D-CALM: A $\boxed{\text{Dynamic}}$ approach $\boxed{\text{Clustering-based Active Learning}}$ Approach for $\boxed{\text{Mitigating Bias}}$

task

Advisor : Jia-Ling, Koh

Speaker : Ting-I, Weng

Source : ACL'23

Date : 2023/08/14

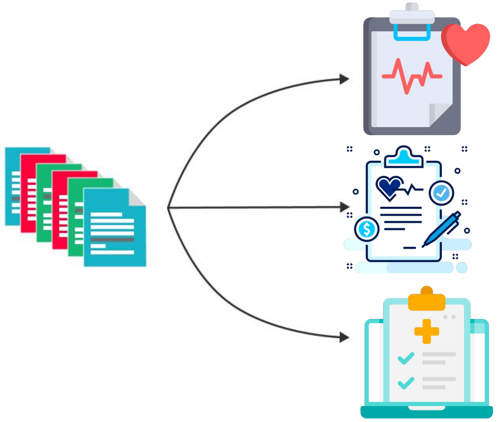
Outline

- Introduction
- Method
- Experiment
- Conclusion

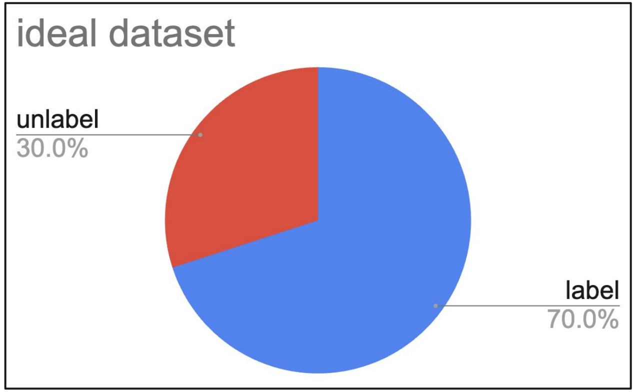
- Background
- Active Learning
- Challenge
  - bias
    - bias induction
  - effective batch selection
- D-CALM

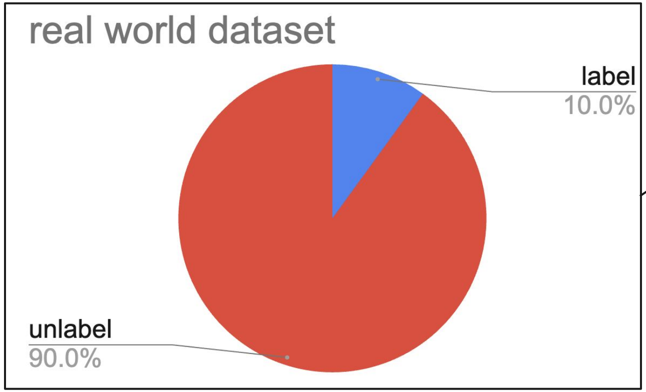
Background

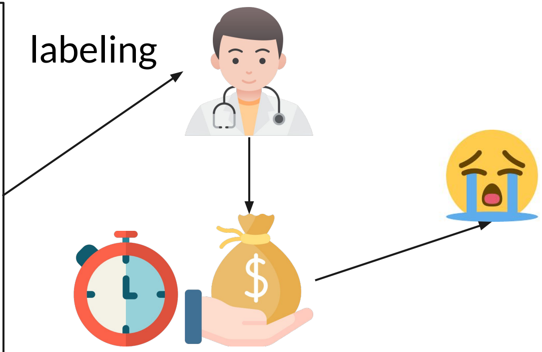
classification task



e.g. Medical text classification

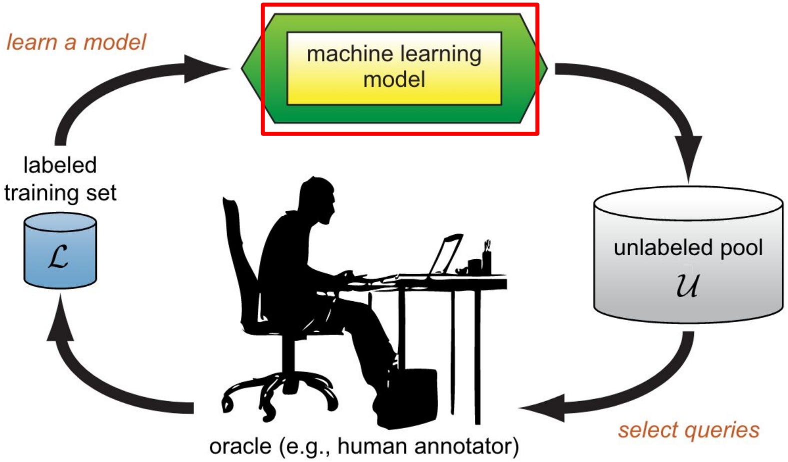






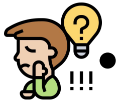
Active Learning

Active Learning : actively select the most valuable samples for labeling



- Goal:
  - reduce labeling costs
  - find more diverse data

- Weakness:
  - rely on the judgment of the classifier



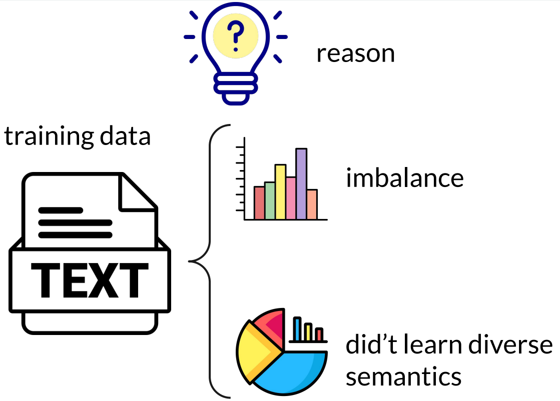
- Issue:
  - whether the data found is helpful for model training

