

Computational Ethics

Yulia Tsvetkov

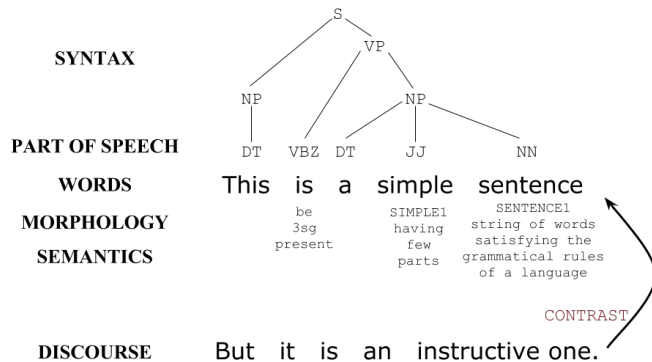
ytsvetko@cs.cmu.edu



Carnegie Mellon University

Language Technologies Institute

What do Language Technologies have to do with Ethics?



Applications

- Machine Translation
- Information Retrieval
- Question Answering
- Dialogue Systems
- Information Extraction
- Summarization
- Sentiment Analysis
- ...

Core technologies

- Language modelling
- Part-of-speech tagging
- Syntactic parsing
- Named-entity recognition
- Coreference resolution
- Word sense disambiguation
- Semantic Role Labelling
- ...

Language and People

The common misconception is that language has to do with **words** and what they mean.

It doesn't.

It has to do with **people** and what ***they*** mean.

Herbert H. Clark & Michael F. Schober, 1992

Decisions we make about our data, methods, and tools are tied up with their impact on people and societies.

What is Ethics?

“Ethics is a study of what are **good and bad** ends to pursue in life and what it is **right and wrong** to do in the conduct of life.

It is therefore, above all, a **practical discipline**.

Its primary aim is to determine how one ought to live and what actions one ought to do in the conduct of one’s life.”

-- Introduction to Ethics, John Deigh

What is Ethics?

It's the **good** things

It's the **right** things



What is Ethics?

It's the **good** things

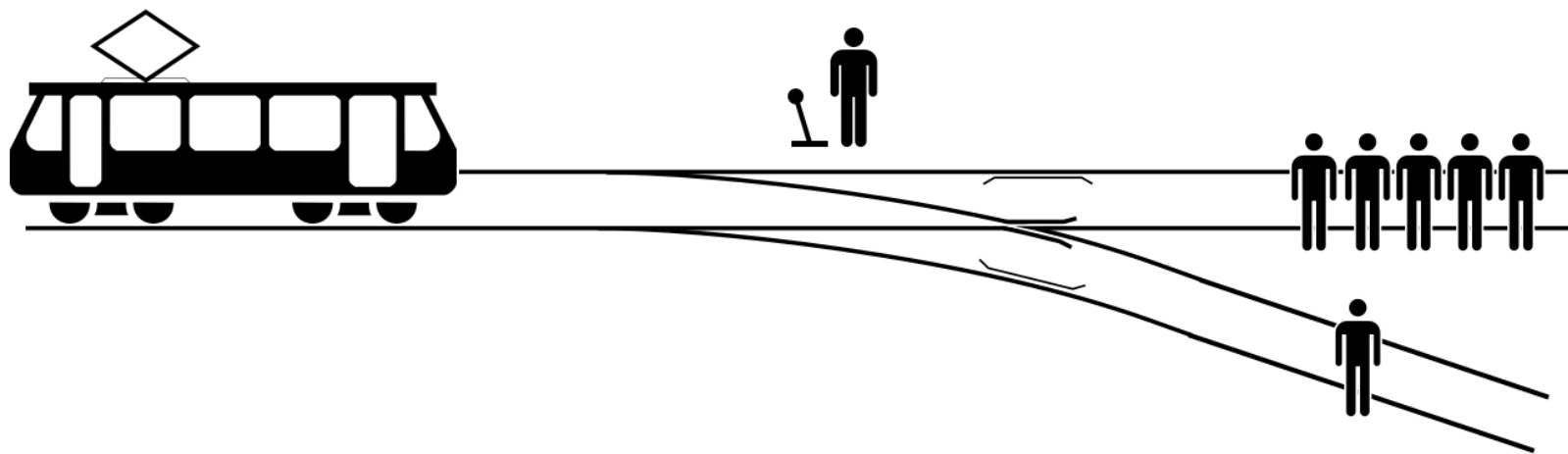
It's the **right** things

How simple is it to define
what's good and what's right?



