# Prototype-Guided Pseudo Labeling for Semi-Supervised Text Classification

**Weiyi Yang[1], Richong Zhang[1,2]*, Junfan Chen[1], Lihong Wang[3], Jaein Kim[1]**
[1]SKLSDE, School of Computer Science and Engineering, Beihang University, Beijing, China
[2]Zhongguancun Laboratory, Beijing, China
[3]CNCERT

**（ACL-2023)**

**code： none**

**Reported by Zhaoze Gao**

# 1. Introduction

# 2. Approach

# 3. Experiments

# Introduction

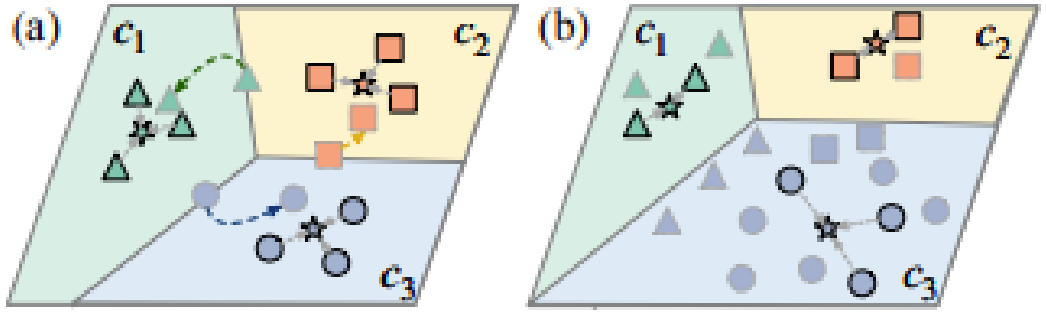



Illustration of **two** cases of pseduo-labeling problems in existing SSTC models:

**(a)** underfitting near decision boundaries;

**(b)** bias from imbalanced data. ⋆ denotes a prototype consisting of labeled data. Different symbols indicate different categories of labeled (bordered) and unlabeled data (unbordered). The areas with different colors indicate different categories.

