AI VIETNAM
All-in-One Course

Extra Class

Imbalanced Data

Nguyen Quoc Thai

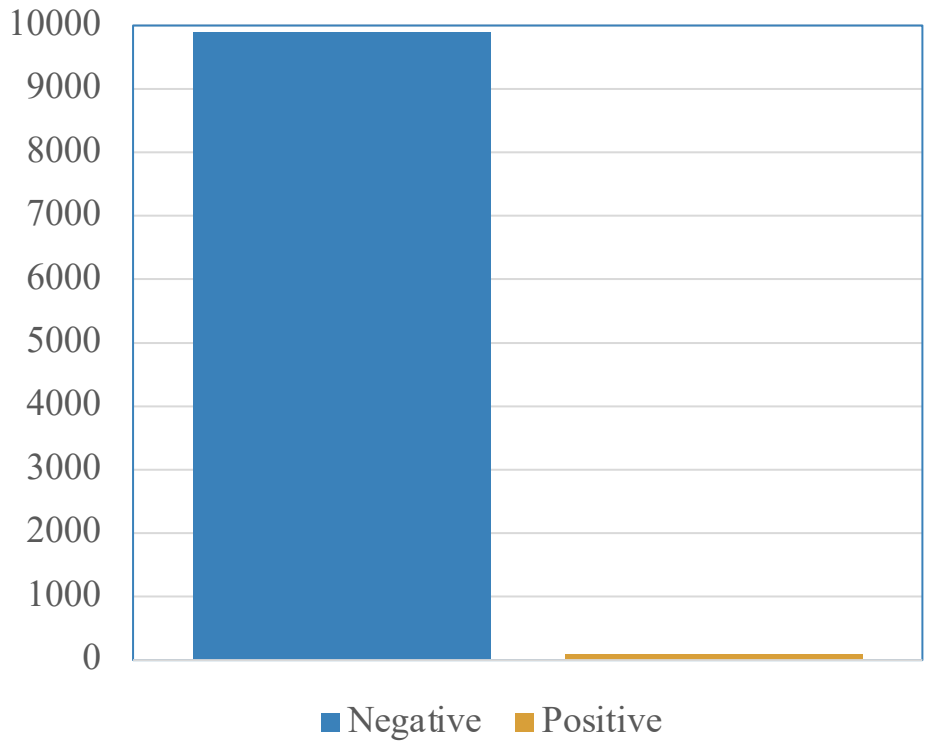
Year 2023

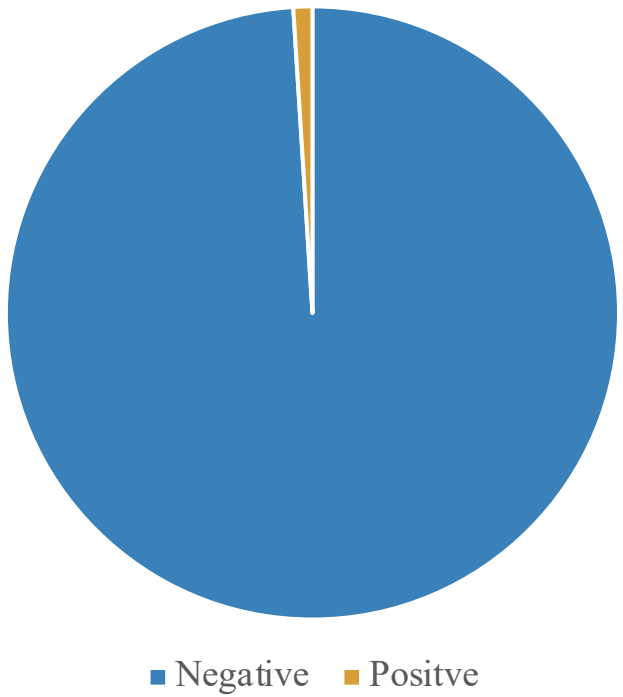
CONTENT

1 – Introduction

Imbalanced Data (Classification)

Negative	9900
Positive	100



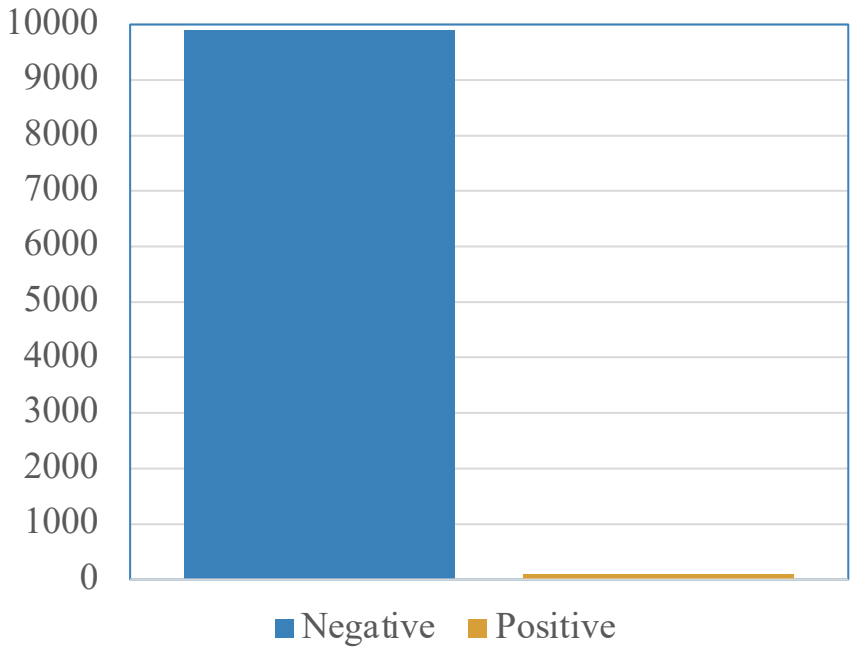


1 – Introduction

Imbalanced Data (Classification)