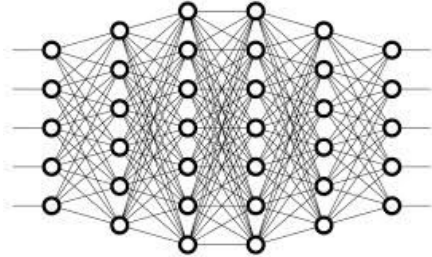
The Trolley Dilemma

Should you pull the lever to divert the trolley?



[From Wikipedia]

Let's Train a Chicken Classifier



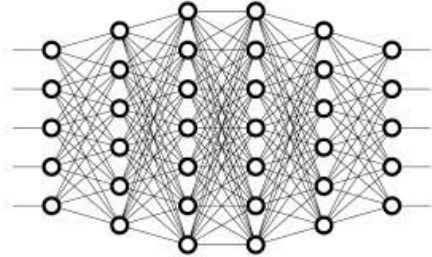
rooster



hen



Let's Train a Chicken Classifier



rooster

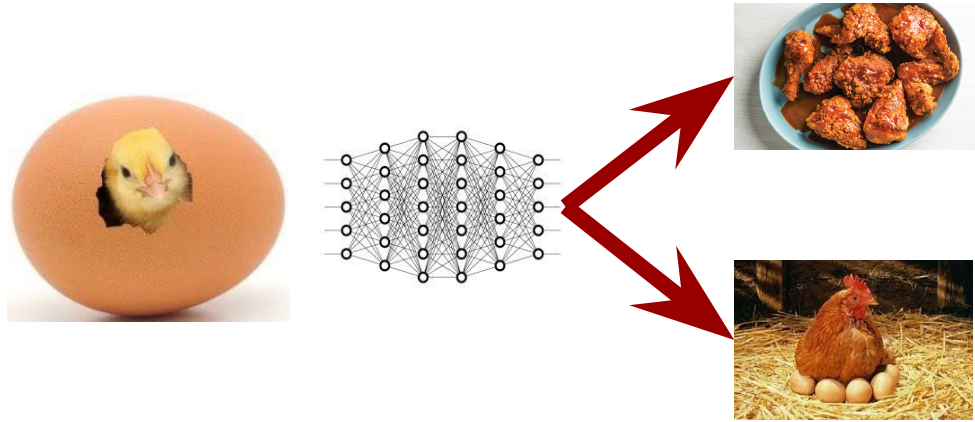


hen



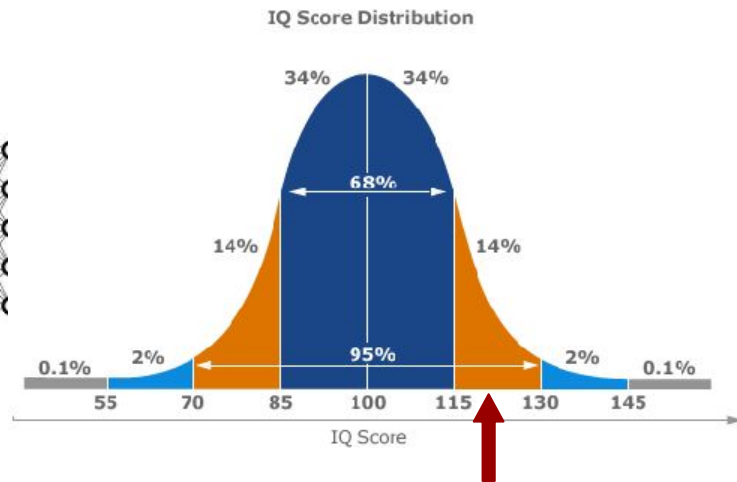
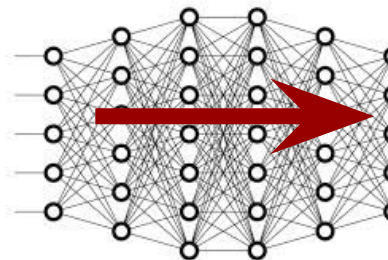
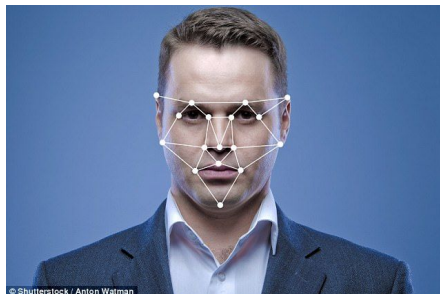
Ethical?