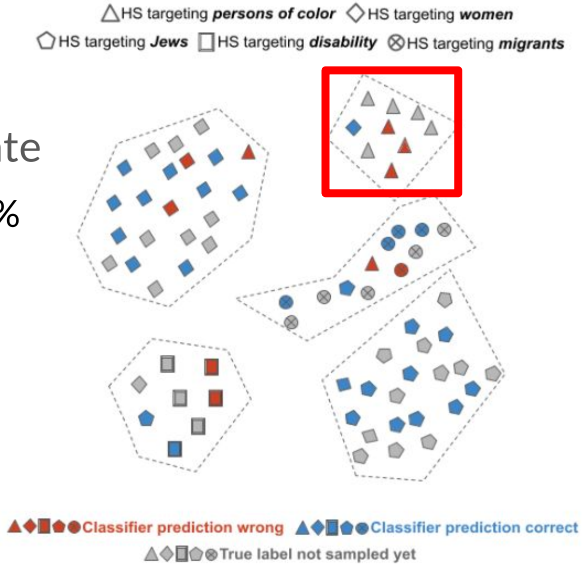
Challenge - bias

- classifiers may not perform well for underrepresented classes in the data
  - category has less data  = 5
  - people of color have a high error rate
     = 100%

- Hope to find more unlabeled data about people of color using active learning

 △ unlabel persons of color



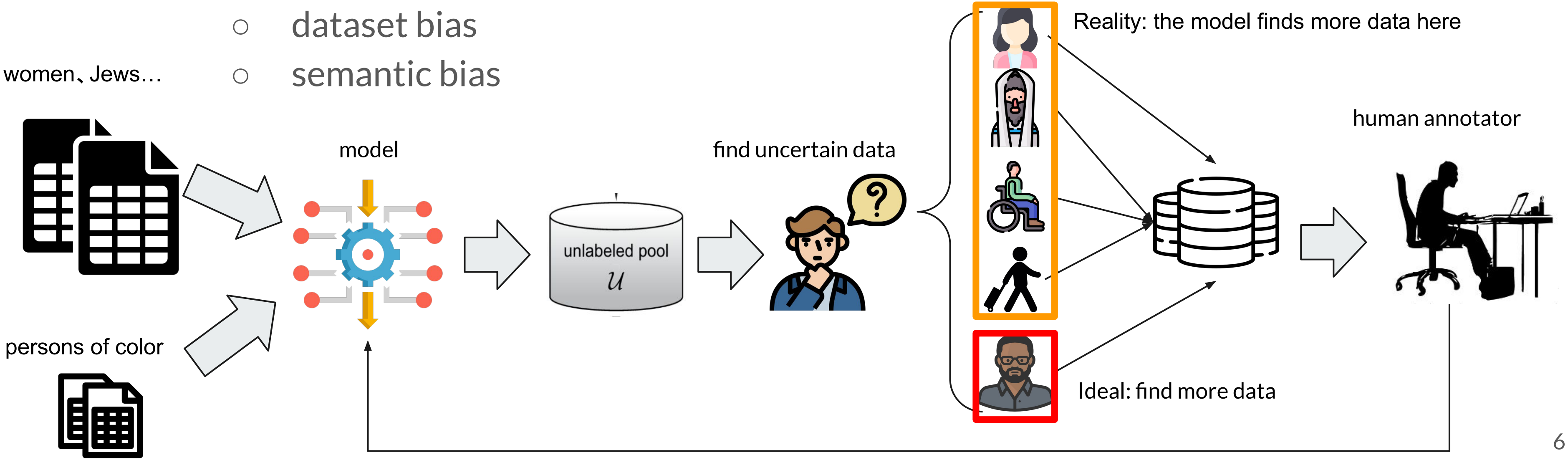


reason

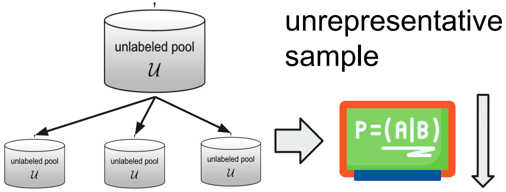
- rely on the judgment of the classifier
- high probability of selecting unrepresentative samples

Challenge - bias induction

- bias induction：active learning can make bias worse
  - dataset bias
  - semantic bias

