Cyberbullying Detection Using Images

<Instructors>

Before the lecture

- Quickly Recall
  - In Lab 1 Cyberbullying Detection Using AI, we have learned:
    - The critical problem of Cyberbullying
    - Processes in AI development
      1. Data Collection
      2. Data Annotation

All experiments are based on the textual model of cyberbullying.

Outline

- Identify cyberbullying in images
- Working approach of fused model
- Evaluation of AI Model
- TP, TN, FP, FN
- Accuracy, Precision, Recall, F1 score
- Q&A

Identify cyberbullying in Text

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 tools

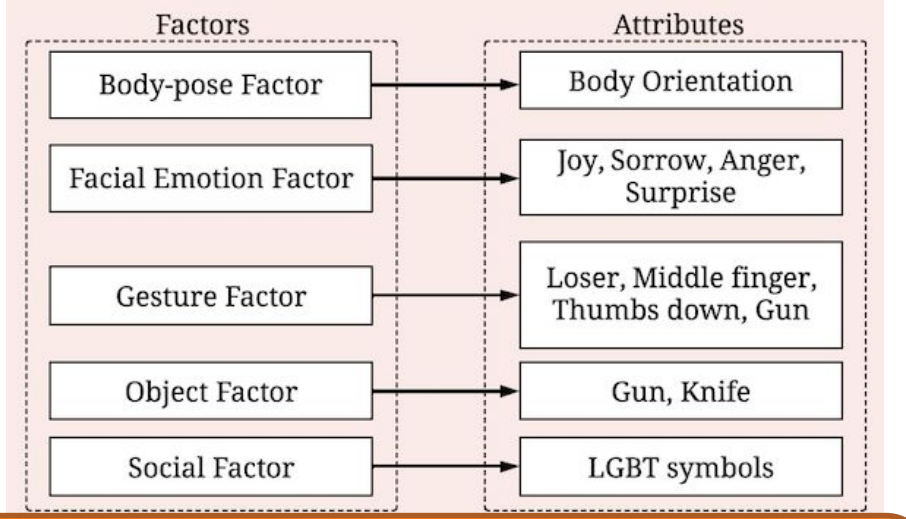
Google clarifai Amazon Rekognition

Identify cyberbullying in images

Based on five visual factors

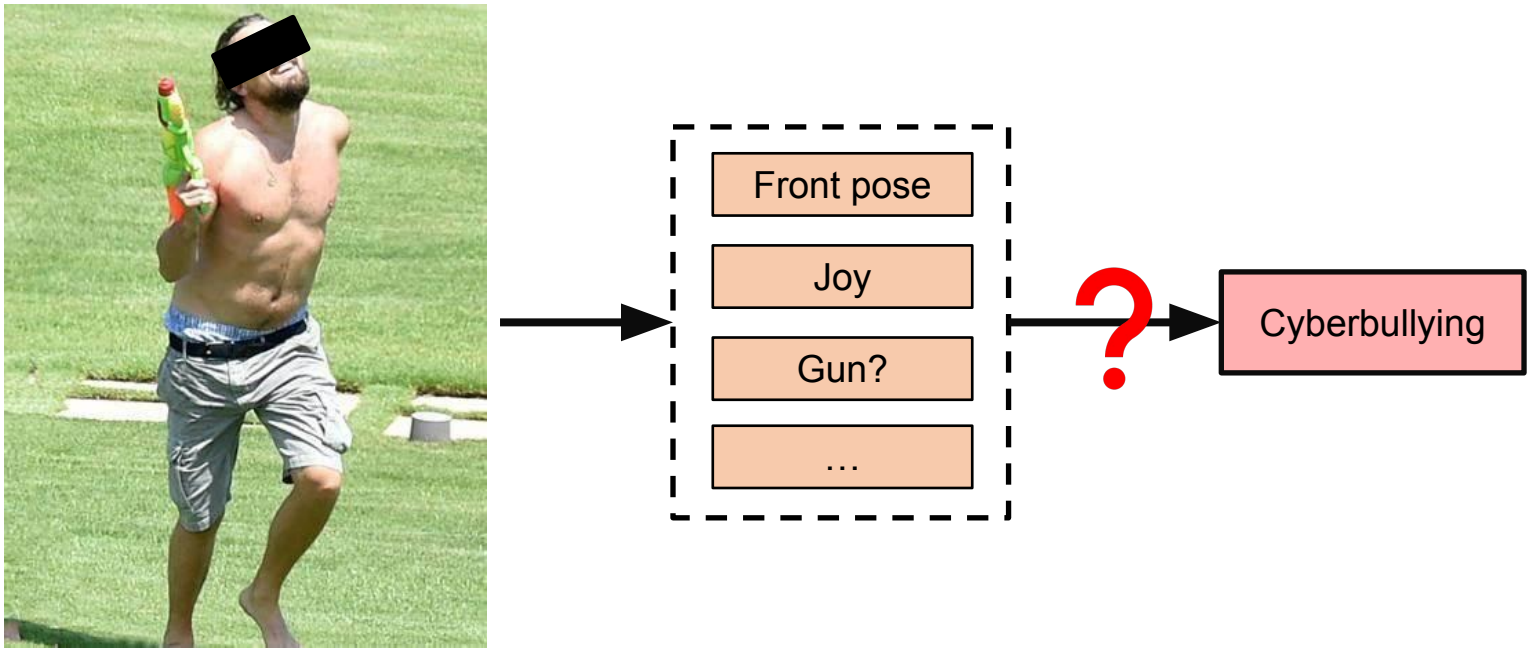
Body-pose, emotion, object, gesture and social factors