Let's Train a Chicken Classifier



- Ethics is inner guiding, moral principles, and values of people and society
- Ethics is not “black and white”, there are grey areas.
We often don't have binary answers.
- Ethics changes over time with values and beliefs of people
- Legal ≠ Ethical

Let's Train an IQ Classifier



- **Intelligence Quotient:** a number used to express the apparent relative intelligence of a person

An IQ Classifier

Let's train a classifier to predict people's IQ from their photos.

- Who could benefit from such a classifier?

An IQ Classifier

Let's train a classifier to predict people's IQ from their photos.

- Who could benefit from such a classifier?
- Assume the classifier is 100% accurate. Who can be harmed from such a classifier? How can such a classifier be misused?



An IQ Classifier

Let's train a classifier to predict people's IQ from their photos.

- Who could benefit from such a classifier?
- Who can be harmed by such a classifier?
- Our test results show 90% accuracy
 - We found out that white females have 95% accuracy
 - People with blond hair under age of 25 have only 60% accuracy

An IQ Classifier

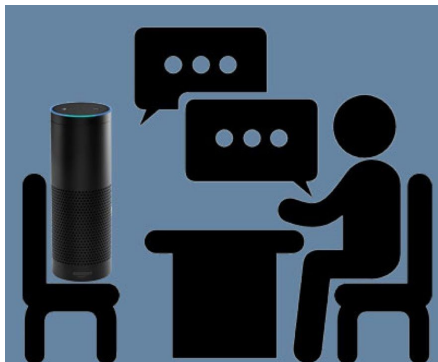
Let's train a classifier to predict people's IQ from their photos.

- Who could benefit from such a classifier?
- Who can be harmed by such a classifier?
- Our test results show 90% accuracy
 - We found out that white females have 95% accuracy
 - People with blond hair under age of 25 have only 60% accuracy
- Who is responsible?
 - Researcher/developer? Reviewer? University? Society?

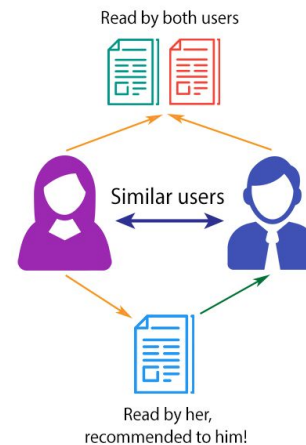
What's the Difference?



