method

Prototype-Guided Pseudo Labeling for Semi-Supervised Text Classification

task

Advisor : Jia-Ling, Koh

Speaker : Ting-I, Weng

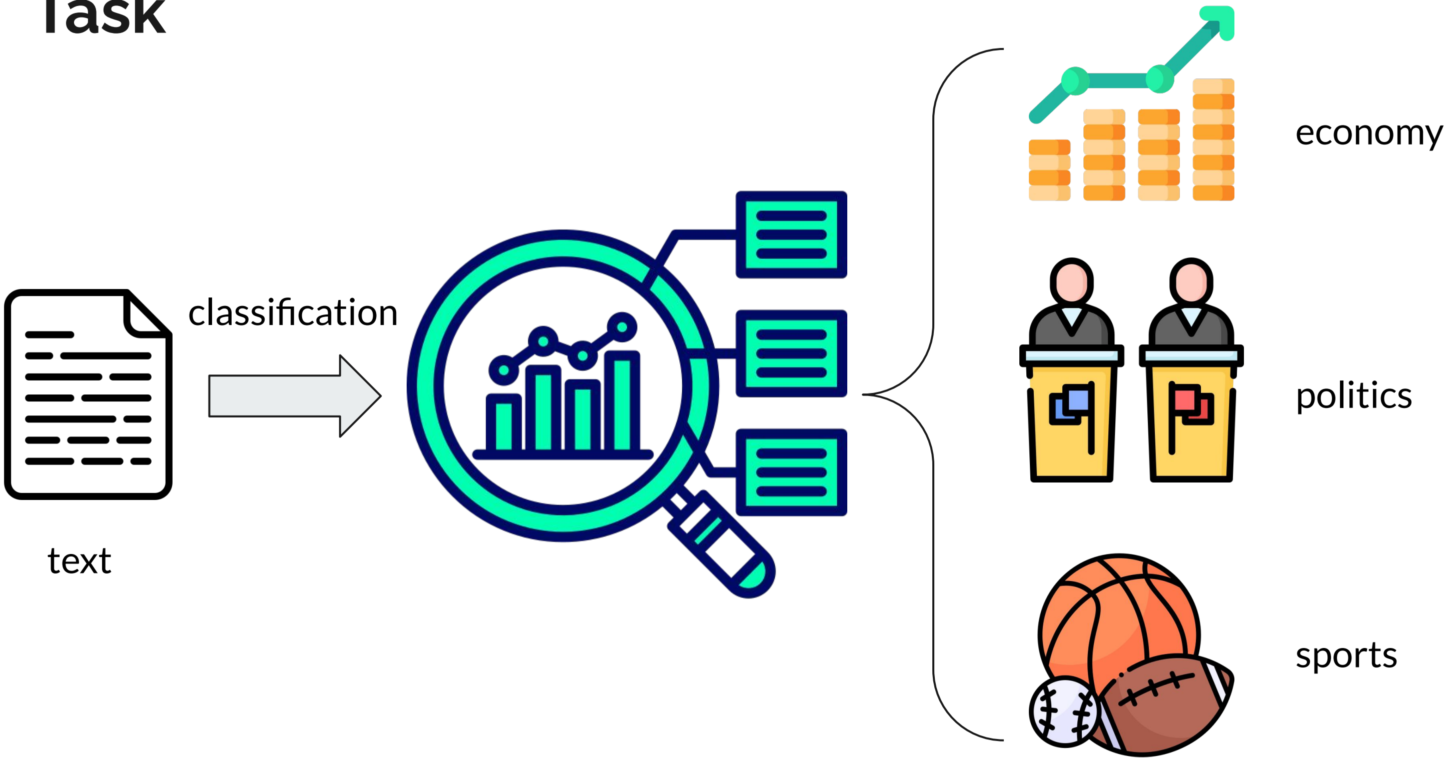
Source : ACL’23

Date : 2023/09/12

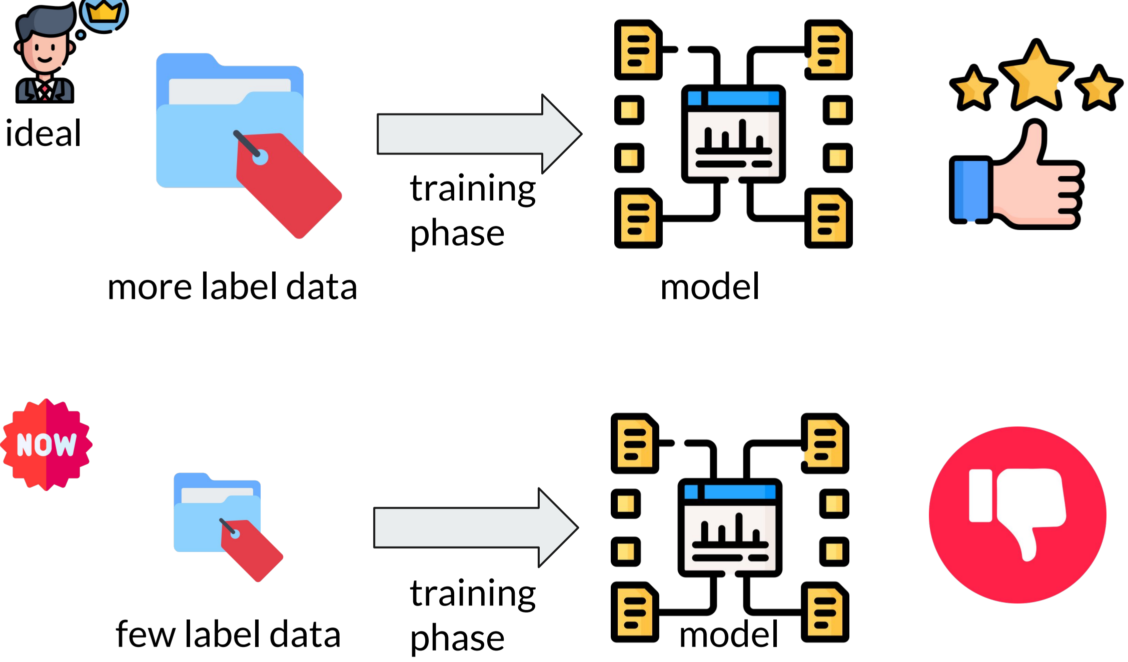
Outline

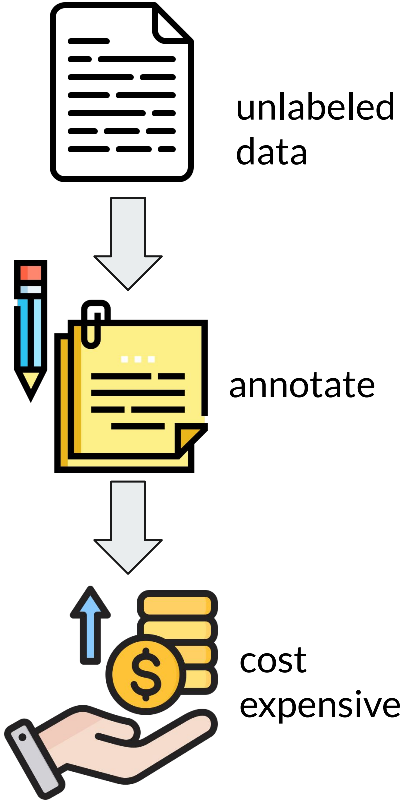
- Introduction
- Method
- Experiment
- Conclusion