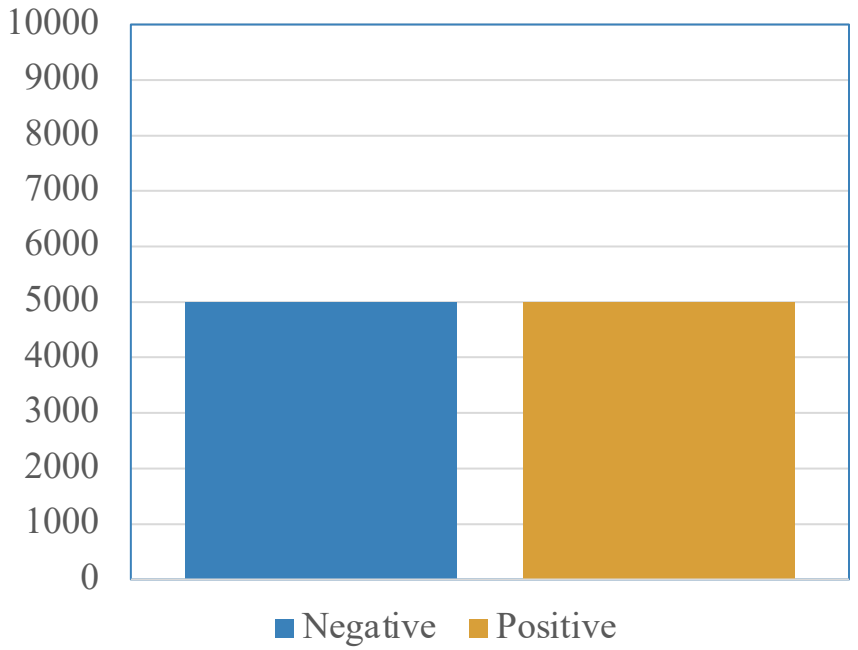
Imbalanced Data

Negative	9900
Positive	100



Balanced Data

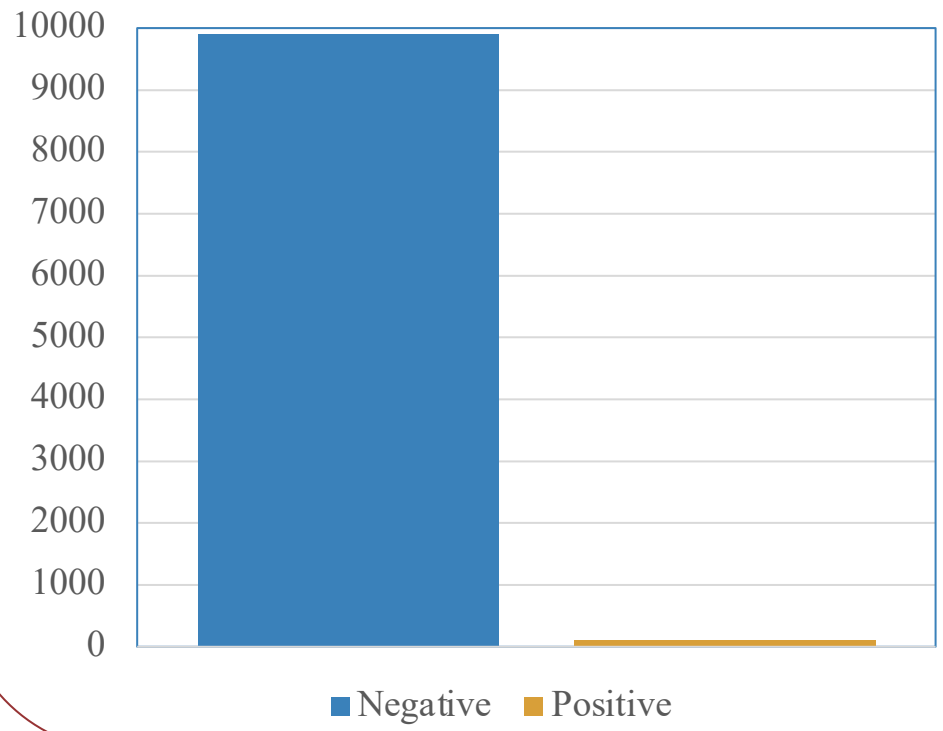
Negative	5000
Positive	5000



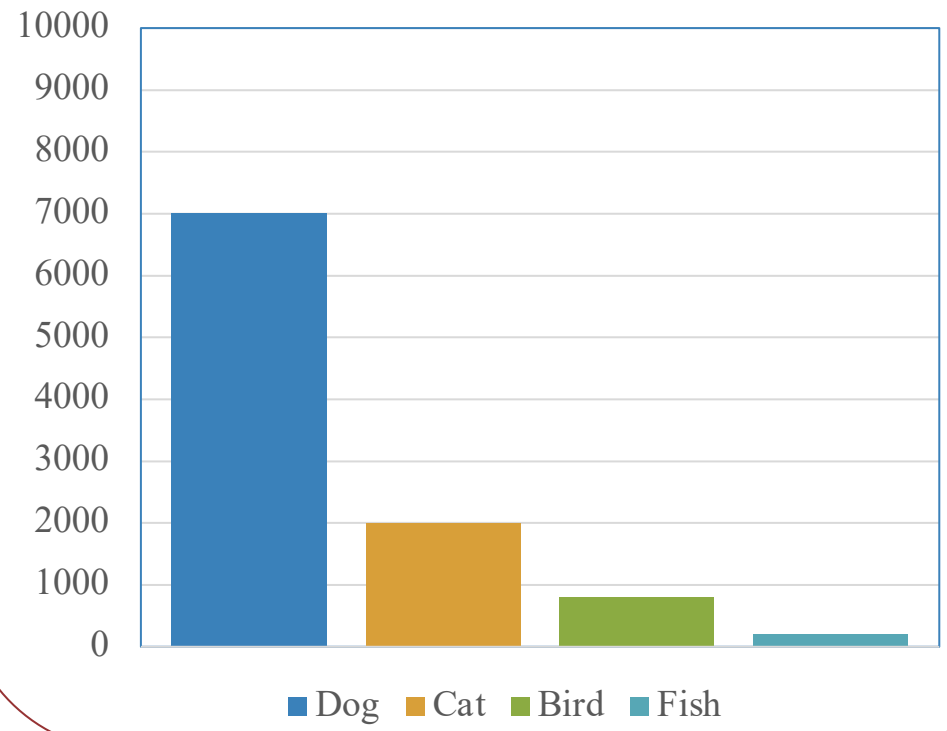
1 – Introduction

Imbalanced Data (Classification)

Binary Classification



Multi-class Classification



1 – Introduction

What happens if dataset is imbalanced ?