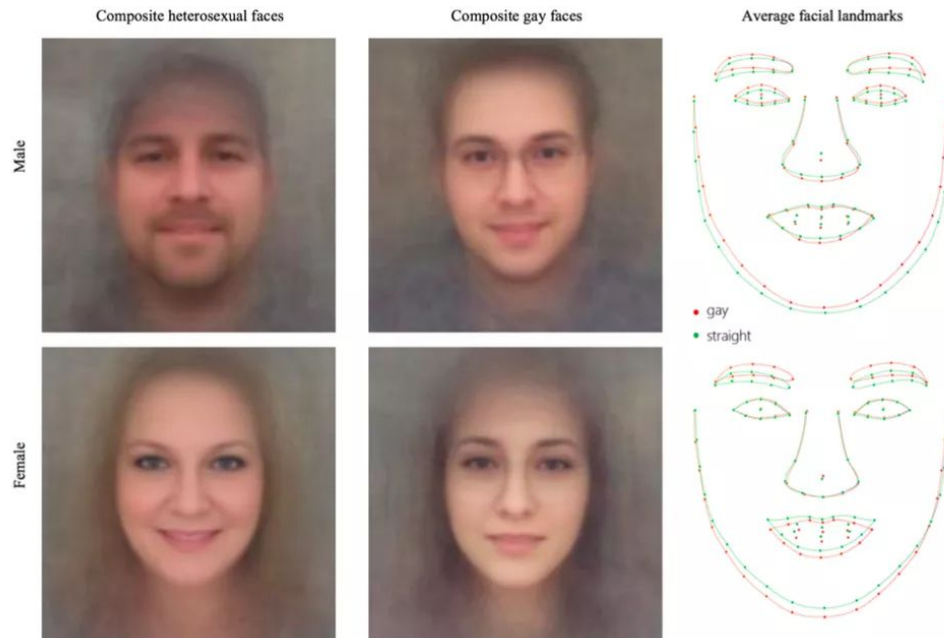
AI and People



PAROLE



A Recent Study: the “A.I. Gaydar”



A Case Study: the “A.I. Gaydar”

Abstract. We show that faces contain much more information about sexual orientation than can be perceived and interpreted by the human brain. We used deep neural networks to extract features from 35,326 facial images. These features were entered into a logistic regression aimed at classifying sexual orientation. Given a single facial image, a classifier could correctly distinguish between gay and heterosexual men in 81% of cases, and in 74% of cases for women. Human judges achieved much lower accuracy: 61% for men and 54% for women. The accuracy of the algorithm increased to 91% and 83%, respectively, given five facial images per person. Facial features employed by the classifier included both fixed (e.g., nose shape) and transient facial features (e.g., grooming style). Consistent with the prenatal hormone theory of sexual orientation, gay men and women tended to have gender-atypical facial morphology, expression, and grooming styles. Prediction models aimed at gender alone allowed for detecting gay males with 57% accuracy and gay females with 58% accuracy. Those findings advance our understanding of the origins of sexual orientation and the limits of human perception. Additionally, given that companies and governments are increasingly using computer vision algorithms to detect people’s intimate traits, our findings expose a threat to the privacy and safety of gay men and women.

Wang & Kosinski. **Deep neural networks are more accurate than humans at detecting sexual orientation from facial images.** *Journal of Personality and Social Psychology (in press)*. September 7,

2017



A Case Study: the “A.I. Gaydar”

- Research question
 - Identification of sexual orientation from facial features
- Data collection
 - Photos downloaded from a popular American dating website
 - 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly
- Method
 - A deep learning model was used to extract facial features + grooming features; then a logistic regression classifier was applied for classification
- Accuracy
 - 81% for men, 74% for women

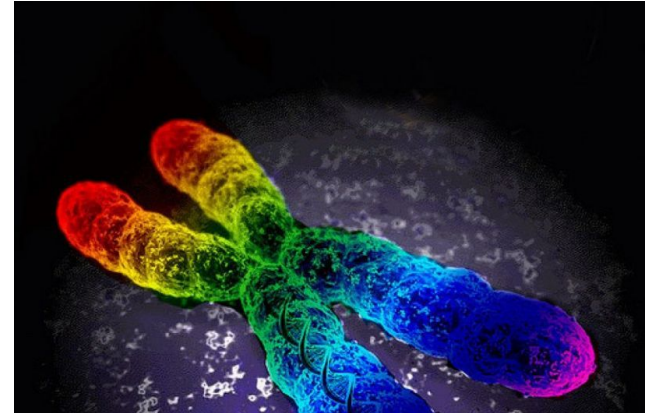
Let's Discuss...

- Research question
 - Identification of sexual orientation from facial features
- Data collection
 - Photos downloaded from a popular American dating website
 - 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented equally
- Method
 - A deep learning model was used to extract facial features + grooming features; then a logistic regression classifier was applied for classification
- Accuracy
 - 81% for men, 74% for women

What went wrong?

Research Question

- Identification of sexual orientation from facial features



Research Question

- Identification of sexual orientation from facial features

How people can be harmed by this research?

- In many countries being gay person is prosecutable (by law or by society) and in some places there is even death penalty for it
- It might affect people's employment; family relationships; health care opportunities;
- Attributes like gender, race, sexual orientation, religion are social constructs. Some may change over time. They can be non-binary. They are private, intimate, often not visible publicly.
- Importantly, these are properties for which people are often discriminated against.