Identify cyberbullying in images

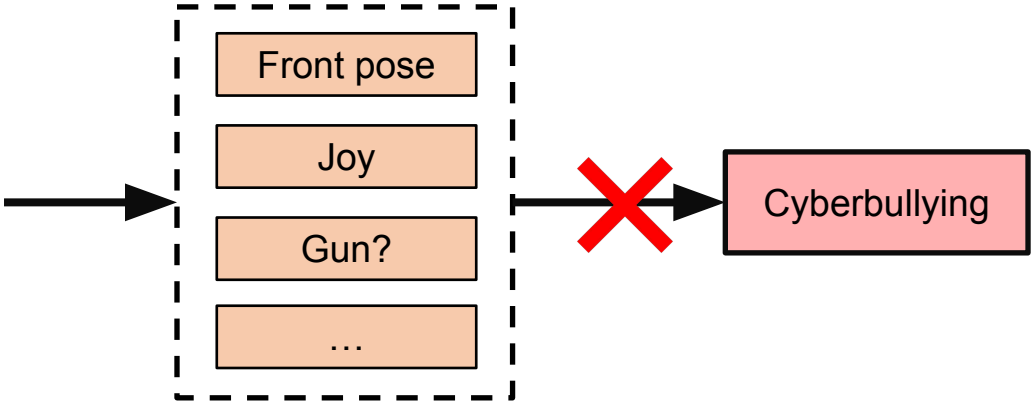


Can we determine whether an image is "cyberbullying" by these factors alone?

