Neighborhood-Regularized Self-Training for Learning with Few Labels

method

task

Advisor : Jia-Ling, Koh

Speaker : Ting-I, Weng

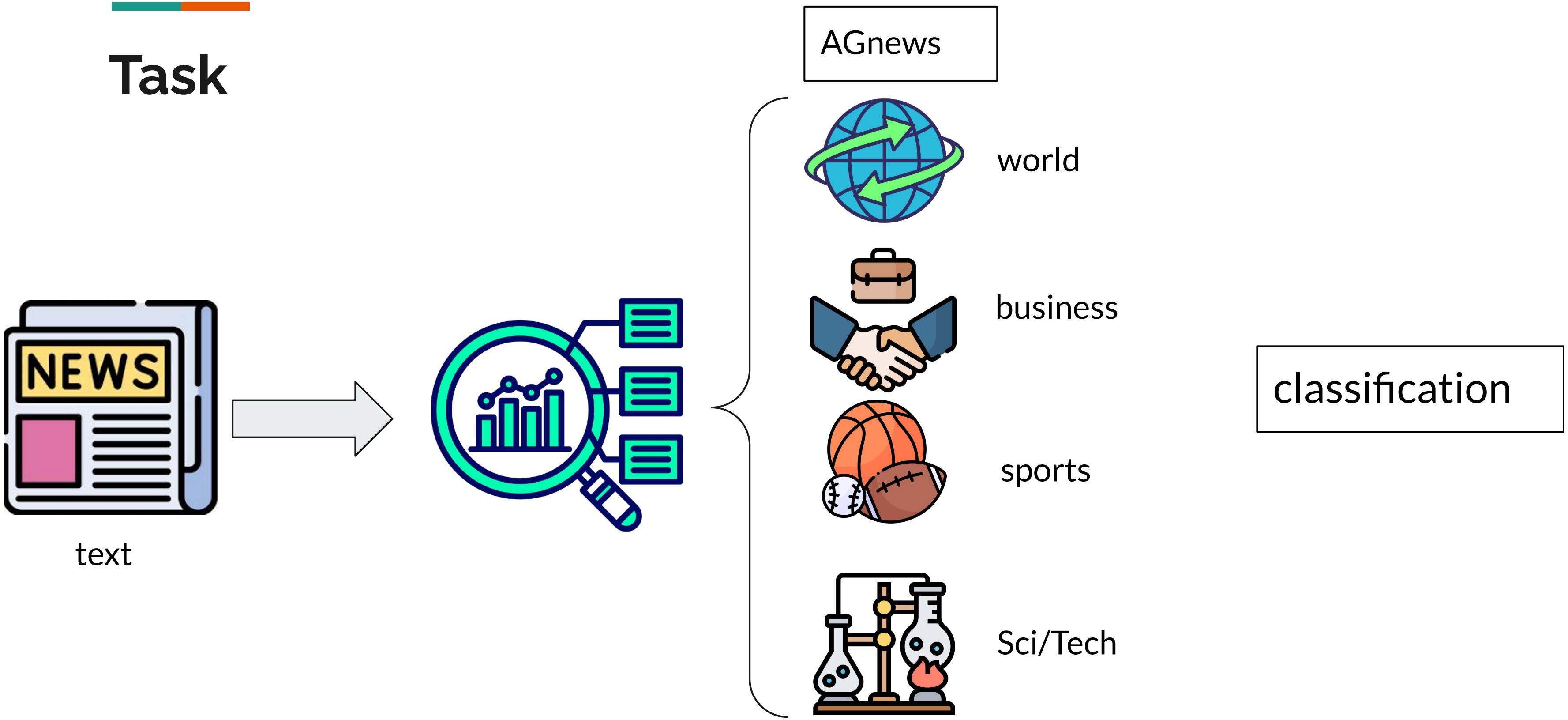
Source : AAAI'23

Date : 2023/12/19

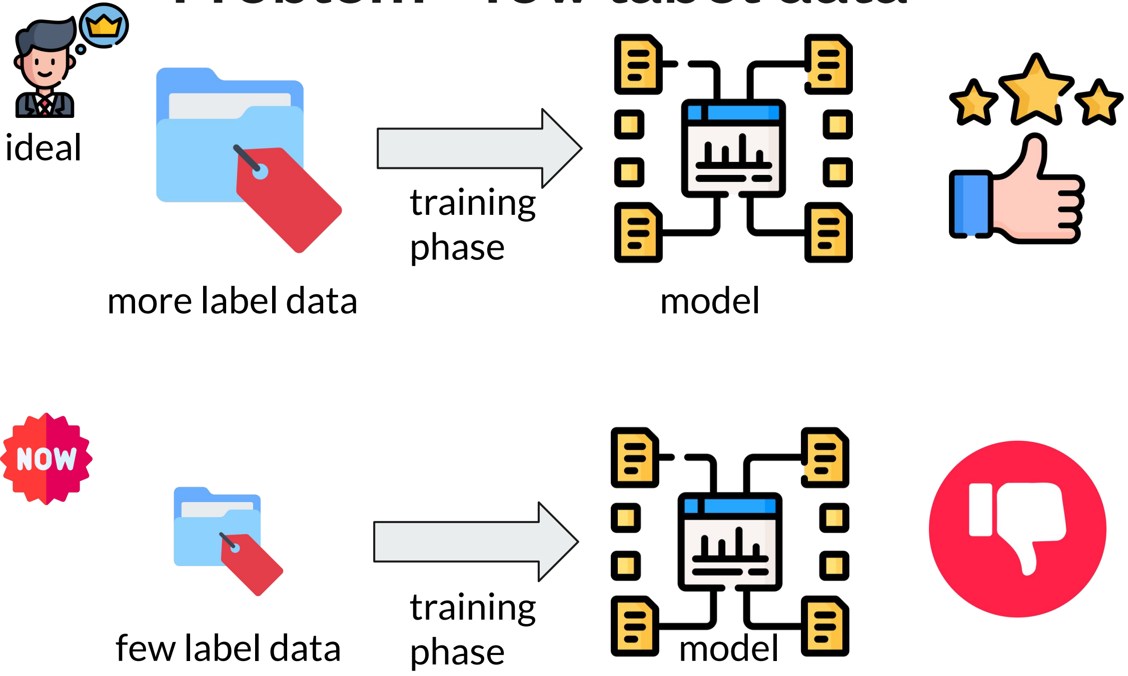
Outline

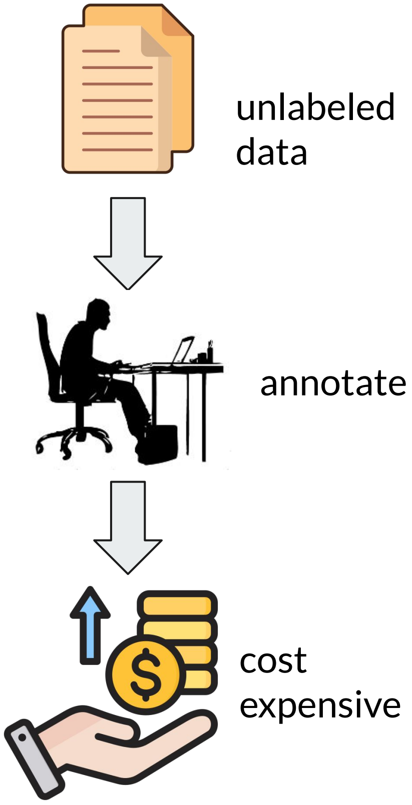
- Introduction
- Method
- Experiment
- Conclusion