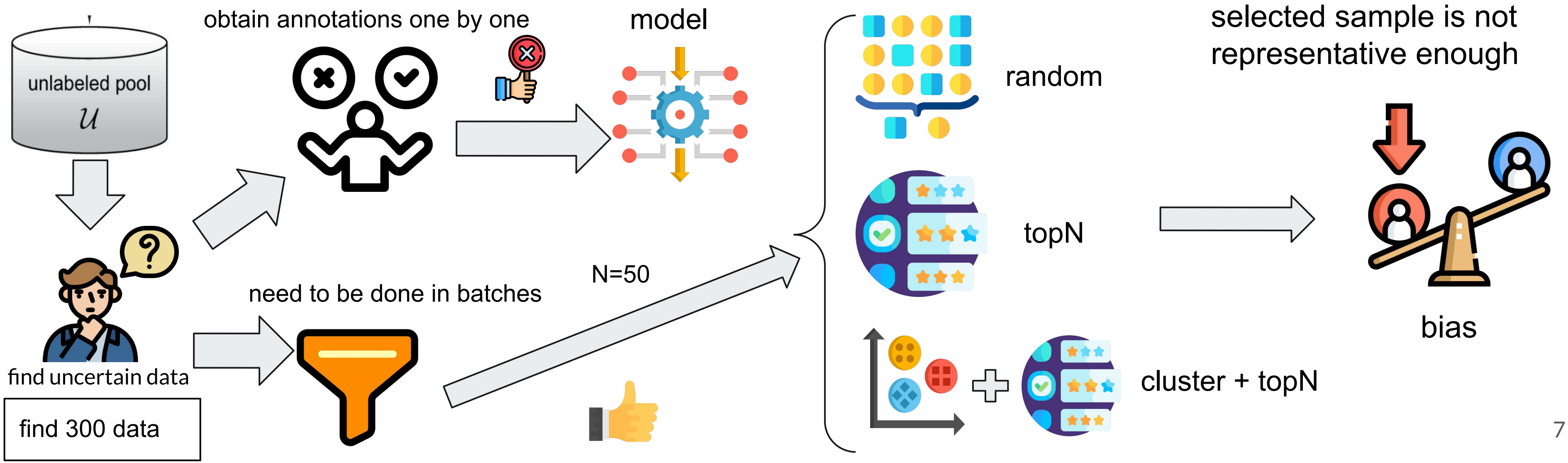


goal :

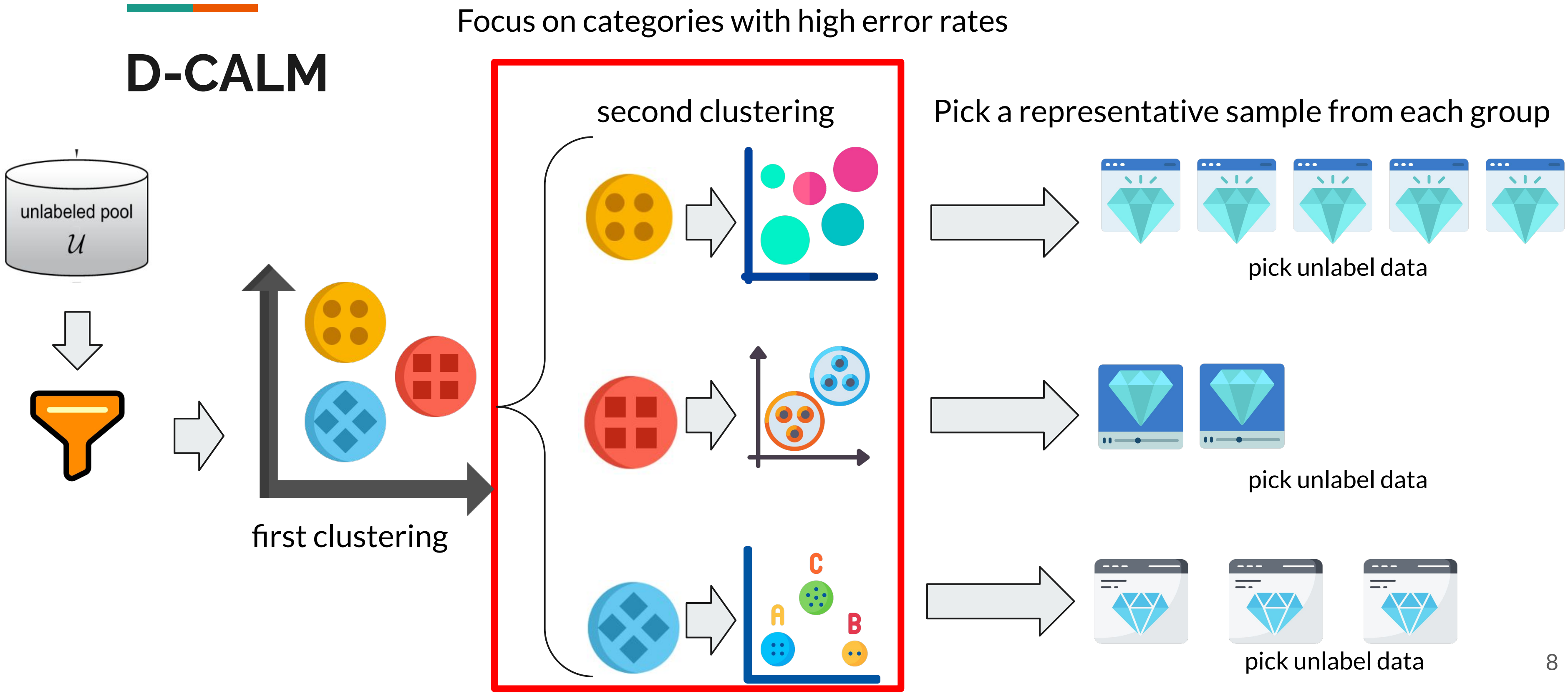


Challenge - effective batch selection

when the batch size is larger, the bias becomes greater



D-CALM



Focus on categories with high error rates
unlabeled pool
$\mathcal{U}$
first clustering
second clustering
A
B
C
Pick a representative sample from each group
pick unlabel data
pick unlabel data
pick unlabel data