# Approach

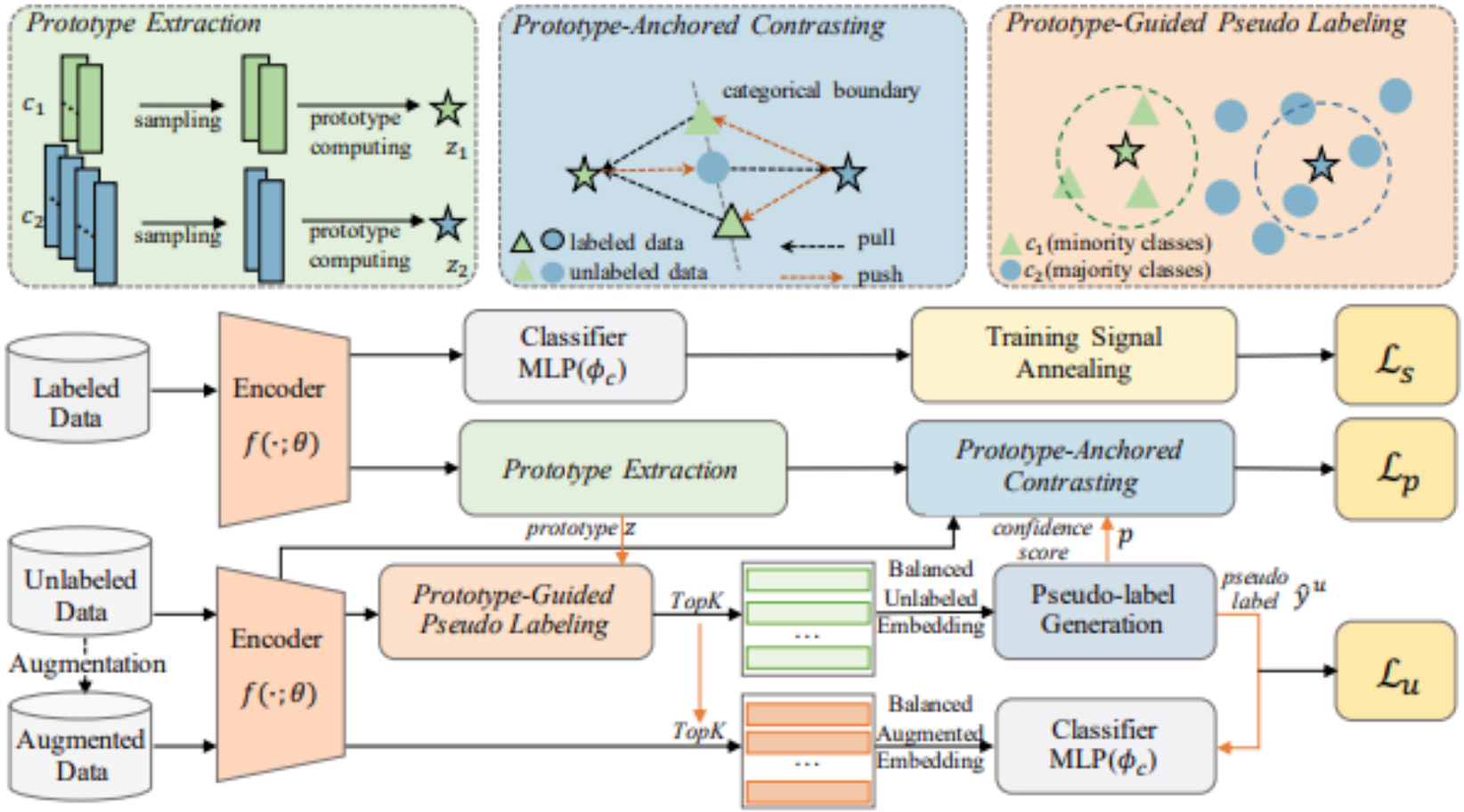



Figure 2: Overall architecture of PGPL. For labeled data, the supervised loss is computed to generate class prototypes for processing the Prototype-Anchored Contrasting module. For unlabeled data, the pseudo-labels are assigned based on Prototype-Guided Pseudo Labeling to augment the text instances for retraining the model.