	Outlook	Temperature	Windy	Humidity	Play
D0	Sunny	70	True	86	No
D1	Rain	80	True	78	No
D2	Sunny	85	False	56	No
D3	Overcast	66	False	87	No
D4	Sunny	77	True	89	No
D5	Sunny	88	False	78	No
D6	Rain	67	False	84	No
D7	Sunny	70	False	90	Yes

1 – Introduction

What happens if dataset is imbalanced ?

CAT

DOG

1 – Introduction

What happens if dataset is imbalanced ?

Documents	Class
Just plain boring	Negative
Entire predictable and lacks energy	Negative
No surprises and very few laughs	Negative
So bad	Negative
Not good	Negative
Don’t like it	Negative
Very powerful	Positive

2 - Metric

Accuracy

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

2 - Metric

Confusion Matrix

Confusion Matrix		Actual Label	
		Positive	Negative
Predicted Label	Positive	TP True Positive	FP False Positive
	Negative	FN False Negative	TN True Negative

True Positive (TP): Observation is positive, and is predicted to be positive
False Negative (FN): Observation is positive, but is predicted negative
True Negative (TN): Observation is negative, and is predicted to be negative
False Positive (FP): Observation is negative, but is predicted positive

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2 - Metric

Accuracy – Example

A

		Actual Label	
		Positive	Negative
Predicted Label	Positive	1	0
	Negative	1	998

$$\text{Acc} = \frac{0 + 998}{1 + 998 + 0 + 1} = 0.999$$

B

		Actual Label	
		Positive	Negative
Predicted Label	Positive	400	200
	Negative	100	300

$$\text{Acc} = \frac{400 + 300}{400 + 300 + 200 + 100} = 0.7$$

2 - Metric

Precision

- Precision: % of items the model labeled as positive that are in fact positive
- Precision attempts to answer the following question: What proportion of positive identifications was actually correct?

Confusion Matrix		Actual Label	
		Positive	Negative
Predicted Label	Positive	TP True Positive	FP False Positive
	Negative	FN False Negative	TN True Negative

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

2 - Metric

Recall

- Precision: % of items actually present in the input that were correctly identified by the model
- Precision attempts to answer the following question: What proportion of actual positive was identified correctly?

Confusion Matrix		Actual Label	
		Positive	Negative
Predicted Label	Positive	TP True Positive	FP False Positive
	Negative	FN False Negative	TN True Negative

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

2 - Metric

Precision – Recall – Example

A

		Actual Label	
		Positive	Negative
Predicted Label	Positive	1	0
	Negative	1	998

$$\text{Acc} = 0.999$$

$$\text{Precision} = \frac{1}{1+0} = 1.0$$

$$\text{Recall} = \frac{1}{1+1} = 0.5$$

B

		Actual Label	
		Positive	Negative
Predicted Label	Positive	400	200
	Negative	100	300

$$\text{Acc} = 0.7$$

$$\text{Precision} = \frac{400}{40+200} = 0.67$$

$$\text{Recall} = \frac{400}{400+100} = 0.8$$

2 - Metric

F Measure

- F Measure

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

- Balance: Precision and Recall

$$F_1 = \frac{2PR}{P + R}$$

2 - Metric

F1 – Example

A

		Actual Label	
		Positive	Negative
Predicted Label	Positive	1	0
	Negative	1	998

$$\text{Acc} = 0.999$$
$$\text{Precision} = 1.0$$
$$\text{Recall} = 0.5$$
$$F_1 = \frac{2PR}{P + R} = \frac{2 * 1 * 0.5}{1 + 0.5} = 0.67$$

B

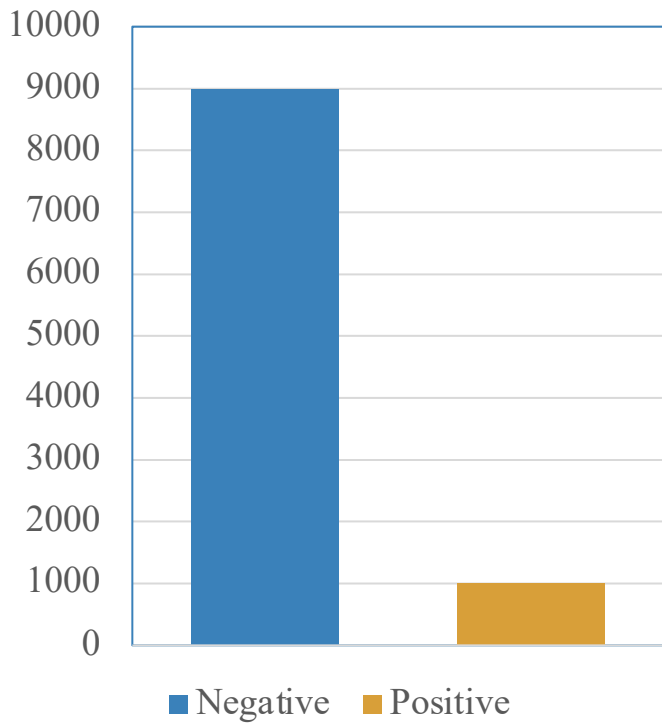
		Actual Label	
		Positive	Negative
Predicted Label	Positive	400	200
	Negative	100	300

$$\text{Acc} = 0.7$$
$$\text{Precision} = 0.67$$
$$\text{Recall} = 0.8$$
$$F_1 = \frac{2PR}{P + R} = \frac{2 * 0.67 * 0.8}{0.67 + 0.8} = 0.73$$