Task



Problem - few label data

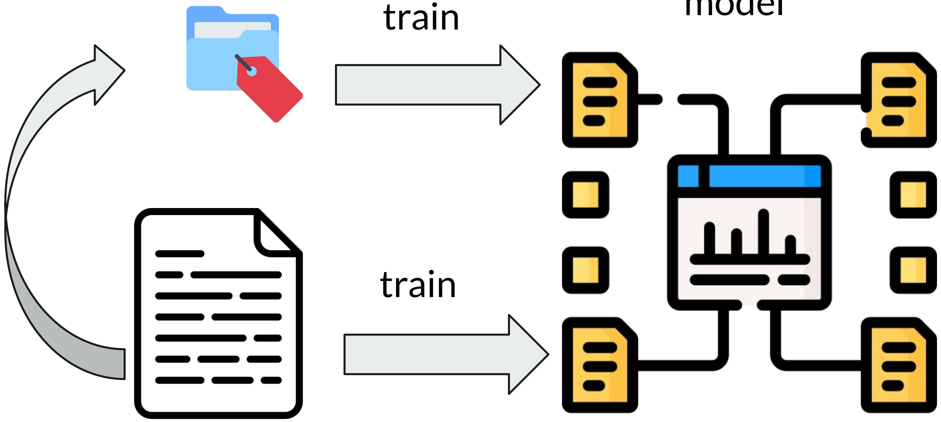




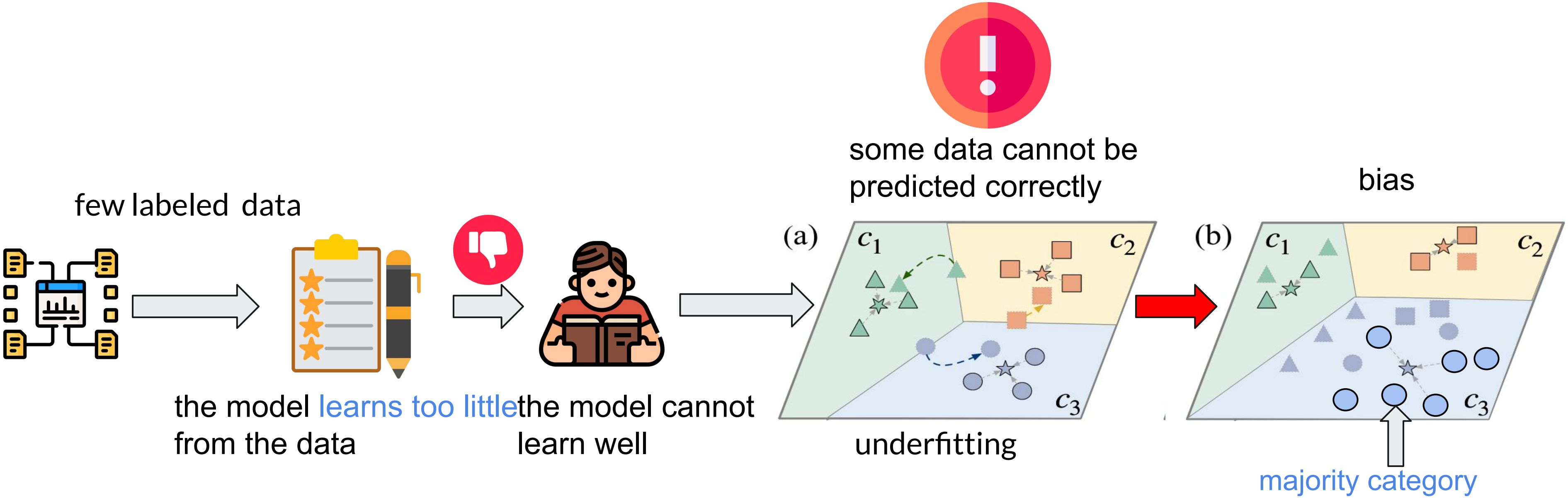
Semi-supervised learning

Solve the problem of too little labeled data

Hope to use unlabeled data to assist training

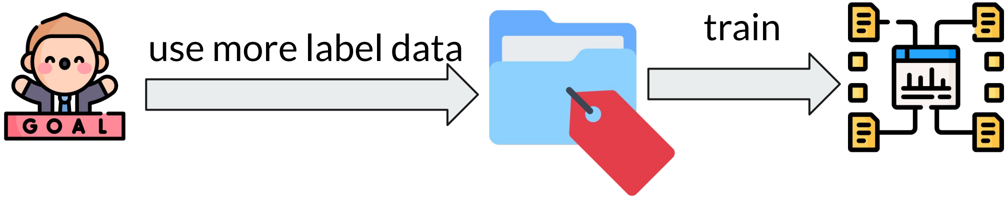


Problem - Semi-supervised learning

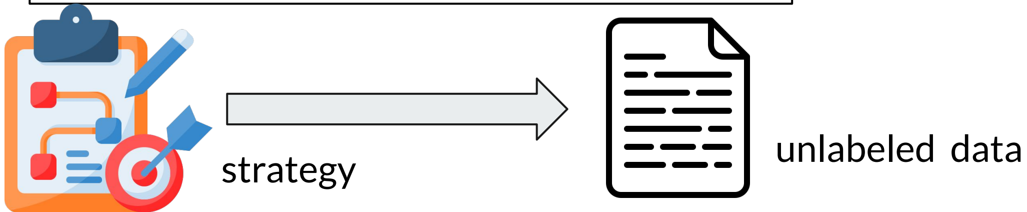


Solution

underfitting

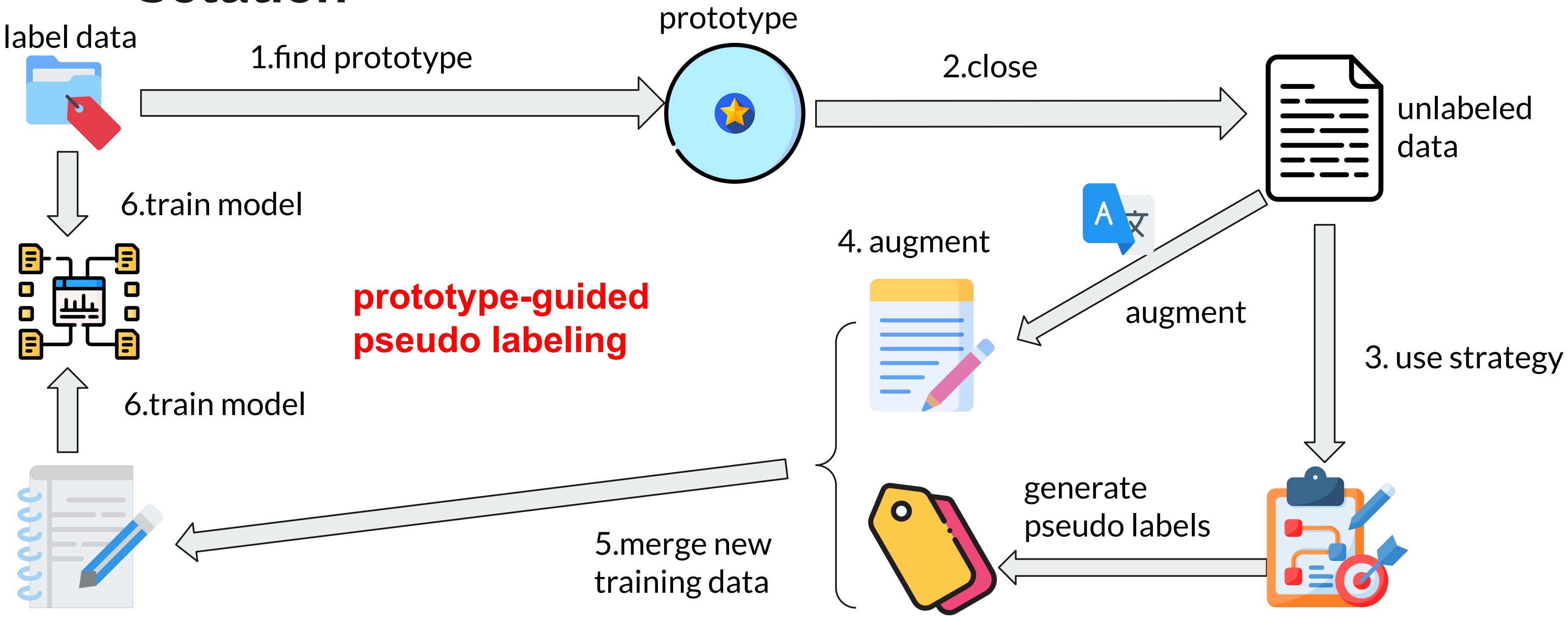


Hope to find representative unlabeled data