# Approach

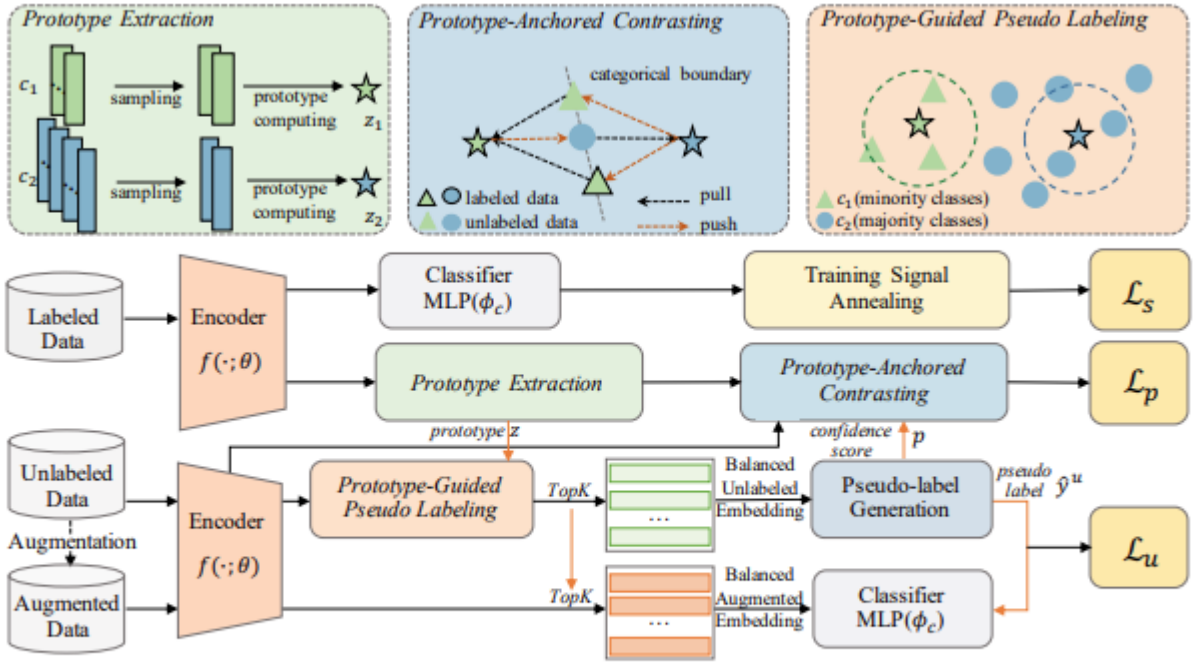



Figure 2: Overall architecture of PGPL. For labeled data, the supervised loss is computed to generate class prototypes for processing the Prototype-Anchored Contrasting module. For unlabeled data, the pseudo-labels are assigned based on Prototype-Guided Pseudo Labeling to augment the text instances for retraining the model.

$$D_l = \{ (x_1^l, y_1^l), \cdots, (x_m^l, y_m^l) \} \qquad D_u = \{ x_1^u, \cdots, x_n^u \}$$

$$g(x, \phi_y, \theta) = \mathrm{MLP}(f(x;\theta);\phi_c), \tag{1}$$

$$\eta_t = \frac{t}{T}(1-\tau) + \tau, \tag{2}$$

$$\mathcal{L}_s = -\frac{1}{m}\sum_{i=1}^{m} \mathbb{I}(p_{y_i^l} < \eta_t) \log \frac{\exp(g(x_i^l, \phi_{y_i^l}, \theta))}{\sum_{c\in\mathcal{C}} \exp(g(x_i^l, \phi_c, \theta))}, \tag{3}$$

$$\hat{y}_i^u = \arg\max_{c\in\mathcal{C}} \exp(g(x_i^u, \phi_c, \theta)), \tag{4}$$

$$\mathcal{L}_u = -\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(x_i^u) \log \frac{\exp(g(\tilde{x_i^u}, \phi_{y_i^u}, \theta))}{\sum_{c\in\mathcal{C}} \exp(g(\tilde{x_i^u}, \phi_c, \theta))}, \tag{5}$$

$$\mathbb{I}(x_i^u) = \mathbb{I}(\hat{y}_i^u = c) \wedge \mathbb{I}(\mathrm{d}(f(x_j^u), z_c) \leq d_c), \tag{6}$$

