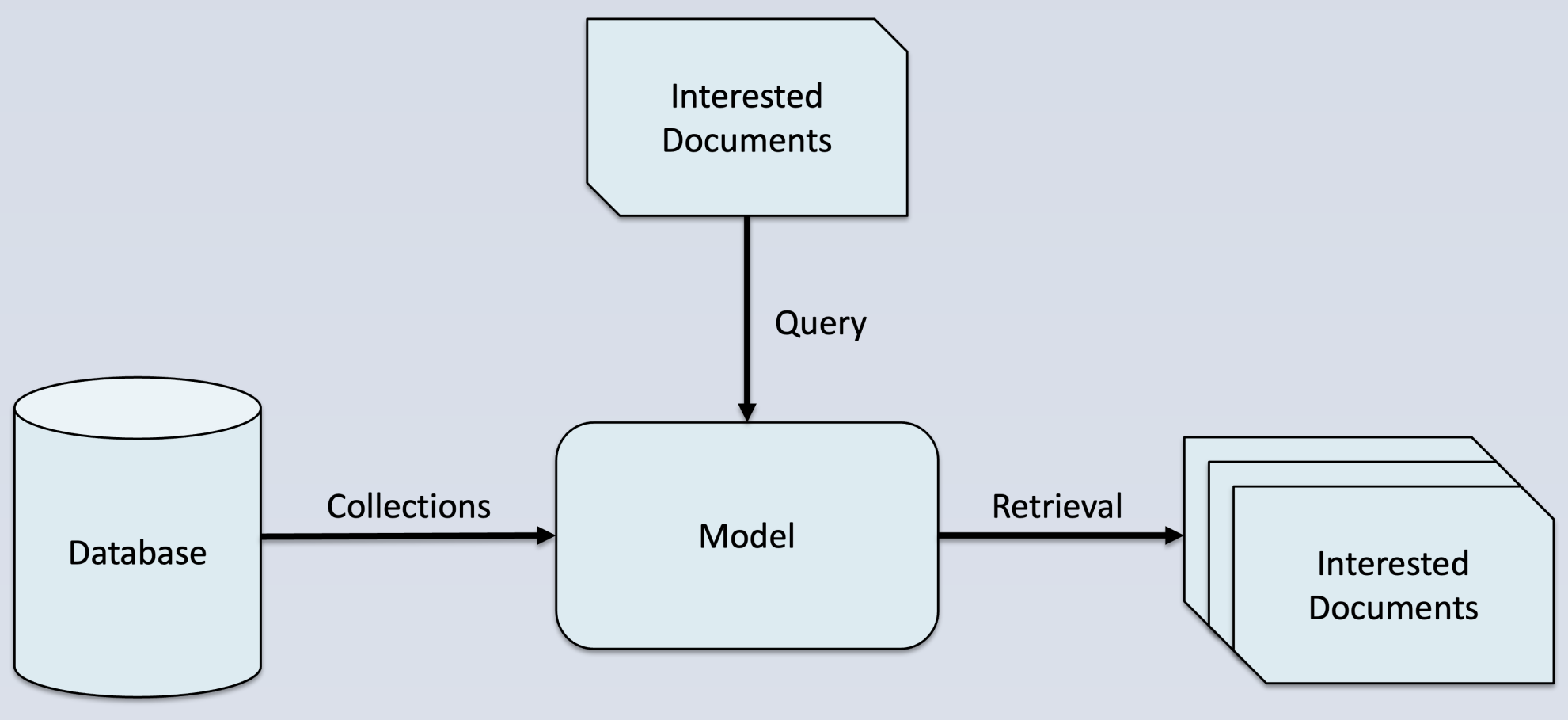
Document Set Expansion with Positive-Unlabeled Learning: A Density Estimation-based Approach

Haiyang Zhang[1], Qiuyi Chen[1], Yuanjie Zou[1], Yushan Pan[1], Jia Wang[1], Mark Stevenson[2]
Xi’an Jiaotong Liverpool University[1], University of Sheffield[2]

Introduction

We address the challenge of document set expansion (DSE). This task involves identifying all documents related to a specific fine-grained topic from a large collection, starting with a small set of known relevant documents, or 'seed studies'. This problem is commonly encountered in literature curation, where experts must sift through extensive databases to find pertinent documents.



