Cyberbullying Detection Using Images

Keyan Guo

University at Buffalo

Outline

- Cyberbullying in Image
- Identify cyberbullying through images
- Factors in Cyberbullying images
- Approach in AI Development
- Working approach of Pre-trained model
- Evaluation of AI Model
- Q&A

Cyberbullying in Images

- Threatening images like these can be sent to a victim to intimidate.
- Detecting such content helps in preventing negative health effects on victim

Cyberbullying in Images

Many companies are trying to solve this problem but they failed by limitations

Google

[ ] DeepAI

clarifai

amazon Rekognition

Identify cyberbullying through images

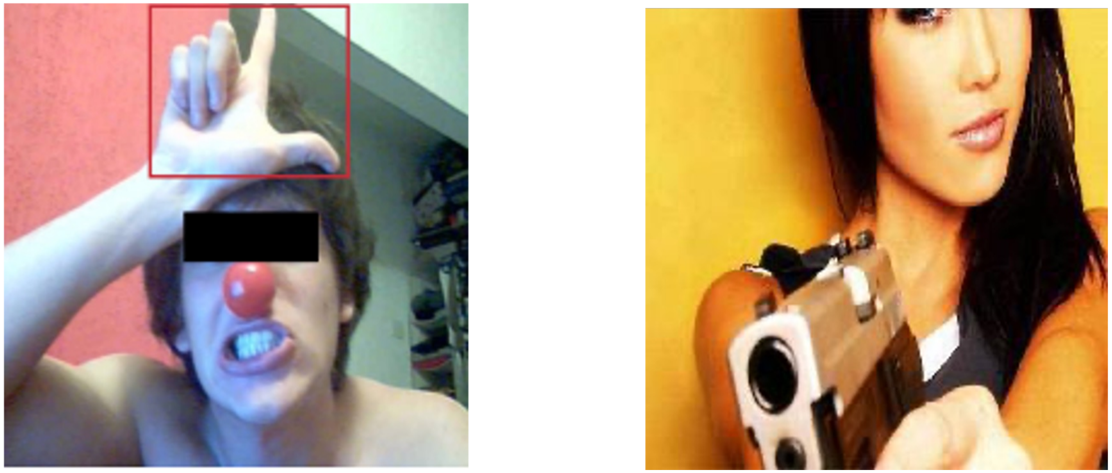


Image with cyberbullying context

Factors in Cyberbullying images

5 Factors to identify cyberbullying in image:

- Body-pose Factor

  → Body Orientation

- Facial Emotion Factor

  → Joy, sorrow, anger, surprise, …

- Gesture Factor

  → Loser, middle finger, thumbs down, gun, …

- Object Factor

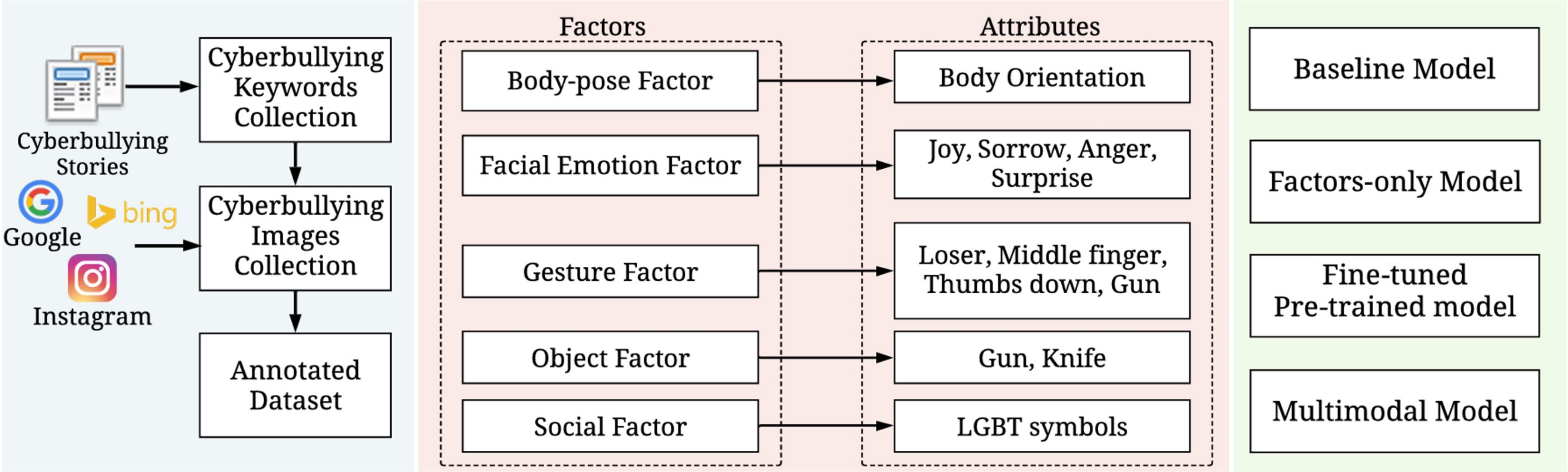
  → Gun, knife, …

- Social Factor

  → LGBT symbols, …

Approach in AI Development

Overview of Approach



Cyberbullying Stories

Google

bing

Instagram

Cyberbullying Keywords Collection

Cyberbullying Images Collection

Annotated Dataset

Factors

Attributes

Body-pose Factor

Facial Emotion Factor

Gesture Factor

Object Factor

Social Factor

