Credit: Pxhere.com

Trigger Warning:

Blocking Offensive Online Comments

Kyle Hayes

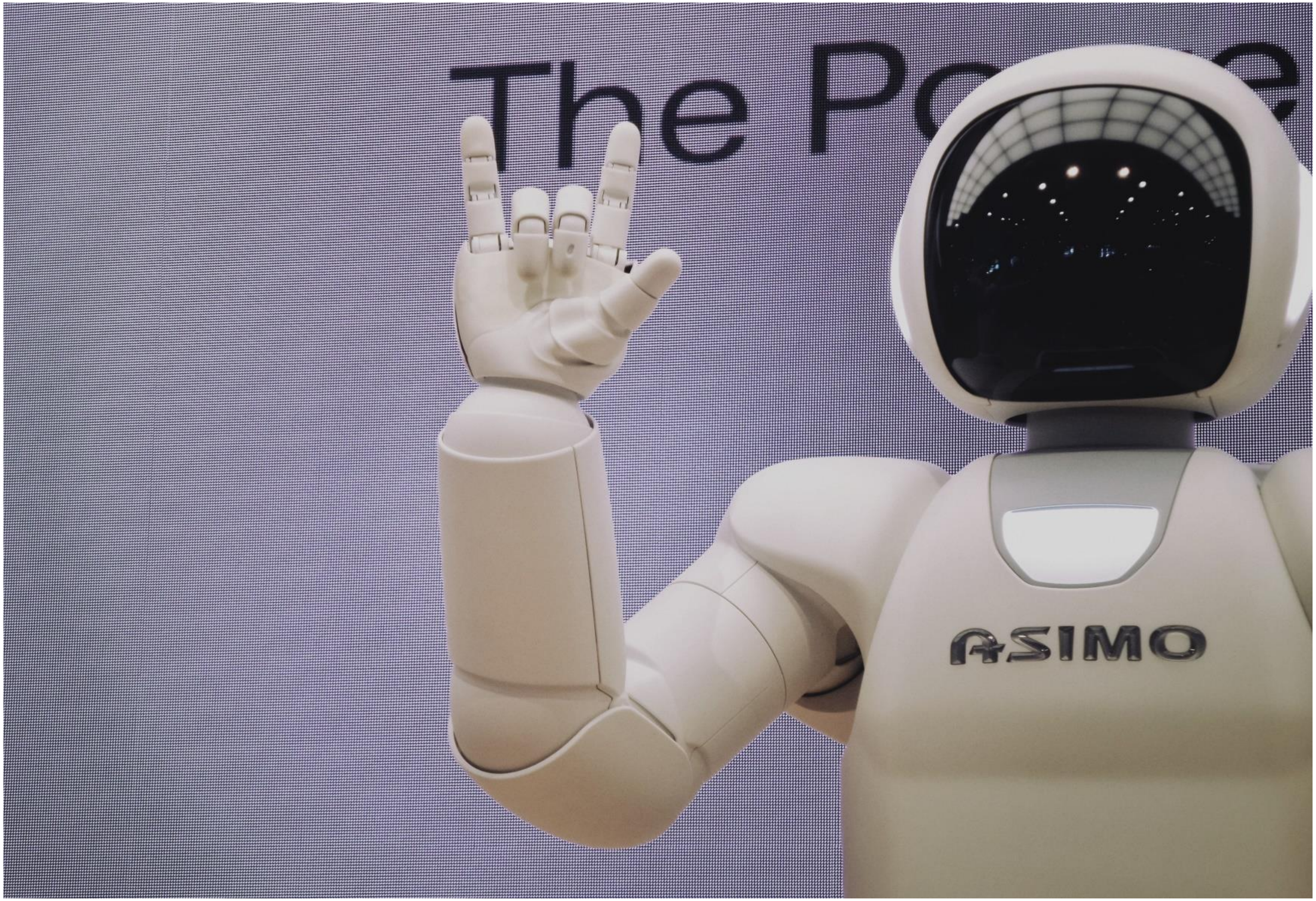
This presentation will address:

- Comment remover problems
- Comment detector criteria
- Data exploration
- Results
- Recommendations