PU Learning with Density Estimation

To overcome the limitations of PU methods relying on the SCAR assumption (Jacovi et al., 2021), we introduce a novel PU learning approach based on density estimation, called puDE. This method does not require knowledge of class prior and avoids SCAR assumptions. The objective of puDE is to learn a function that approximates $P(Y=+1|x)$ by leveraging Bayesian rule. If we can estimate the probability density ratio of $f_p(x)$ and $f(x)$, π will be a constant for each x and can be ignored in training. So, we train two separately model for predicting label:

$$\mathrm{g}(\mathbf{x}) = \frac{p(\mathbf{x})}{q(\mathbf{x})}\pi \approx \mathbb{P}(Y = 1 \mid X)$$

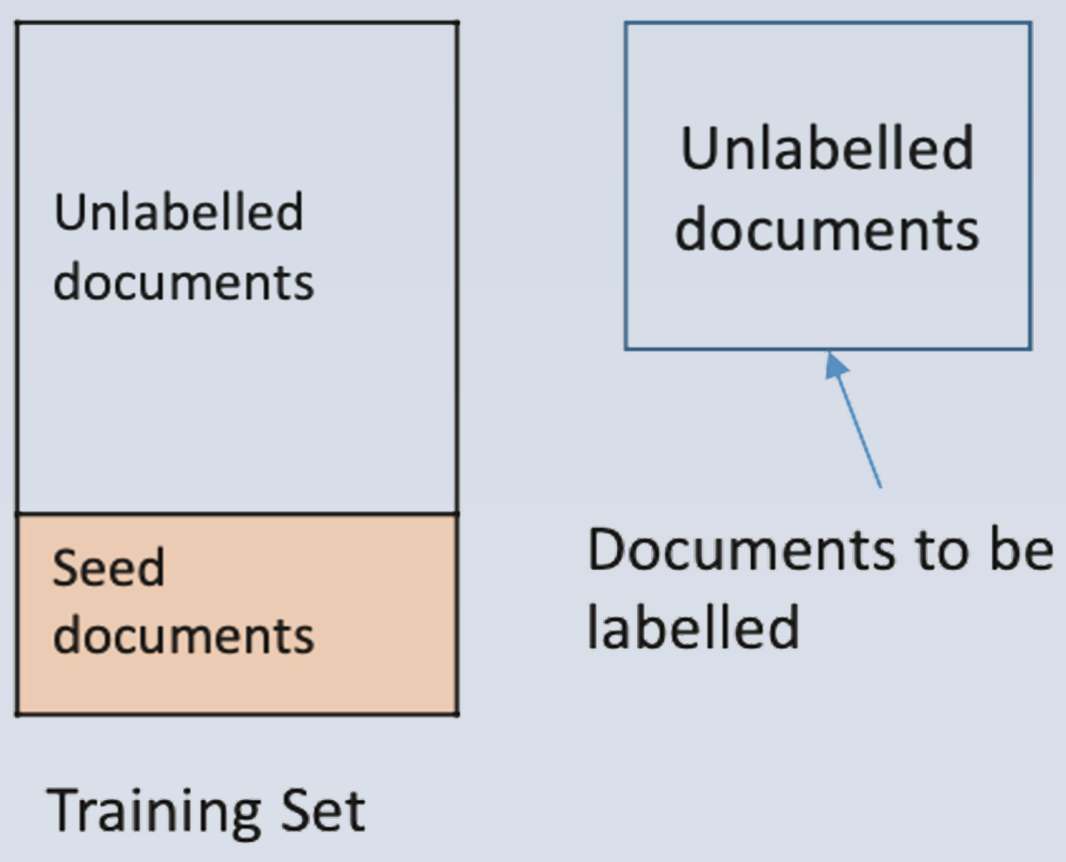
Nonparametric Density Estimation

We employ Kernel Density Estimation (KDE) for nonparametric density estimation, which is effective as it does not assume a specific data distribution. Given a dataset, KDE estimates density $\hat{f}$ using a kernel function and a bandwidth parameter. In practice, due to high-dimensional data challenges, we reduce dimensionality using Variational Autoencoders (VAE) before applying KDE.

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h})$$

PU Learning

Jacovi et al. (2021) proposed modeling DSE as a positive and unlabeled (PU) learning problem, which involves training a binary classifier with only positive and unlabeled data (Bekker and Davis, 2020).



Parametric Density Estimation