# Approach

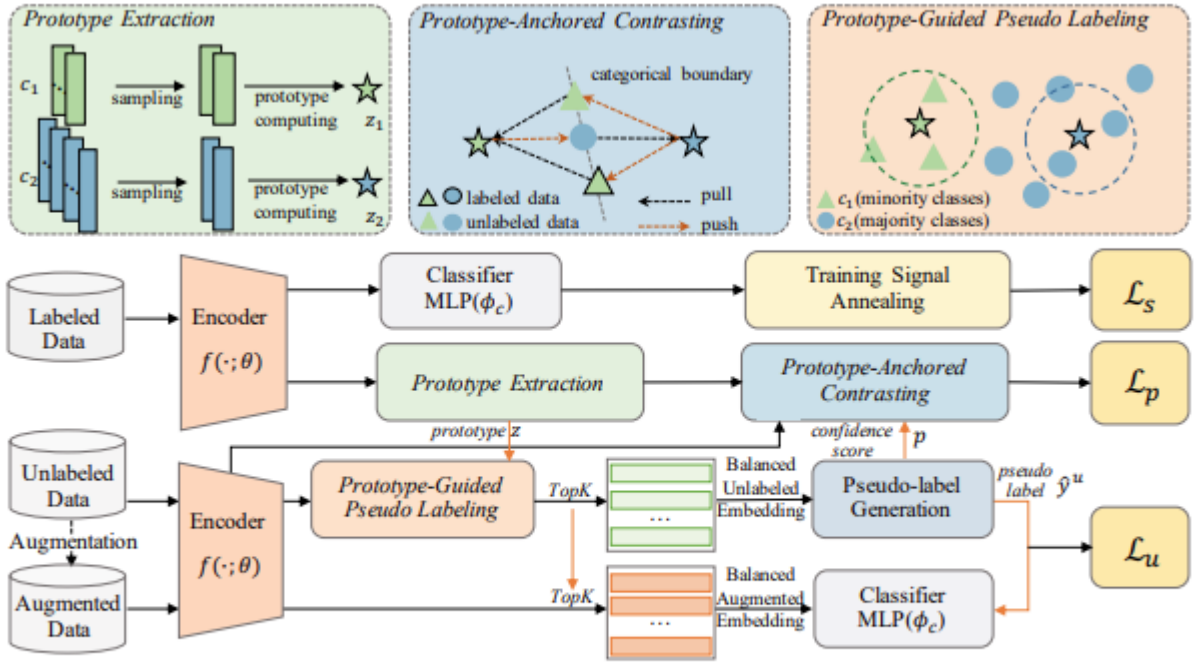



Figure 2: Overall architecture of PGPL. For labeled data, the supervised loss is computed to generate class prototypes for processing the Prototype-Anchored Contrasting module. For unlabeled data, the pseudo-labels are assigned based on Prototype-Guided Pseudo Labeling to augment the text instances for retraining the model.

$$z_c = \frac{1}{n_c} \sum_{y_i^l = c} f(x_i^l), \tag{7}$$

$$n_c = \sum_{y_i^l \in D_l} \mathbb{I}(y_i^l = c), \tag{8}$$

$$\mathcal{L}_p = -\frac{1}{m} \sum_{i}^{m} \sum_{c} \mathbb{I}(y_i^l = c) \log \frac{\exp\left(-\mathrm{d}(f(\bar{x_i^l}), z_c)\right)}{\sum_{k \in \mathcal{C}} \exp(-\mathrm{d}(f(\bar{x_i^l}), z_k))}$$
$$- \frac{\lambda}{n} \sum_{j}^{n} \sum_{c} p_{\hat{y}_j^u} \mathbb{I}(\hat{y}_j^u = c) \log \frac{\exp\left(-\mathrm{d}(f(\bar{x_j^u}), z_c)\right)}{\sum_{k \in \mathcal{C}} \exp(-\mathrm{d}(f(\bar{x_j^u}), z_k))}, \tag{9}$$

