

Cyberbullying Detection Using AI

Keyan Guo

University at Buffalo



Outline

- The Critical Problem of Cyberbullying
- Processes in AI Development
- Data Collection
- Annotation
- Training
- Evaluation
- Deployment
- Q&A

What is Cyberbullying?

- Different from bullying, CYBERBULLYING is bullying with the use of **Digital Technologies**.



What is Cyberbullying?

- Cyberbullying is bullying with the use of **Digital Technologies**
 - Social media
 - Messaging platforms
 - Gaming platforms
 - Mobile phones

What is Cyberbullying?

- Cyberbullying is bullying with the use of **Digital Technologies**
- It is repeated behavior, aimed at **scaring**, **angering** or **shaming** those who are targeted.

Common types of cyberbullying

- **Spreading** lies about or posting embarrassing photos or videos of someone
- **Sending** hurtful, abusive or threatening messages, images or videos to someone
- **Impersonating** someone and sending mean messages to others

The Critical Problem of Cyberbullying

I hate you! I dislike these people because...
You are an idiot, ... Get out from my
house! You are not my friend anymore...



Based on words and phrases

hate, dislike, ugly, get out ...

State-of-the-art detectors



Amazon Rekognition

Actively researched problem

Psychology

Sociology

Computer science

Process of AI Development

- Data Collection
 - We need dataset for training AI
- Annotation
 - We need to label dataset
- Training
 - AI training process
- Evaluation
 - How good are we doing?
- Deployment
 - Detect on real-world samples