prototype → the most representative points in each category

Solution



Outline

- Introduction
- Method
- Experiment
- Conclusion

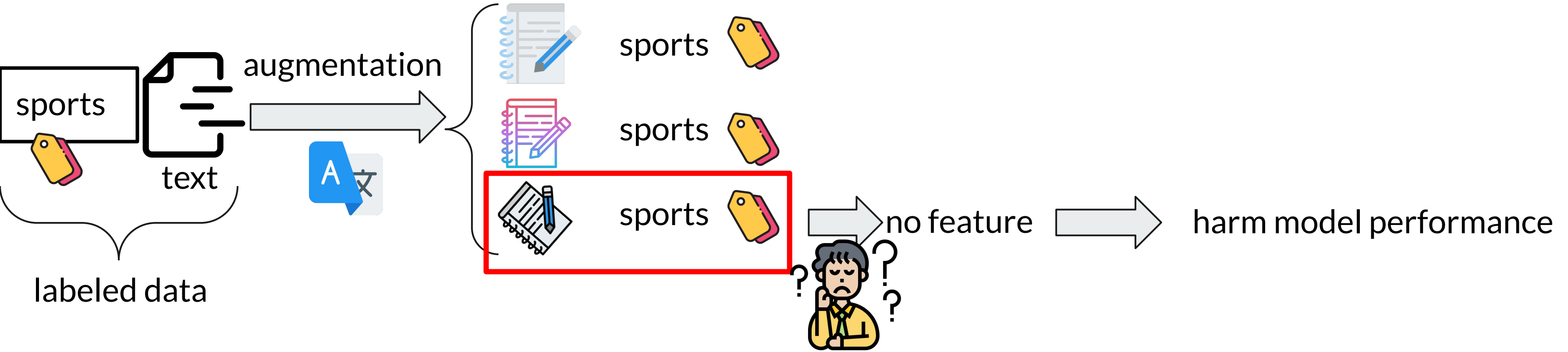
https://arxiv.org/abs/1904.12848

Unsupervised Data Augmentation

normal data Augmentation

reason : few label data

hope to get more labeled data to train the model

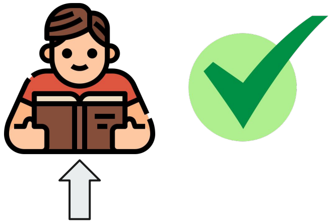


https://arxiv.org/abs/1904.12848

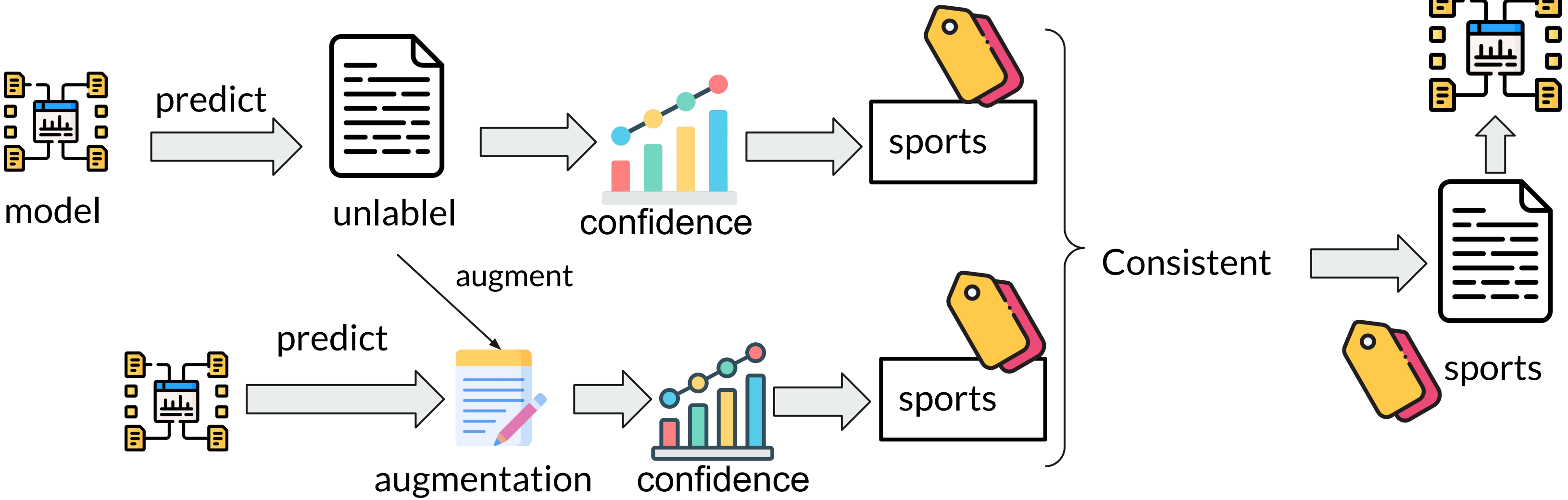
at the beginning,
model trained with label data