Identify cyberbullying in images

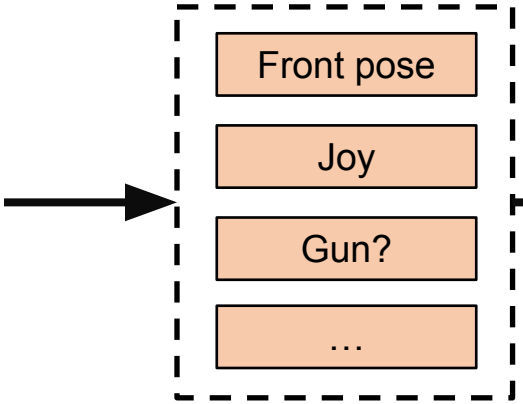


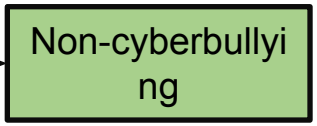
Identify cyberbullying in images

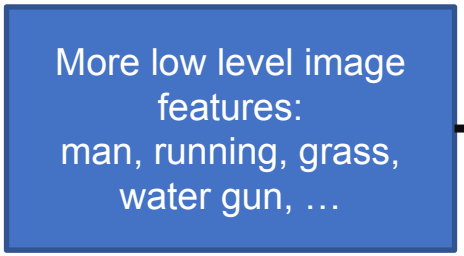


More low-level image features:
man, running, grass, water gun, …

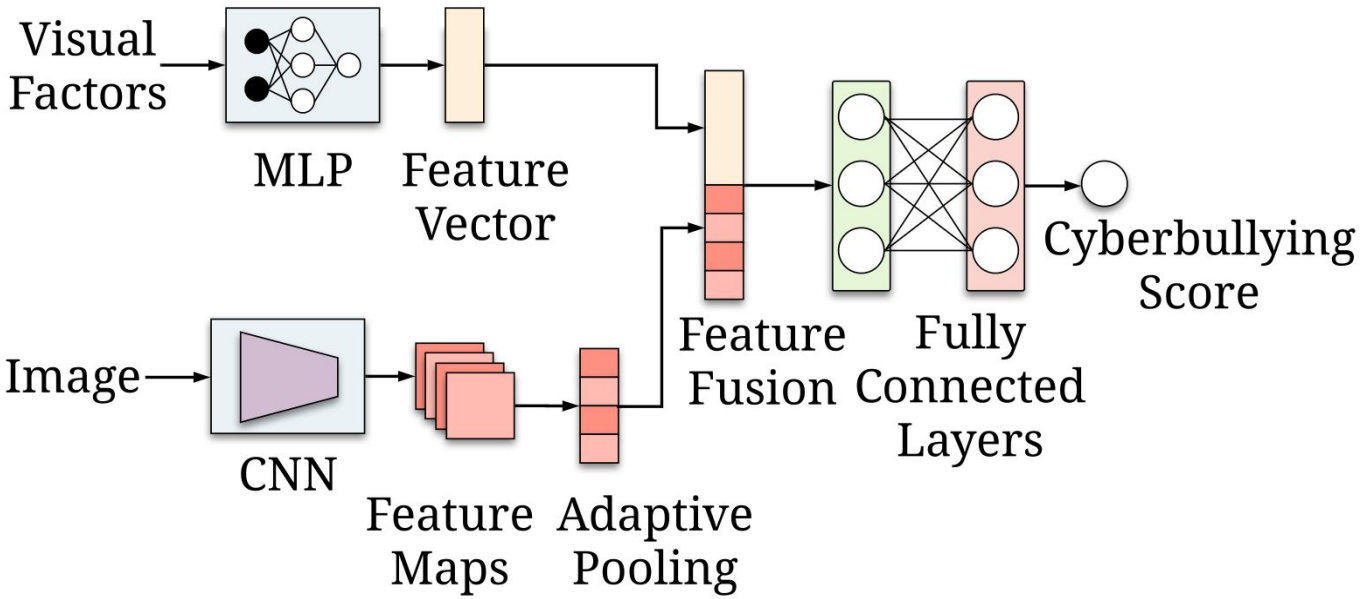
Identify cyberbullying in images







Working approach of fused model



The multimodal model used in the approach

More details can be found in 2021 NDSS paper: Towards Understanding and Detecting Cyberbullying in Real-world Images [link]

Capabilities of the model

- Understanding the effectiveness of factors of cyberbullying in images by using exploratory factors analysis
- Demonstrating the effectiveness of our factors in accurately predicting cyberbullying in images, using classifier model.
- Evaluating the false positives of our model on the images depicting the American Sign Language.
- Validation of our cyberbullying factors with a wider audience.
- Analyzing the capabilities of the state-of-the-art offensive image detectors with respect to the cyberbullying factors.

Evaluation of AI Model

- Accuracy

$$Accuracy = \frac{Number\ of\ correct\ prediction}{Number\ of\ all\ prediction}$$

Is accuracy a satisfactory evaluation metric?

Evaluation of AI Model

- Accuracy
  - How about the dataset is not “balanced”, e.g., 99% of the data is “non-cyberbullying”

- Can we say that the model is good at detecting "cyberbullying" samples?

Evaluation of AI Model

True Positive: ○ **Reality: Cyberbullying** ○ **Model Prediction: Cyberbullying**	**False Positive:** ○ **Reality: Non-cyberbullying** ○ **Model Prediction: Cyberbullying**
False Negative: ○ **Reality: Cyberbullying** ○ **Model Prediction: Non-cyberbullying**	**True Negative:** ○ **Reality: Non-cyberbullying** ○ **Model Prediction: Non-cyberbullying**

Accuracy, Precision, Recall, and F1 score

- Accuracy

$$Accuracy = \frac{Number\ of\ correct\ prediction}{Number\ of\ all\ prediction}$$

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

Accuracy, Precision, Recall, and F1 score

- Precision

$$Precision = \frac{\#\ correct\ predicted\ positive\ samples}{\#\ all\ samples\ predicted\ as\ positive}$$