Dataset Collection

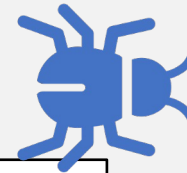
Search the Internet



Search Engine



flickr



Social Media

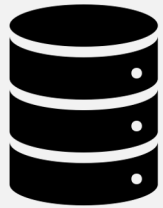


Dataset

Cyberbullying dataset is from Formspring

Annotation

Humans Label Dataset



Dataset



Annotators



Cyberbullying

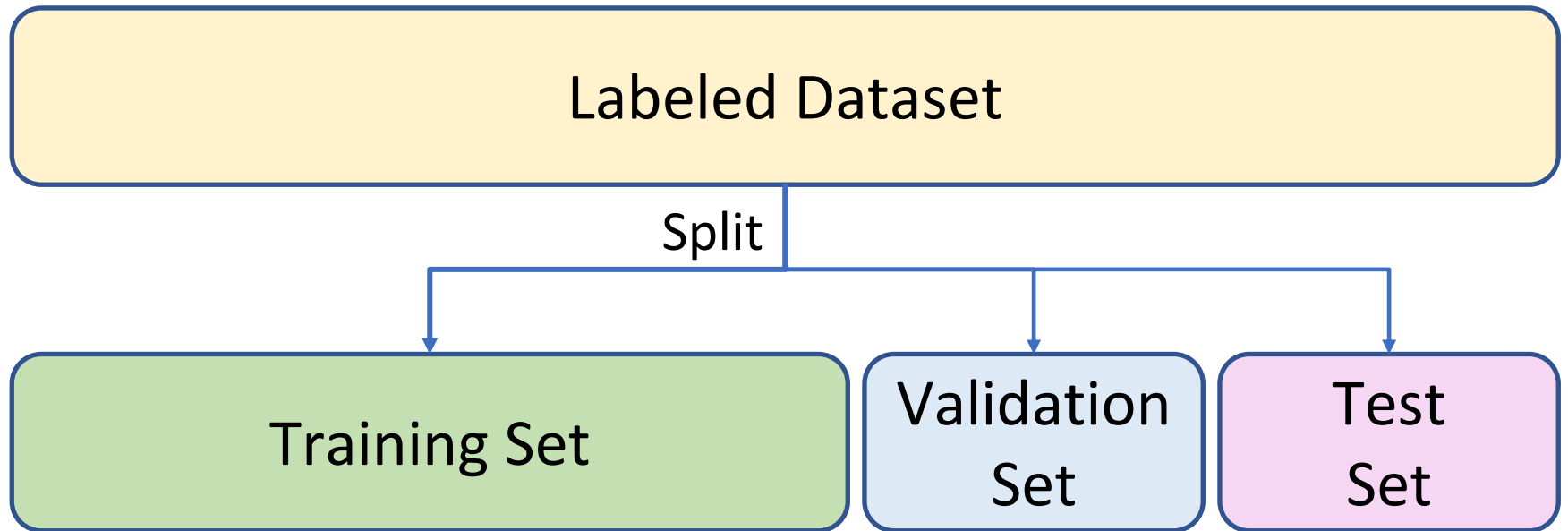


Non-cyberbullying

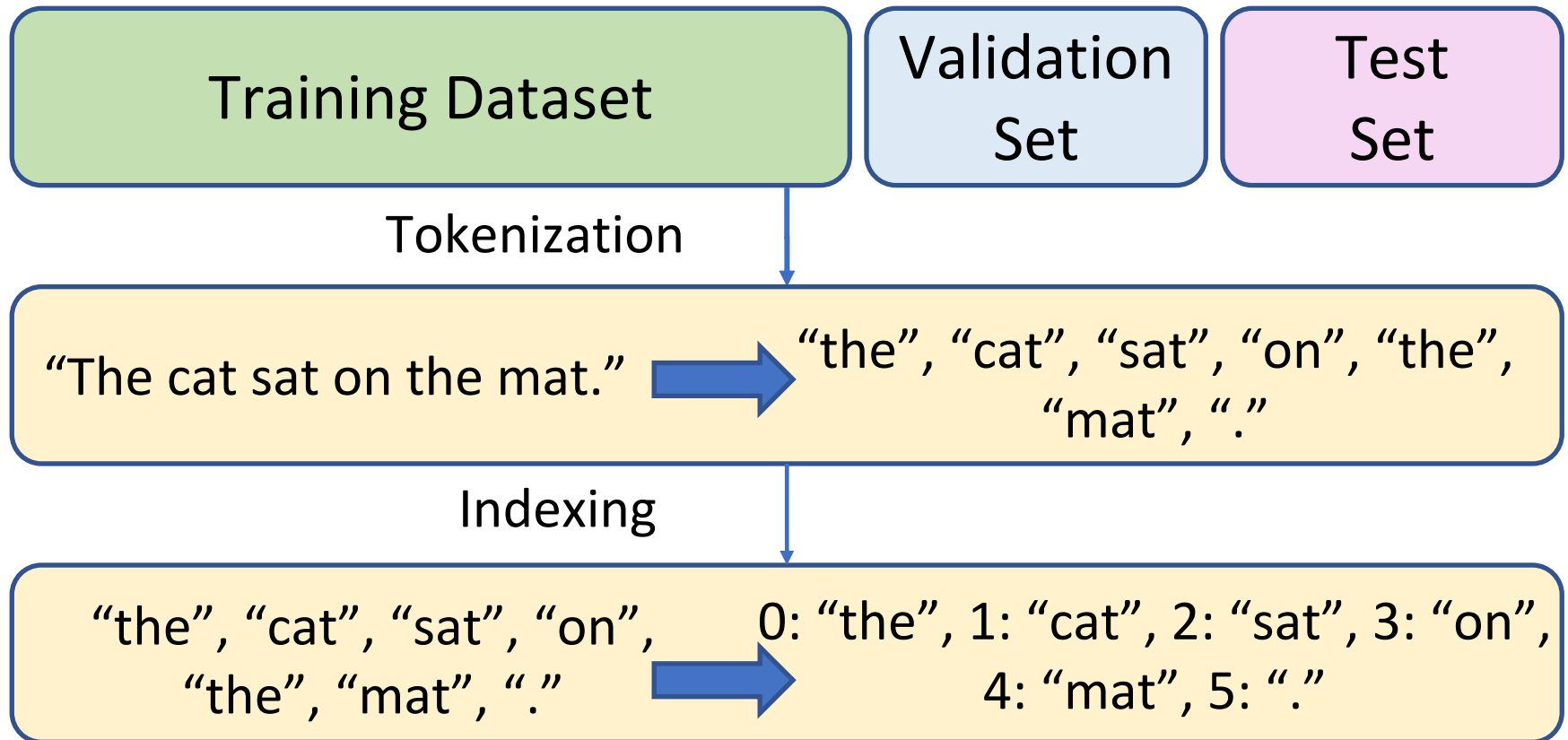


Labeled Dataset

Training



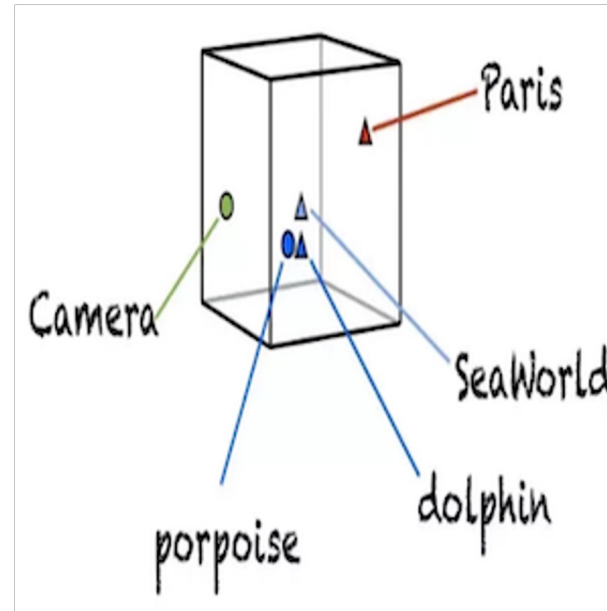
Training Cont...



Training Cont...