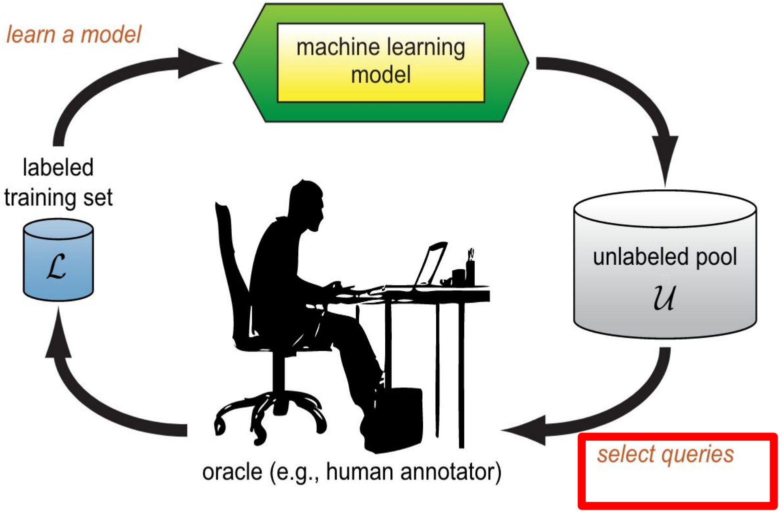
Outline

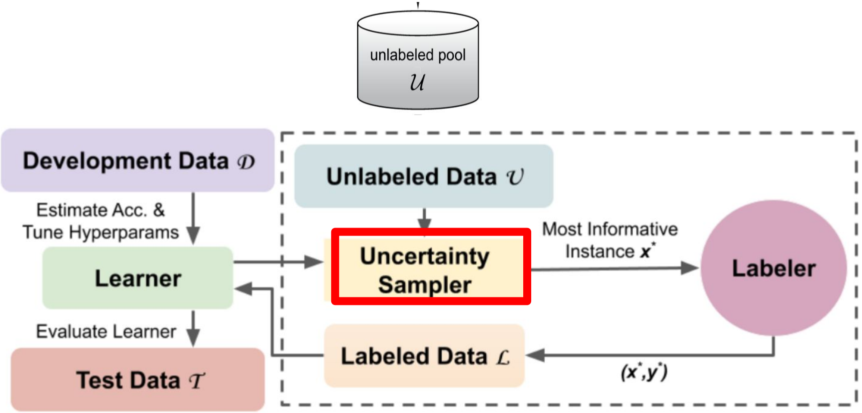
- Introduction
- Method
- Experiment
- Conclusion

- Active learning - Query Strategy
- Query-Strategy
  - Least Confident
  - Smallest margin
  - Entropy
- Active Learning v.s. Clustering-based Active Learning
- Dynamic Cluster AL v.s. Clustering-based AL

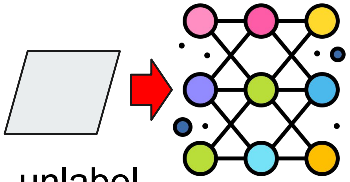
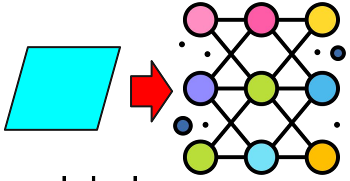
Active learning - Query Strategy

- how to pick？
- How many ways?

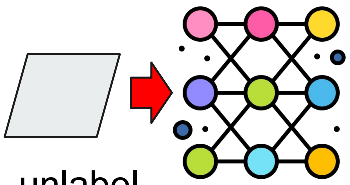
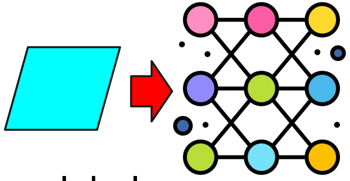




Query-Strategy - Least Confident

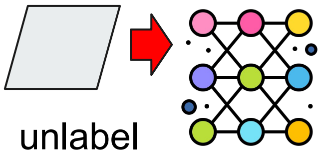
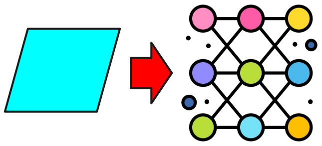
	$\hat{y} = argmax_y P_\theta(y\|x)$	$x^*_{LC} = \underset{x}{argmax}\ 1 - P_\theta(\hat{y}\|x)$
 unlabel — class A: 0.93 class B: 0.05 class C: 0.02	$\hat{y} = 0.93$	$x^*_{LC} = 1 - 0.93 = 0.07$
 unlabel — class A: 0.55 class B: 0.35 class C: 0.1	$\hat{y} = 0.55$ Confident ⇩ uncertain ⇧	$\boxed{x^*_{LC} = 1 - 0.55 = 0.45}$

Query-Strategy - Smallest margin

		$\hat{y} = argmax_y P_\theta(y\|x)$	$x^*_{MS} = \underset{x}{argmin}\ P_\theta(\hat{y}_1\|x) - P_\theta(\hat{y}_2\|x)$
 unlabel	class A: 0.93 class B: 0.05 class C: 0.02	$\hat{y_1} = 0.93$ $\hat{y_2} = 0.05$	$x^*_{MS} = 0.93 - 0.05 = 0.88$
 unlabel	class A: 0.55 class B: 0.35 class C: 0.1	$\hat{y_1} = 0.55$ $\hat{y_2} = 0.35$	$\boxed{x^*_{MS} = 0.55 - 0.35 = 0.2}$