Body Orientation

Joy, Sorrow, Anger, Surprise

Loser, Middle finger, Thumbs down, Gun

Gun, Knife

LGBT symbols

Baseline Model

Factors-only Model

Fine-tuned Pre-trained model

Multimodal Model

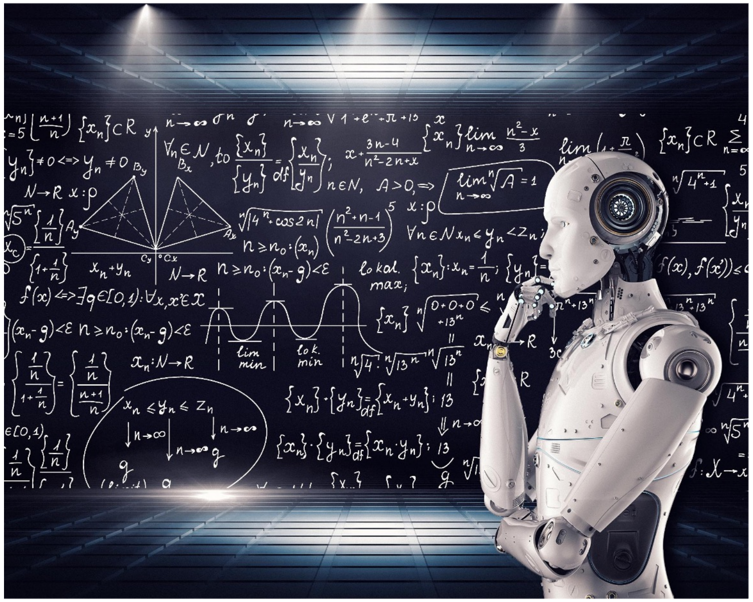
1. Data collection and annotation

2. Factor identification and extraction

3. Classifier model measurement

Working approach of Pre-trained model

AI model was trained well for specific tasks and ready to be deployed…



Evaluation of AI Model

Accuracy

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

Is accuracy a satisfactory evaluation method?

Evaluation of AI Model

Accuracy

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

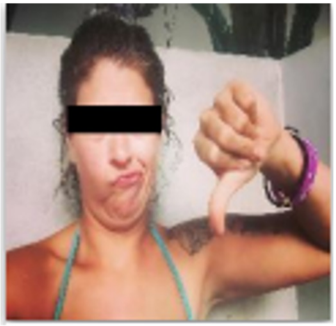
- Can we say that the model is good at detecting "cyberbullying" samples?

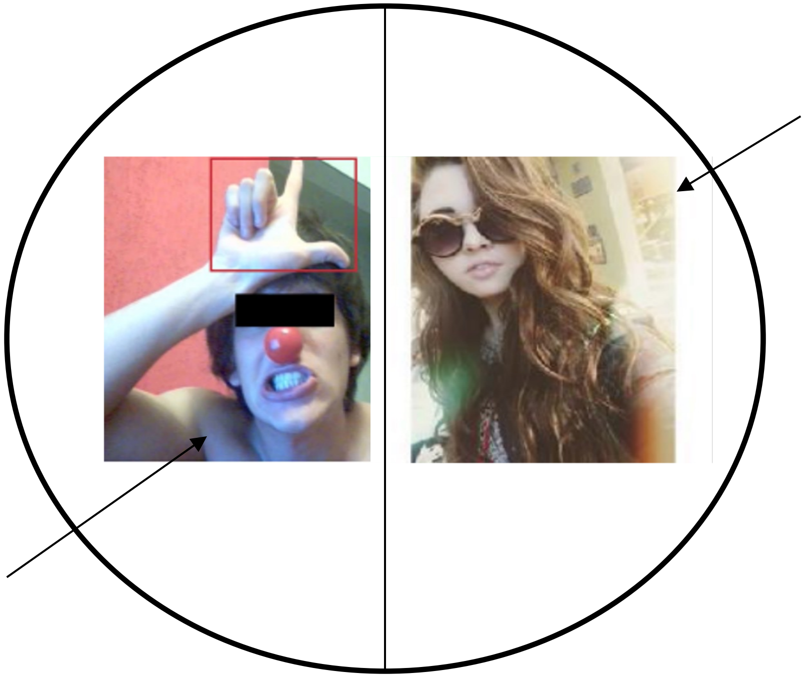
Evaluation of AI Model

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

False Negative

Prediction:
non-cyberbullying





False Positive

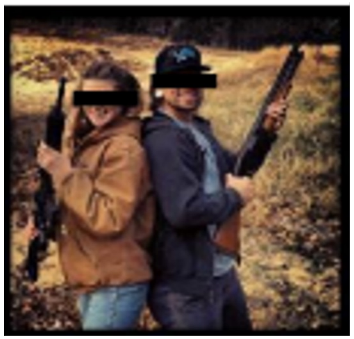
Prediction:
cyberbullying

True Positive

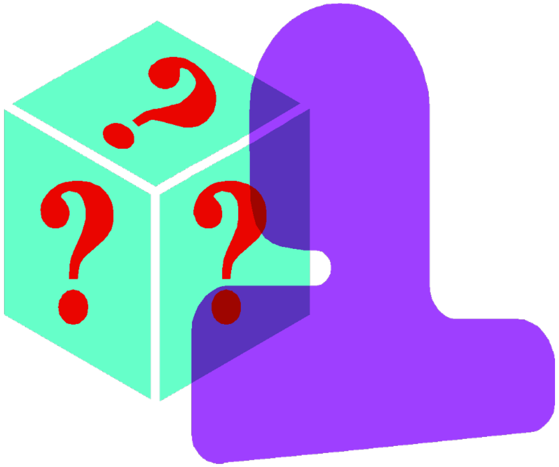
Prediction:
cyberbullying

True Negative

Prediction:
non-cyberbullying



Q & A



Experiment

Let’s jump into our Lab2

https://colab.research.google.com/github/cuadvancelab/cuadvancelab.github.io/blob/main/instructions/lab2/social-science/lab2_interactive_non_cs.ipynb