data diversity
Increase model robustness

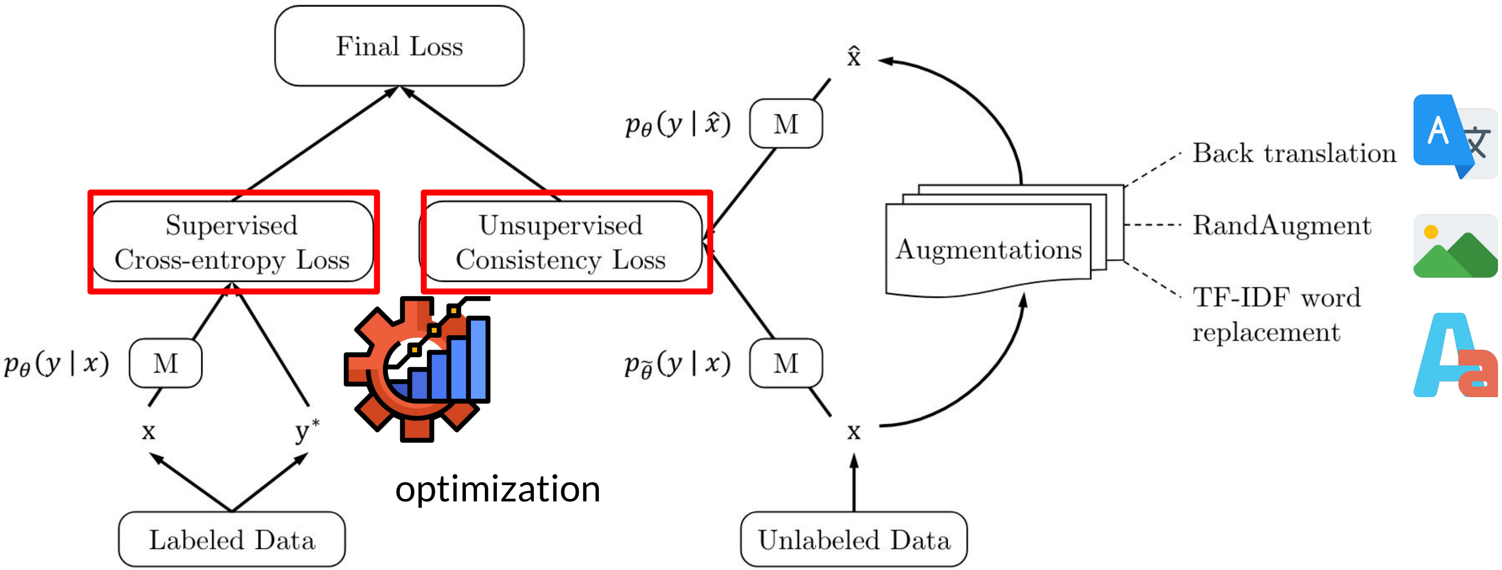
Unsupervised Data Augmentation

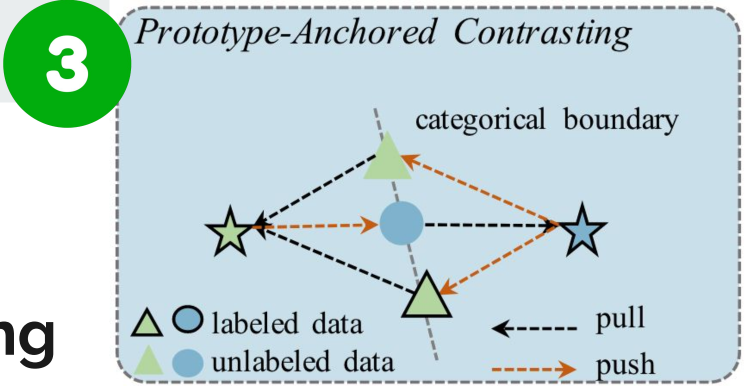


reason : label data has limited effect after augmentation

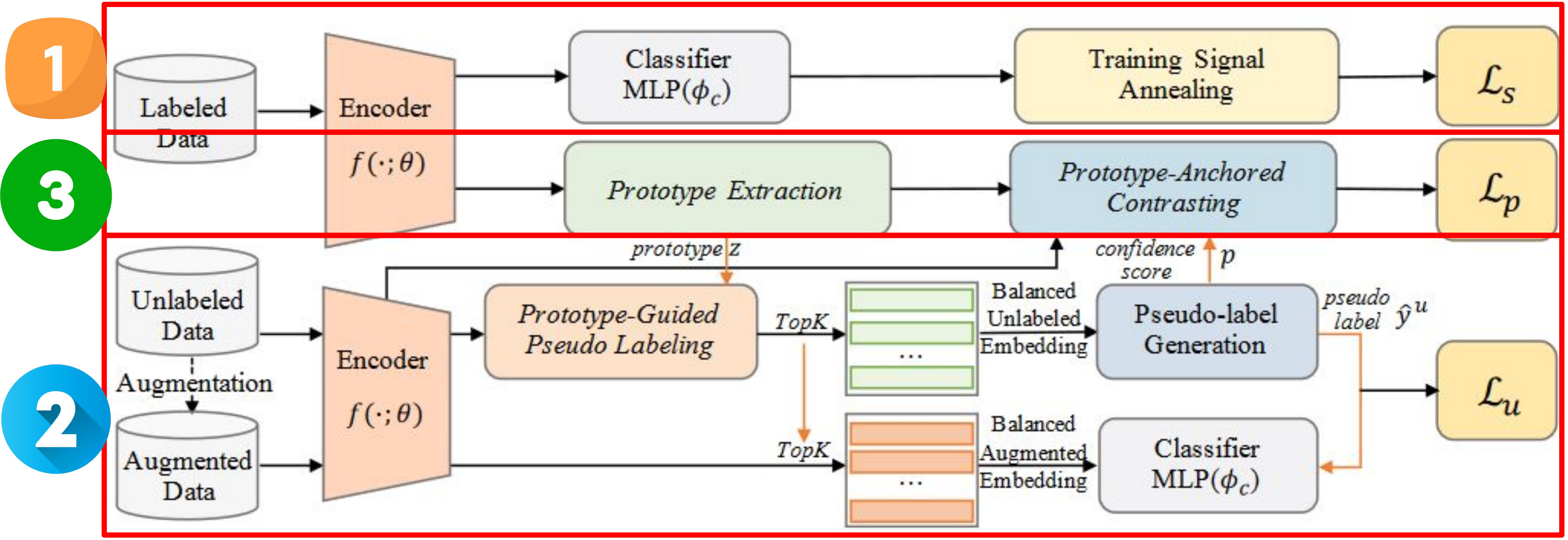


Unsupervised Data Augmentation

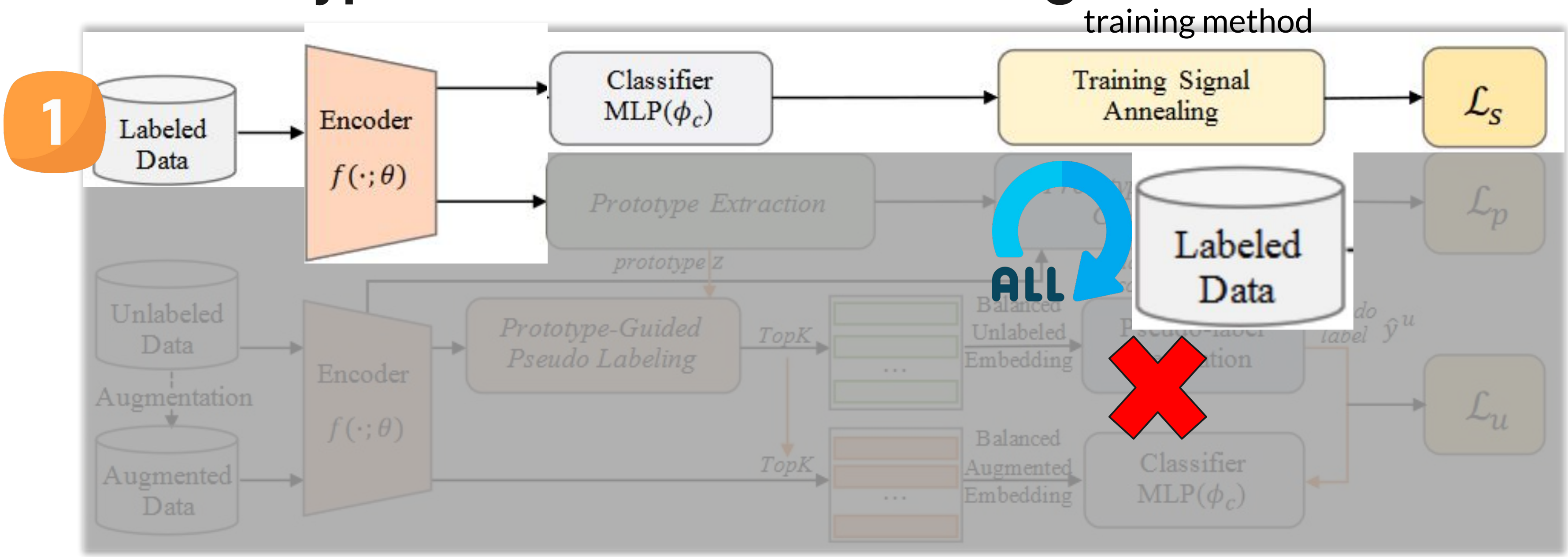




Prototype-Guided Pseudo Labeling



Prototype-Guided Pseudo Labeling(Partition1)



1 Training Signal Annealing

strategy

releases labeled instances gradually during the training phase

prevent the model from overfitting too quickly

GOAL

Training signal annealing

few labeled data
count : 100

many unlabeled data
count : 10000

training phase

overfitting

underfitting

difficult threshold

few labeled data
count : 100

label data
count : 5

t = 1

label data
count : 20

t = 5

label data
count : 100

t = 10

Training signal annealing

current iterations

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

threshold

total iterations

initial threshold

don't use all labeled data at once

e.g. number of categories = 10
In this paper, tau need to > 1/(number of categories)=> 1/10, so suppose tau = 5/10 = 0.5

t

number of categories	T(total iterations)	t	tau	threshold
10	10	1	0.5	1/10 x (1-0.5) + 0.5 = 0.55
10	10	5	0.5	5/10 x (1-0.5) + 0.5 = 0.75
10	10	10	0.5	10/10 x (1-0.5) + 0.5 = 1

threshold

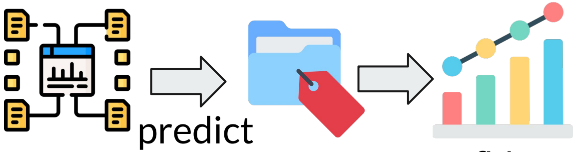
1 $\mathcal{L}_s$

	$P_{y_i^l}$	threshold	$\mathbb{I}(p_{y_i^l} < \eta_t)$
labeled data 1	0.4	0.55	1
labeled data 2	0.8	0.55	0