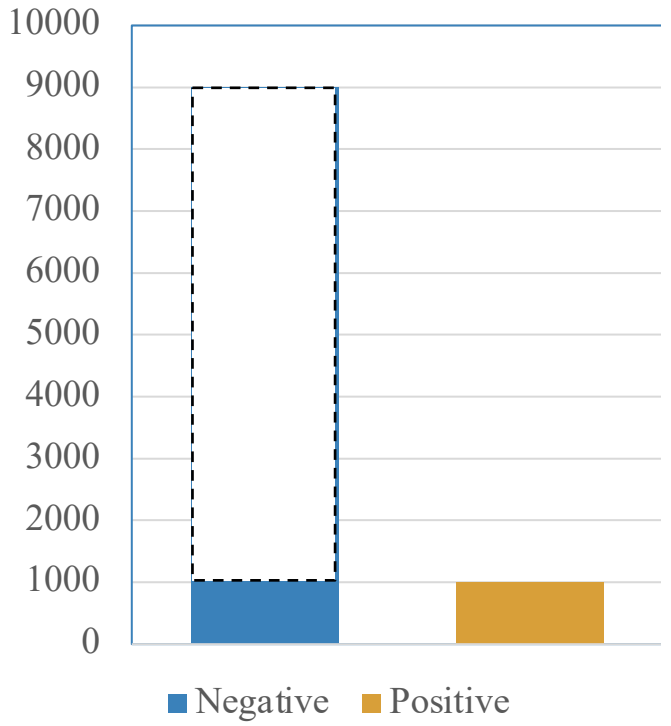
3 - Appoarches

Approach 1: Data Manipulation

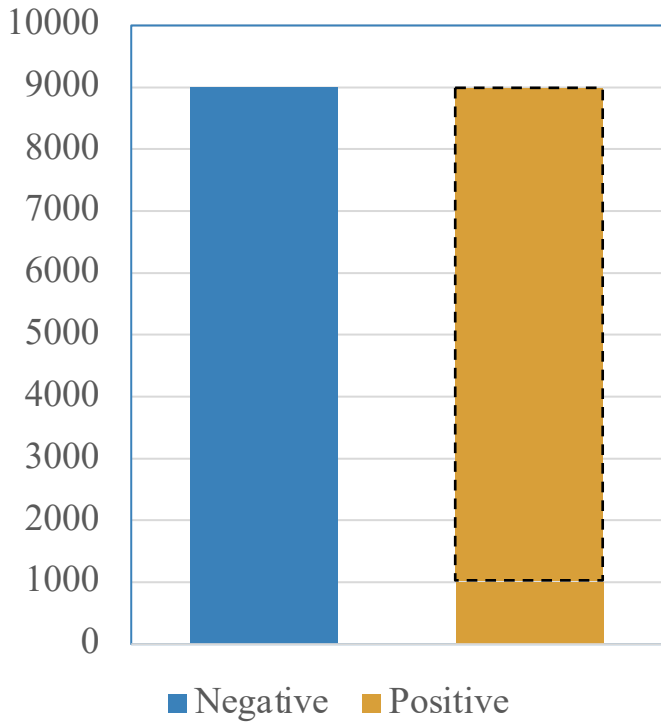
Original Data



Undersampling Data



Oversampling Data

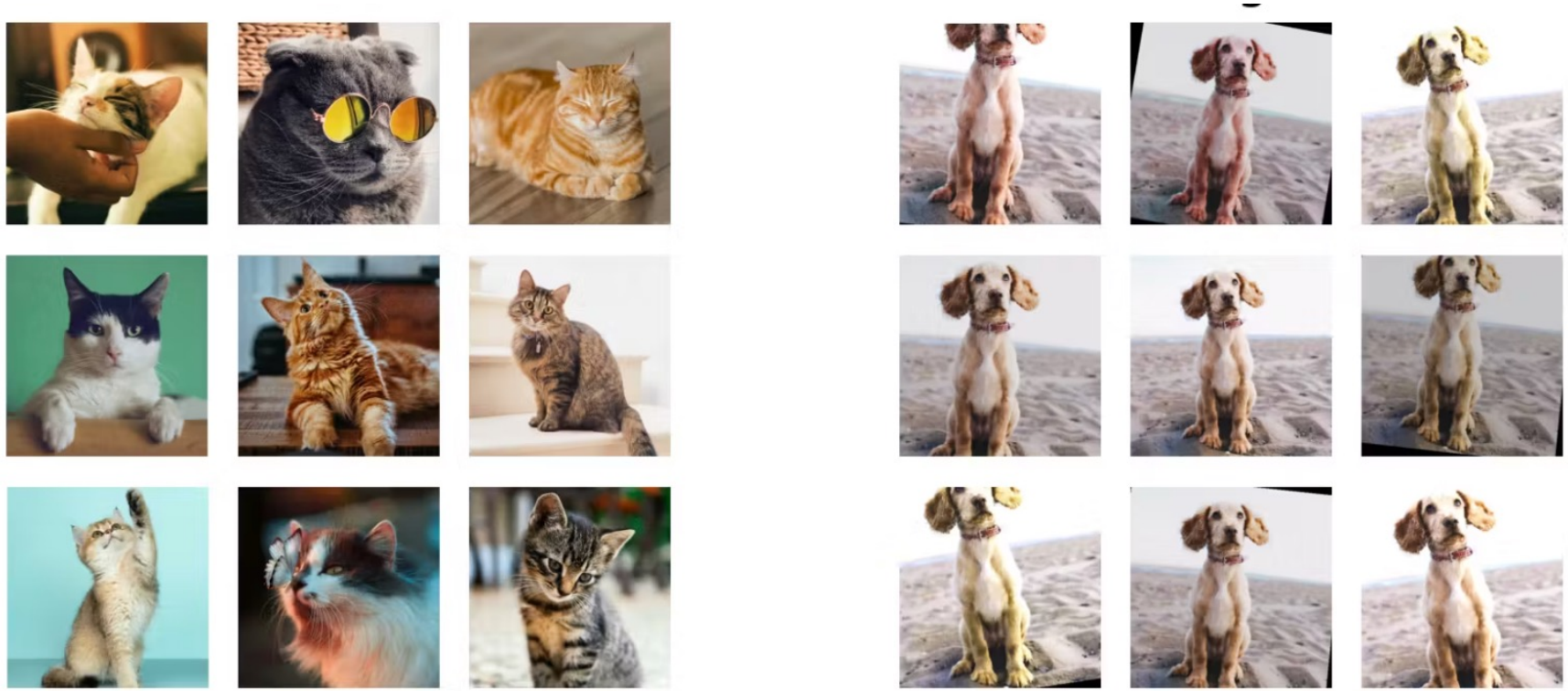


AI VIET NAM
@aivietnam.edu.vn

3 - Appoarches

Approach 1: Data Manipulation

Augmentation (Oversampling)