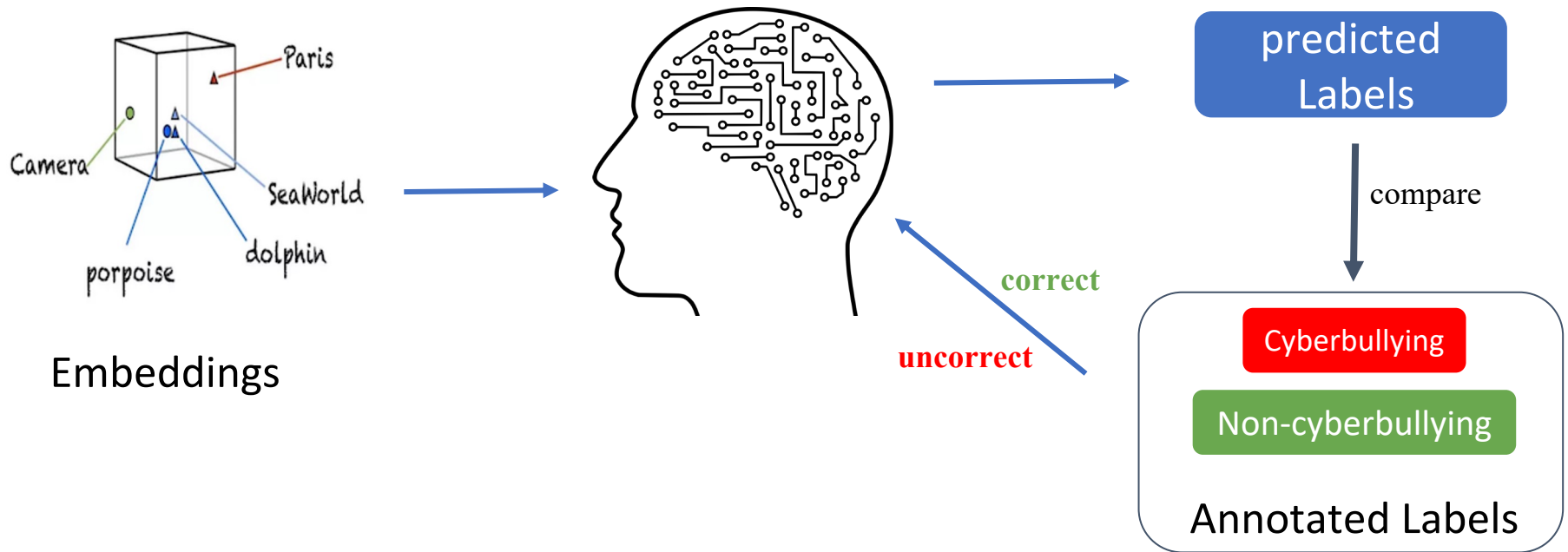
Paris
Camera
dolphin
SeaWorld
Porpoise

Tokens



Embeddings

Training Cont...



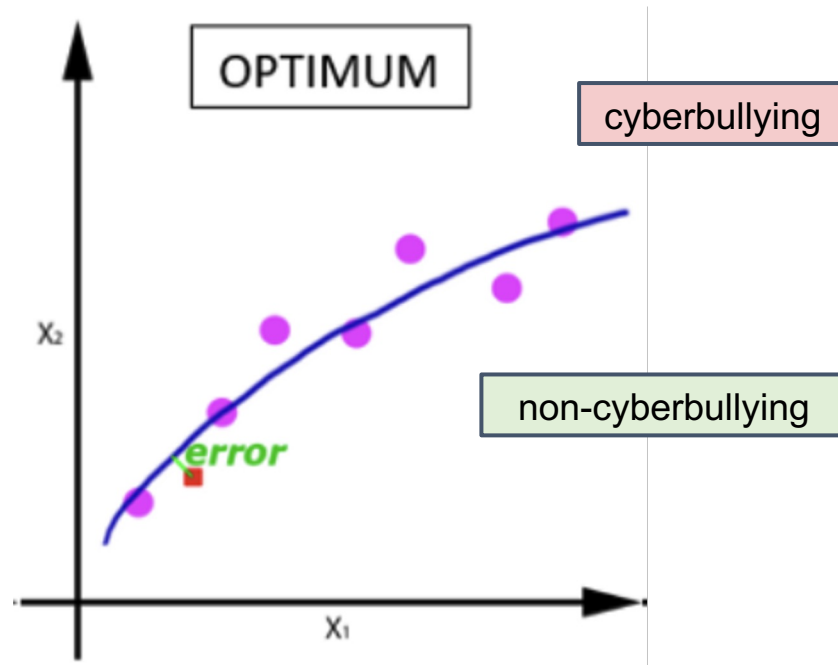
AI is Learning ...

Training Cont...

- epoch
 - When all training data has been learned by the AI once, this process is called an **epoch**
- How many epochs should we use?
 - **Unfortunately**, there is no correct answer to this question.
 - For different datasets, the answer is not the same. We still need to try different epochs number to figure it out.