Annealing Supervised Loss

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

labeled data



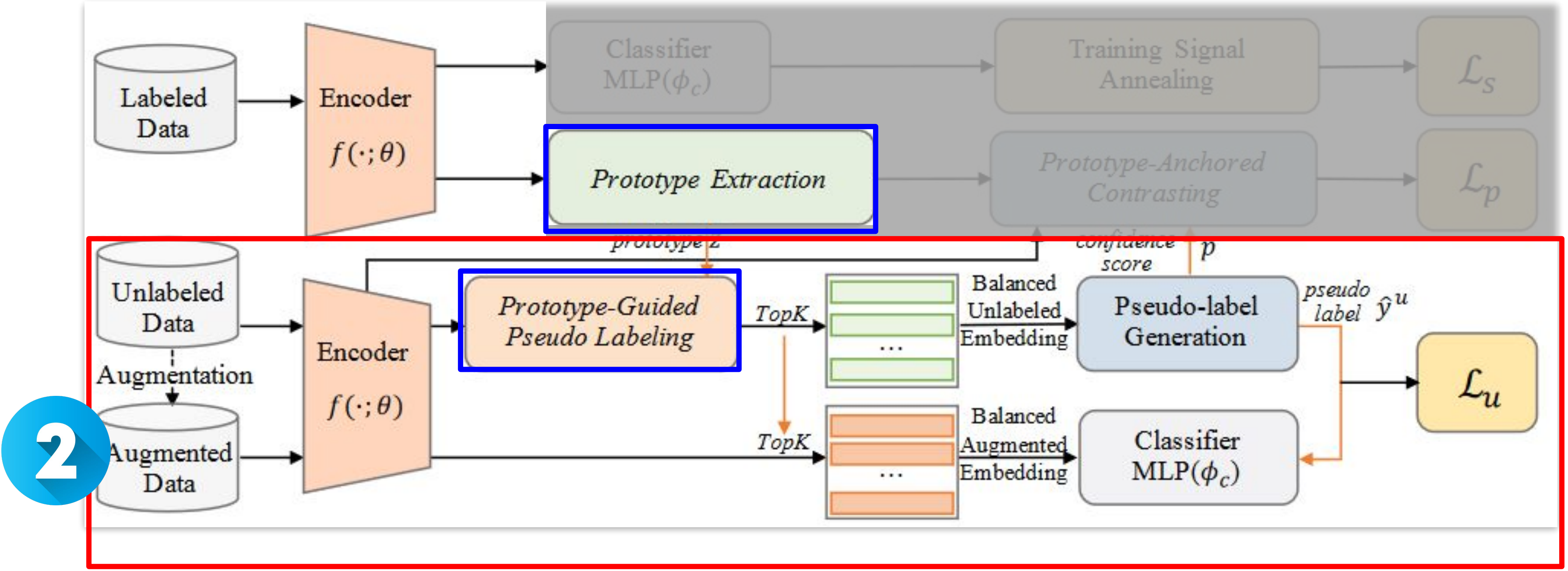
predict

confidence

$$\begin{cases} P_{y_i^l} < \text{threshold} \Rightarrow 0 \\ P_{y_i^l} > \text{threshold} \Rightarrow 1 \end{cases}$$

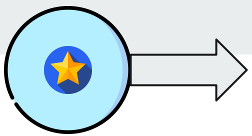
$\Rightarrow$ encourage the model learn the data with relatively lower confidence

Prototype-Guided Pseudo Labeling(Partition2)

