Ko-Active Learning

Jaeyoeng Lee, Mose Park

Department of Statistical Data Science
University of Seoul

June 20, 2024

Index

1

Introduction

Main concepts

Imbalanced data

Active Learning

KcBERT

Our Task

KoBERT	SKT	위키피디아 50M	Sentence-Piece	8,002	92M

* review text, Yanolja

* Active Learning

* Assumption : Don't know

▶ KoBERT → KcBERT, Because it was trained on online news comments.

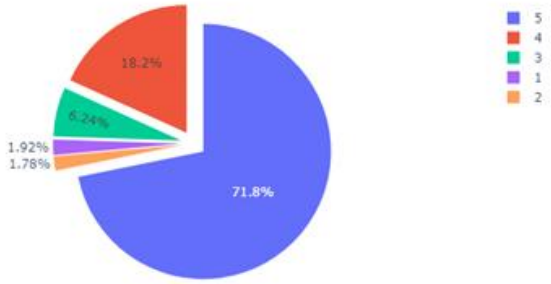
▶ User review texts are similar to online news comments.

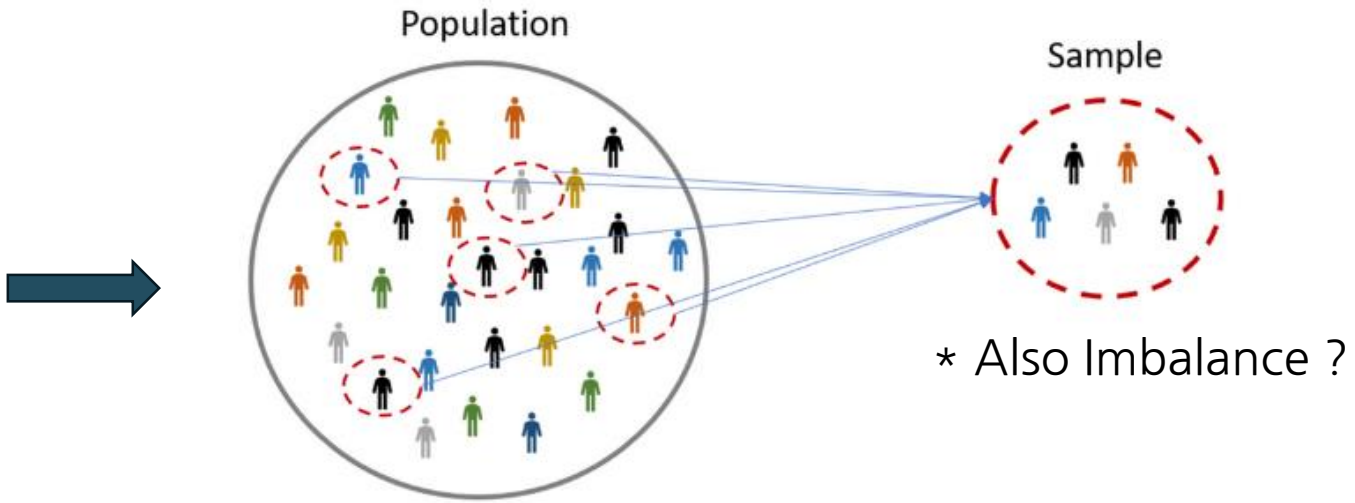
2

Problem

Problem

* rating ratio





* Also Imbalance ?

* Imbalance

- Label imbalance
- There are many high scores and few low scores.