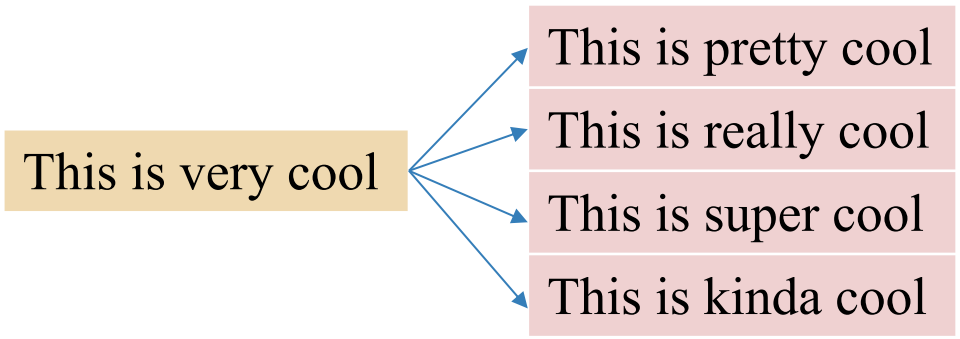


3 - Appoarches

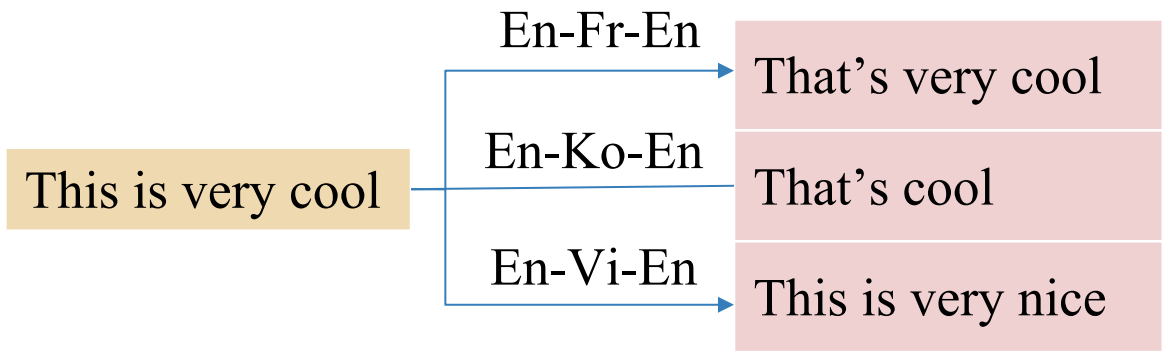
Approach 1: Data Manipulation

Augmentation (Oversampling)

Easy Data Augmentation	Short Example
Random Swap	I am jogging => I jogging am
Random Deletion	I am jogging => I jogging
Random Insertion	I am jogging => I am a jogging