Research Question

“... Additionally, given that companies and governments are increasingly using computer vision algorithms to detect people’s intimate traits, our findings expose a threat to the privacy and safety of gay men and women.”

→ your thoughts on this?

Data

- Photos downloaded from a popular American dating website
- 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly



Data & Privacy

- Photos downloaded from a popular American dating website
- 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly

Legal ≠ Ethical

Public ≠ Publicized

Did these people agree to participate in the study?

→ Violation of social contract

Data

- Photos downloaded from a popular American dating website
- 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly



Data & Bias

- Photos downloaded from a popular American dating website
- 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly

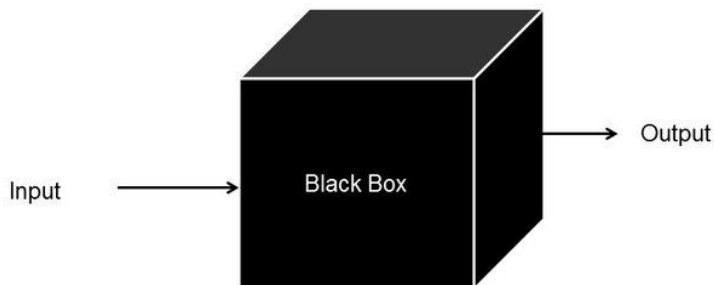
Only white people, who self-disclose their orientation, certain social groups, certain age groups, certain time range/fashion;

the photos were carefully selected by subjects to be attractive so there is even self-selection bias...

The dataset is balanced, which does not represent true class distribution.

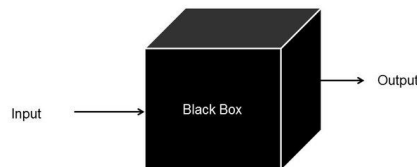
Method

- A deep learning model was used to extract facial features + grooming features; then a logistic regression classifier was applied for classification



Method & Human Biases in Models + Interpretability

- A deep learning model was used to extract facial features + grooming features; then a logistic regression classifier was applied for classification



- **can we use not interpretable models when we make predictions about sensitive attributes, about complex experimental conditions that require broader world knowledge?**
- **how to analyze errors and bias amplification?**

Evaluation

- Accuracy: 81% for men, 74% for women



The Cost of Misclassification and The Importance of Social Context



The Cost of Misclassification and The Importance of Social Context



Dual Use Dual Framing



“We live in a dangerous world, where harm doers and criminals easily mingle with the general population; the vast majority of them are unknown to the authorities.

As a result, it is becoming ever more challenging to detect anonymous threats in public places such as airports, train stations, government and public buildings and border control. Public Safety agencies, city police department, smart city service providers and other law enforcement entities are increasingly strive for Predictive Screening solutions, that can monitor, prevent, and forecast criminal events and public disorder without direct investigation or innocent people interrogations. “

Learn to Assess AI Systems Adversarially

- Ethics of the research question
- Impact: Who could benefit from such a technology? Who can be harmed by such a technology? Could sharing this data have major effect on people's lives?
- Privacy: Who owns the data? Published vs. publicized? User consent and implicit assumptions of users how the data will be used.
- Bias in data: Artifacts in data, population-specific distributions, representativeness of data.
- Algorithmic bias: How to control for confounding variables and corner cases? Does the system optimize for the “right” objective? Does the system amplify bias?
- Utility-based evaluation beyond accuracy: FP & FN rates, “the cost” of misclassification, fault tolerance.

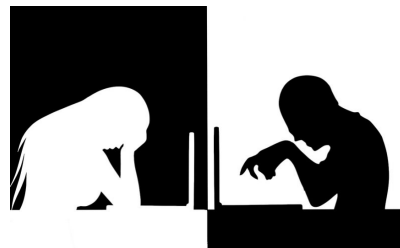
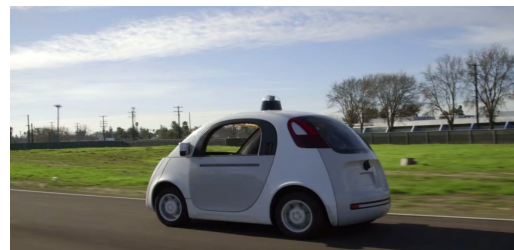