For parametric density estimation, we use Energy-Based Models (EBM). Unlike other methods, EBMs do not assume a specific probability density form. EBMs learn an energy function assigning low values to observed data and high values to others. We estimate $p(x)$and q(x)using parameterized neural networks

$$p_{(\mathbf{x})} \approx p_\theta(\mathbf{x}) = \frac{e^{-g_{p_\theta}(\mathbf{x})}}{Z_{p_\theta}},\ q_{(\mathbf{x})} \approx q_\theta(\mathbf{x}) = \frac{e^{-g_{q_\theta}(\mathbf{x})}}{Z_{q_\theta}}$$

Hence the classifier is then rewritten as:

$$f(\mathbf{x}) = \frac{e^{-g_{p_\theta}(\mathbf{x})}}{Z_{p_\theta}} / \frac{e^{-g_{q_\theta}(\mathbf{x})}}{Z_{q_\theta}}\pi = e^{\left(g_{q_\theta}(\mathbf{x}) - g_{p_\theta}(\mathbf{x})\right)}\left(\frac{Z_{q_\theta}}{Z_{p_\theta}}\pi\right)$$

The term $\left(\frac{Z_{q_\theta}}{Z_{p_\theta}}\pi\right)$ in the above is a constant for each $\mathbf{x}$ and can be ignored in practice. Hence, the classifier can be approximated by the exponent:

$$f(\mathbf{x}) := g_{q_\theta}(\mathbf{x}) - g_{p_\theta}(\mathbf{x})$$

The standard maximum likelihood training algorithm with Markov Chain Monte Carlo (MCMC) sampling is employed to train the neural networks g_{p_θ} and g_{q_θ}. It should be noted that Langevin dynamics can be **unreliable in high-intensity areas** for high-dimensional datasets, which leads to low-performance models. To address this concern, we add a normal PU loss component in the early stages of training. The total loss function is defined as:

$$\alpha(\nabla_\theta \log p_\theta(\mathbf{x})) + \beta(\nabla_\theta \log q_\theta(\mathbf{x})) + \gamma\left(R_{\ell_{0-1}}(f(\mathbf{x}))\right)$$