Task


AGnews
NEWS
text
world
business
sports
Sci/Tech
classification

Problem - few label data

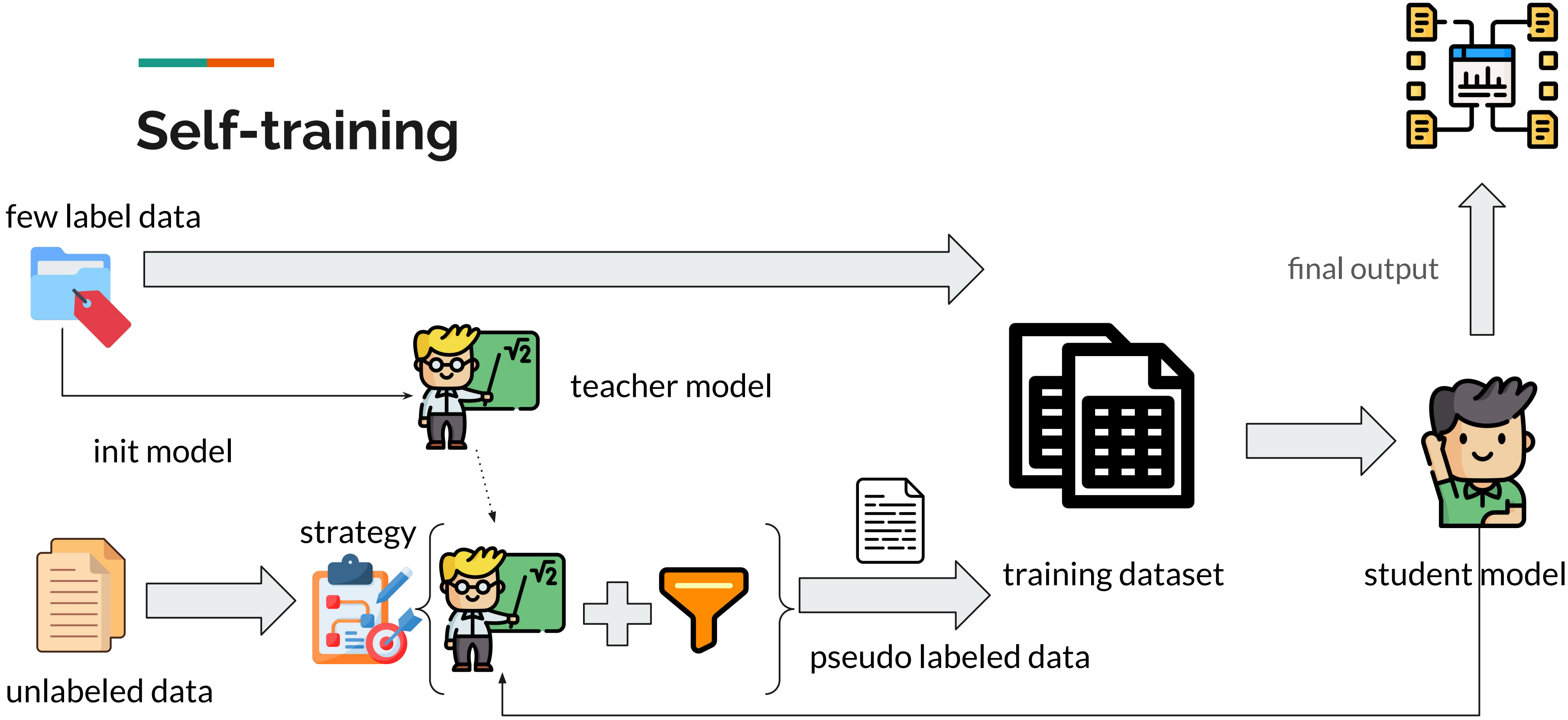




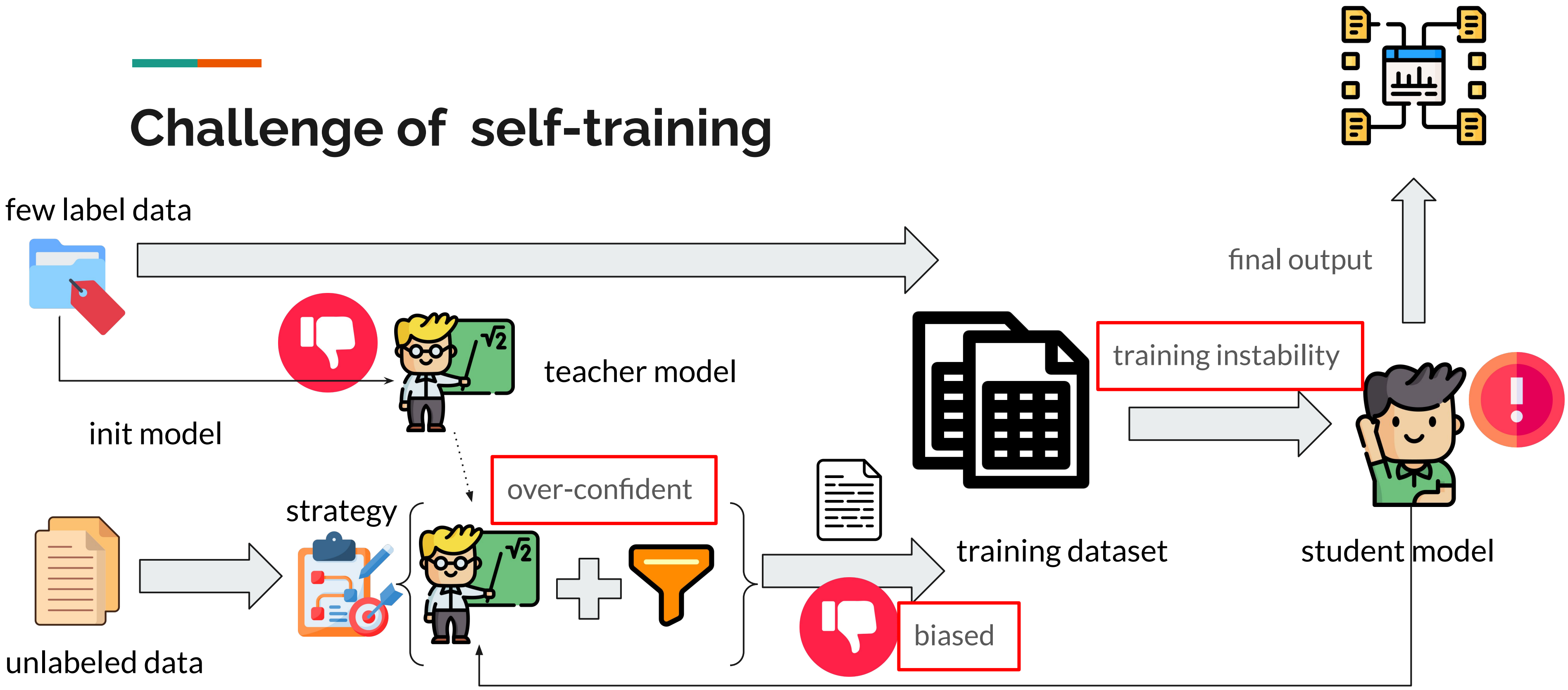
GOAL

Hope to use unlabeled data to assist training

Self-training



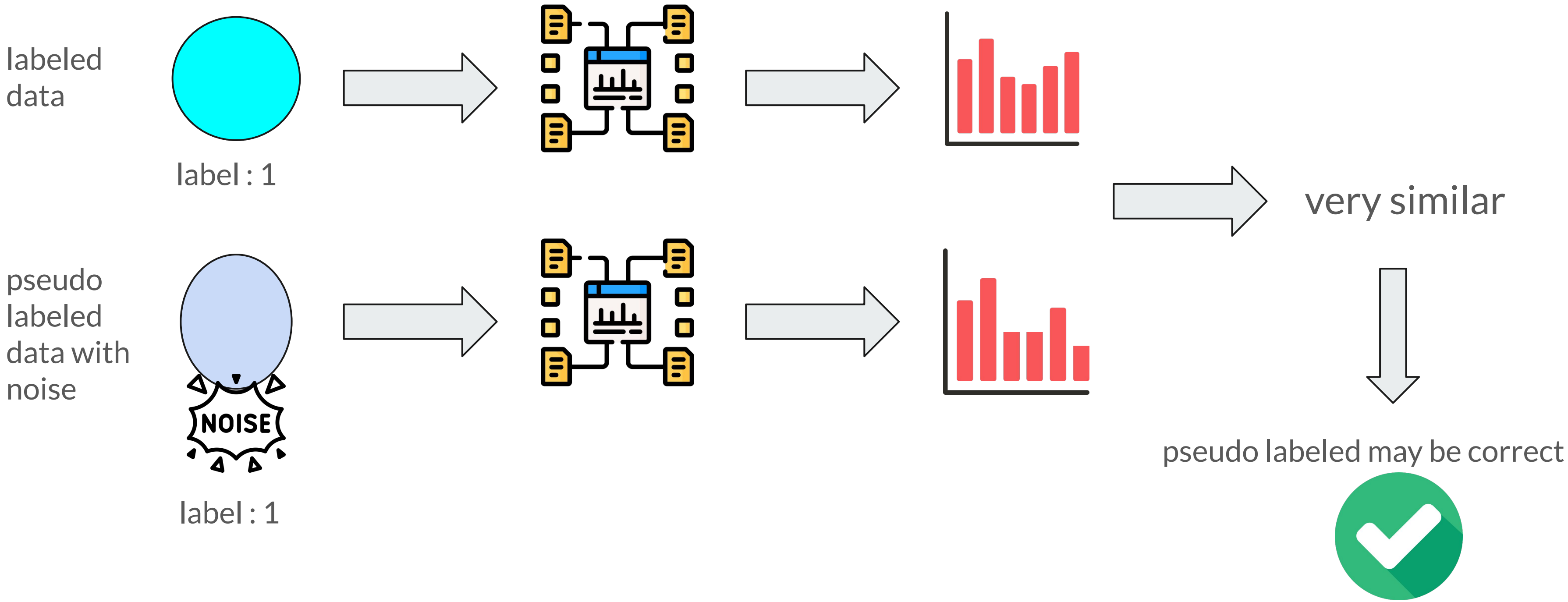
Challenge of self-training



Discriminative Topic Mining via Category-Name Guided Text Embedding(https://arxiv.org/pdf/1908.07162.pdf)

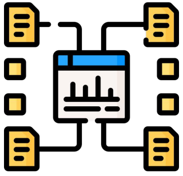
samples with similar labels tend to share similar representations