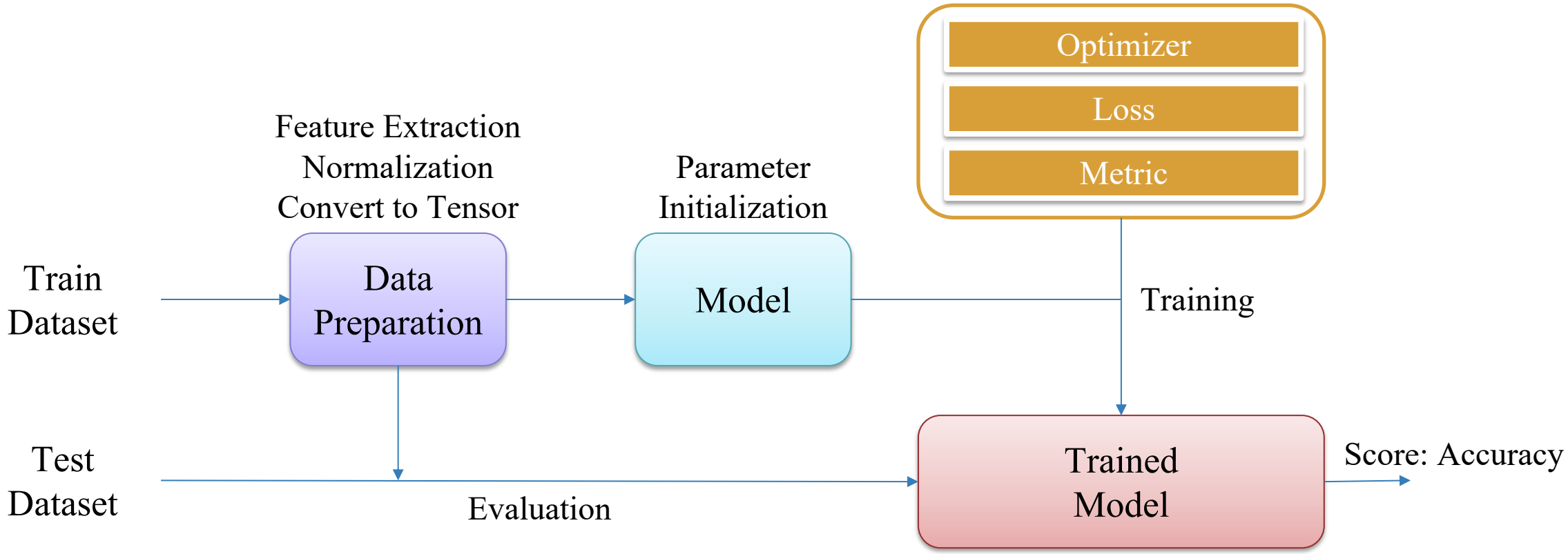
Synonym Replacement



Back-Translation

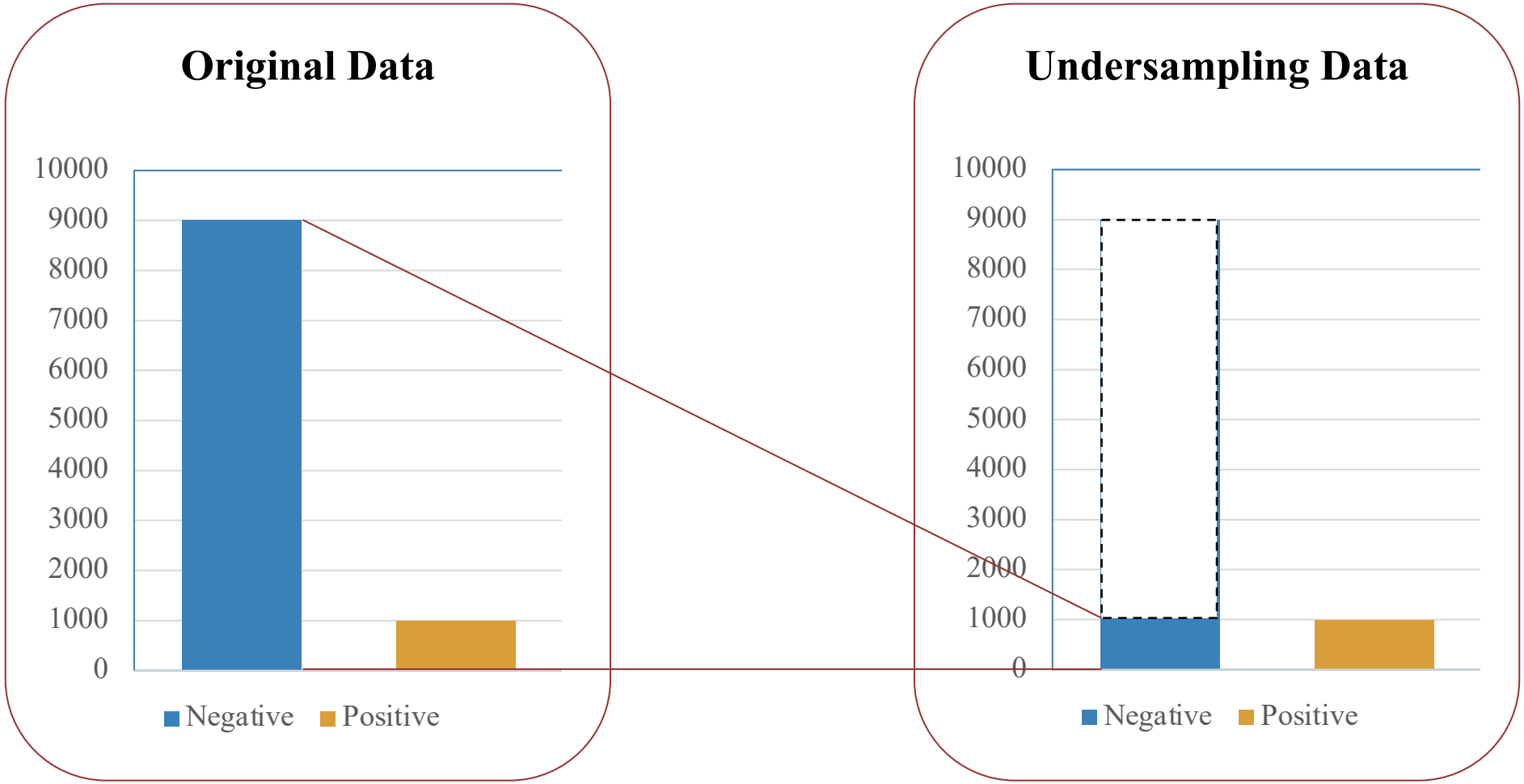
3 - Appoarches

Approach 2: Loss Function and Optimization



4 – Undersampling

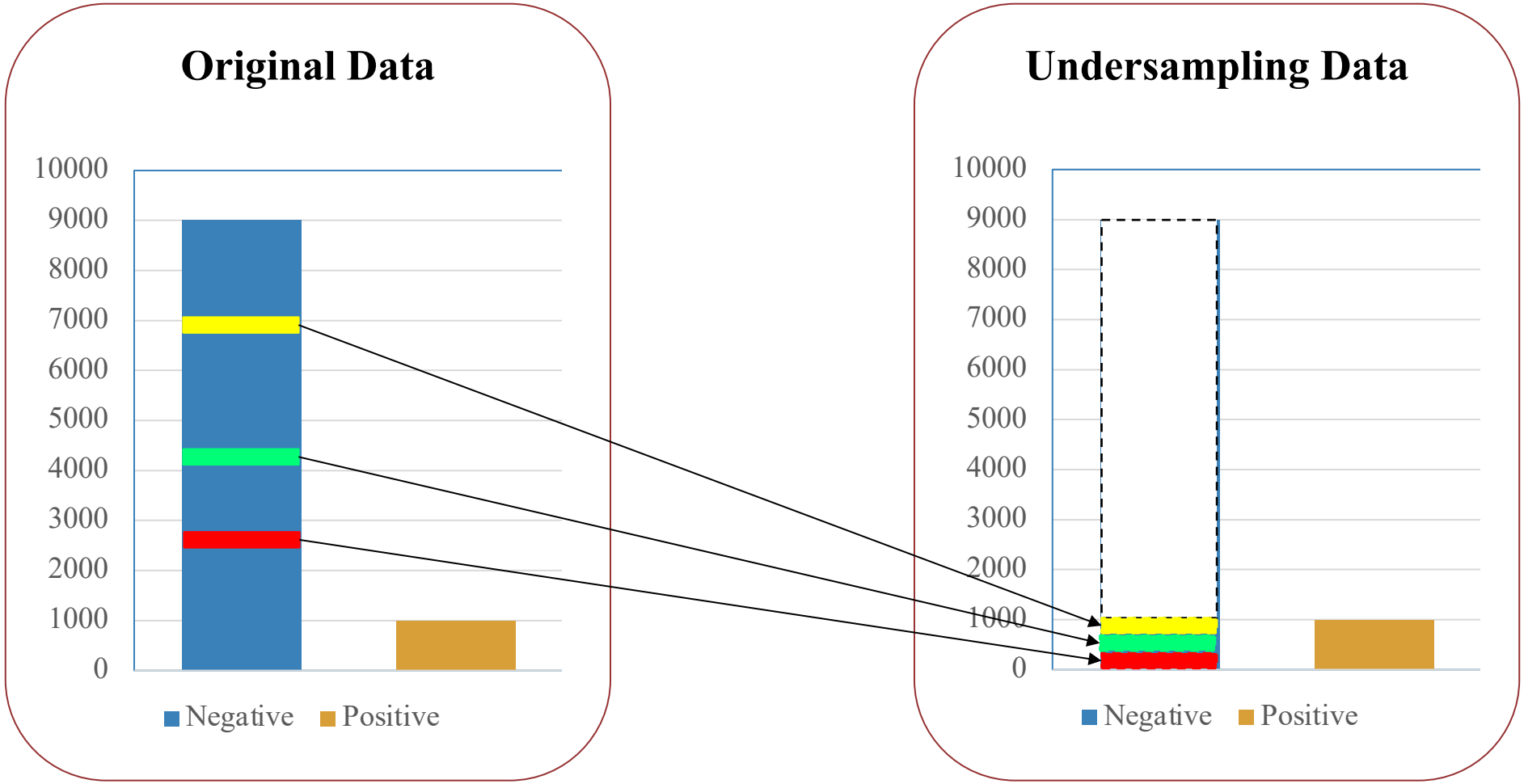
Overview