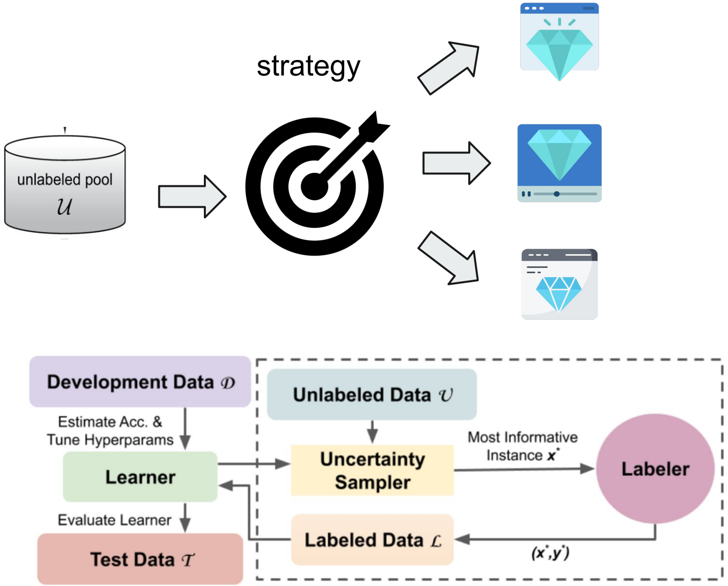
smallest margin ⇩ uncertain ⇧

Query-Strategy - Entropy

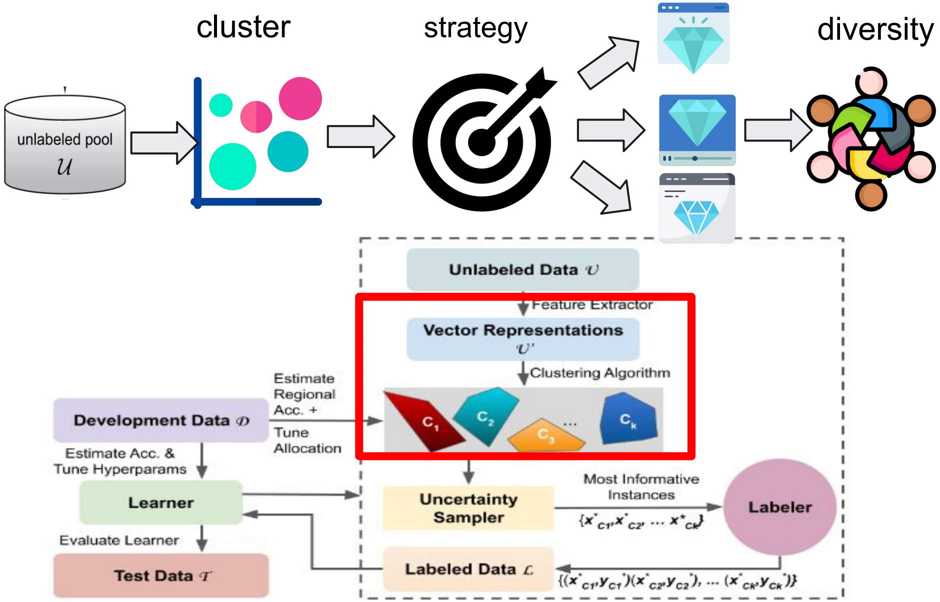
	$logP_\theta(y_i\|x)$	$P_\theta(y_i\|x)logP_\theta(y_i\|x)$	$x_E^* = \underset{x}{argmax} \; -\sum_i P_\theta(y_i\|x)logP_\theta(y_i\|x)$
 unlabel class A: 0.93 class B: 0.05 class C: 0.02	class A: -0.104 class B: -4.321 class C: -5.6438	class A: -0.09672 class B: -0.21605 class C: -0.11287	-(0.09672+0.21605+ 0.11287) = -0.4256 $x_E^* = -(-0.4256) = 0.4256$
 unlabel 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 $\boxed{x_E^* = -(-1.33658) = 1.33658}$