3

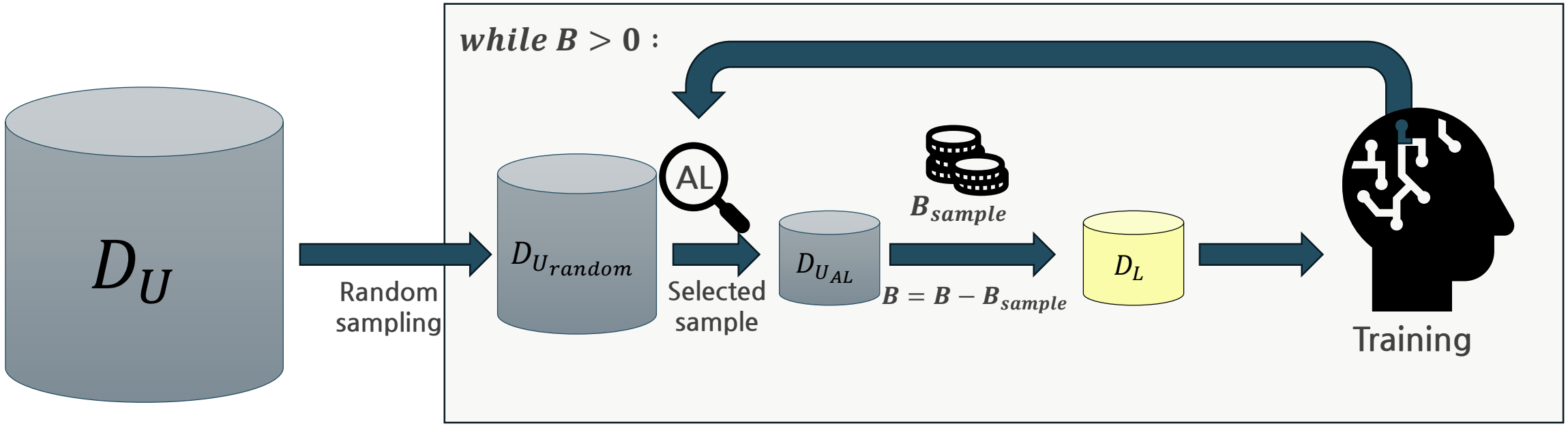
Methodology

Benchmark Paper

Class-Balanced Active Learning for Image Classification

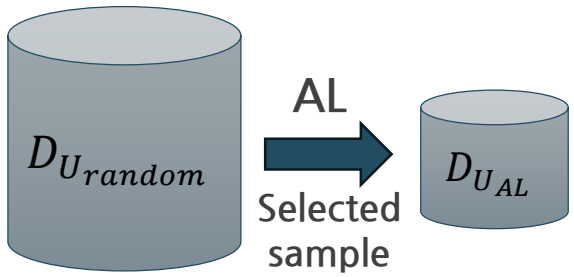
Javad Zolfaghari Bengar[1,2] Joost van de Weijer[1,2] Laura Lopez Fuentes[1]
Bogdan Raducanu[1,2]
Computer Vision Center (CVC)[1], Univ. Autònoma of Barcelona (UAB)[2]
{jzolfaghari,joost,llopez,bogdan}@cvc.uab.es

Active Learning

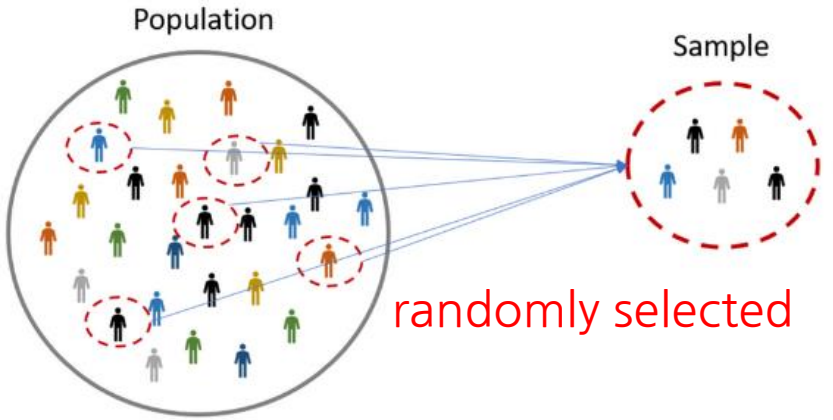


- Label imbalance
- There are many high scores and few low scores.