$$\mu_t^c = \sum_{\bar{x_j^u} \in B_u} \mathbb{I}(\hat{y}_j = c), \tag{10}$$

$$\gamma_t = \arg\min_{c \in \mathcal{C}} \mu_{<t}^c. \tag{11}$$

# Approach

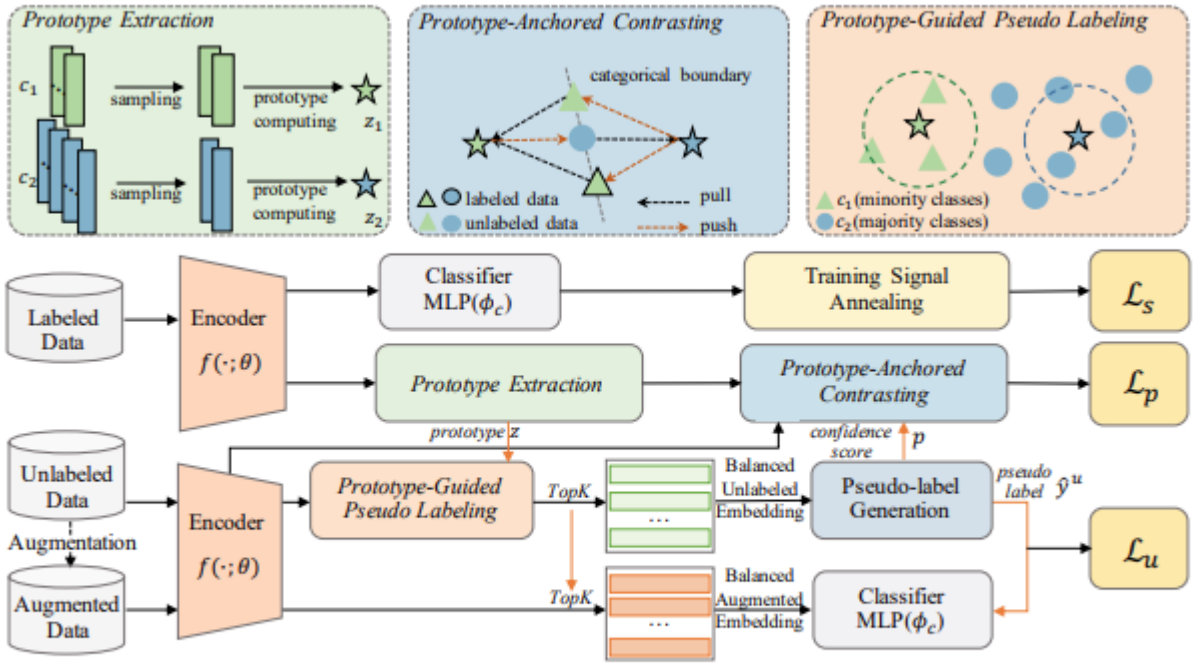



Figure 2: Overall architecture of PGPL. For labeled data, the supervised loss is computed to generate class prototypes for processing the Prototype-Anchored Contrasting module. For unlabeled data, the pseudo-labels are assigned based on Prototype-Guided Pseudo Labeling to augment the text instances for retraining the model.

$$k_c = \begin{cases} \mu_t^c & \text{if } \mu_{<t}^c - \gamma_t = 0 \\ \mu_t^c - (\mu_{<t}^c - \gamma_t) & \text{if } 0 \le \mu_{<t}^c - \gamma_t < \mu_t^c \\ 0 & \text{if } \mu_{<t}^c - \gamma_t \ge \mu_t^c \end{cases}, \tag{12}$$

$$d_c = \text{TopK}(\text{d}(f(\tilde{x_j^u}), z_c), k_c), \tag{13}$$

$$\mathcal{L} = \mathcal{L}_s + \beta_1 \mathcal{L}_u + \beta_2 \mathcal{L}_p, \tag{14}$$

# Experiments

| Dataset | Classification Type | Class | Train | Unlabeled | Dev | Test |
|---|---|---|---|---|---|---|
| AG News | News Topic | 4 | 200 | 5000 | 2000 | 1900 |
| DBpedia | Wikipedia Topic | 14 | 200 | 5000 | 2000 | 5000 |
| Yahoo! Answer | QA Topic | 10 | 200 | 5000 | 2000 | 6000 |
| IMDB | Movie Review Sentiment | 2 | 200 | 5000 | 2000 | 12500 |

Table 1: The dataset statistics for the per-class number of unlabeled, dev and test data.

# Experiments

| Model | AG News | | | IMDB | | | Yahoo! Answer | | | DBpedia | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 10 | 30 | 200 | 10 | 30 | 200 | 10 | 30 | 200 | 10 | 30 | 200 |
| Bert | 81.0 | 84.3 | 87.2 | 70.6 | 73.3 | 86.1 | 60.1 | 64.1 | 69.3 | 96.6 | 98.2 | 98.6 |
| UDA | 86.4 | 86.4 | 88.3 | 86.4 | 86.4 | 88.7 | 64.3 | 68.3 | 70.2 | 97.8 | 98.3 | 98.8 |
| Mixtext | 87.3 | 87.4 | 88.2 | 74.2 | 85.3 | 89.1 | **67.7** | 68.5 | 70.6 | 98.5 | 98.8 | 98.9 |
| PGPL | **87.8** | **88.5** | **89.2** | **88.9** | **90.2** | **90.3** | 67.4 | **69.1** | **70.7** | **98.7** | **99.0** | **99.0** |

Table 2: Comparison with state-of-the-arts on the AG News, DBpedia, Yahoo! Answer and IMDB test set under different partition protocols. The Wilcoxons test shows significant difference ($p<0.05$) between our model and baselines averaged after five runs except on Yahoo! Answer (10 labels).