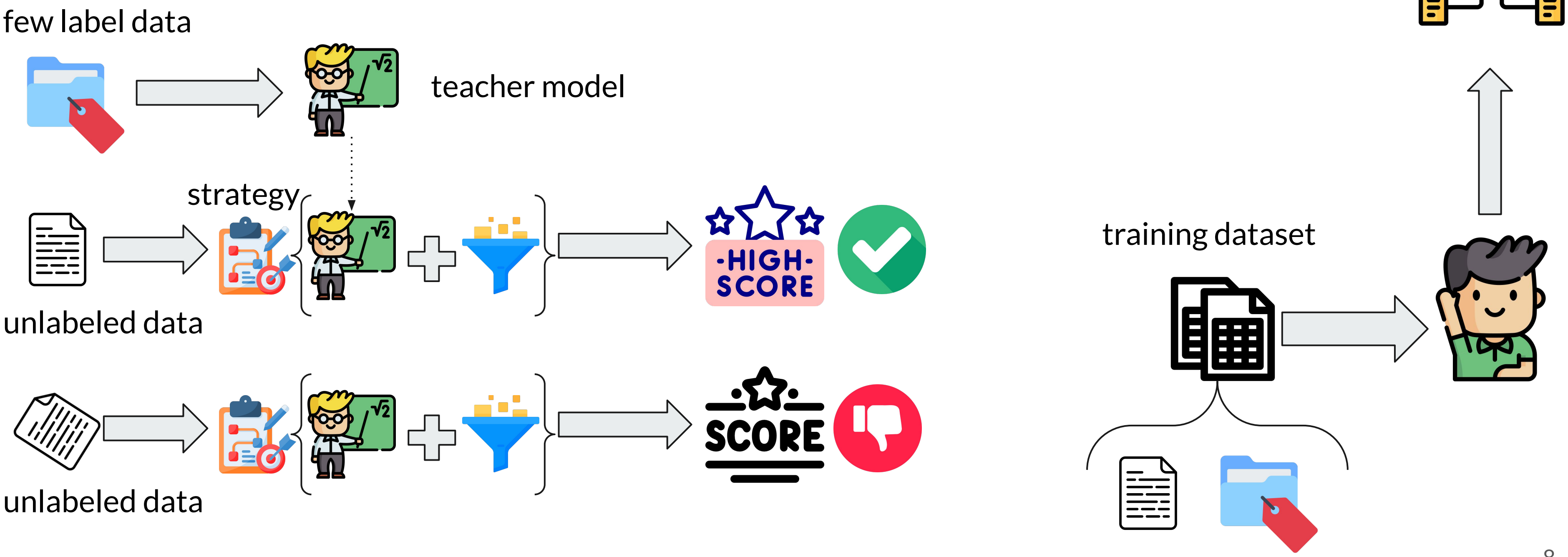
Hypothesize



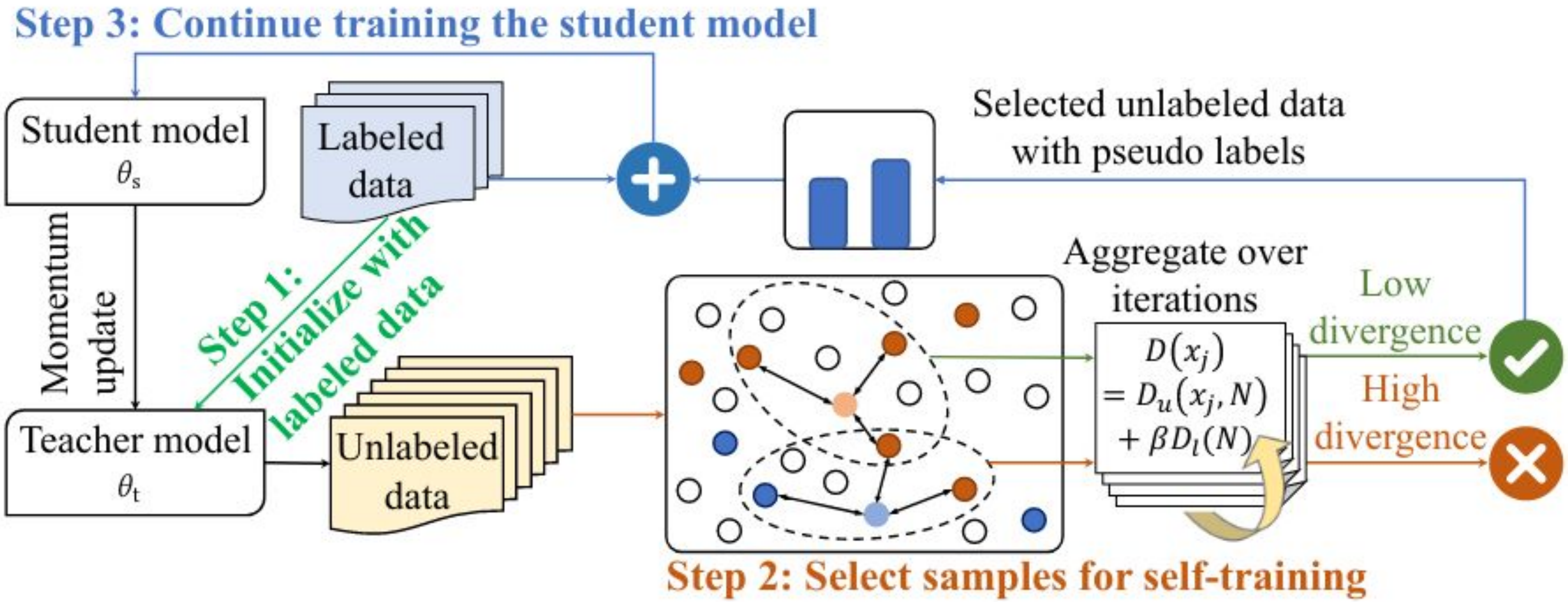
- over-confident
- biased
- training instability

Solution





NeST(Neighborhood Regularized Self-Training)



Outline

- Introduction
- Method
- Experiment
- Conclusion

class:
- black : 0
- green: 1

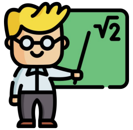
samples with similar labels tend to share similar representations

- Xl: labeled data
  - black : 5
  - green: 5

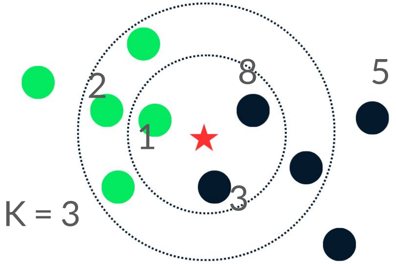
Neighborhood-Regularized Sample Selection

$$\mathcal{N}_j = \{x_i \mid x_i \in \mathcal{X}_l \cap \mathrm{KNN}(\boldsymbol{v}_j, \mathcal{X}_l, k)\}$$

unlabeled data



embedding



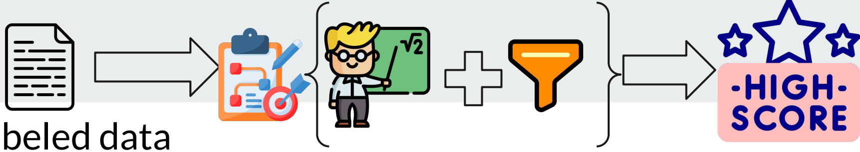
Number close to unlabeled data

$$N = \{X_1, X_3, X_8\}$$

- black : 2
- green: 1

Neighborhood

https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence
https://pytorch.org/docs/stable/generated/torch.nn.KLDivLoss.html

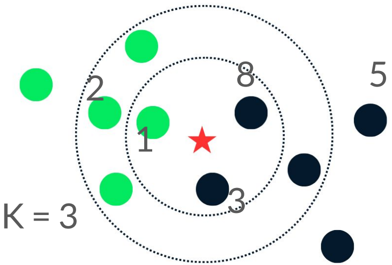


Divergence-based Sample Selection

P(x) unlabeled data

$$D_{\mathrm{KL}}(Q \parallel P) = \sum_{x \in \mathcal{X}} Q(x) \ln\left(\frac{Q(x)}{P(x)}\right)$$

Q(x)



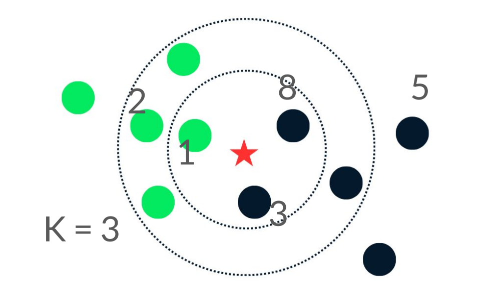
Neighborhood

$$\mathcal{D}(x_j) = \mathcal{D}_{\mathrm{u}}(x_j, \mathcal{N}) + \beta \mathcal{D}_{\mathrm{l}}(\mathcal{N})$$

unlabeled data

$$N = \{X_1, X_3, X_8\}$$

class:
black : 0
green: 1



Unlabeled Divergence Du

$$\boxed{\mathcal{D}_{\mathrm{u}}(x_j, \mathcal{N}) = \sum_{(x_i, y_i) \in \mathcal{N}} d(f(x_j; \theta_t), y_i)}$$

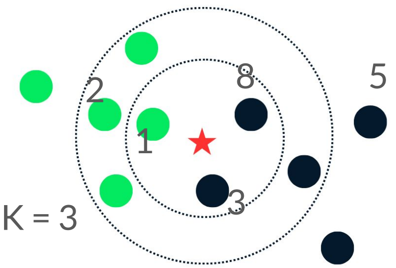
unlabeled data

$$f(x_1^u; \theta_t) = [0.2, 0.8]$$

$$y_i \begin{cases} y_1 = [0, 1] \\ y_2 = [1, 0] \\ y_3 = [1, 0] \end{cases}$$

$$\begin{aligned} D_u(x_1^u, N_1) &= d(f(x_1^u; \theta_t), y_1) \\ &\quad + d(f(x_1^u; \theta_t), y_2) \\ &\quad + d(f(x_1^u; \theta_t), y_3) \\ &= 0 log_e \frac{0}{0.2} + 1 log_e \frac{1}{0.8} \\ &\quad + 1 log_e \frac{1}{0.2} + 0 log_e \frac{0}{0.8} \\ &\quad + 1 log_e \frac{1}{0.2} + 0 log_e \frac{0}{0.8} \\ &= 0.2231 + 1.6094 + 1.6094 = 3.442 \end{aligned}$$

class:
black : 0
green: 1



K = 3

- black : 2
- green: 1

$$N = \{X_1, X_3, X_8\}$$

Labeled Divergence Dl

$$\mathcal{D}_l(\mathcal{N}) = \sum_{(x_i, y_i) \in \mathcal{N}} d(\bar{y}, y_i)$$

$$\bar{y} = \Sigma \frac{y_i}{|N|} = \frac{[0,1] + [1,0] + [1,0]}{3} = [\frac{2}{3}, \frac{1}{3}]$$

$$y_i \begin{cases} y_1 = [0,1] \\ y_2 = [1,0] \\ y_3 = [1,0] \end{cases}$$

$$\begin{aligned} D_l(N_1) &= \Sigma d(\bar{y}, y_i) \\ &= d([\frac{2}{3}, \frac{1}{3}], [0,1]) \\ &+ d([\frac{2}{3}, \frac{1}{3}], [1,0]) \\ &+ d([\frac{2}{3}, \frac{1}{3}], [1,0) \\ &= 1.9125 \end{aligned}$$