Sampling method - Random

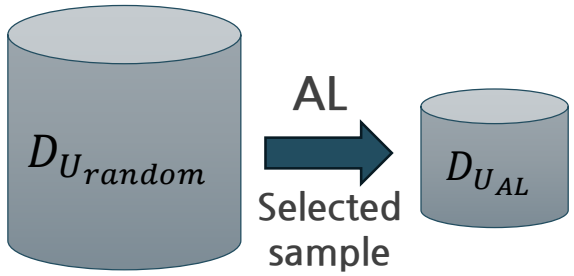


❖ Random

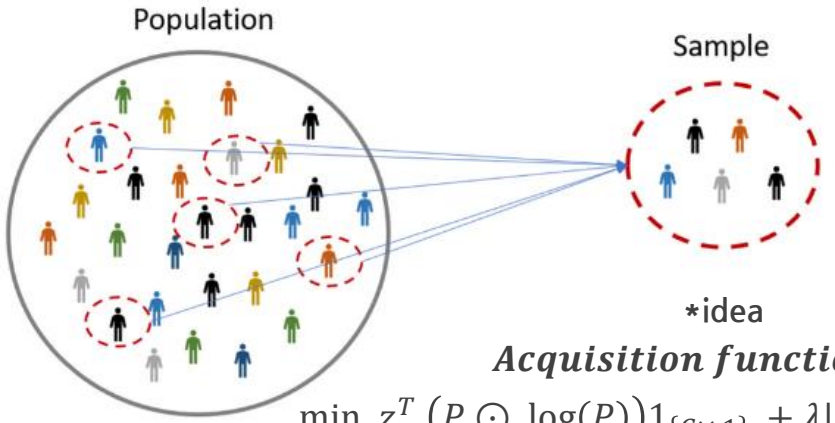


▶ Random sampling selects samples purely at random without considering the data imbalance.

Sampling method - Entropy



❖ Entropy



*idea

Acquisition function :