entropy ⇧ uncertain ⇧

Active Learning v.s. Clustering-based Active Learning

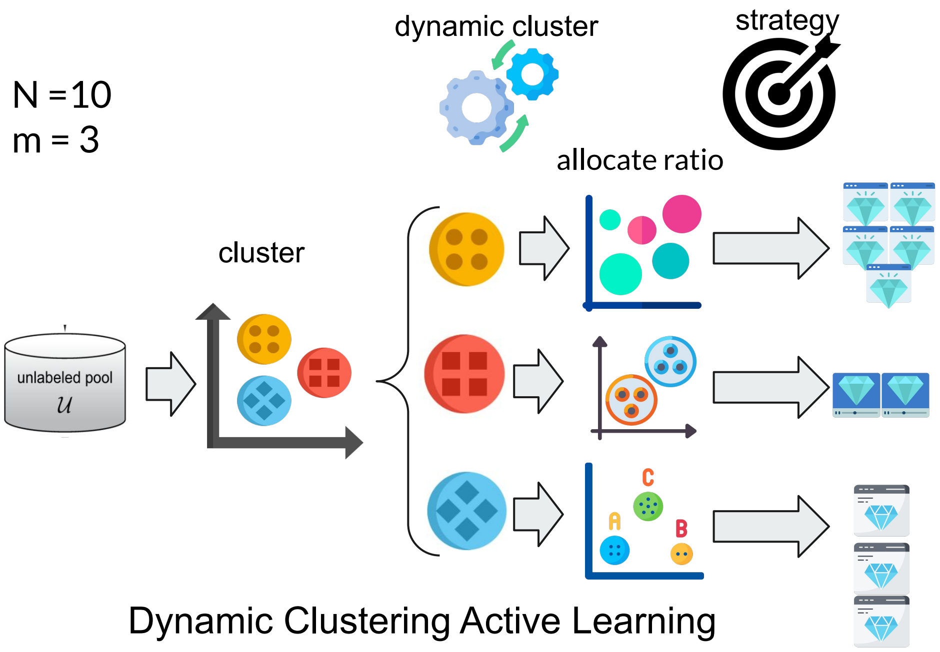


Active Learning

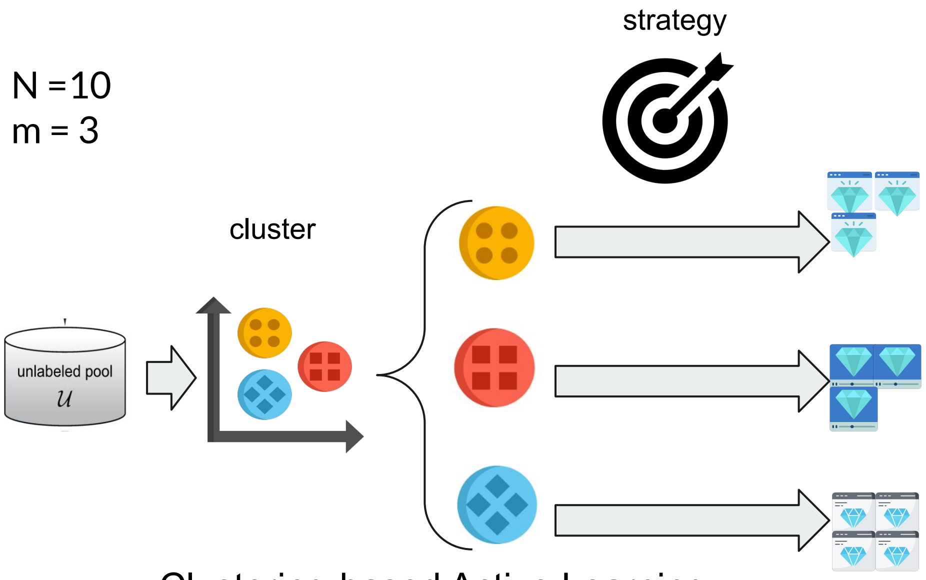


Clustering-based Active Learning

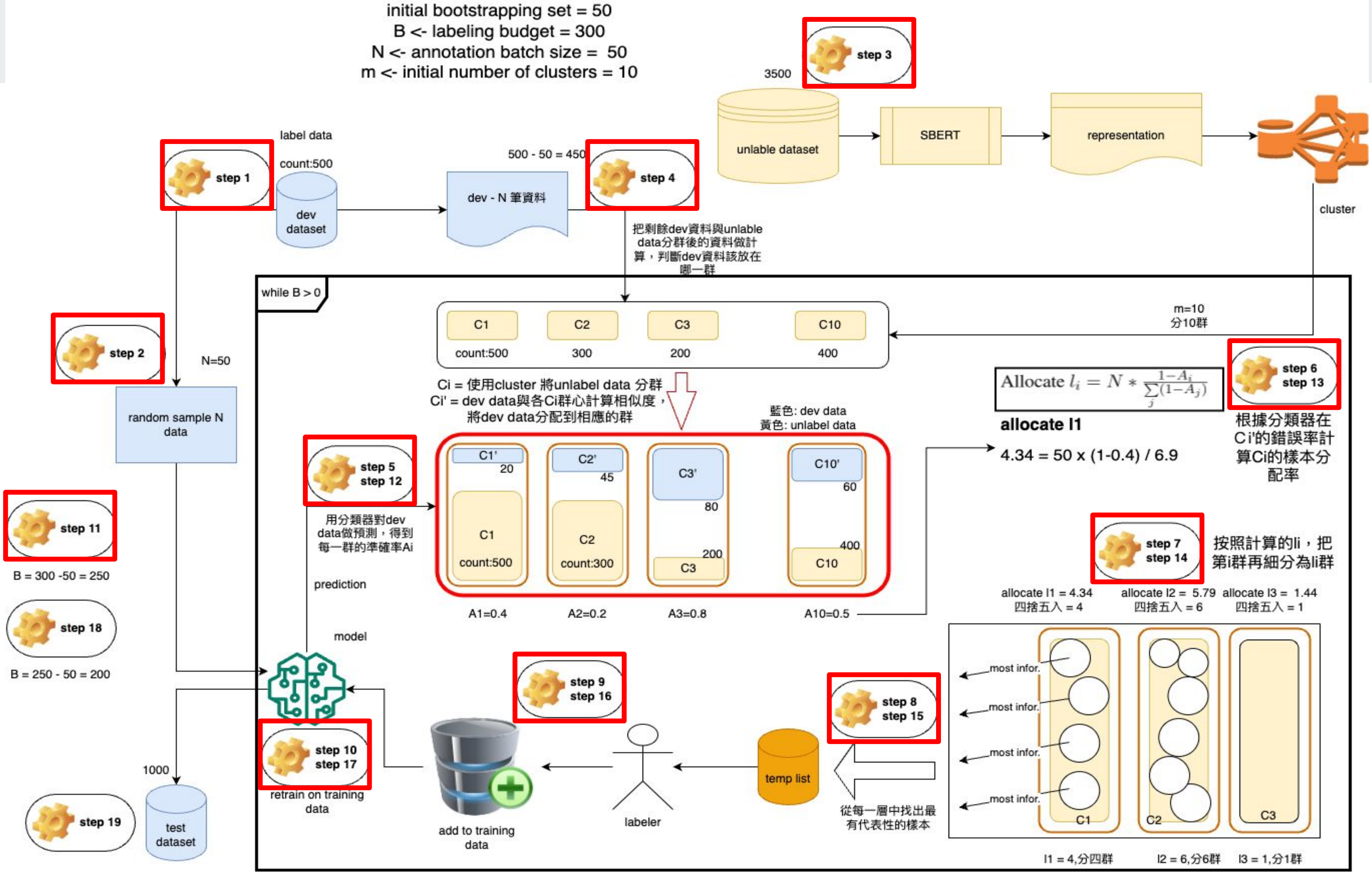
Dynamic Cluster AL v.s. Clustering-based AL