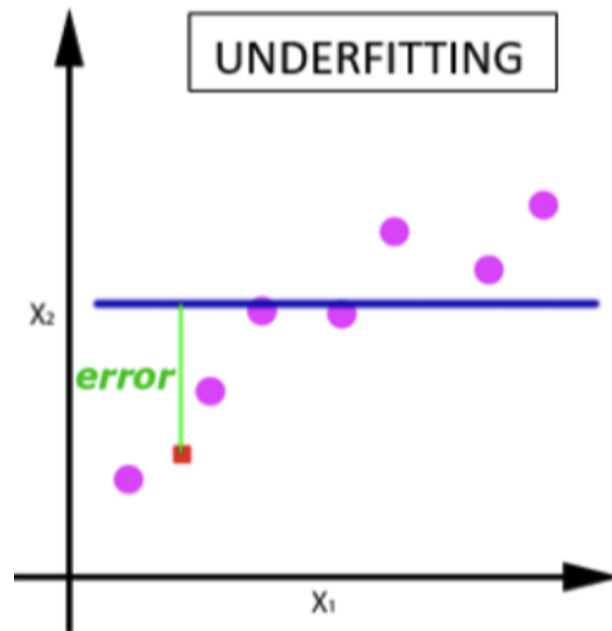
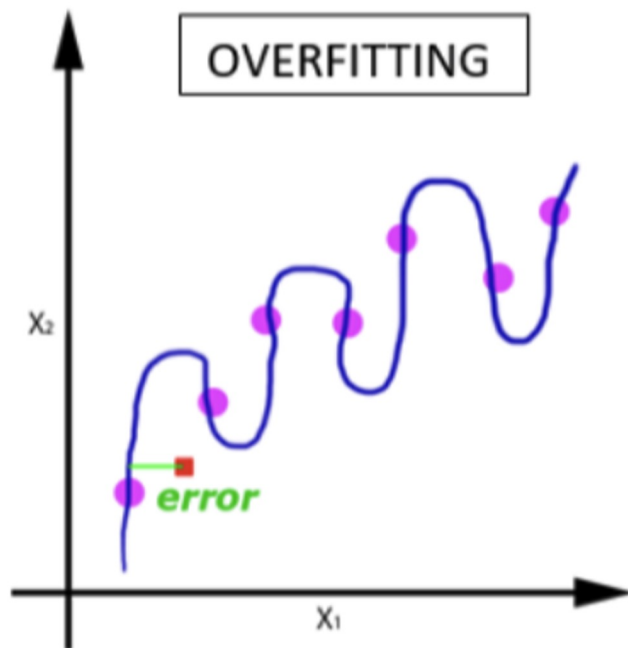
Training Cont...