$$\min_{z} \; z^T \left(P \odot \log(P)\right) 1_{\{C \times 1\}} + \lambda || \, \Omega(c) - P^T z ||_1$$

$\Omega(c)$: The data distribution to be sampled in the current cycle.

$P^T z$: The distribution of the sampled data

$|| \, \Omega(c) - P^T z \, ||_1$: $L1 \; norm$

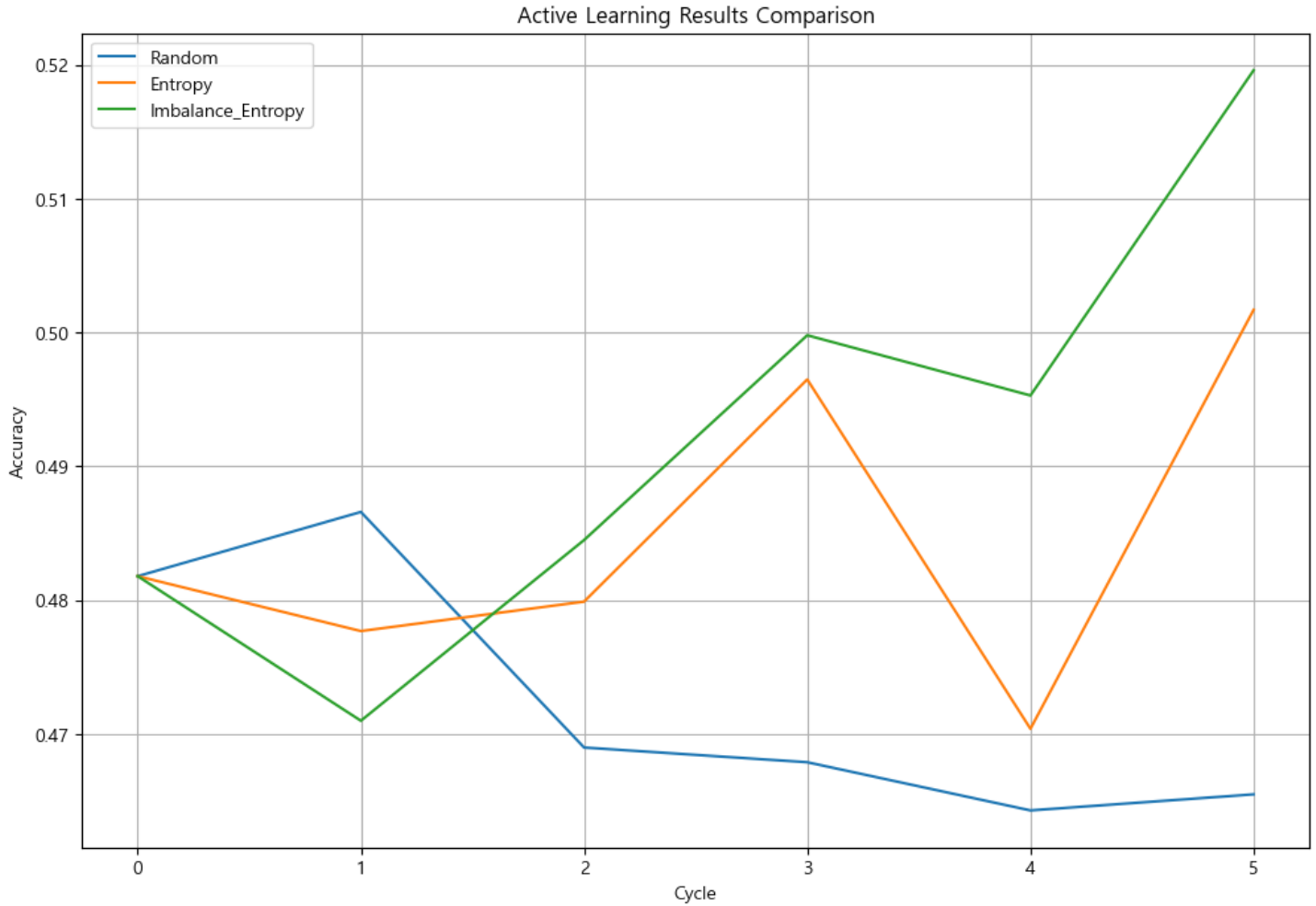
▶ Only entropy → Not penalty term in objective function.