The Dual Use of A.I. Technologies

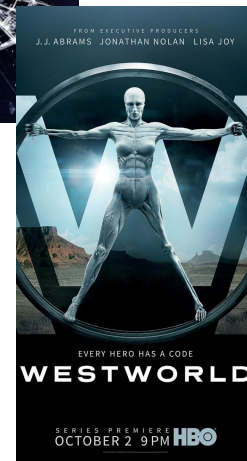
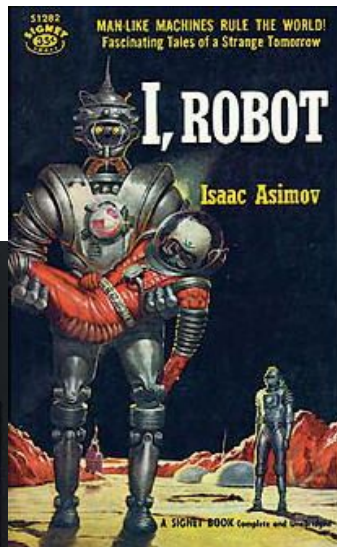
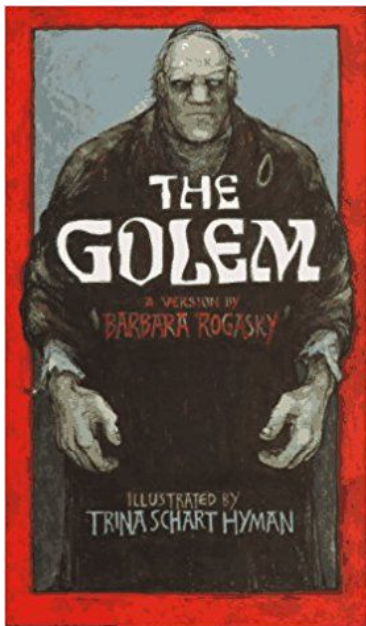
- Who should be responsible?
 - The person who uses the technology?
 - The researcher/developer?
 - Paper reviewers?
 - University?
 - Society as a whole?

We need to be aware of real-world impact of our research and understand the relationship between ideas and consequences

We Cannot Foresee All Possible Uses of Technology



The Long History of Ethics and AI



Eliza

- A rule-based dialogue system which mimicked a Rogerian psychotherapist, built at the MIT Artificial Intelligence Laboratory by Joseph Weizenbaum
- A simple rule-based algorithm that “reflects” what human is saying
- One of the first “AI systems” (1964) that raised ethical questions

```
Human:  Men are all alike.
ELIZA:  In what way?
Human:  They're always bugging us about something.
ELIZA:  Can you think of a special example?
Human:  Well, my boyfriend made me come here.
ELIZA:  Your boyfriend made you come here?
Human:  He says I'm depressed much of the time.
ELIZA:  I'm sorry to hear that you are depressed much of the time.
Human:  It's true.  I'm unhappy.
ELIZA:  Do you think coming here will help you not to be unhappy?
Human:  I need some help, that much seems certain.
```



“Smart” Toys: Barbie

Barbara Grosz, NYT 2015: Barbie Wants to Get to Know Your Child

“Hey, new question,” Barbie said. “Do you have any sisters?”

“Yeah,” Tiara said. “I only have one.”

“What’s something nice that your sister does for you?” Barbie asked.

“She does nothing nice to me,” Tiara said tensely.

Barbie forged ahead. “Well, what is the last nice thing your sister did?”

“She helped me with my project — and then she *destroyed* it.”

“Oh, yeah, tell me more!” Barbie said, oblivious to Tiara’s unhappiness.

“That’s it, Barbie,” Tiara said.

“Have you told your sister lately how cool she is?”

“No. She is *not* cool,” Tiara said, gritting her teeth.

“You never know, she might appreciate hearing it,” Barbie said.



Watch This Talk

“Hey, new question,” Barbie said. “Do you have any sisters?”

“Yeah,” Tiara said. “I only have one.”

Barbara Grosz NYT article: Barbie Wants to Get to Know Your Child

Intelligent Systems: Design & Ethical Challenges

“She does nothing nice to me,” Tiara said tensely.