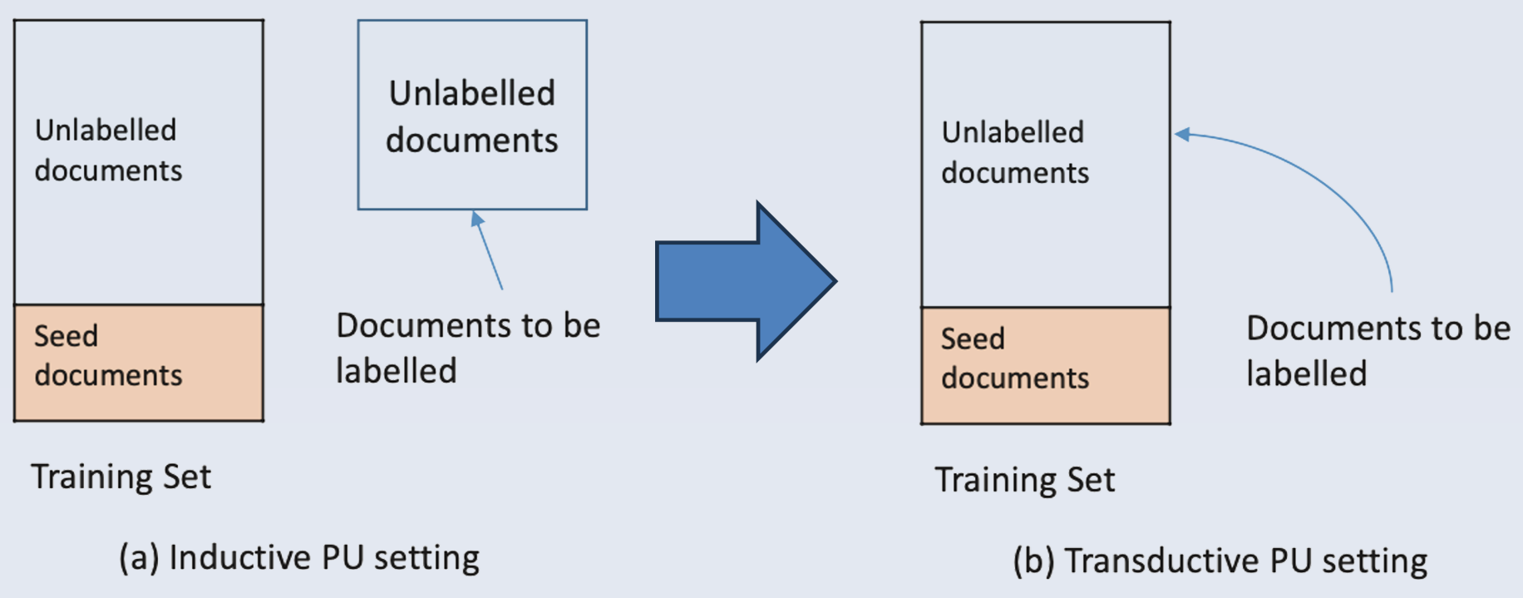
Main Contribution

Despite their advancements, key issues remain unresolved:

- The empirical improvements achieved still rely heavily on model-specific optimizations rather than addressing fundamental PU method limitations.
- The experimental setup used by Jacovi et al. does not accurately reflect the transductive nature of the DSE task, where the entire unlabeled set should be used for both training and testing.
- PU methods like nnPU assume a Selected Completely At Random (SCAR) labeling mechanism, which does not hold in DSE scenarios due to biased expert selection.

To address these challenges, we propose a novel PU learning framework, puDE, that operates without SCAR assumptions and does not require prior knowledge of class distribution. Our contributions include:

- Highlighting the limitations of current PU solutions for DSE.
- Introducing a new PU learning framework based on density estimation.
- Demonstrating the superior performance of our method on real-world datasets, establishing it as a better solution for the DSE task.



(a) Inductive PU setting

(b) Transductive PU setting

Experiment Setting

For experiment, we chose 4 dataset:

- PubMed datasets with 3 fine-grained topics, generated by Jacovi et al. (2021) for the DSE task.
- A single dataset used for Covid-19 study classification Shemilt et al.

The Training data consist Labeled data + Unlabeled data. All test data are Unlabeled data.

dataset	|LP|	N_U	N_{UP}	N_{UN}
Pubmed-*topic1*	20	10012	1844	8168
	50	10027	2568	7459
Pubmed-*topic2*	20	10012	2881	7131
	50	10027	3001	7026
Pubmed-*topic3*	20	7198	1201	5997
	50	10025	1916	8109
Covid	{47..4722}	4722	2310	2412

We selected four baseline model which are commonly use in previous research.