- How many epochs should we use?
 - optimum epochs

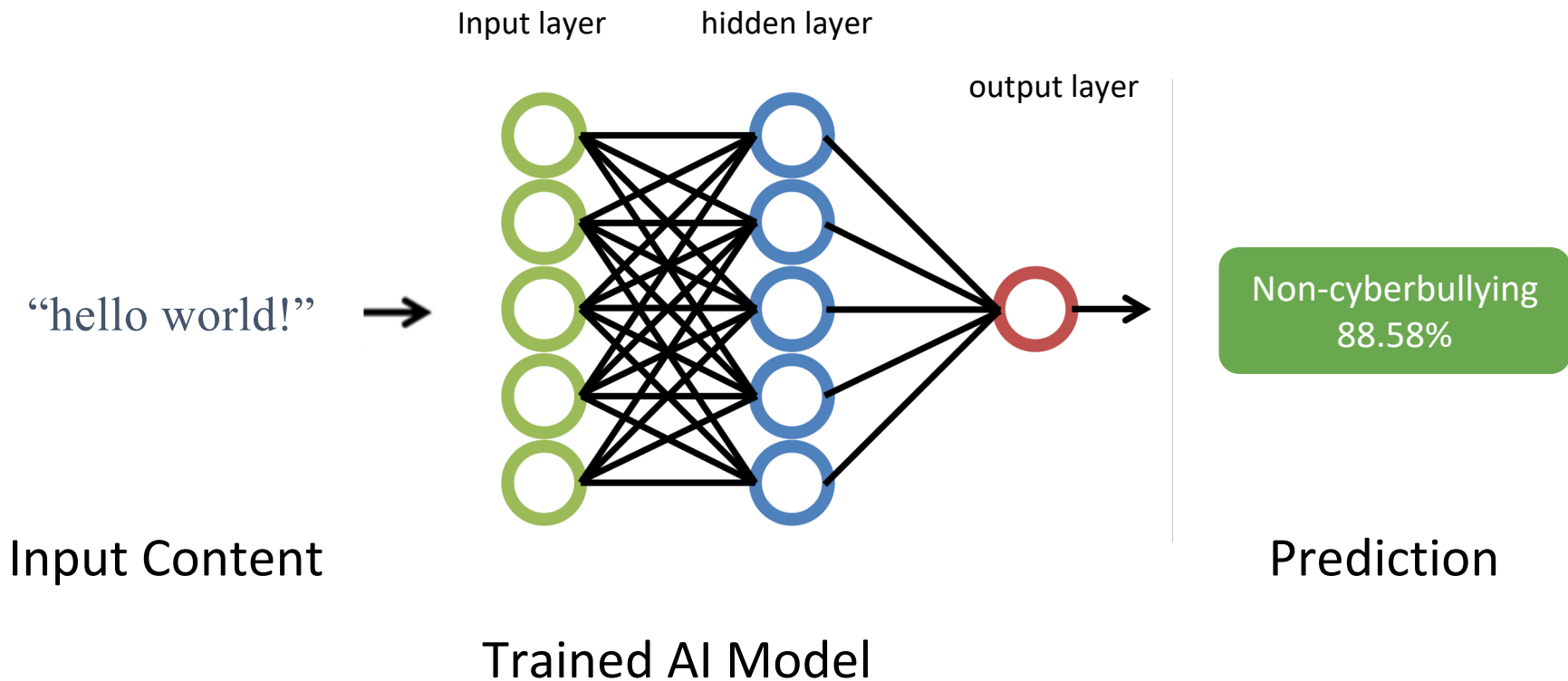


Training Cont...

- How many epochs should we use?
 - too many or too few epochs



Evaluation



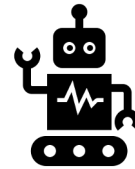
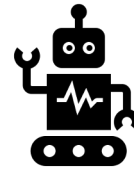
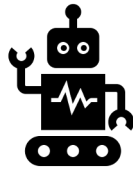
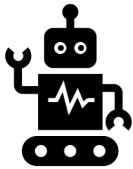
Evaluation Cont...

- Accuracy

- One simple way we measure an AI model's performance with test dataset
- Accuracy = the percentage of the correct predictions

$$Accuracy = \frac{\text{\textit{\# of correct predictions}}}{\text{\textit{\# of all predictions}}}$$

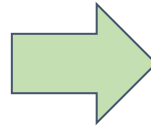
Deployment



Even More Labs!

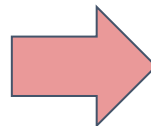
Go check Link: <https://cuadvancelab.github.io/labs.html>

- Cyberbullying Detection on Images



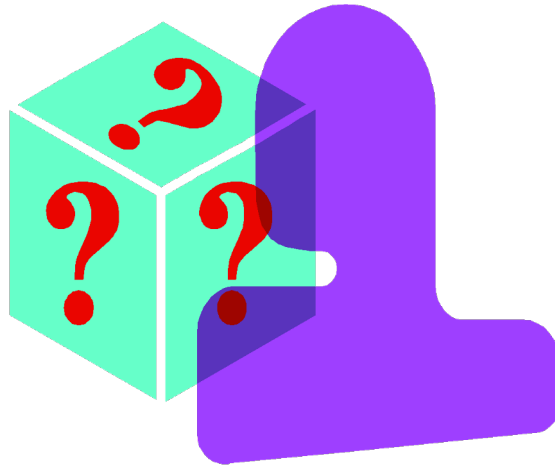
Cyberbullying detected!

- Adversarial attack on Cyberbully detection models



Cyberbullying not detected.

Q & A



Jump to the Lab

Let's get our hands dirty!

Link for our lab:

https://colab.research.google.com/github/cuadvancelab/cuadvancelab.github.io/blob/main/instructions/lab1/social-science/lab1_interactive.ipynb