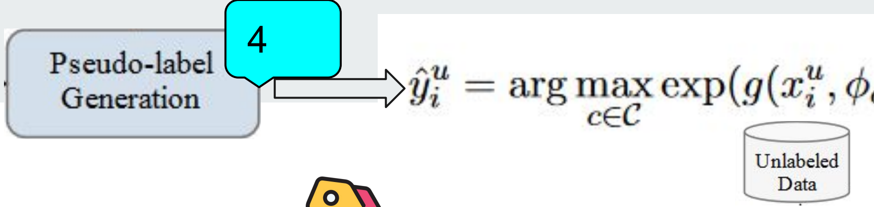


2

prototype

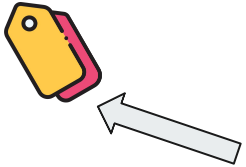


the most representative points in each category

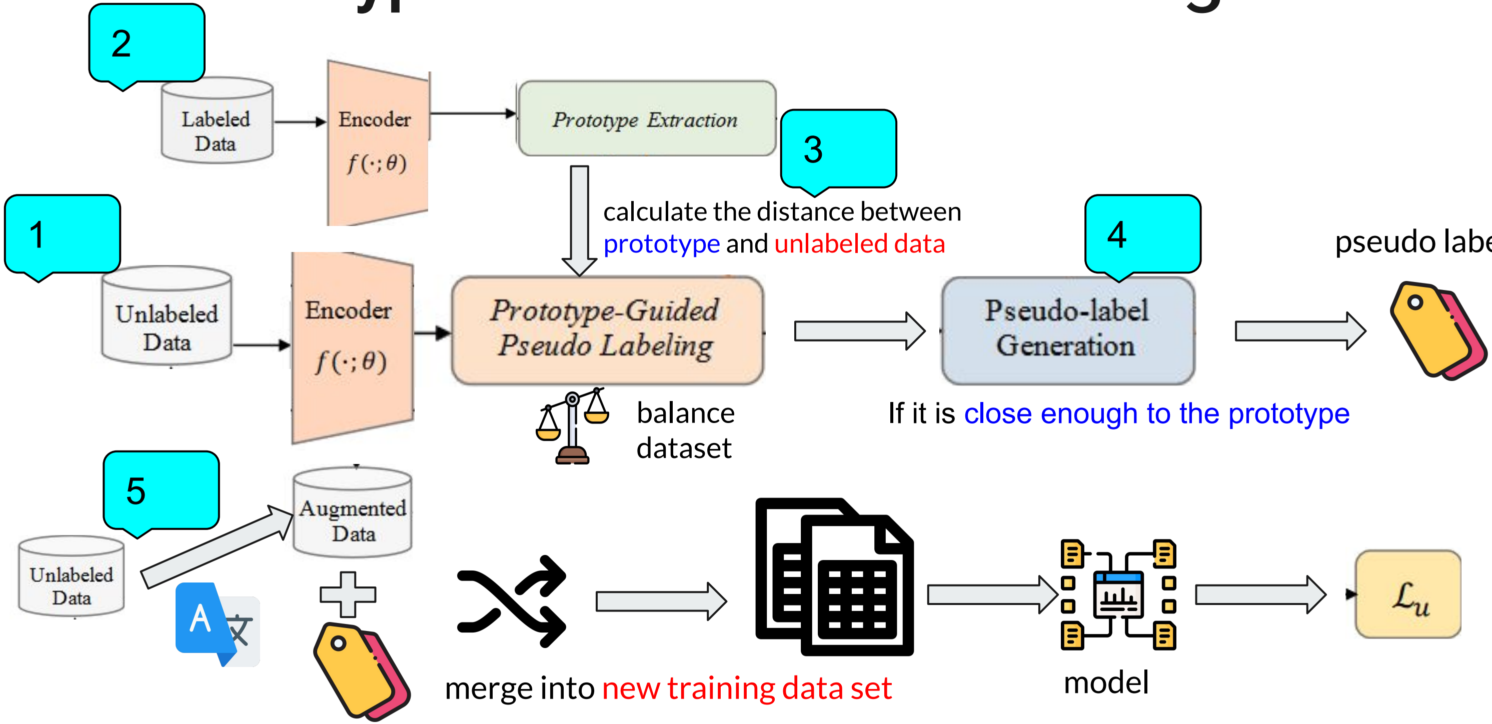


$$\hat{y}_i^u = \arg\max_{c \in C} \exp(g(x_i^u, \phi_c, \theta))$$

Prototype-Guided Pseudo Labeling

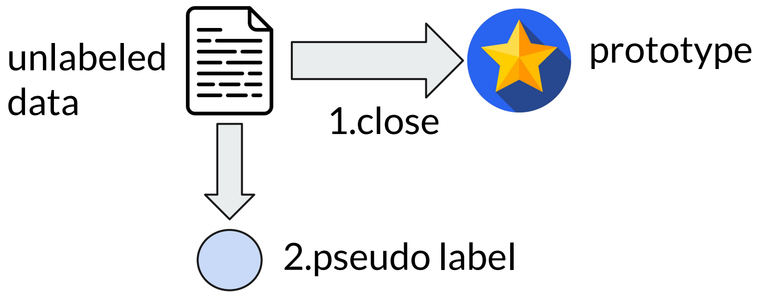


	relevance score
A	0.8
B	0.15
C	0.05



Q&A why use prototype

- bring each unlabeled data's embedding closer to its associated prototype



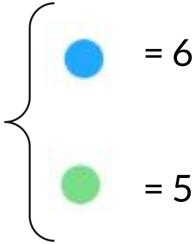
GOAL

Find the most representative data from each category

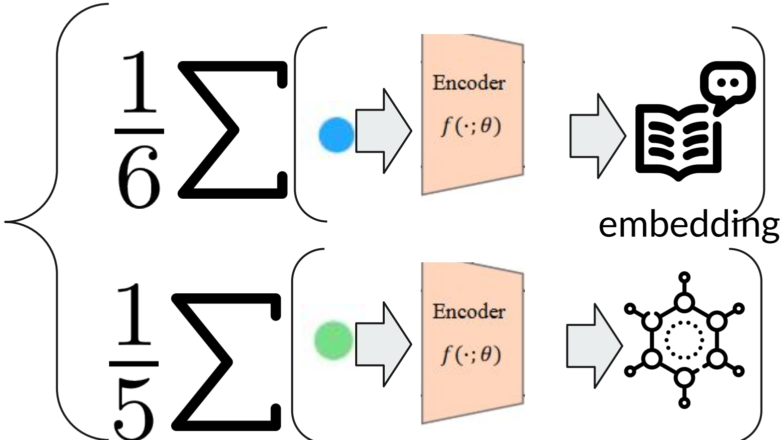
Prototype Extraction

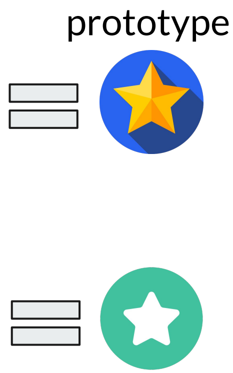
count

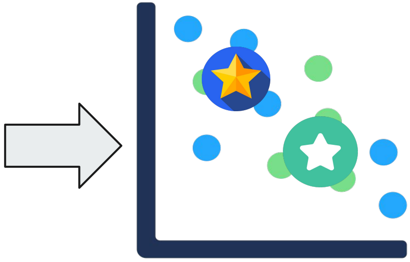
$$n_c = \sum_{y_i^l \in D_l} \mathbb{I}(y_i^l = c)$$



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



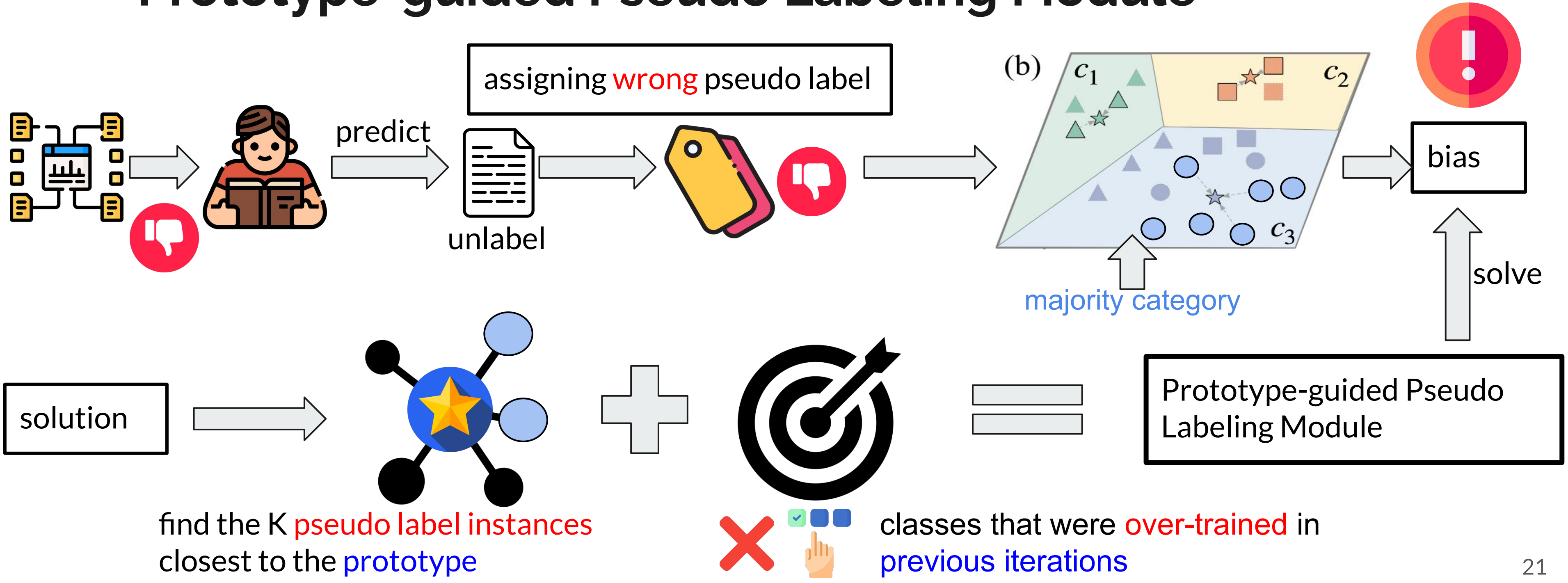




over-trained

class	<t	t	final
A	50	5	50
B	20	4	22
C	5	10	15

Prototype-guided Pseudo Labeling Module



2 *Prototype-Guided Pseudo Labeling*

GOAL calculate how many pseudo labels each category should receive in this iteration

Prototype-guided Pseudo Labeling Module

Augment data

count:10

$$\mu_t^c = \sum_{\tilde{x}_j^u \in B_u} \mathbb{I}(\hat{y}_j = c)$$

Augment data batches of unlabeled data

	count
economy	2
politics	3
sports	5

returns the number of selected pseudo-label for each class at the t iteration

Prototype-guided Pseudo Labeling Module

$$\boxed{\mu^c_{<t}}$$

Get the total number of pseudo label for each category from the previous iteration

$$\boxed{\gamma_t = \arg\min_{c \in C} \mu^c_{<t}.}$$

The previous iteration finds the smallest number of pseudo-labels in each category

Assign the number of pseudo labels to each category

$$k_c = \begin{cases} \mu^c_t & \text{if } \mu^c_{<t} - \gamma_t = 0 \\ \mu^c_t - (\mu^c_{<t} - \gamma_t) & \text{if } 0 \leq \mu^c_{<t} - \gamma_t < \mu^c_t \\ 0 & \text{if } \mu^c_{<t} - \gamma_t \geq \mu^c_t \end{cases}$$

2 *Prototype-Guided Pseudo Labeling*

γ_t =2 (sports) $\gamma_t = \arg\min_{c \in C} \mu^c_{<t}.$

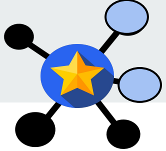
	μ^c_t
economy	2
politics	3
sports	5

Prototype-guided Pseudo Labeling Module

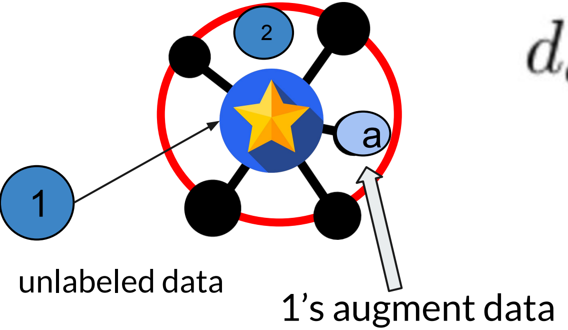
$\mu^c_{<t}$ μ^c_t

	Previous t times	pseudo label currently found	$\mu^c_{<t} - \gamma_t$	condition	Kc
economy	6	2	6-2 = 4	$\mu^c_{<t} - \gamma_t \geq \mu^c_t$	0
politics	4	3	4-2=2	$0 \leq \mu^c_{<t} - \gamma_t < \mu^c_t$	3-(4-2) = 1 $\mu^c_t - (\mu^c_{<t} - \gamma_t)$
sports	2 γ_t	5	2-2=0	$\mu^c_{<t} - \gamma_t = 0$	5 μ^c_t