Offensive Comment Removal

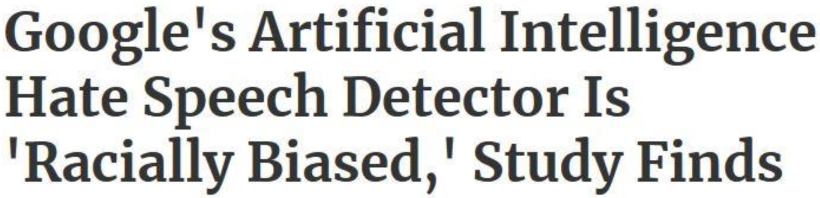
- 510,000 Facebook comments per minute
  (Forbes 2018/5/21)
- 50% of businesses use online communities
  (Greenbook Research Industry Trends, 2015)
- 37% of Americans have been harassed online
  (USA Today, 2019/2/13)
- Average comment moderator: 900 per day
  (Buzzfeed, 2019/3/4)

AI and hate speech: what could possibly go wrong?

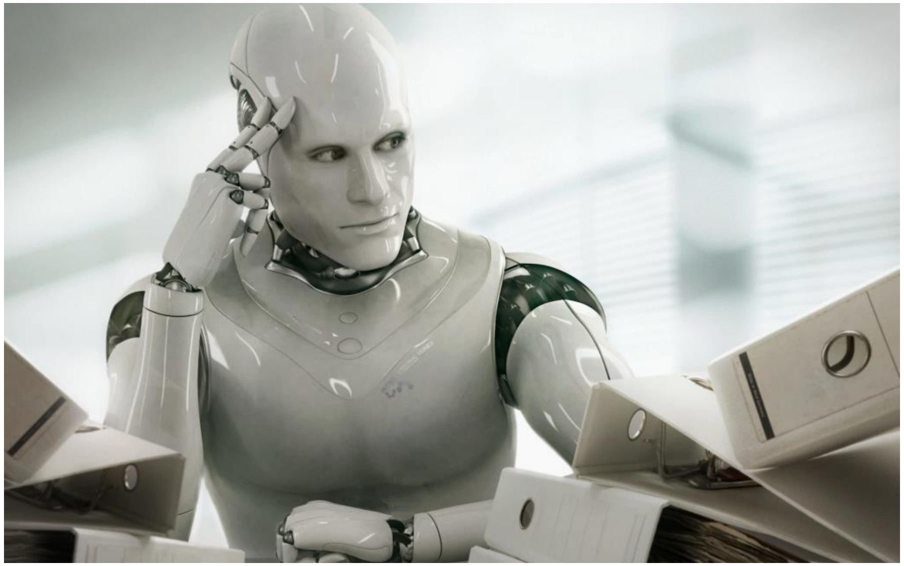


Credit: Unsplash.com

Oh, right.



Google's Artificial Intelligence Hate Speech Detector Is 'Racially Biased,' Study Finds



Credit: BlutGruppe

- 2x more likely to flag African-Americans posters *(Forbes, 2019/8/13)*
- Speech about groups v. hate speech
- Nuance-free

Ideal detector:

Shouldn’t:

- Be ‘color-blind’
- Measure only correct assignations
- Err on the ‘block’ side

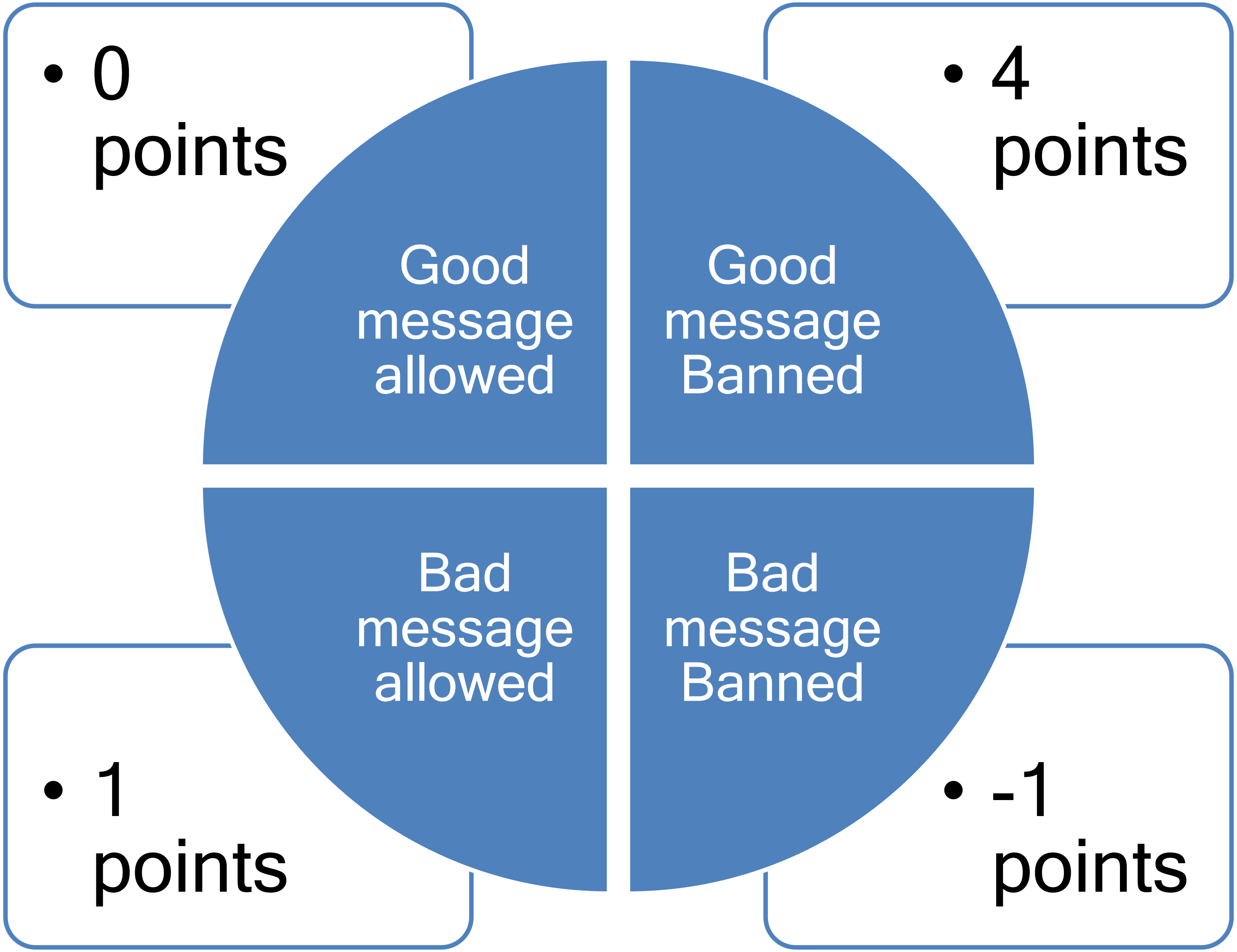
✗

Should:

- Notice ‘group’ content
- Measure incorrect assignations
- Minimize incorrect blocks