Dynamic Clustering Active Learning



Clustering-based Active Learning


initial bootstrapping set = 50
B <- labeling budget = 300
N <- annotation batch size = 50
m <- initial number of clusters = 10
step 3
3500
unlable dataset
SBERT
representation
cluster
step 1
label data
count:500
dev dataset
dev - N 筆資料
500 - 50 = 450
step 4
把剩餘dev資料與unlable data分群後的資料做計算，判斷dev資料該放在哪一群
while B > 0
step 2
N=50
random sample N data
m=10
分10群
C1
count:500
C2
300
C3
200
C10
400
Ci = 使用cluster 將unlabel data 分群
Ci' = dev data與各Ci群心計算相似度，將dev data分配到相應的群
藍色: dev data
黃色: unlabel data
Allocate $l_i = N * \frac{1-A_i}{\sum_j (1-A_j)}$
allocate l1
4.34 = 50 x (1-0.4) / 6.9
step 6
step 13
根據分類器在Ci'的錯誤率計算Ci的樣本分配率
step 5
step 12
用分類器對dev data做預測，得到每一群的準確率Ai
C1'
20
C1
count:500
C2'
45
C2
count:300
C3'
80
200
C3
C10'
60
400
C10
step 11
B = 300 -50 = 250
step 18
B = 250 - 50 = 200
prediction
A1=0.4
A2=0.2
A3=0.8
A10=0.5
step 7
step 14
按照計算的li，把第i群再細分為li群
allocate l1 = 4.34
四捨五入 = 4
allocate l2 = 5.79
四捨五入 = 6
allocate l3 = 1.44
四捨五入 = 1
model
step 9
step 16
step 8
step 15
most infor.
most infor.
most infor.
most infor.
C1
C2
C3
step 10
step 17
retrain on training data
add to training data
labeler
temp list
從每一層中找出最有代表性的樣本
1000
step 19
test dataset
l1 = 4,分四群
l2 = 6,分6群
l3 = 1,分1群

Outline

- Introduction
- Method
- Experiment
  - Dataset
  - D-CALM v.s. Baseline with BERT
  - D-CALM v.s. Baseline with SVM
- Conclusion

Dataset

- book title
- hatespeech
- emotion detection
- acceptable sentence
- tweets for hatespeech
- dialog act
- question
- snippets from IMDB reviews

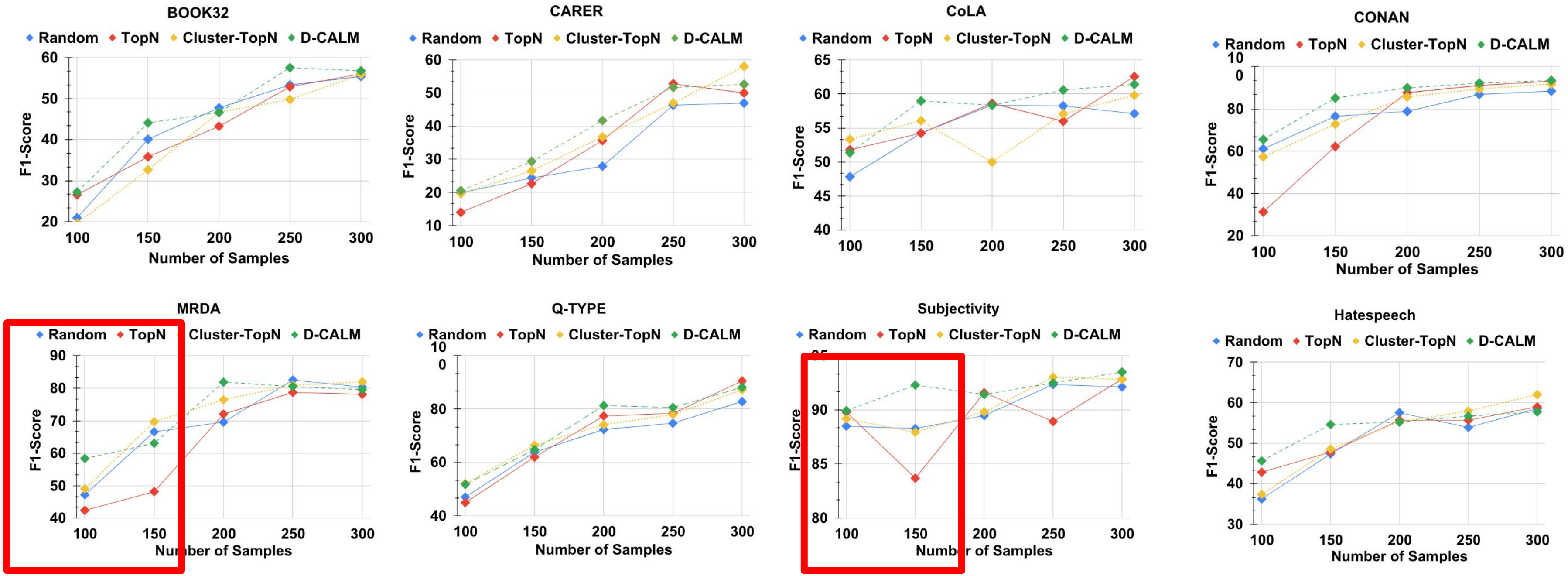
count : 70% : 10% : 20%

Dataset	classes	Pool	Dev	Test
BOOK32	32	14K	2K	4K
CONAN	8	3.5K	0.5K	1K
CARER	6	16K	2K	4K
CoLA	2	8.5K	0.5K	0.5K
Hatespeech	3	17.2K	2.4K	4.9K
MRDA	5	14K	2K	4K
Q-Type	6	4.9K	0.5K	0.5K
Subjectivity	2	7K	1K	2K