▶ Imbalance + Entropy → There is a penalty term in the objective function.

4

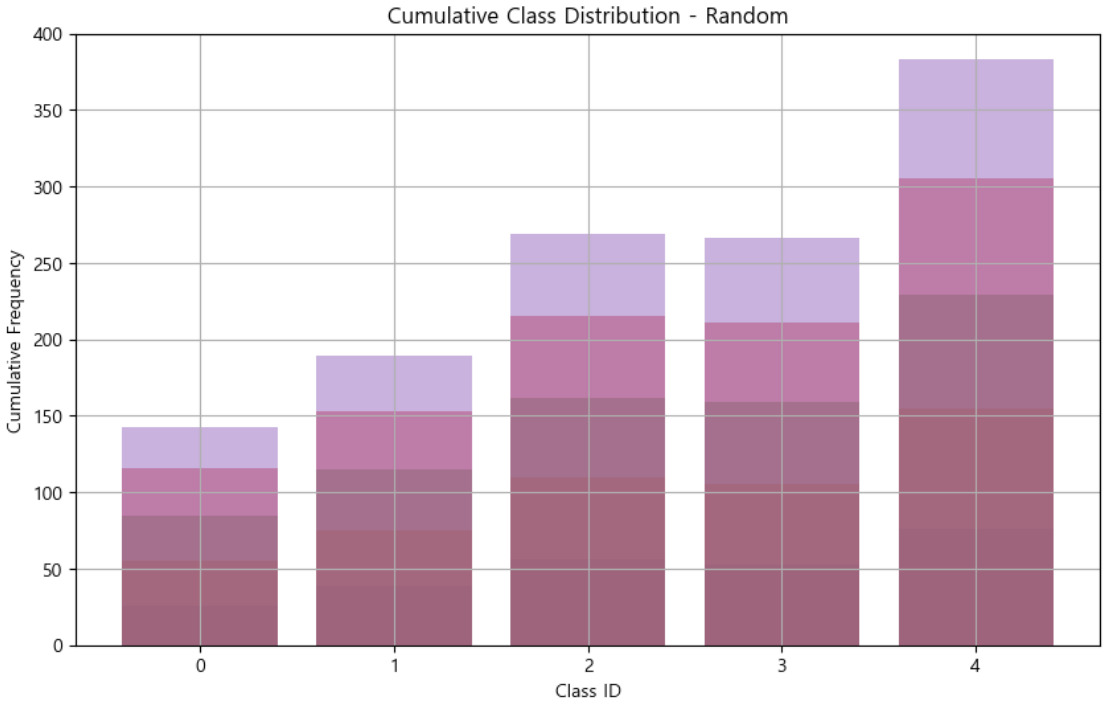
Experiment

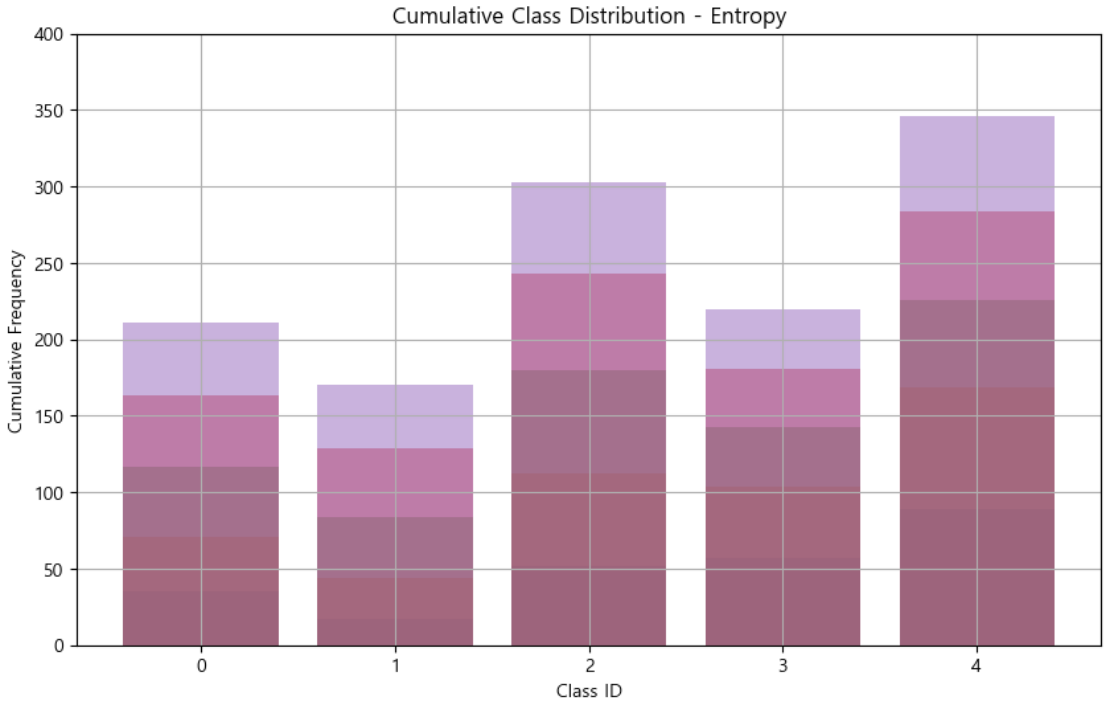
Evaluation



- There was a very slight improvement.