AI VIET NAM
@aivietnam.edu.vn

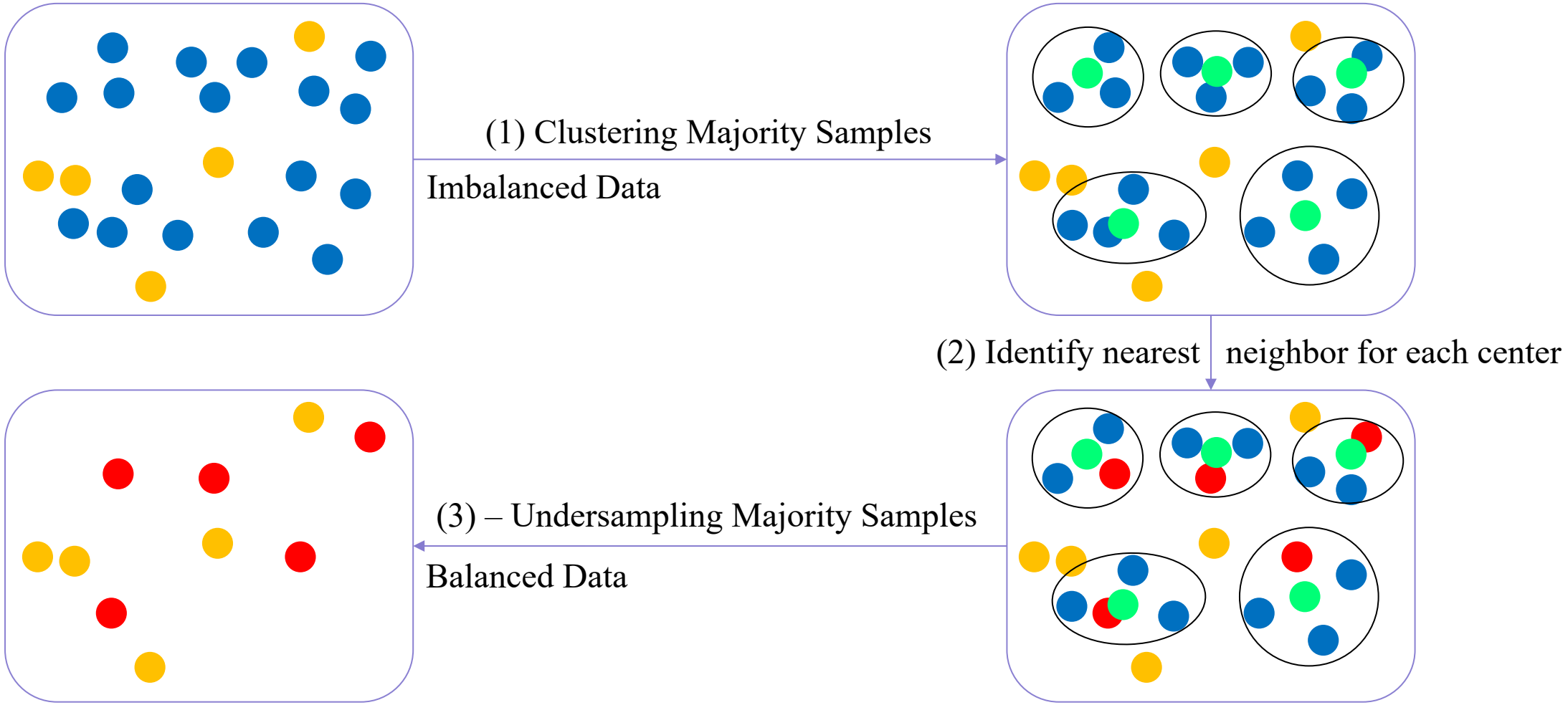
4 – Undersampling

Random Undersampling



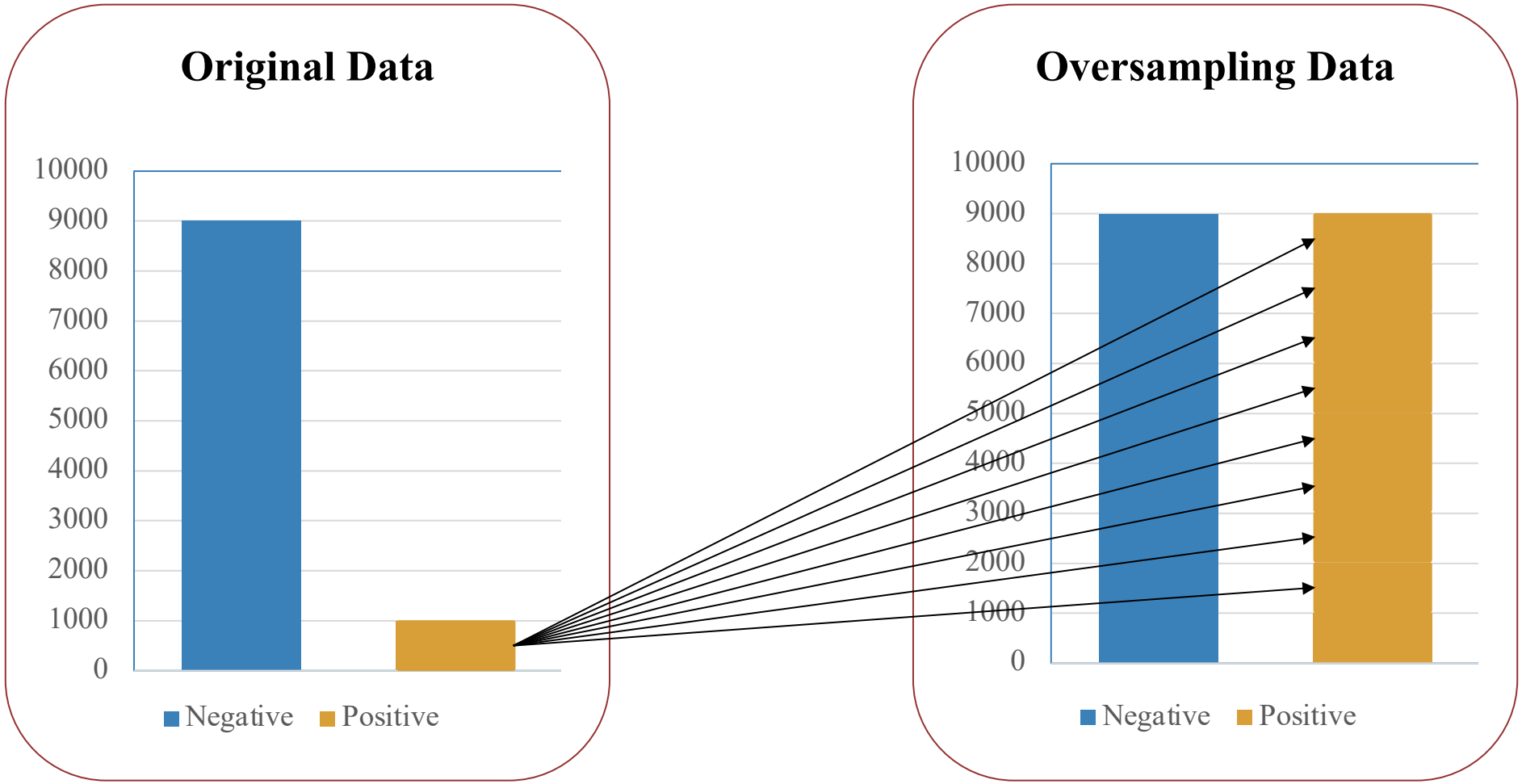
4 – Undersampling

Clustering-based Undersampling



5 – Oversampling

Duplicate



AI VIET NAM
@aivietnam.edu.vn