unlabeled data

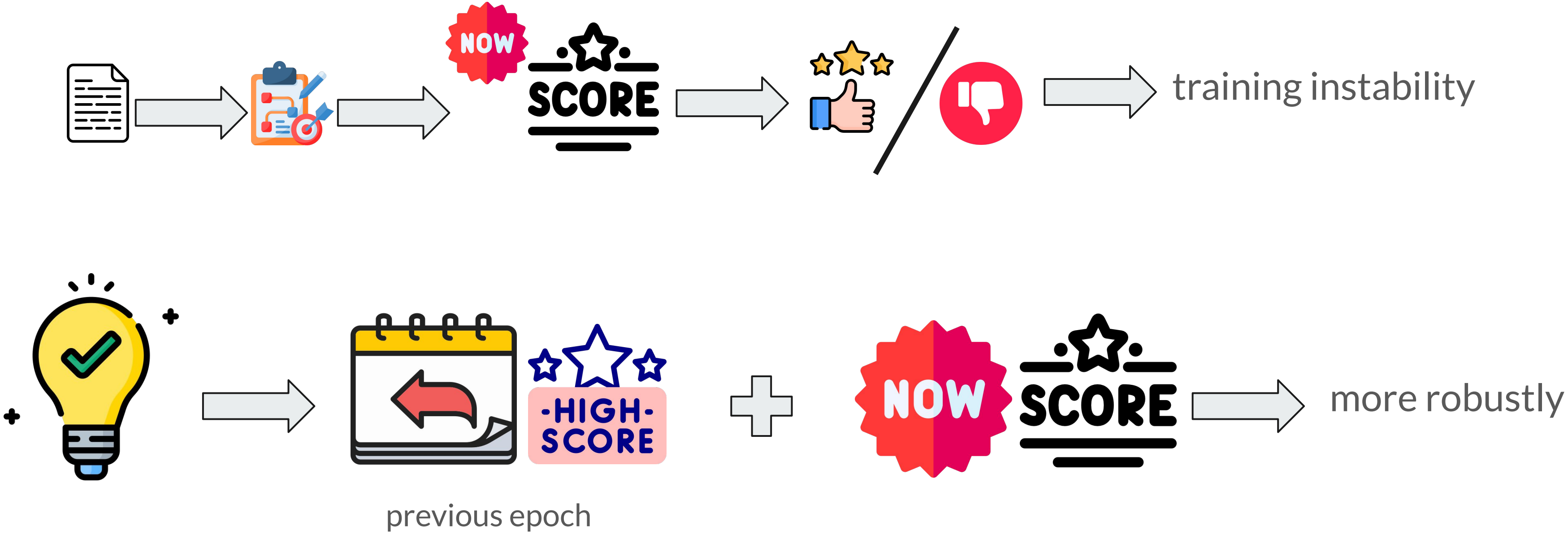
Divergence-based Sample Selection

$$\mathcal{D}(x_j) = \boxed{\mathcal{D}_{\mathrm{u}}(x_j, \mathcal{N})} + \beta \boxed{\mathcal{D}_{\mathrm{l}}(\mathcal{N})}$$

$$D(x_1^u) = \boxed{3.442} + \beta * \boxed{1.9125}$$

$$= 3.442 + 0.1 * 1.9125 = 3.6332$$

Aggregation of Predictions from Different Iterations

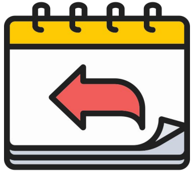


Aggregation of Predictions from Different Iterations

$$\mu^{(t)}(x_j) = (1-m)\times \boxed{\mu^{(t-1)}(x_j)} + m \times \boxed{\left(\mathcal{D}^{(t)}(x_j)\right)}$$

unlabeled data

previous



$$\mu^{(t)}(x_j) = (1-m)\times\mu^{(t-1)}(x_j) + m\times\left(\mathcal{D}^{(t)}(x_j)\right)$$ ⇨ more robustly

previous epoch

m = 0.6

t = 1

Aggregation of Predictions from Different Iterations

$$\mu^1(x_1^u) = (1-m)*\mu^{1-1}(x_1^u) + m*(D^1(x_1^u)) = 0.6*3.442 = 2.0652$$

$$\mu^1(x_2^u) = (1-m)*\mu^{1-1}(x_2^u) + m*(D^1(x_2^u)) = 0.6*1.9125 = 1.1475$$

$$\mu^1(x_3^u) = (1-m)*\mu^{1-1}(x_3^u) + m*(D^1(x_3^u)) = 0.6*0.9 = 0.54$$

label	$D^{t=1}(x_j)$	$D^{t=2}(x_j)$	$D^{t=3}(x_j)$
x_1^u	3.442	2.5	1.5
x_2^u	1.9125	1.6	2.7
x_3^u	0.9	0.7	0.5

t = 2

m = 0.6

	μ^1
x_1^u	2.0652
x_2^u	1.1475
x_3^u	0.54

$$\mu^2(x_1^u) = (1-m) * \mu^{2-1}(x_1^u) + m * (D^2(x_1^u)) = 0.4 * 2.0652 + 0.6 * 2.5 = 2.326$$

$$\mu^2(x_2^u) = (1-m) * \mu^{2-1}(x_2^u) + m * (D^2(x_2^u)) = 0.4 * 1.1475 + 0.6 * 1.6 = 1.149$$

$$\mu^2(x_3^u) = (1-m) * \mu^{2-1}(x_3^u) + m * (D^2(x_3^u)) = 0.4 * 0.54 + 0.6 * 0.6243 = 0.5905$$

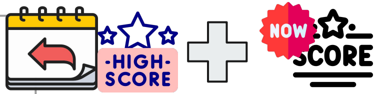
label	$D^{t=1}(x_j)$	$D^{t=2}(x_j)$	$D^{t=3}(x_j)$
x_1^u	3.442	2.5	1.5
x_2^u	1.9125	1.6	2.7
x_3^u	0.9	0.7	0.5

SCORE

Robust Aggregation of Predictions from Different Iterations

x_1^u model gives inconsistent predictions in different iterations → hurt model

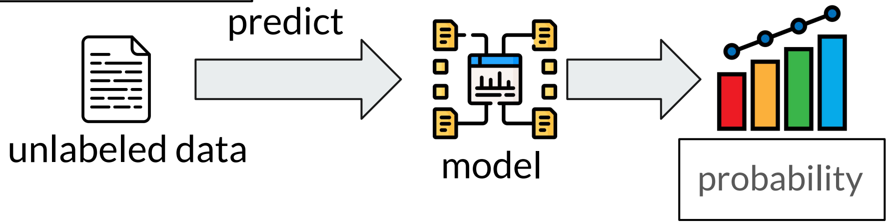
x_3^u model output **consistently low scores** in different iterations → model more certain



	μ^1	μ^2	μ^3 suppose	μ^4 suppose
x_1^u	2.0652	2.326	0.9	2.2
x_2^u	1.1475	1.149	1.2	1.05
x_3^u	0.54	0.5905	0.4	0.3

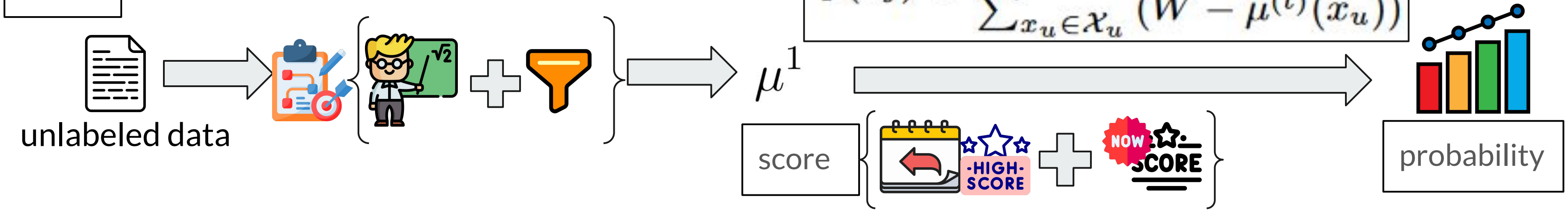
Robust Aggregation of Predictions from Different Iterations

general model



NeST

$$p(x_j) \propto \frac{W - \mu^{(t)}(x_j)}{\sum_{x_u \in \mathcal{X}_u} (W - \mu^{(t)}(x_u))}$$



t = 1

Robust Aggregation of Predictions from Different Iterations

$$p(x_j) \propto \frac{W - \mu^{(t)}(x_j)}{\sum_{x_u \in \mathcal{X}_u} (W - \mu^{(t)}(x_u))}$$

$W = \max_{x \in \mathcal{X}_u}(\mu^{(t)}(x))$ is the normalizing factor

$$W = max_x(\mu^t(x)) = 2.0652$$

	μ^1	μ^2
x_1^u	2.0652	2.326
x_2^u	1.1475	1.149
x_3^u	0.54	0.5905

$$p(x_1) = \frac{2.0652 - 2.0652}{0 + (2.0652 - 1.1475) + (2.0652 - 0.54)} = 0$$

$$p(x_2) = \frac{2.0652 - 1.1475}{0 + (2.0652 - 1.1475) + (2.0652 - 0.54)} = 0.3756$$

$$p(x_3) = \frac{2.0652 - 0.54}{0 + (2.0652 - 1.1475) + (2.0652 - 0.54)} = 0.6243$$

t = 2

Robust Aggregation of Predictions from Different Iterations

$$p(x_j) \propto \frac{W - \mu^{(t)}(x_j)}{\sum_{x_u \in \mathcal{X}_u} (W - \mu^{(t)}(x_u))}$$