2 *Prototype-Guided Pseudo Labeling*

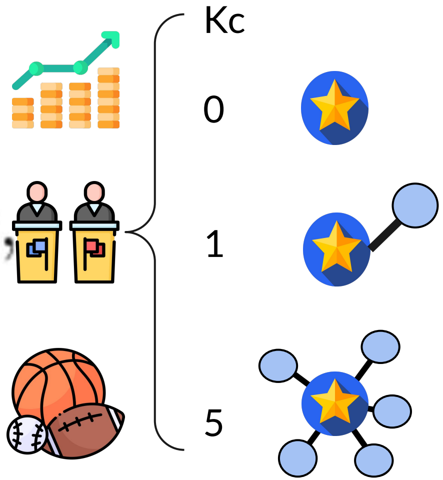


Prototype-guided Pseudo Labeling Module



$$d_c = \mathrm{TopK}(\mathrm{d}(f(\tilde{x}_j^u), z_c), k_c),$$

augment data



Choose the length closest to the prototype

2 $\mathcal{L}_u$

Selective Unsupervised Loss

entropy

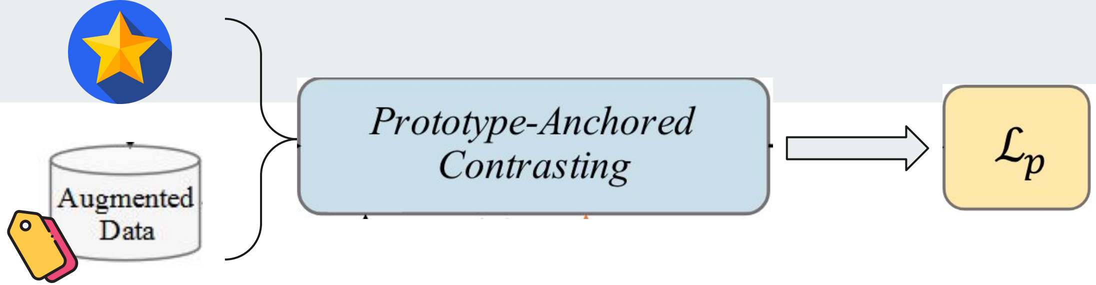
$$\mathcal{L}_u = -\frac{1}{n}\sum_{i=1}^{n} \boxed{\mathbb{I}(x_i^u)} \boxed{\log \frac{\exp(g(\tilde{x_i^u}, \phi_{y_i^u}, \theta))}{\sum_{c \in \mathcal{C}} \exp(g(\tilde{x_i^u}, \phi_c, \theta))}}$$

if pseudo label = sports
& true label = sports

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

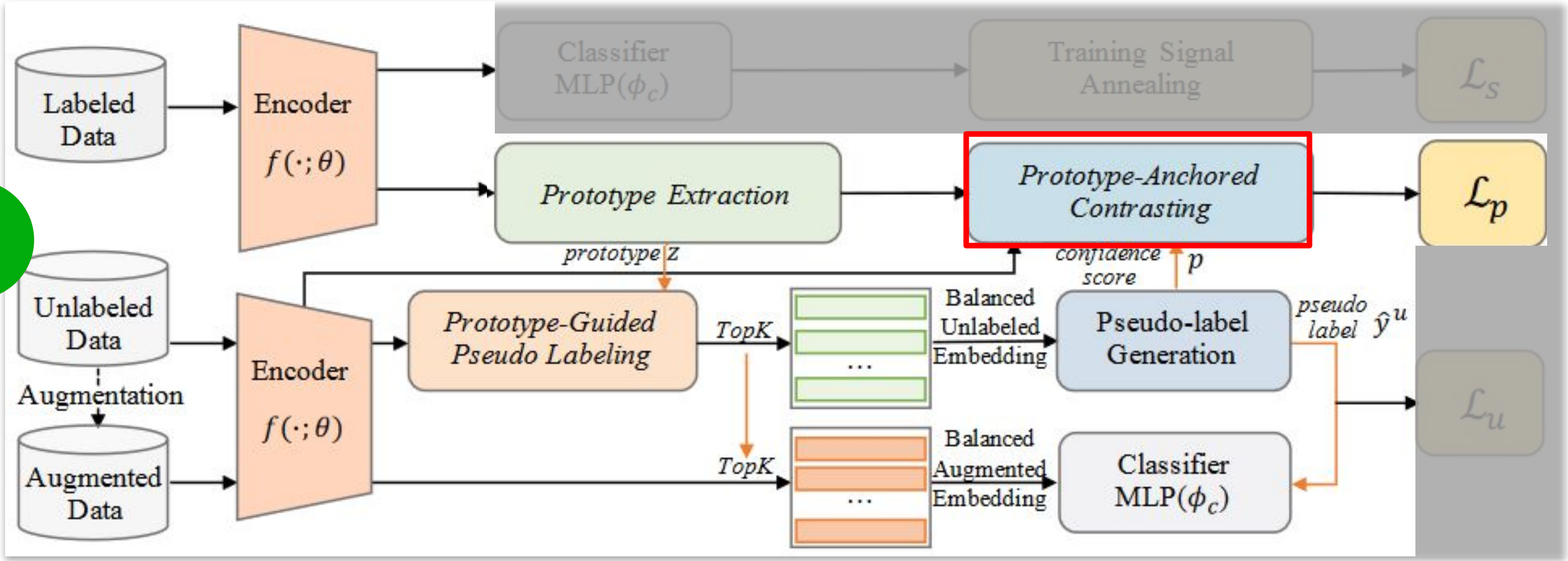
1 or 0

The distance between unlabeled data and prototype is smaller than the distance between augmentation and prototype.



Prototype-Guided Pseudo Labeling(Partition3)

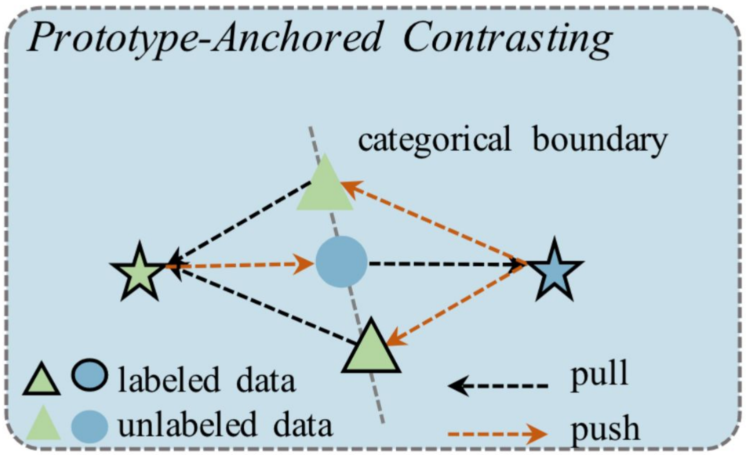
3



3 $\mathcal{L}_p$

make each instance closer to the prototype of the category

Prototype-Anchored Contrasting (PAC)



$$\mathcal{L}_p = -\frac{1}{m}\sum_{i}^{m}\sum_{c}\mathbb{I}(y_i^l = c)\log\frac{\exp\left(-\mathrm{d}(f(\tilde{x_i^l}), z_c)\right)}{\sum_{k\in\mathcal{C}}\exp(-\mathrm{d}(f(\tilde{x_i^l}), z_k))}$$

loss of label data

$$-\frac{\lambda}{n}\sum_{j}^{n}\sum_{c}\boxed{p_{\hat{y}_j^u}}\mathbb{I}(\hat{y}_j^u = c)\log\frac{\exp\left(-\mathrm{d}(f(\tilde{x_j^u}), z_c)\right)}{\sum_{k\in\mathcal{C}}\exp(-\mathrm{d}(f(\tilde{x_j^u}), z_k))},$$