✓

Cost Function



- 0 points

Good message allowed

- 4 points

Good message Banned

Bad message allowed

- 1 points

Bad message Banned

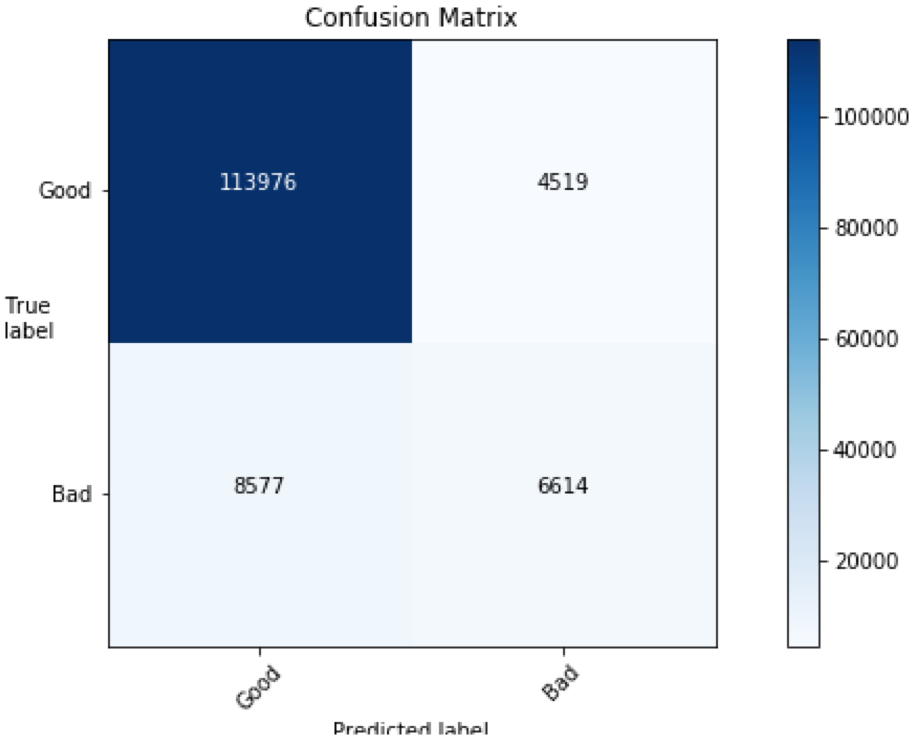
- -1 points

Data Exploration:

- Google Jigsaw: AI and Online Harassment
- 400,000 Quora comments
- Annotated toxicity and ‘group’ subject labels

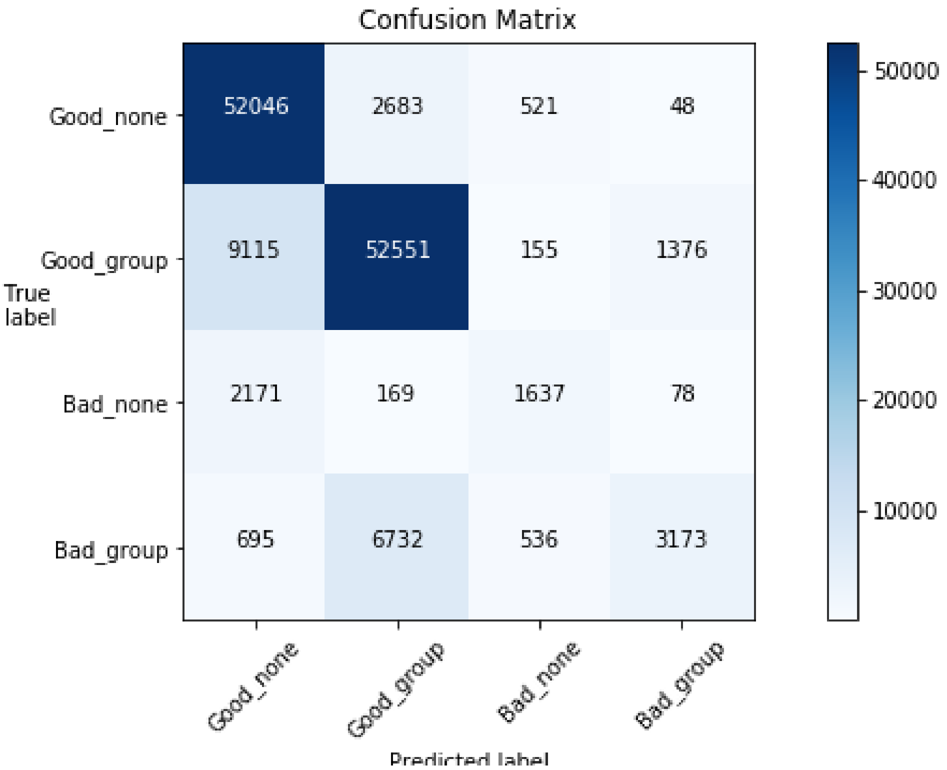
No group classification

- Correct bans: 6614
- Good comments banned: 4519
- Bad comments missed: 8577
- Score: 20,039