$W = \max_{x \in \mathcal{X}_u}(\mu^{(t)}(x))$ is the normalizing factor

	μ^1	μ^2
x_1^u	2.0652	2.326
x_2^u	1.1475	1.149
x_3^u	0.54	0.5905

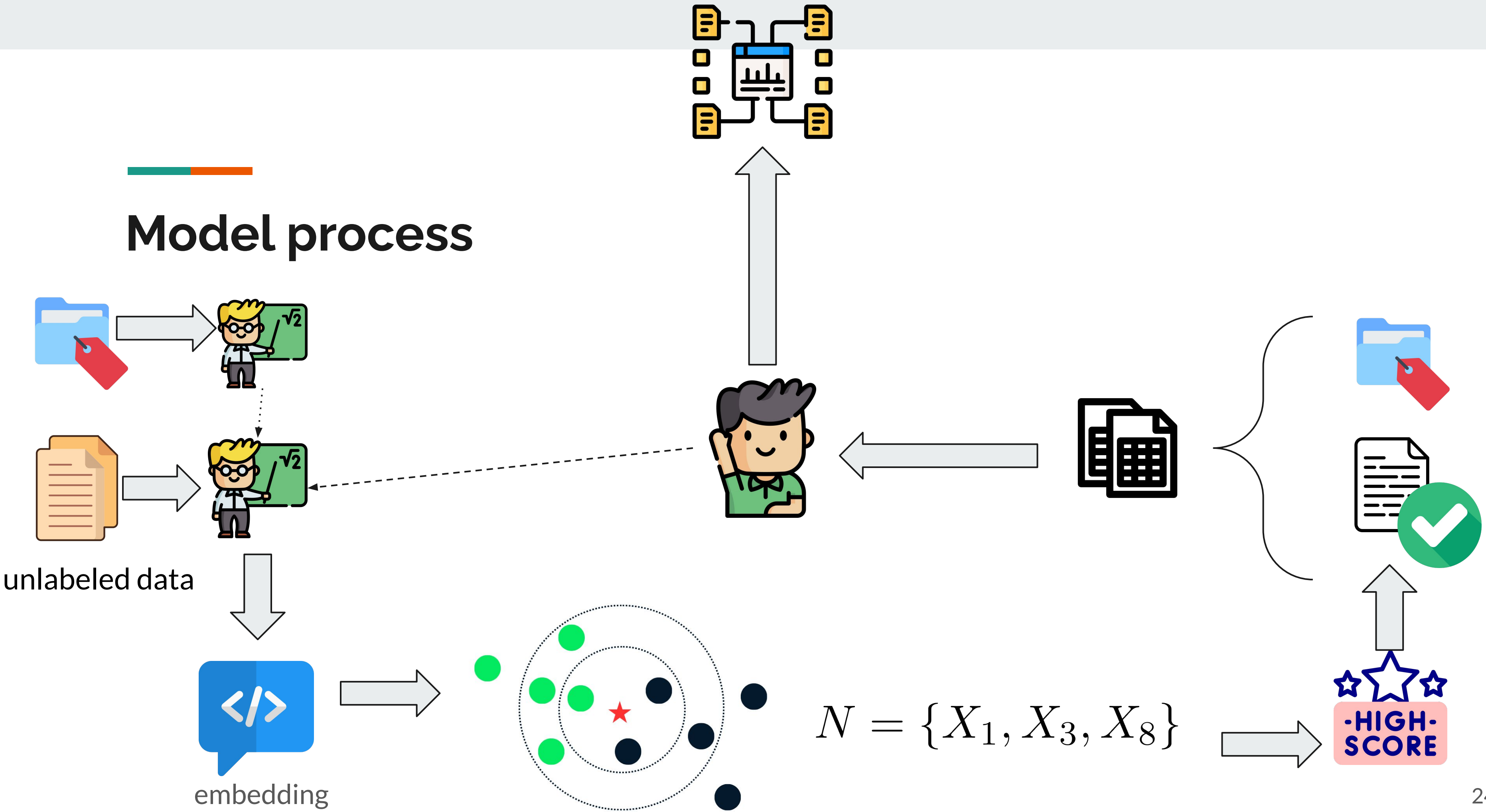
$$W = max_x(\mu^t(x)) = 2.236$$

$$p(x_1) = \frac{2.326 - 2.326}{(2.326 - 1.149) + (2.326 - 0.5905)} = 0$$

$$p(x_2) = \frac{2.326 - 1.149}{(2.326 - 1.149) + (2.326 - 0.5905)} = 0.4041$$

$$p(x_3) = \frac{2.326 - 0.5905}{(2.326 - 1.149) + (2.326 - 0.5905)} = 0.5958$$

Model process



Outline

- Introduction
- Method
- Experiment
- Conclusion

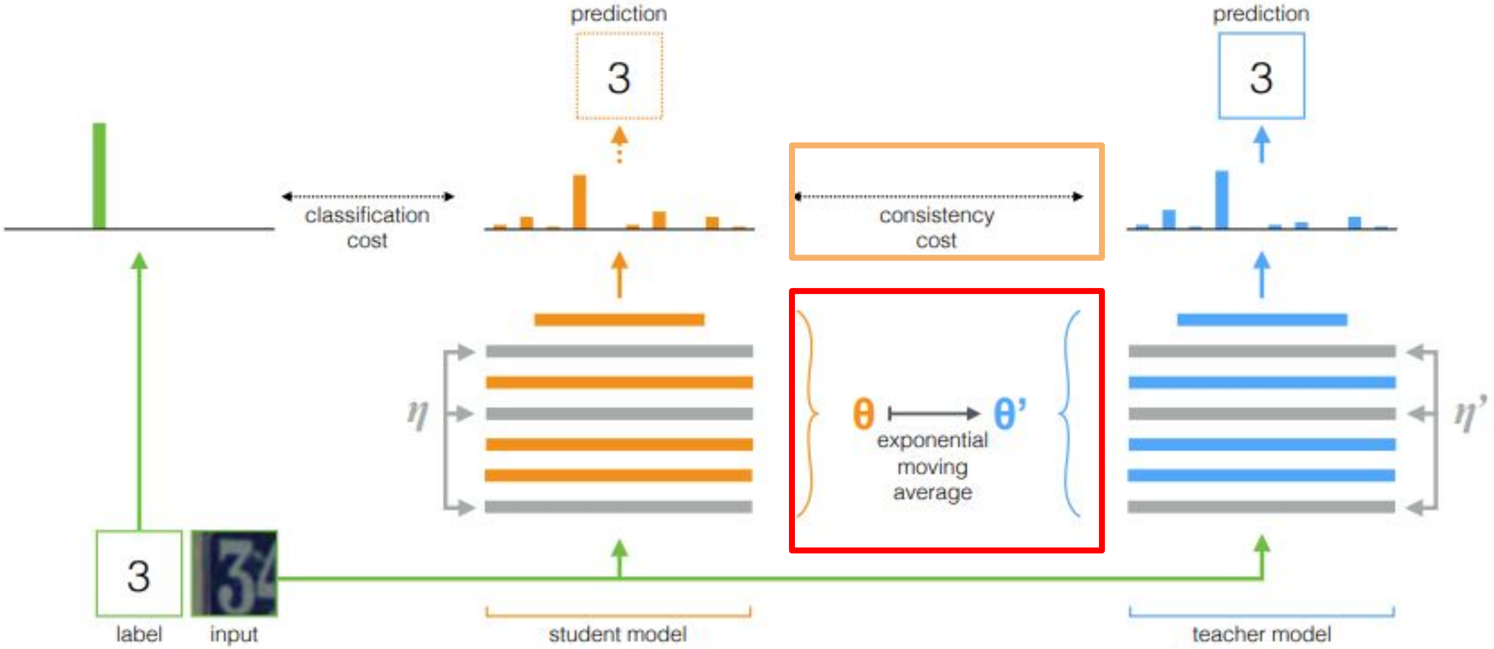
PubMed : Free search engine for life sciences and biomedical references and indexes (https://pubmed.ncbi.nlm.nih.gov/)

Dataset

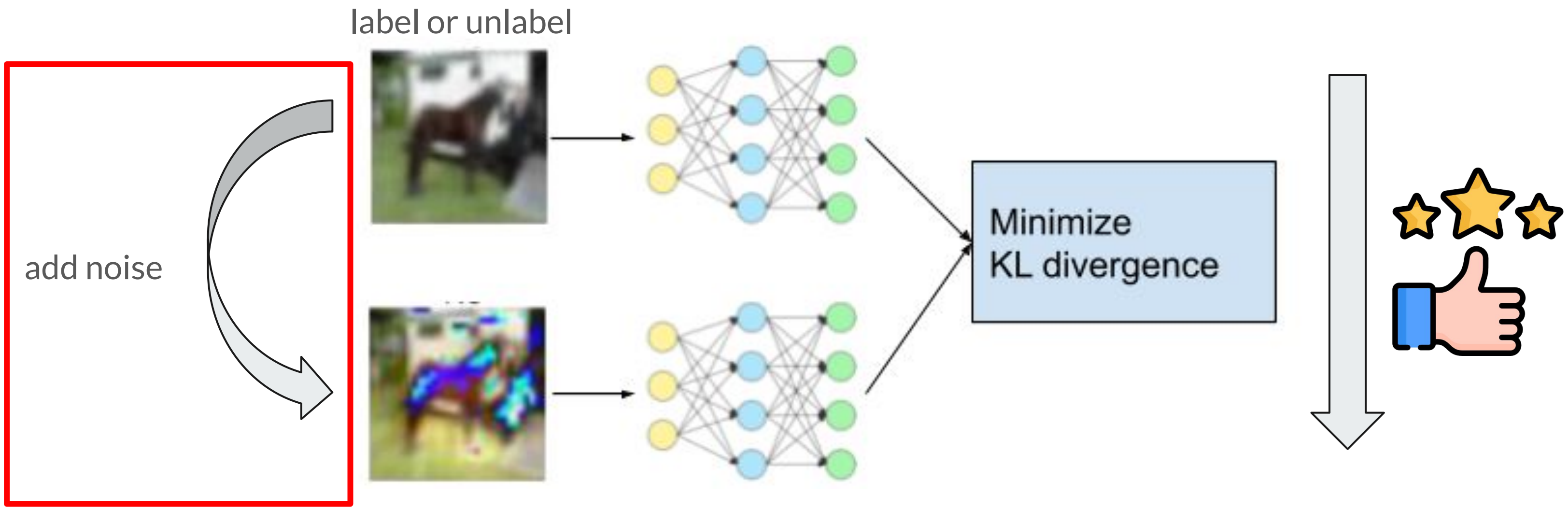
Dataset	Domain	Task	# Train / Test	# Class	Metric
Elec	Reviews	Sentiment Analysis	25K / 25K	2	Acc.
AG News	News	Topic Classification	120K / 7.6K	4	Acc.
NYT	News	Topic Classification	30K / 3.0K	9	Acc.
Chemprot	Chemical	Relation Classification	12K / 1.6K	10	F1

Dataset	Elec	AG News	NYT	Chemprot
description	Amazon shopping review	collection of news	New York Times	Contains PubMed abstracts containing chemical-protein interactions annotated by experts
category	positive、negative	World、Sports、Business、Sci/Tech	science、sports、music…	upregulator上調劑、 downregulator下調劑 agonist激動劑 antagonist拮抗劑