loss of augment data

predict

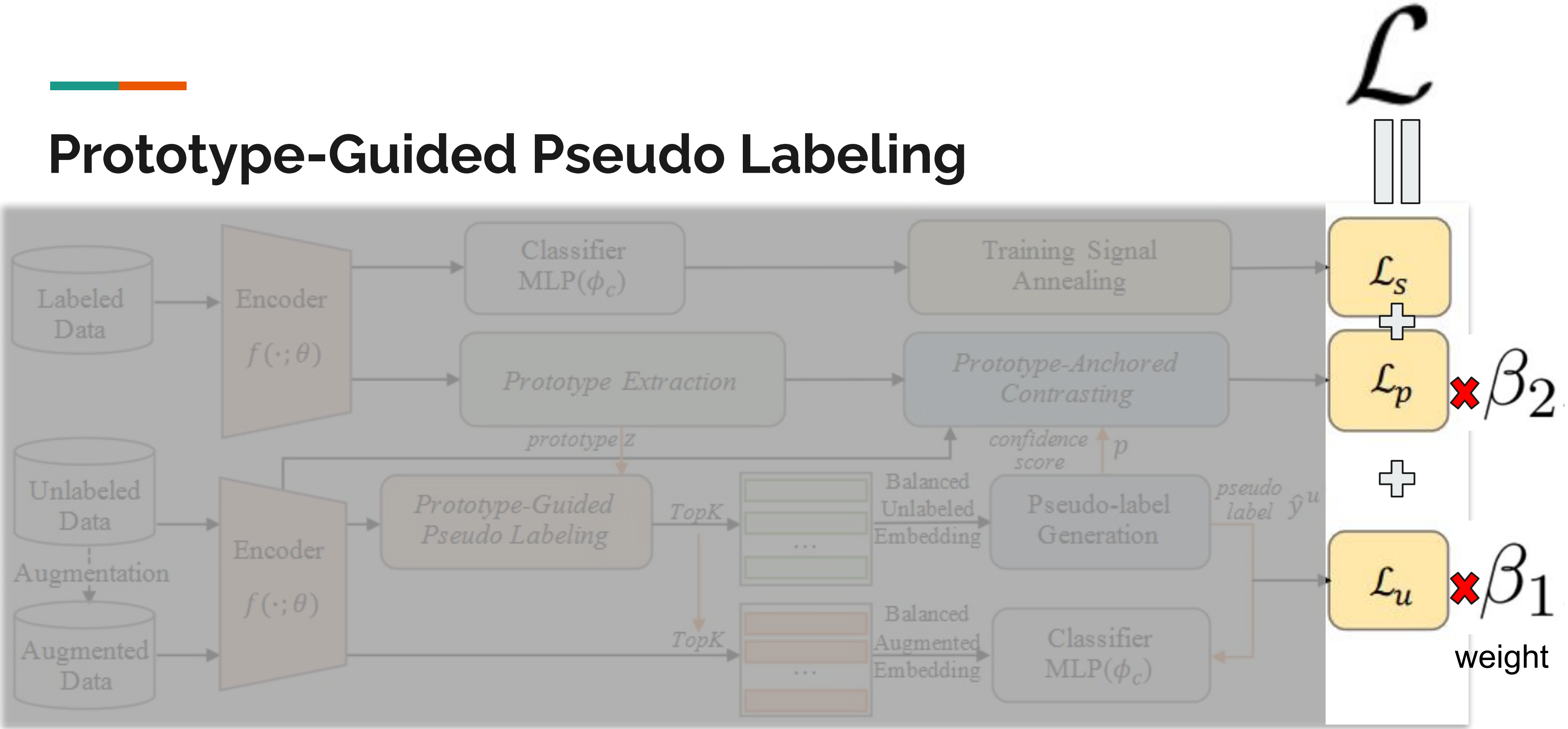
unlabeled

confidence

Use the confidence of unlabeled data as weight

balance

Prototype-Guided Pseudo Labeling



Outline

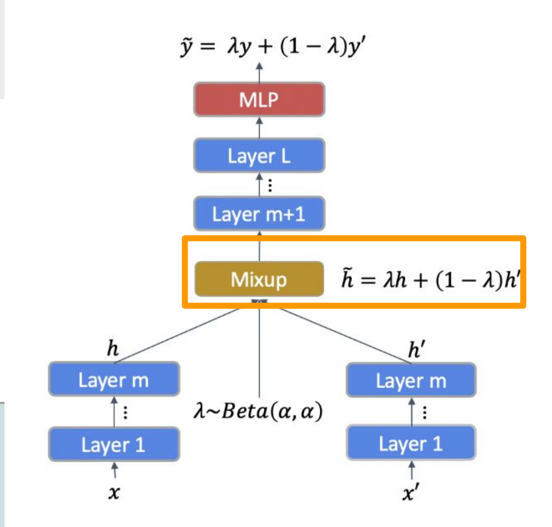
- Introduction
- Method
- Experiment
- Conclusion

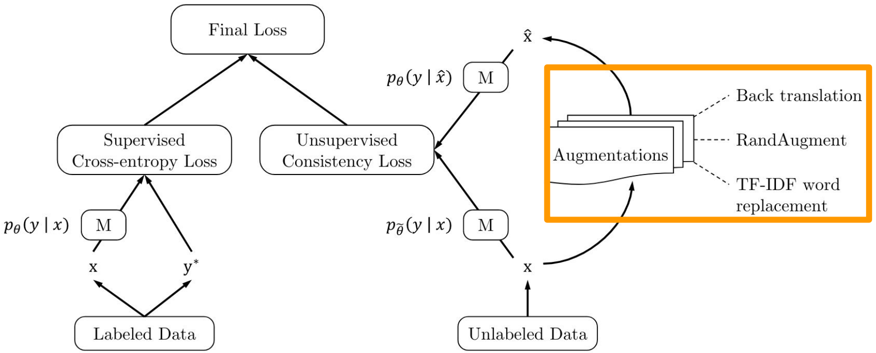
Dataset

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

Dataset	AG News	IMDB	Yahoo! Answer	DBpedia
description	collection of news	• movie reviews • Binary emotion classification	topic classification	extract structured content from the information created in Wikipedia
category	• World • Sports • Business • Sci/Tech	• positive • negative	• Society & Culture • Science & Mathematics • Health • Education & Reference • Sports • Business & Finance • ……	14 class • company • education • ……

Baseline





Model		learning method
BERT	trained with only labled data	supervised
UDA	make different augment to unlabeled data	semi-supervised
Mixtext	TMix: A new data augmentation method that interpolates two input vectors in hidden space to generate a new vector	semi-supervised

UDA:unsupervised data augmentation

- metric : accuracy

Experiment Results

the number of
labeled data used

	Model	AG News			IMDB			Yahoo! Answer			DBpedia		
		10	30	200	10	30	200	10	30	200	10	30	200
supervised	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
semi-supervised	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
this paper	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**

unsupervised > supervised

the gap between unsupervised and supervised narrows

PGPL is stable in all aspects except Yahoo

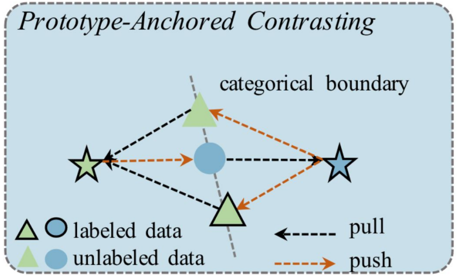
- metric : accuracy

Evaluation Results with other pre-trained models

this paper

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

Prototype-Guided Pseudo Labeling



- metric : accuracy

Ablation Study

1. PGP and PAC can independently improve model performance

2. TSA helps too, besides DBpedia

Training Signal Annealing

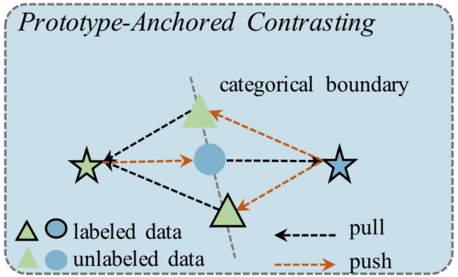
Data	AG News	IMDB
PGPL	**88.3**	**89.7**
w/o PGP	86.5	88.2
w/o PAC	86.2	88.9
w/o TSA	87.2	87.9

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**

Outline

- Introduction
- Method
- Experiment
- Conclusion

PGP



PAC

Conclusion

1. A semi-supervised model PGPL that combines PAC and PGP strategies is proposed for semi-supervised text classification tasks.

2. After constructing the prototype, use PAC to group text embeddings belonging to the same category together to alleviate the problem of underfitting.

3. PGP selects reliable pseudo-labeled data nearby prototypes to address the training bias from the imbalanced data