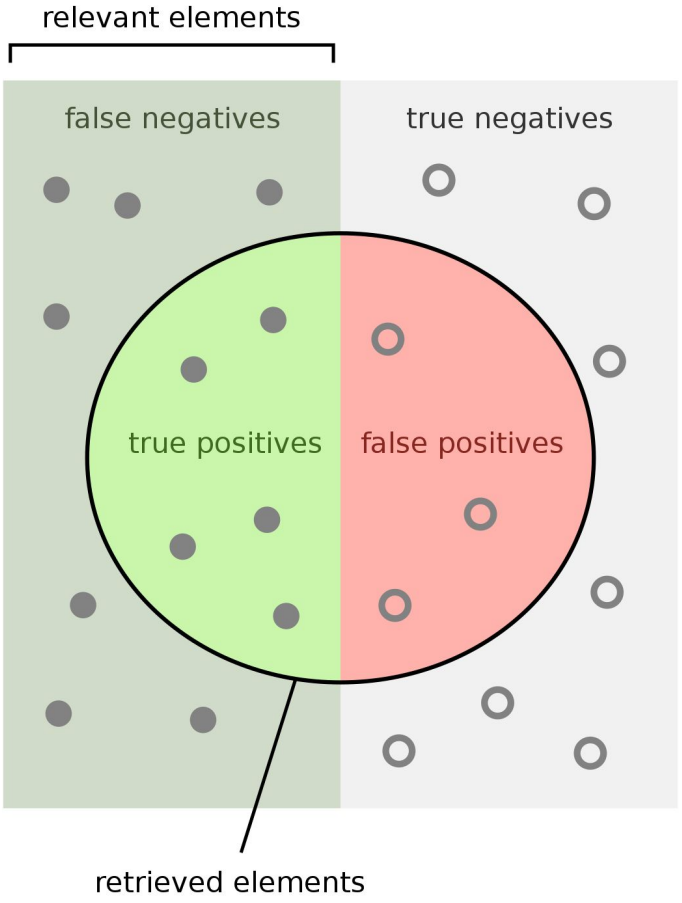
$$= \frac{TP}{TP + FP}$$

Accuracy, Precision, Recall, and F1 score

- Recall

$$Recall = \frac{\#\ correct\ predicted\ positive\ samples}{\#\ all\ positive\ samples}$$

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



How many retrieved items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{retrieved elements}}$$

How many relevant items are retrieved?

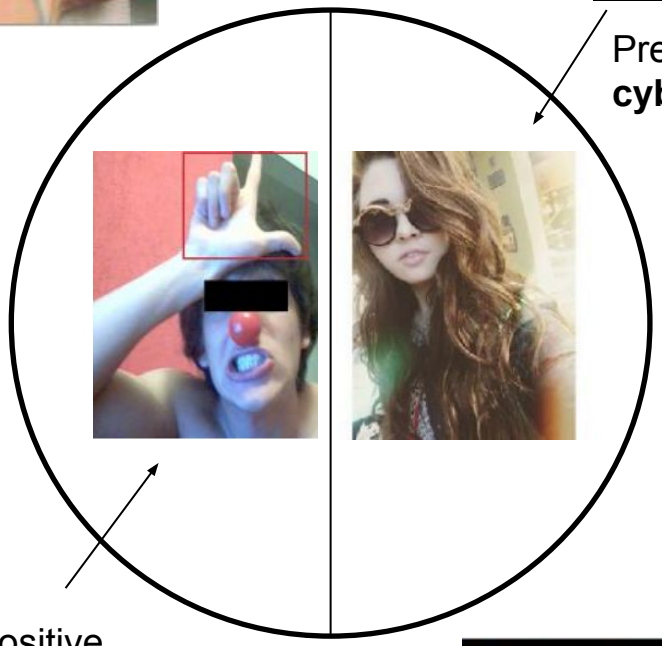
$$\text{Recall} = \frac{\text{true positives}}{\text{relevant elements}}$$

False Negative

Prediction:
non-cyberbullying

False Positive

Prediction:
cyberbullying



True Positive

Prediction:
cyberbullying

True Negative

Prediction:
non-cyberbullying

Accuracy, Precision, Recall, and F1 score

- F1 Score

$$F1\ Score = \frac{2\ *(Precision * Recall)}{Precision + Recall}$$

  - A good evaluation metric can work both on **balanced** and **imbalanced** datasets

Experiment

- Let’s jump into our Lab2
- https://colab.research.google.com/github/cuadvancelab/cuadvancelab.github.io/blob/main/instructions/lab2/computer-science/lab2_interactive_cs.ipynb

Questions

- Answer the following question in the chat
  - What other gestures you think can be taken in account to find cyberbullying?

Q & A