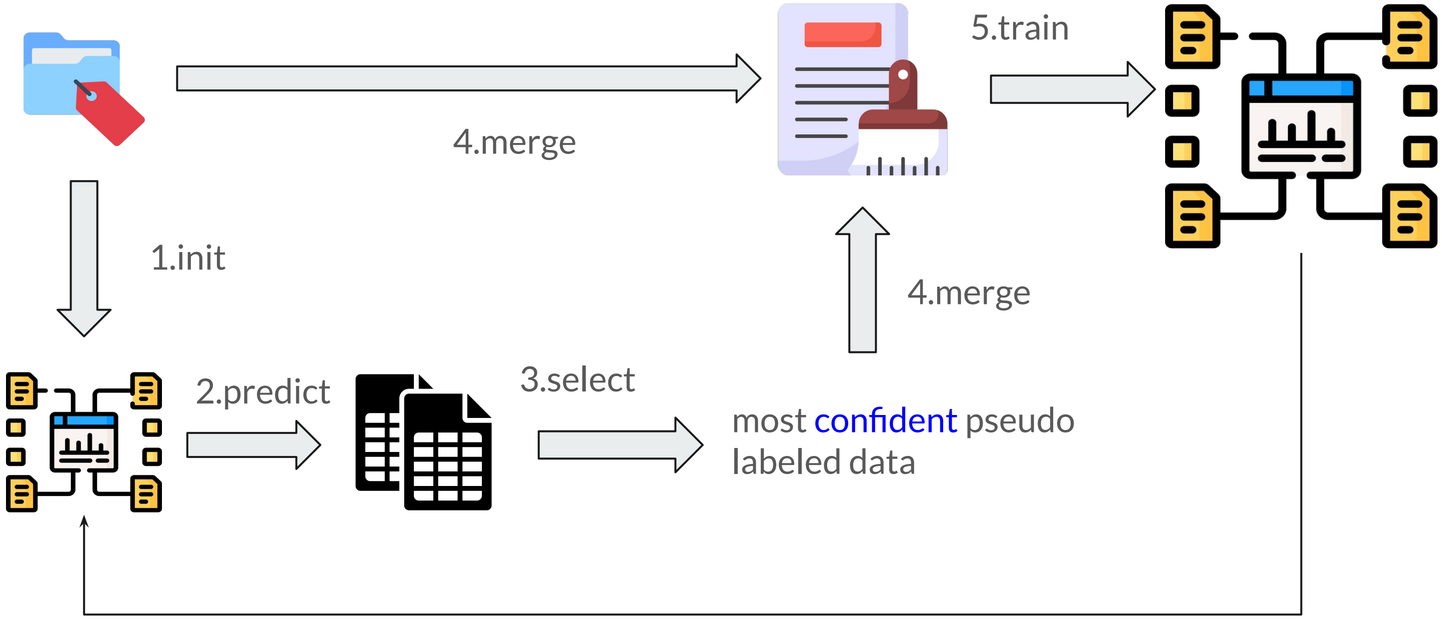
Baseline - Mean-Teacher



Baseline - Virtual Adversarial Training(VAT)

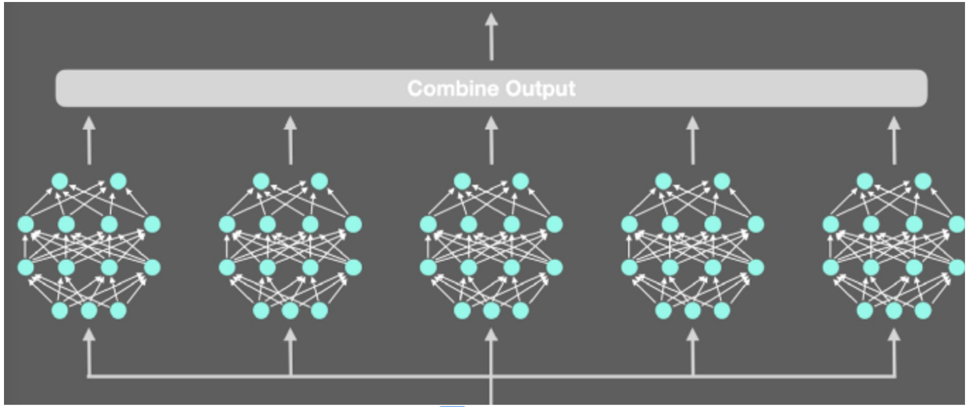


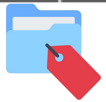
Baseline - Self-training(ST)



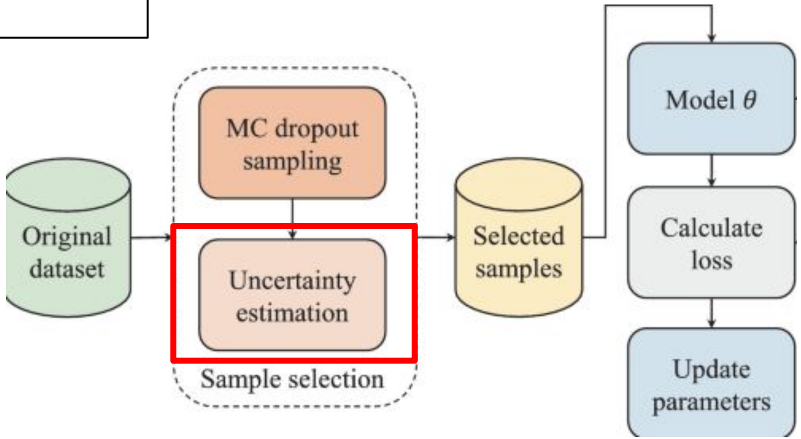
Baseline - Uncertainty-aware Self-training(UST)

MC dropout

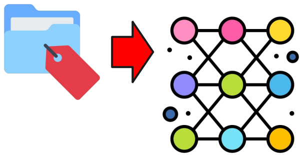
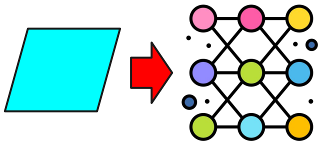




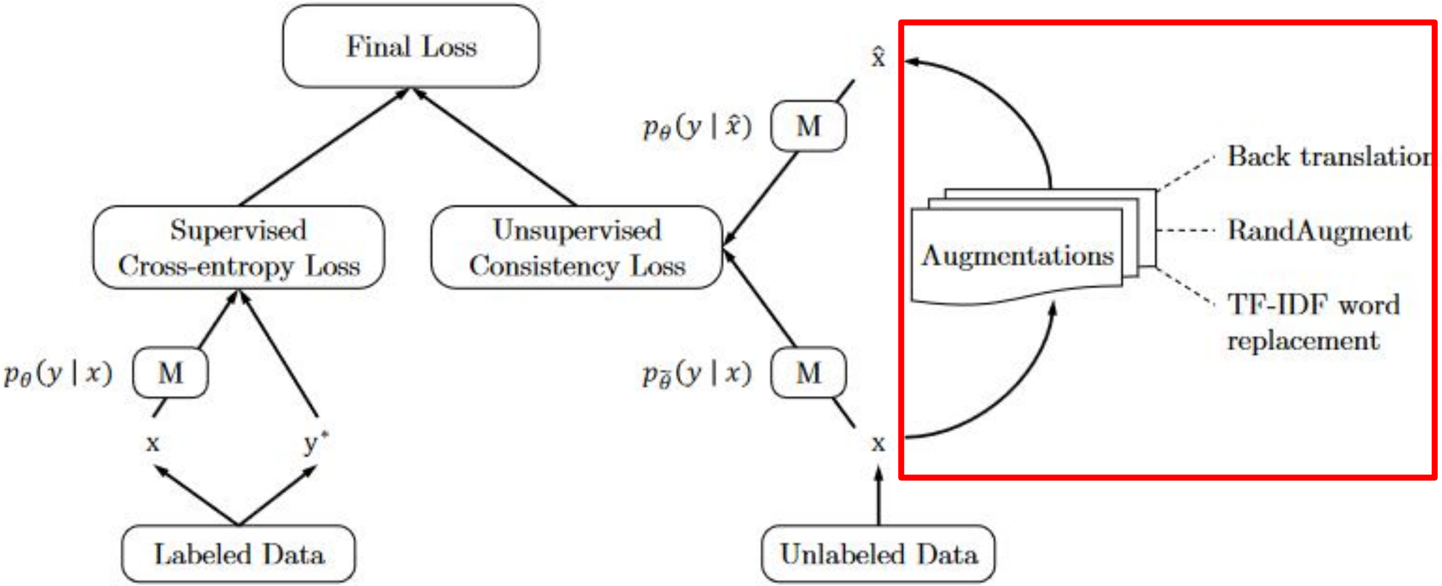
UST



Uncertainty estimation - Entropy

		$logP_\theta(y_i\|x)$	$P_\theta(y_i\|x)logP_\theta(y_i\|x)$	$x^*_E = \underset{x}{argmax} \; -\sum_i P_\theta(y_i\|x)logP_\theta(y_i\|x)$
	class A: 0.93 class B: 0.05 class C: 0.02	class A: -0.104 class B: -4.321 class C: -5.6438	class A: -0.09672 class B: -0.21605 class C: -0.11287	-(0.09672+0.21605+ 0.11287) = -0.4256 $\boxed{x^*_E = -(-0.4256) = 0.4256}$ entropy ⇩ uncertain ⇩
	class A: 0.55 class B: 0.35 class C: 0.1	class A: -0.8624 class B: -1.5145 class C: -3.3219	class A: -0.47432 class B: -0.53007 class C: -0.33219	-(0.47432+0.53007+ 0.33219) = -1.33658 $x^*_E = -(-1.33658) = 1.33658$

Baseline - Unsupervised Data Augmentation(UDA)