Results

Grouped

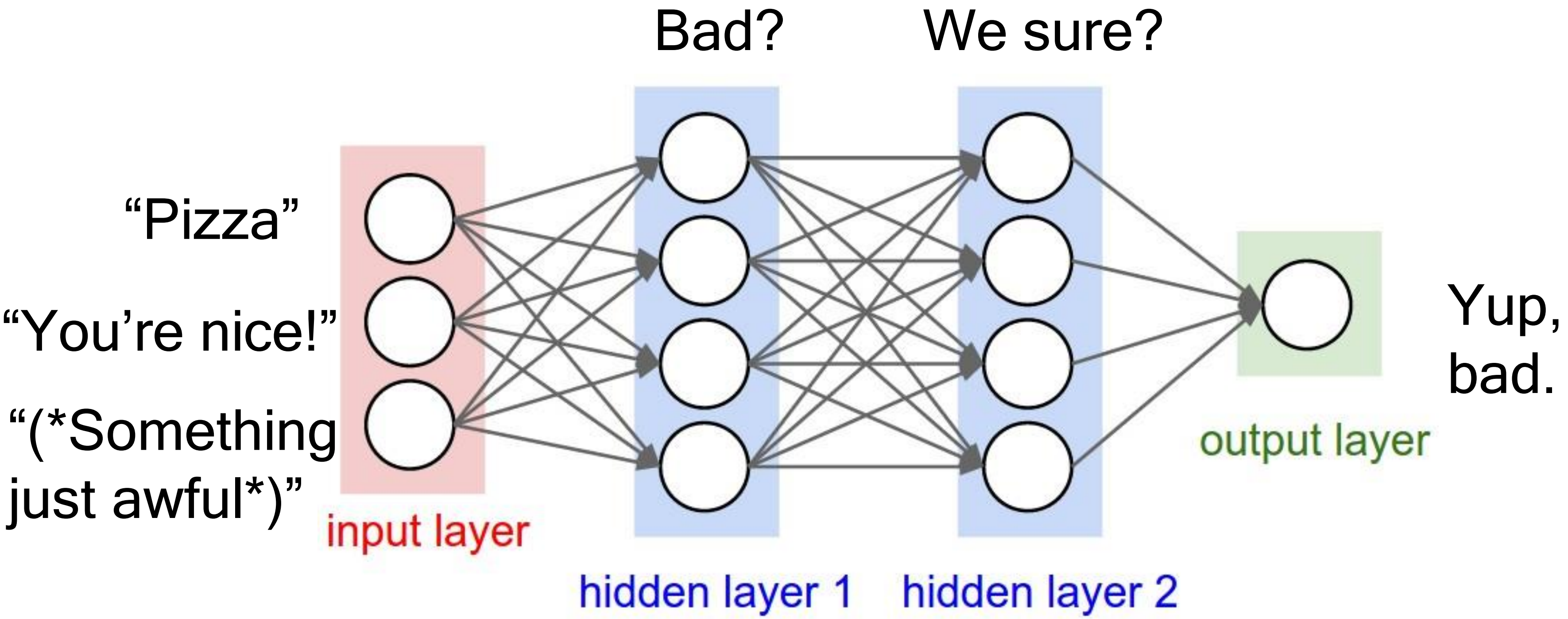


True label \ Predicted label	Good_none	Good_group	Bad_none	Bad_group
Good_none	52046	2683	521	48
Good_group	9115	52551	155	1376
Bad_none	2171	169	1637	78
Bad_group	695	6732	536	3173

Grouped (simplified)

- Good comments banned: 53.5% less
- Bad comments missed: 13.9% more
- Correct bans: 18% less
- Score: 12,743

Deep Learning Network



Problems with classification:

Not Toxic:

- **“This nation is afflicted with an epidemic of Black-on-White… violence…** unreported by the mainstream media.”

Toxic:

- “(I don’t see) racial discrimination in favor of black people… **Do you have any studies that don't include cherry-picked items?“**

Problems with classification:

Not a group comment

- "I think **Native Americans** should get a pass for being suspicious of… this government."

Group comment

- "Haha *[sic]* you guys are a bunch of **losers**."

Findings

- Offensive speech categorization decreases false bans significantly
- Deep learning decreases levels of false bans
- Anti-group speech is harder to recognize and to correctly classify

Recommendations

- Use together with human process
- Introduce Convolutional Neural Network
- Use unsupervised learning to investigate incorrect assignments
- Use diverse group of annotators to better define offensiveness