5 – Oversampling

Data Augmentation

Original

Flip

Color

Blur

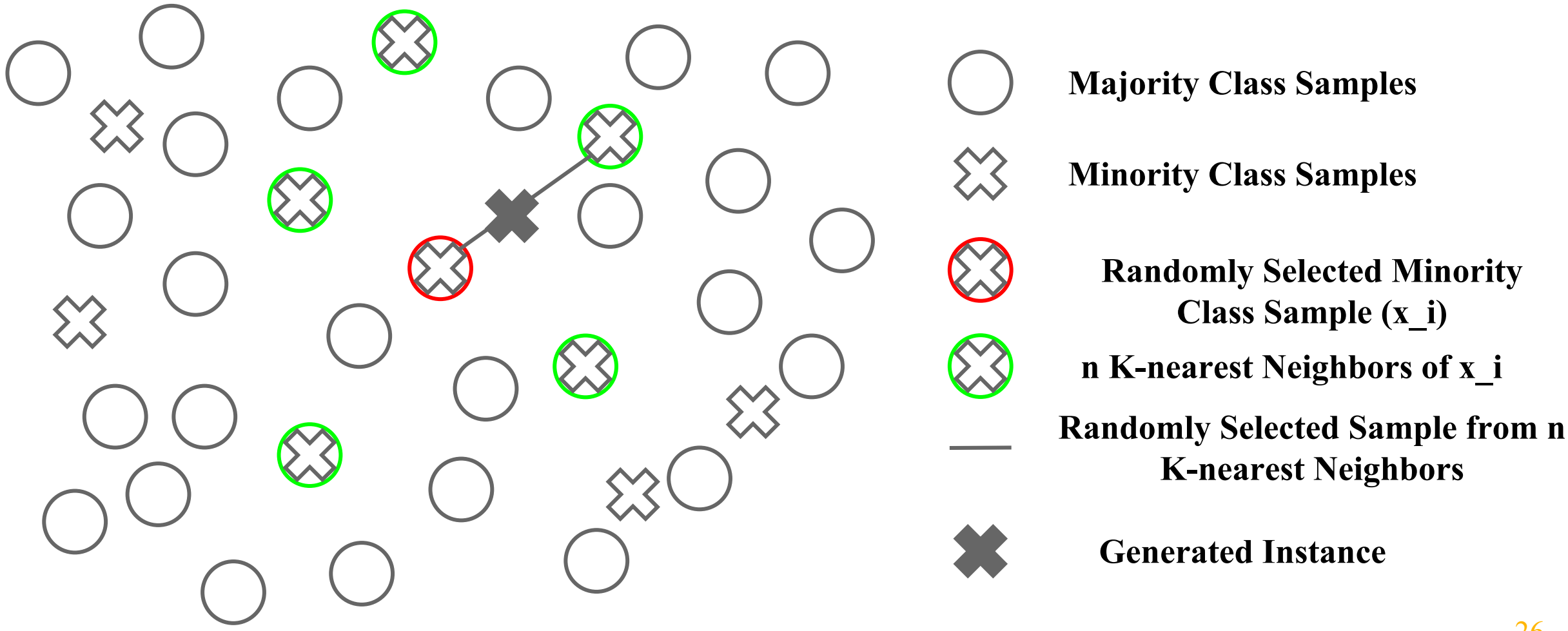
Brightness

Rotate

Noise

5 – Oversampling

SMOTE (Synthetic Minority Over-sampling TEchnique)



AI VIET NAM

@aivietnam.edu.vn

Thanks!

Any questions?