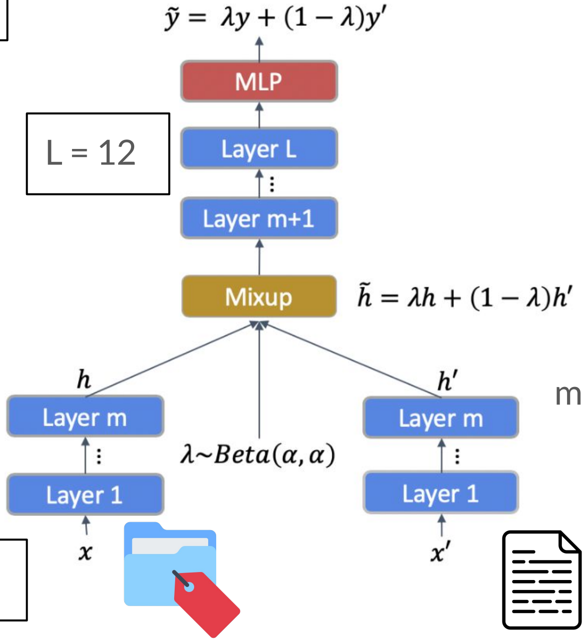


Baseline - MixText

Data Augmentation by interpolating

BERT Base

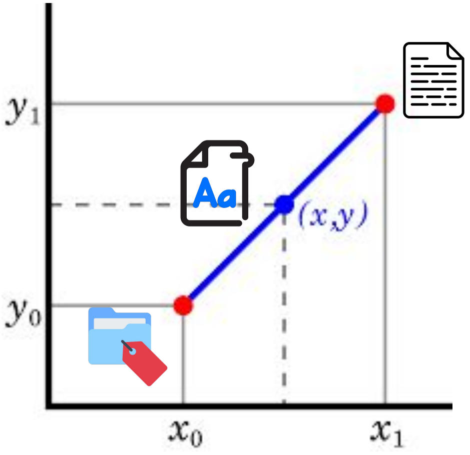
L = 12



m = 7

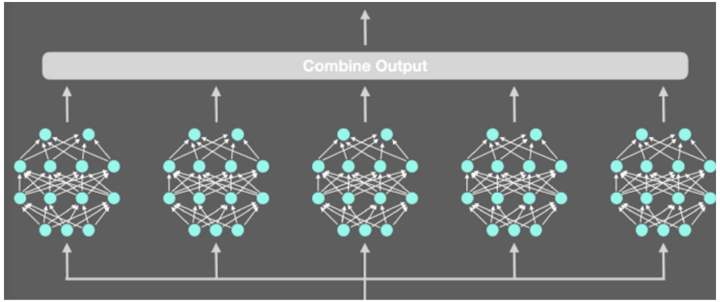
labeled

pseudo label

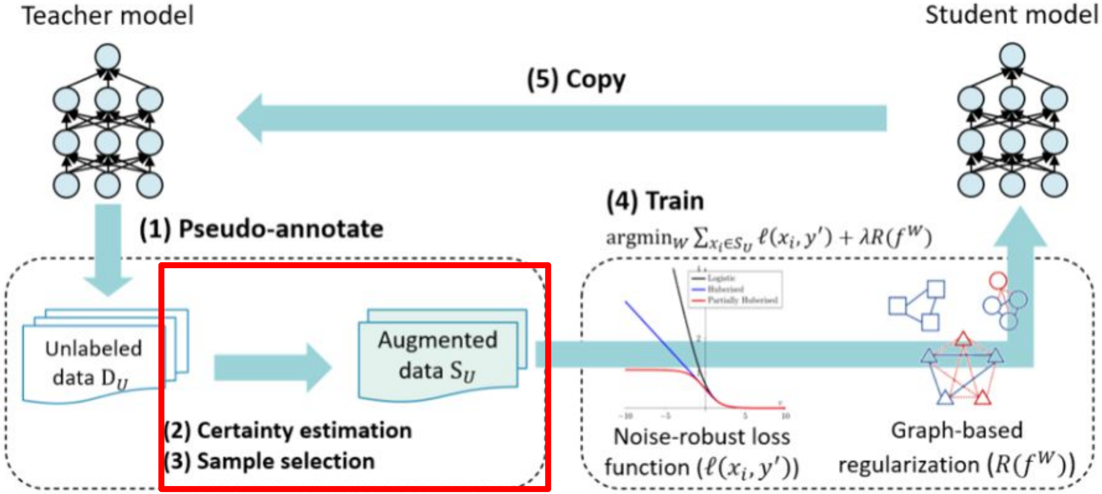


Baseline - Contrast-Enhanced Semi-supervised Text Classifcation(CEST)

MC dropout



CEST



Experiment

Method	AG News (Accuracy, ↑)		
	30	50	100
BERT	80.6±1.4		
MT	81.8±1.2		
VAT	82.1±1.2		
UDA	86.5±0.9		
MixText†	87.0±1.2		
NeST	**87.8±0.8**		
Superv.		93.0*	

all unlabels are marked as pseudo labels and used to train the model

Method	Name	Description
MT	Mean Teacher	average model weight
VAT	Virtual Adversarial Training	add noise with unlabel
UDA	Unsupervised Data Augmentation	data augmentation with unlabel
MixText	MixText	data augmentation + interpolating with unlabel
ST	self-training	use strategy to select unlabel
UST	Uncertainty-aware Self-training	MCdropout + uncertainty to select unlabel
CEST	Contrast-Enhanced Semi-supervised	MCdropout + certainty + Graph-based Contrast
Nest	Neighborhood-Regularized Self-Training	KNN + self-training

- use all unlabel
  - data augmentation methods are more effective than BERT , e.g. UDA、MixText

Experiment

Method	AG News (Accuracy, ↑)		
	30	50	100
UDA	86.5±0.9		
MixText†	87.0±1.2		
NeST	**87.8±0.8**		
Superv.		93.0*	

all unlabels are marked as pseudo labels and used to train the model

Method	Name	Description
MT	Mean Teacher	average model weight
VAT	Virtual Adversarial Training	add noise with unlabel
UDA	Unsupervised Data Augmentation	data augmentation with unlabel
MixText	MixText	data augmentation + interpolating with unlabel
ST	self-training	use strategy to select unlabel
UST	Uncertainty-aware Self-training	MCdropout + uncertainty to select unlabel
CEST	Contrast-Enhanced Semi-supervised	MCdropout + certainty + Graph-based Contrast
Nest	Neighborhood-Regularized Self-Training	KNN + self-training

- wrong pseudo label causes model confusion

Experiment

Method	AG News (Accuracy, ↑)		
	30	50	100
ST	86.0±1.4		
UST	86.9*		
CEST‡	86.5*		
NeST	**87.8±0.8**		
Superv.		93.0*	

use strategy to select unlabel

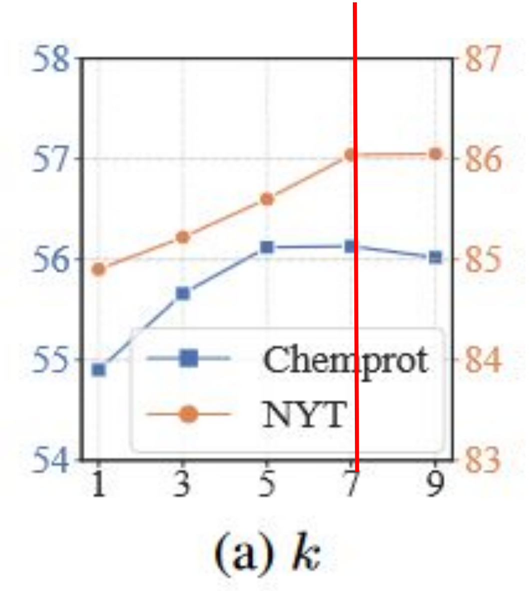
Method	Name	Description
MT	Mean Teacher	average model weight
VAT	Virtual Adversarial Training	add noise with unlabel
UDA	Unsupervised Data Augmentation	data augmentation with unlabel
MixText	MixText	data augmentation + interpolating with unlabel
ST	self-training	use strategy to select unlabel
UST	Uncertainty-aware Self-training	MCdropout + uncertainty to select unlabel
CEST	Contrast-Enhanced Semi-supervised	MCdropout + certainty + Graph-based Contrast
Nest	Neighborhood-Regularized Self-Training	KNN + self-training

- too much reliance on model predictions
- NeST is selected by aggregating the scores from the previous iteration

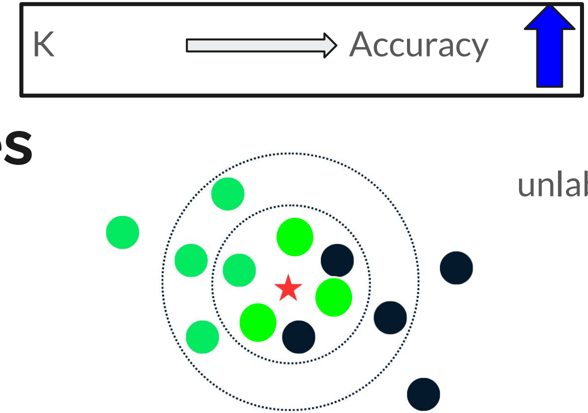
Experiment

Method	AG News (Accuracy, ↑)			Elec (Accuracy, ↑)			NYT (Accuracy, ↑)			Chemprot (F1, ↑)		
	30	50	100	30	50	100	30	50	100	30	50	100
BERT	80.6±1.4	83.1±1.6	86.0±1.1	85.0±1.9	87.2±1.0	90.2±1.2	79.4±1.6	83.0±1.1	85.7±0.5	49.1±2.3	51.2±1.7	54.9±1.4
MT	81.8±1.2	83.9±1.4	86.9±1.1	87.6±0.9	88.5±1.0	91.7±0.7	80.2±1.1	83.5±1.3	86.1±1.1	50.0±0.7	54.1±0.8	56.8±0.4
VAT	82.1±1.2	85.0±0.8	87.5±0.9	87.9±0.8	89.8±0.5	91.5±0.4	80.7±0.7	84.4±0.9	86.5±0.6	50.7±0.7	53.8±0.4	57.0±0.5
UDA	86.5±0.9	87.1±1.2	87.8±1.2	89.6±1.1	91.2±0.6	92.3±1.0	—	—	—	—	—	—
MixText†	87.0±1.2	87.7±0.9	88.2±1.0	91.0±0.9	91.8±0.4	92.4±0.5	—	—	—	—	—	—
ST	86.0±1.4	86.9±1.0	87.8±0.6	89.6±1.2	91.4±0.4	92.1±0.5	85.4±0.9	86.9±0.5	87.5±0.5	54.1±1.1	55.3±0.7	59.3±0.5
UST	86.9*	87.4*	87.9*	90.0*	91.6*	91.9*	85.0±0.6	86.7±0.4	87.1±0.3	53.5±1.3	55.7±0.4	59.5±0.7
CEST‡	86.5*	87.0*	88.4*	91.5*	92.1*	92.5*	—	—	—	—	—	—
NeST	**87.8±0.8**	**88.4±0.7**	**89.5±0.3**	**92.0±0.3**	**92.4±0.2**	**93.0±0.2**	**86.5±0.7**	**88.2±0.7**	**88.6±0.6**	**56.5±0.7**	**57.2±0.4**	**62.0±0.5**
Superv.	93.0*			95.3*			93.6±0.5			82.5±0.4		