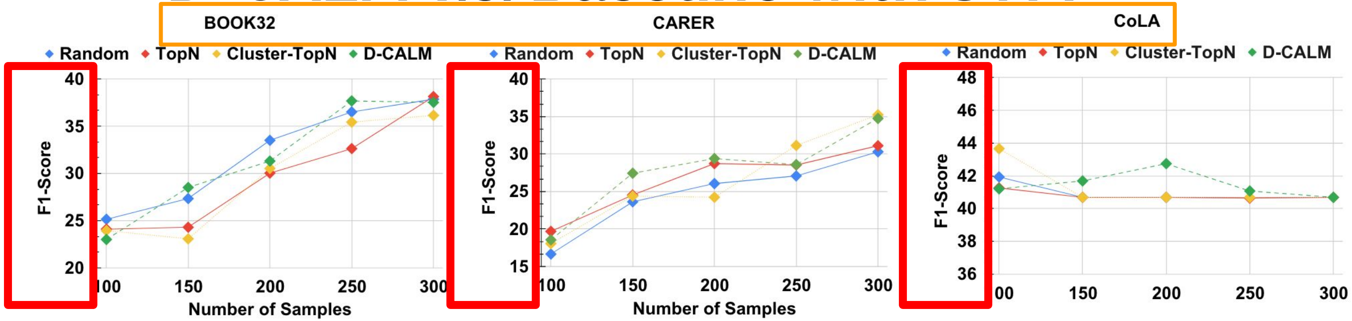
unlabel : label : unlabel

- metric : F1-score
- Strategy : entropy

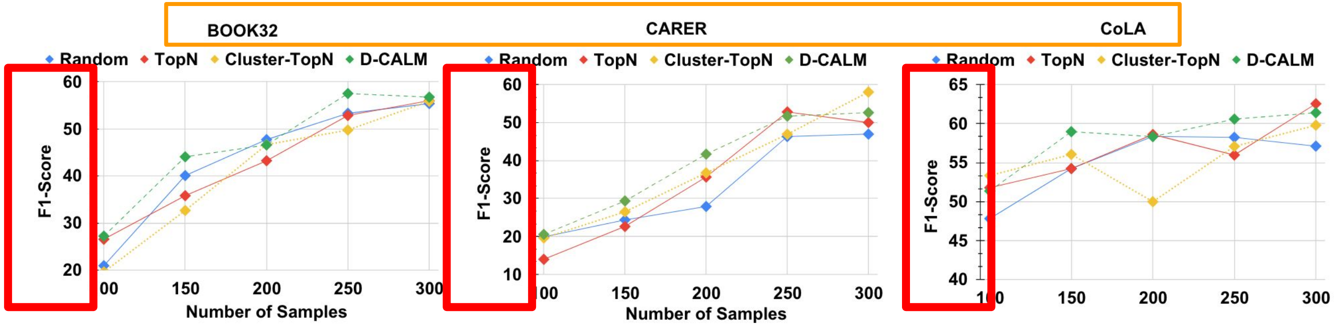
D-CALM v.s. Baseline with BERT



D-CALM v.s. Baseline with SVM



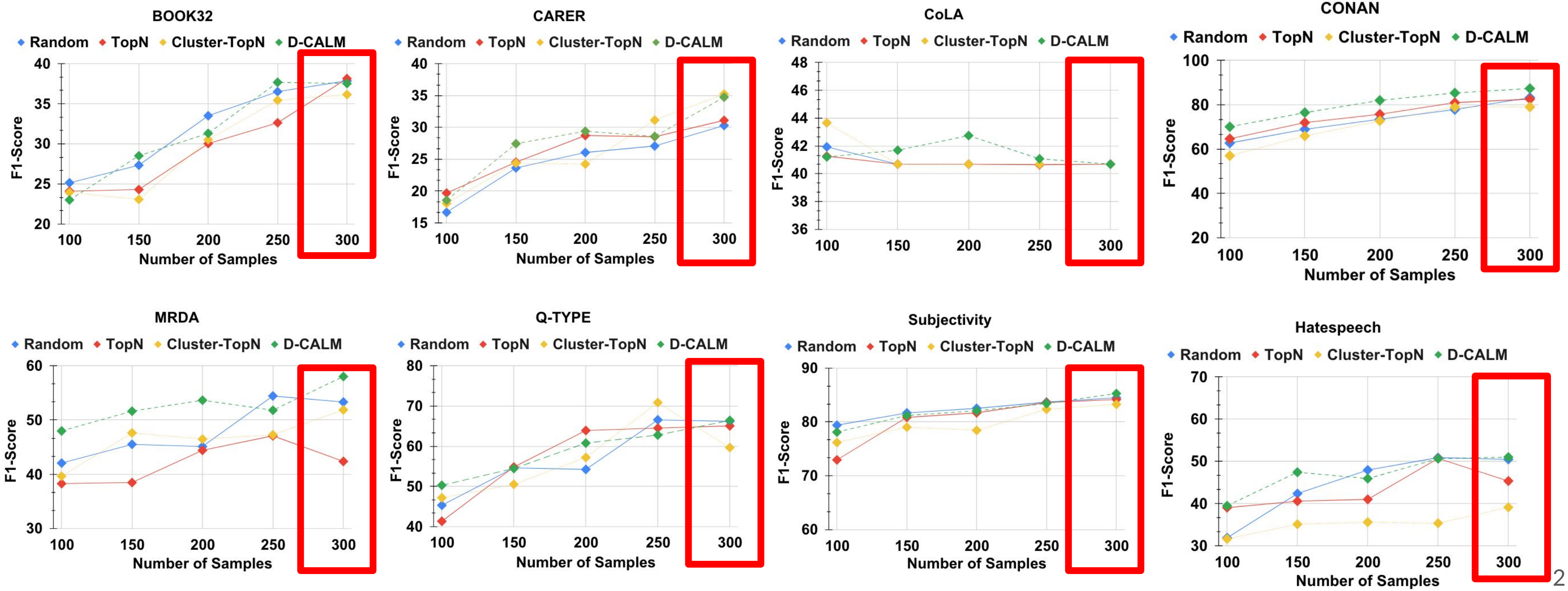
learner : SVM



learner: BERT

- Strategy : entropy

D-CALM v.s. Baseline with SVM



Outline

- Introduction
- Method
- Experiment
- Conclusion

Conclusion

1. The model trained by DCALM can second clustering the unlabeled data through the error rate, reduce the bias against underrepresented groups in the unlabeled data.

2. DCALM improvements are model-independent