# Experiments

| Datasets | Model | $K$ | Acc. |
|---|---|---|---|
| AG News | PGPL | 10 | **87.8** |
| | SAT | 20 | 86.4 |
| | CEST | 30 | 87.1 |
| | UST | 30 | 87.7 |
| | VAMPIRE | 200 | 83.9 |
| IMDB | PGPL | 30 | **90.2** |
| | SAT | 10 | 69.0 |
| | CEST | 30 | **90.2** |
| | SALNet | 21 | 75.7 |
| | VAMPIRE | 200 | 82.2 |
| | Delta-training | 212 | 75.0 |
| Yahoo! Answer | PGPL | 10 | **67.4** |
| | SAT | 20 | 61.5 |
| | SALNet | 34 | 53.7 |
| | VAMPIRE | 200 | 59.9 |
| | FLiText | 500 | 65.1 |
| DBpedia | PGPL | 10 | **98.7** |
| | SALNet | 20 | 98.2 |
| | UST | 30 | 98.6 |
| | CEST | 30 | 98.6 |

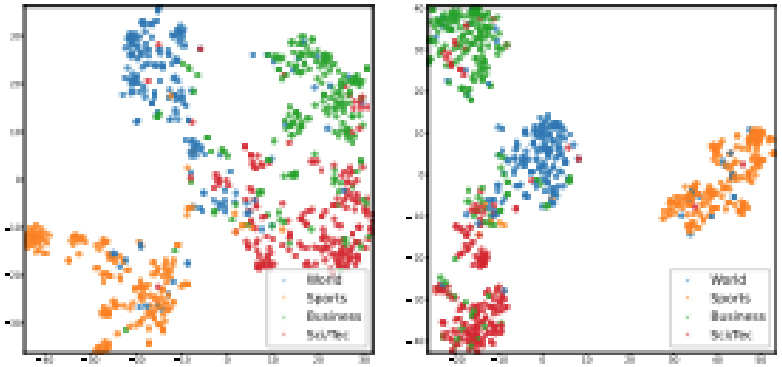
Table 3: SSTC methods with $K$ training labels per class (UST, VAMPIRE, SALNet, FLiText, Delta-training, CEST, SAT).

# Experiments

| Data setting (labeled/unlabeled) | Model | Result |
|---|---|---|
| balanced / balanced | UDA | 64.3 |
| | Mixtext | 68.1 |
| | PGPL | **68.3** |
| imbalanced / balanced | UDA | 64.8 |
| | Mixtext | 67.7 |
| | PGPL | **68.8** |
| balanced / imbalanced | UDA | 57.9 |
| | Mixtext | 67.6 |
| | PGPL | **67.8** |
| imbalanced / imbalanced | UDA | 64.1 |
| | Mixtext | 65.5 |
| | PGPL | **67.3** |

Table 4: Experimental results on different data balance settings, grouped by labeled/unlabeled data.

# Experiments




(a) PGPL$_{w/oPAC}$ (b) Full PGPL

Figure 3: Comparison of t-SNE visualization results with and without PAC module on the AG News dataset.

# Experiments

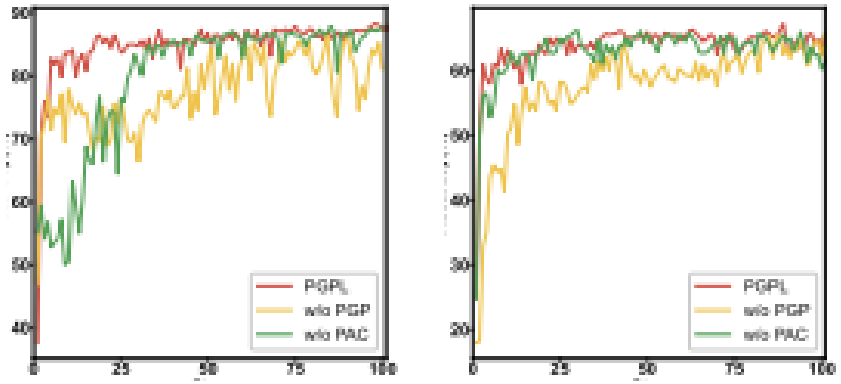
| Class | Model | Accuracy |
|---|---|---|
| World | PGPL | **91.2** |
| | -w/o PGP | 88.6 |
| Sports | PGPL | **84.0** |
| | -w/o PGP | 81.1 |
| Business | PGPL | **96.1** |
| | -w/o PGP | 96.0 |
| Sci/Tech | PGPL | **81.9** |
| | -w/o PGP | 80.3 |
| All categories | PGPL | **88.3** |
| | -w/o PGP | 86.5 |

