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/1183614.1183709

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 – May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Sentence Transformers. 2024. paraphrase-mpnet-base-v2 (revision e6981e5). Hugging Face.

Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Eyal Shnarch, Ariel Gera, Alon Halfon, Lena Dankin, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2022. Cluster & tune: Boost cold start performance in text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7639–7653, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.526

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2023. Selective annotation makes language models better few-shot learners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1–5, 2023*. OpenReview.net. https://doi.org/10.1109/ICASSP49357.2023.10095738

Katrin Tomanek, Florian Laws, Udo Hahn, and Hinrich Schütze. 2009. On proper unit selection in active learning: Co-selection effects for named entity recognition. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 9–17, Boulder, Colorado. Association for Computational Linguistics. https://doi.org/10.3115/1564131.1564135

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288v2*.

Cheng Wang. 2024. Calibration in deep learning: A survey of the state-of-the-art. *arXiv preprint arXiv:2308.01222v2*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381. https://doi.org/10.1016/j.future.2022.05.014

Bartosz Wójcik, Jacek Grela, Marek Smieja, Krzysztof Misztal, and Jacek Tabor. 2022. SLOVA: Uncertainty estimation using single label one-vs-all classifier. *Applied Soft Computing*, 126:109219. https://doi.org/10.1016/j.asoc.2022.109219

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Hualong Yu, Xibei Yang, Shang Zheng, and Changyin Sun. 2019. Active learning from imbalanced data: A solution of online weighted extreme learning machine. *IEEE Transactions on Neural Networks and Learning Systems*, 30(4):1088–1103. https://doi.org/10.1109/TNNLS.2018.2855446, PubMed: 30137013

Yue Yu, Rongzhi Zhang, Ran Xu, Jieyu Zhang, Jiaming Shen, and Chao Zhang. 2023. Cold-start data selection for better few-shot language model fine-tuning: A prompt-based uncertainty propagation approach. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2499–2521, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.141

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020a. Cold-start active learning through self-supervised language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7935–7948, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.637

Mu Yuan, Lan Zhang, Xiang-Yang Li, and Hui Xiong. 2020b. Comprehensive and efficient data labeling via adaptive model scheduling. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1858–1861. https://doi.org/10.1109/ICDE48307.2020.00188

Shiwei Zhang, Mingfang Wu, and Xiuzhen Zhang. 2023. Utilising a large language model to annotate subject metadata: A case study in an Australian national research data catalogue. *arXiv preprint arXiv:2310.11318v1*.

Tong Zhang, Xinrong Gong, and C. L. Philip Chen. 2022a. BMT-Net: Broad multitask transformer network for sentiment analysis. *IEEE Transactions on Cybernetics*, 52(7):6232–6243. https://doi.org/10.1109/TCYB.2021.3050508, PubMed: 33661741

Tong Zhang, Guoxi Su, Chunmei Qing, Xiangmin Xu, Bolun Cai, and Xiaofen Xing. 2021. Hierarchical lifelong learning by sharing representations and integrating hypothesis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 51(2):1004–1014. https://doi.org/10.1109/TSMC.2018.2884996

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2022b. A survey of active learning for natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6166–6190, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.414

Jingbo Zhu, Huizhen Wang, Tianshun Yao, and Benjamin K. Tsou. 2008. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1137–1144, Manchester, UK. Coling 2008 Organizing Committee. https://doi.org/10.3115/1599081.1599224