Parameter Studies

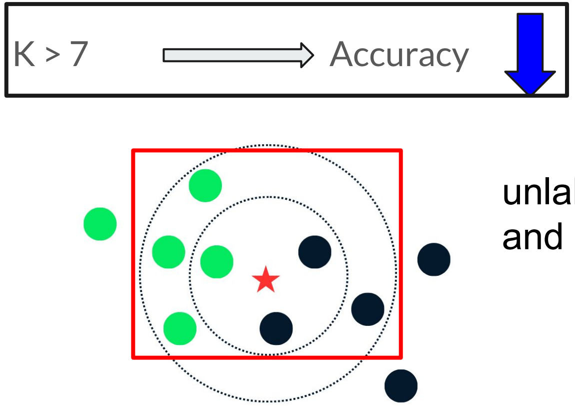


K = count of KNN



unlabel can find more labeled data

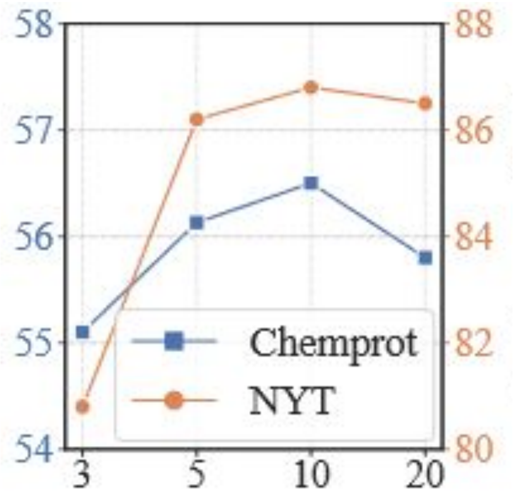
stable



unlabel is too far away from label and has high divergence.

unstable

Parameter Studies

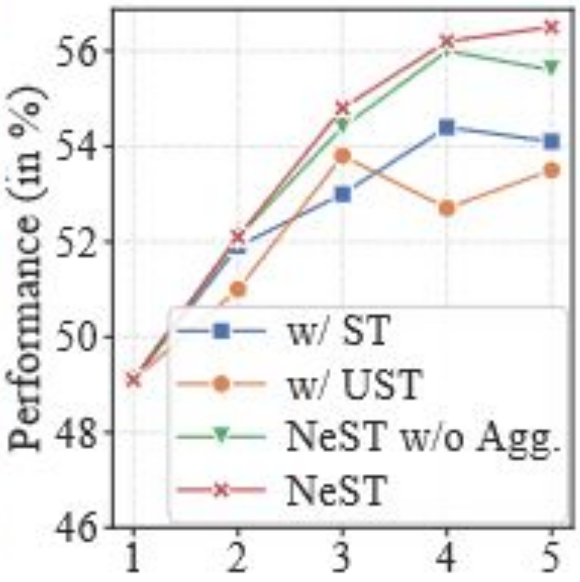


(b) c

- if c = 3 , labeled data = 120, b=c|xi| = 3 * 120 = 360
    - the number of pseudo labels is not enough
    - accuracy is not high
- if c = 20 , labeled data = 400, b=c|xi| = 20 * 400 = 8000
    - pseudo data selected is too messy and poor quality
    - disrupt model learning

c : multiple of how many samples to select in an epoch

Ablation Studies



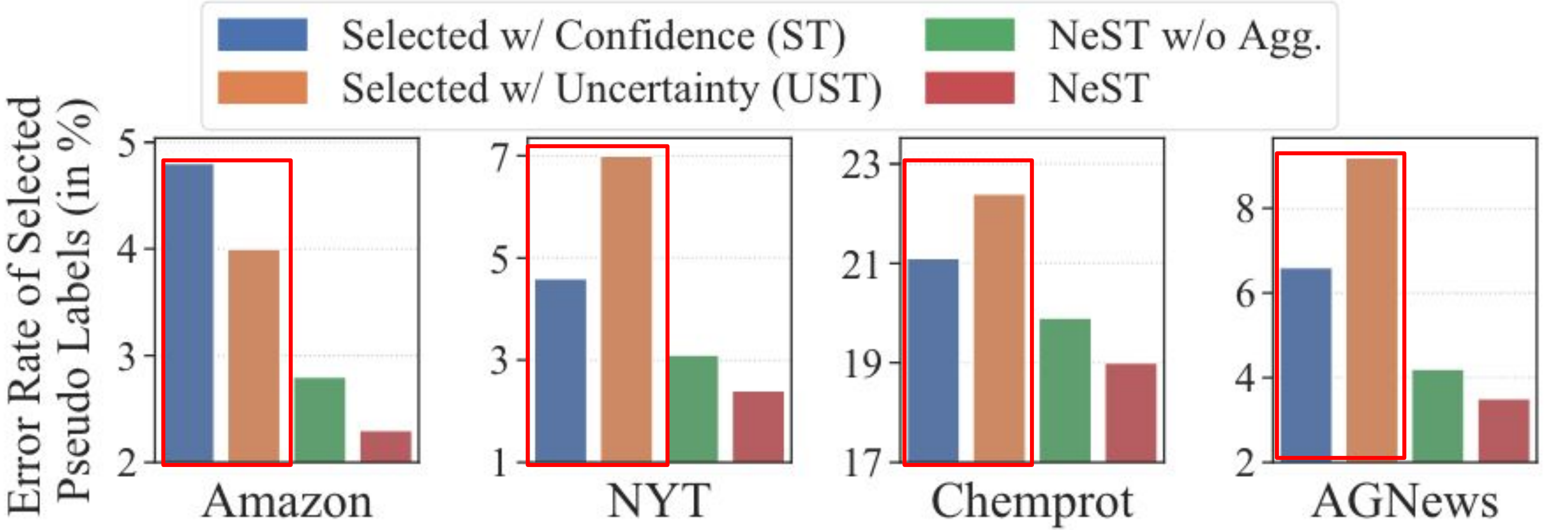
(c) Acc. over iters.

epochs

- in the early stage
  - selected pseudo labels that can help the model learn.

ST	self-training	use strategy to select unlabel
UST	Uncertainty-aware Self-training	MCdropout + uncertainty to select unlabel
Nest	Neighborhood-Regularized Self-Training	KNN + self-training

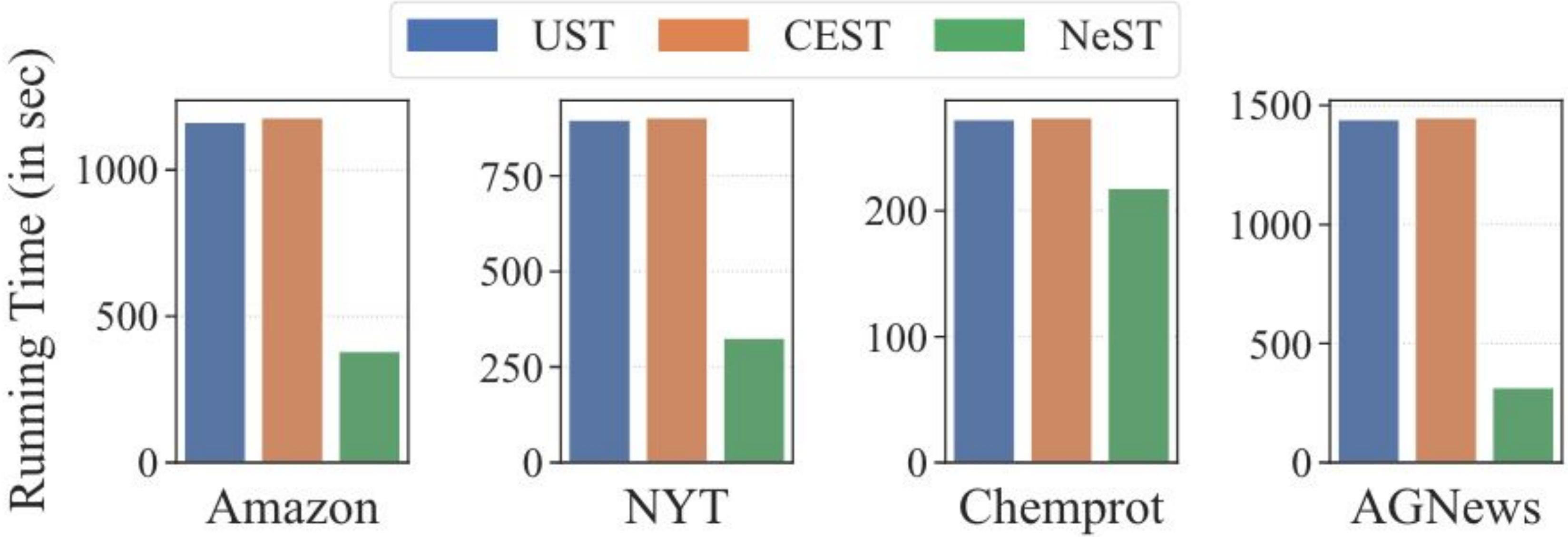
Error of Pseudo Labels



https://github.com/facebookresearch/faiss

UST	Uncertainty-aware Self-training	MCdropout + uncertainty to select unlabel
CEST	Contrast-Enhanced Semi-supervised	MCdropout + certainty + Graph-based Contrast
Nest	Neighborhood-Regularized Self-Training	KNN + self-training

Running time of different methods



Outline

- Introduction
- Method
- Experiment
- Conclusion

Conclusion

- propose NeST to improve sample selection in self-training for robust label efficient learning
- design a neighborhood-regularized approach to select more reliable samples based on representations for self-training
- propose to aggregate the predictions on different iterations to stabilize self-training