- nnPU (5-layer neural network)
- BM25
- puDE-kde (VAE with 50 latent space)
- puDE-em (5-layer neural network)

Result

- Performance of nnPU is much worse than that reported by Jacovi et al. (2021) and is similar to BM25, which indicate that the PU solutions proposed in (Jacovi et al., 2021) is not as effective as they stated for the DSE task in transductive setting.
- Both puDE methods outperform other methods, with one exception where BM25 get the best result on the last topic. It should be noticed that result reported for BM25 is the average across 5000-|LP| cases, which is not the direct classification result, and it serves as references to the state-of-the-art (Jacovi et al., 2021)
- nnPU and VPU get stable results only when more than 20% of labelled data is available
- puDE methods perform well with less data (<10%) and consistently shown significant improvements over other methods with the increase of labelled data
- When the number of labelled positives is small, the performance of vPU is poor since its training strategy needs equal batch size of unlabeled (U) and labelled (LP) samples feeding into the model to calculate the variational loss

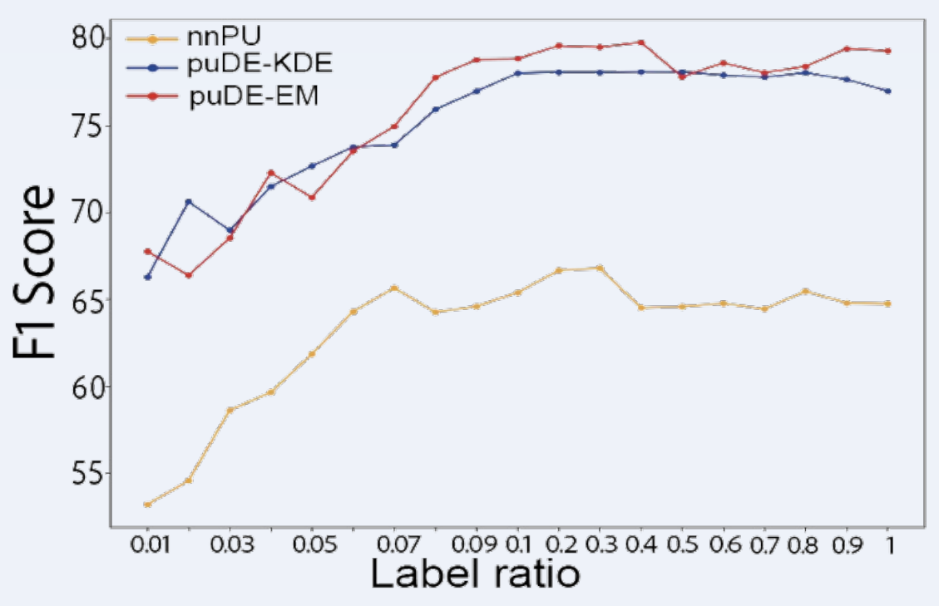


Figure 1: F1 comparison on covid dataset with respect to the ratio of |LP| over |U| ranging from 0.01 to 1.

|LP|	Topic	BM25	nnPU-trans.	puDE-*kde*	puDE-*em*
20	*topic1*	32.25	33.03	37.31	40.59
	topic2	26.75	31.30	36.18	39.67
	topic3	41.23	27.76	36.63	35.59
50	*topic1*	32.80	38.76	44.65	44.91
	topic2	31.85	34.16	44.03	46.22
	topic3	35.78	32.84	36.63	36.57

Table 2: F1 comparison against baseline and state-of-the-art DES methods.

Result

Jacovi, Alon, et al. "Scalable evaluation and improvement of document set expansion via neural positive-unlabeled learning." arXiv preprint arXiv:1910.13339 (2019).

Bekker, Jessa, and Jesse Davis. "Learning from positive and unlabeled data: A survey." Machine Learning 109.4 (2020): 719-760.

Shemilt, Ian, et al. "Machine learning reduced workload for the Cochrane COVID-19 Study Register: development and evaluation of the Cochrane COVID-19 Study Classifier." Systematic Reviews 11.1 (2022): 15.