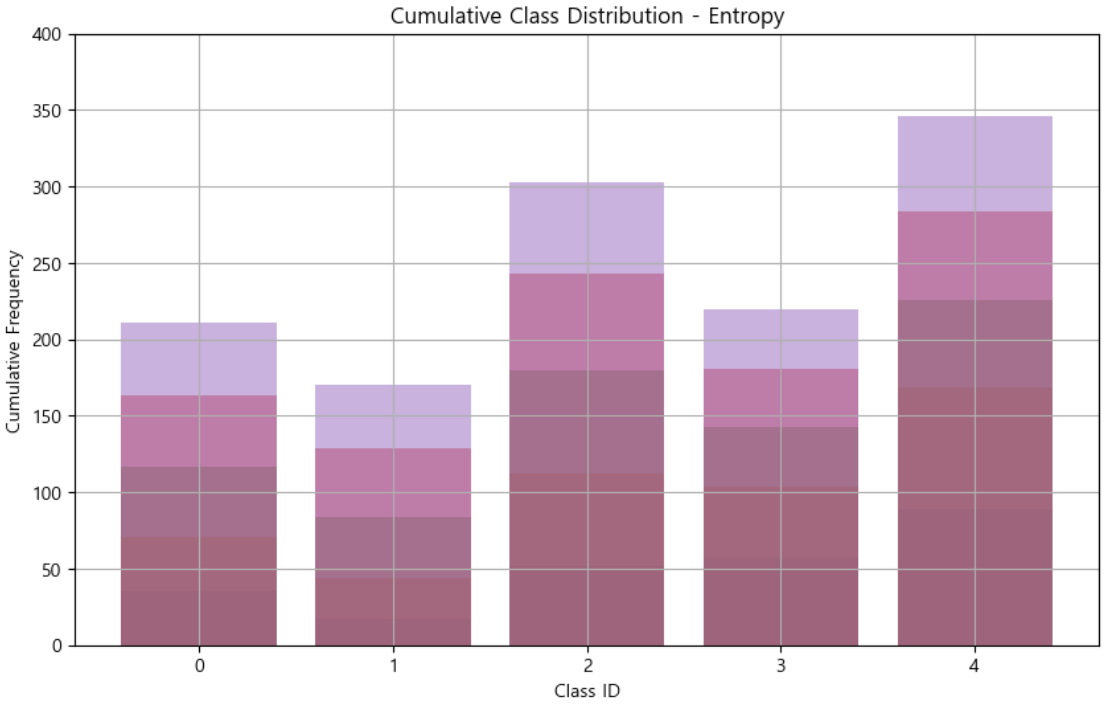
Random vs Entropy

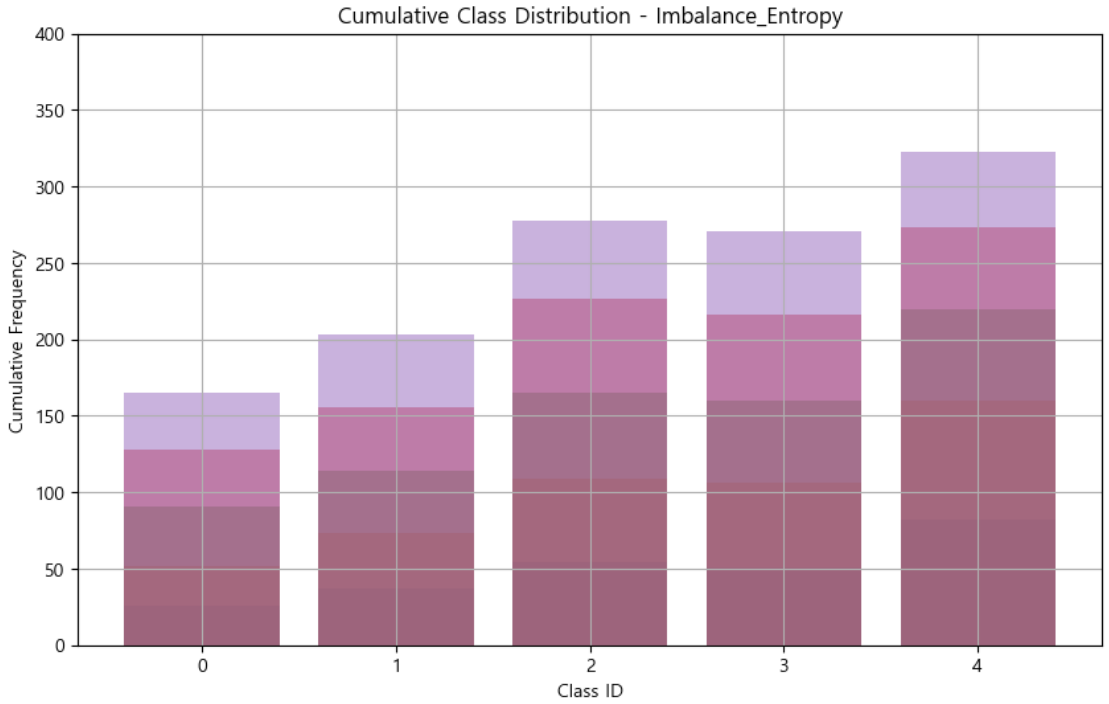




- Random : Maintain a balanced distribution of data across all classes.
- Entropy : Focus on samples where the model is uncertain.

Entropy vs Imbalance Entropy





▶ Imbalance + Entropy : Effective at reducing model uncertainty while maintaining data balance.