Table 5: Experimental results on different data balance settings, grouped by labeled/unlabeled data.

| Data | AG News | IMDB |
|---|---|---|
| PGPL | **88.3** | **89.7** |
| w/o PGP | 86.5 | 88.2 |
| w/o PAC | 86.2 | 88.9 |
| w/o TSA | 87.2 | 87.9 |
| Data | Yahoo!Answer | DBpedia |
| PGPL | **68.3** | 98.4 |
| w/o PGP | 65.7 | 98.2 |
| w/o PAC | 67.4 | 98.2 |
| w/o TSA | 68.0 | **98.6** |

Table 6: Ablation analysis of PGPL with different modules on valid data with 10 labeled data.

# Experiments




(a) AG News (b) Yahoo!Answer

Figure 4: The comparison of training stability and convergence speed among different models..

# Experiments

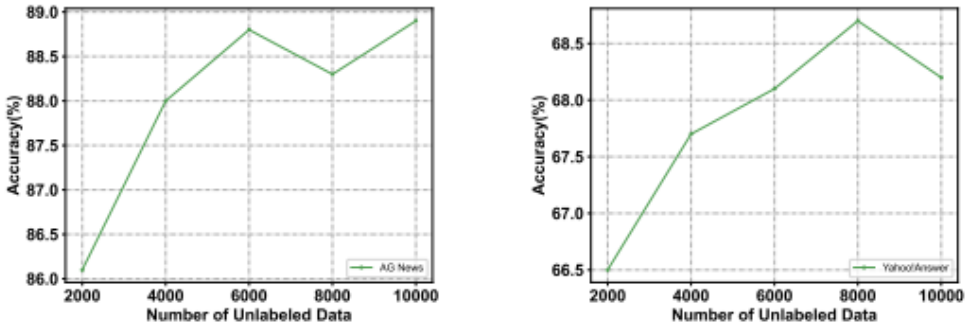
| Dataset | Supervised (all labels) | | Supervised (10 labels) | | Semi-supervised (10 labels) | |
|---|---|---|---|---|---|---|
| | BERT | RoBERTa | BERT | RoBERTa | PGPL(BERT) | PGPL(RoBERTa) |
| AG News | 91.2 | 92.4 | 80.2 | 80.7 | 87.8 | 88.4 |
| IMDB | 90.4 | 93.5 | 70.9 | 71.2 | 88.9 | 91.2 |
| Yahoo!Answer | 73.7 | 74.2 | 60.1 | 61.0 | 67.4 | 67.8 |
| DbPedia | 99.1 | 99.1 | 96.6 | 96.1 | 98.7 | 98.8 |
| Average | 88.6 | 89.8 | 76.9 | 77.3 | 85.7 | 86.6 |

Table 7: Comparison with state-of-the-arts on the AG News, DBpedia, Yahoo! Answer and IMDB test set under different partition protocols. The results are averaged after three runs.

# Experiments

| Hyper-parameter | PGPL |
|---|---|
| type embedding dimension $d$ | 768 |
| Bert attention dropout | 0.1 |
| Bert hidden dropout | 0.1 |
| MLP hidden dimension | 128 |
| Sequence Length | 256 |
| batch size on labeled data | class numbers |
| batch size on unlabeled data | 16 |
| training epoch | 20 |
| initial learning rate of BERT | $1e^{-5}$ |
| learning rate of MLP | $1e^{-3}$ |
| threshold $\lambda$ (AG News,IMDB) | 0.2 |
| threshold $\lambda$(DBpedia,Yahoo!Answer) | 0.5 |
| threshold $\beta_1$ | 1 |
| threshold $\beta_2$ | 0.5 |
| t memory set | 5 |

Table 8: Hyper-parameter settings of PGPL.




(a) AG News (b) Yahoo!Answer

Figure 5: The accuracy of PGPL when varying the number of unlabeled data on AG News and Yahoo!Answer Datasets.

# Thank you !