Barbie forged ahead. “Well, what is the last nice thing your sister did?”

“She helped me with my project – and then she *destroyed* it.”

<https://goo.gl/8tBho8>

“Oh, yeah, tell me more!” Barbie said, oblivious to Tiara’s unhappiness.

“That’s it, Barbie,” Tiara said.

“Have you told your sister lately how cool she is?”

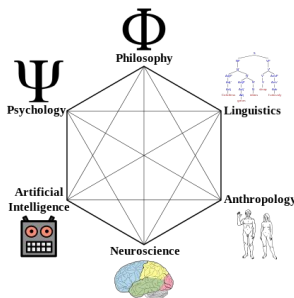
“No. She is *not* cool,” Tiara said, gritting her teeth.

“You never know, she might appreciate hearing it,” Barbie said.



Topics in ethical language technologies

- Philosophical foundations
- Algorithmic bias
- Civility in communication, hate speech
- Privacy and profiling
- The language of manipulation: fake news, propaganda, polarization in online media
- LT for social good



Project example: Veiled aggression on social media



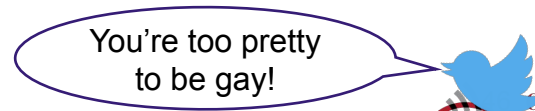
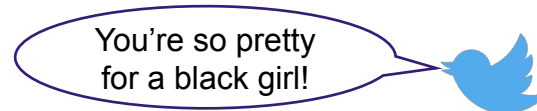
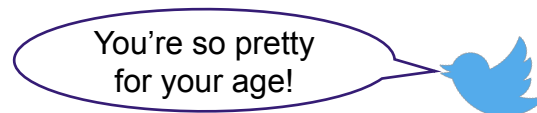
Positive or negative?



Positive or negative?



Positive or negative?



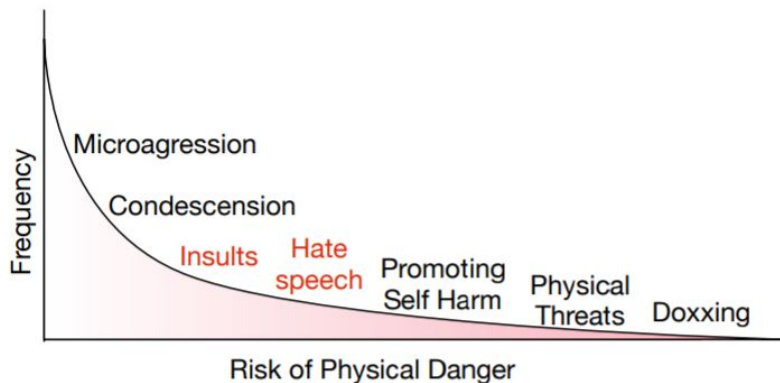
Detection of toxic/offensive/biased comments

- Recent NLP advances have focused on **overt** toxic language (e.g. hate speech)
 - Explicit bias
 - Profane / violent



Toxicity in disguise

- Little focus on **veiled** negativity that is not directly encoded in lexicons
 - subtle toxic language, where bias is implicit and contextual
 - codewords, spelling variations of hate lexicons



What is a microaggression?

“A comment or action that **subtly and often unconsciously or unintentionally** expresses a prejudiced attitude towards a member of a marginalized group”

- Merriam Webster

What is a microaggression?

“A comment or action that **subtly and often unconsciously or unintentionally** expresses a prejudiced attitude towards a member of a marginalized group”

- Merriam Webster

Surface-level sentiment can be negative, neutral, or positive. For example:

- “Girls just **aren’t good** at math.”

What is a microaggression?

“A comment or action that **subtly and often unconsciously or unintentionally** expresses a prejudiced attitude towards a member of a marginalized group”

- Merriam Webster

Surface-level sentiment can be negative, neutral, or positive. For example:

- “Girls just **aren’t good** at math.”
- “Don’t you people **like** tamales?”

What is a microaggression?

“A comment or action that **subtly and often unconsciously or unintentionally** expresses a prejudiced attitude towards a member of a marginalized group”

- Merriam Webster

Surface-level sentiment can be negative, neutral, or positive. For example:

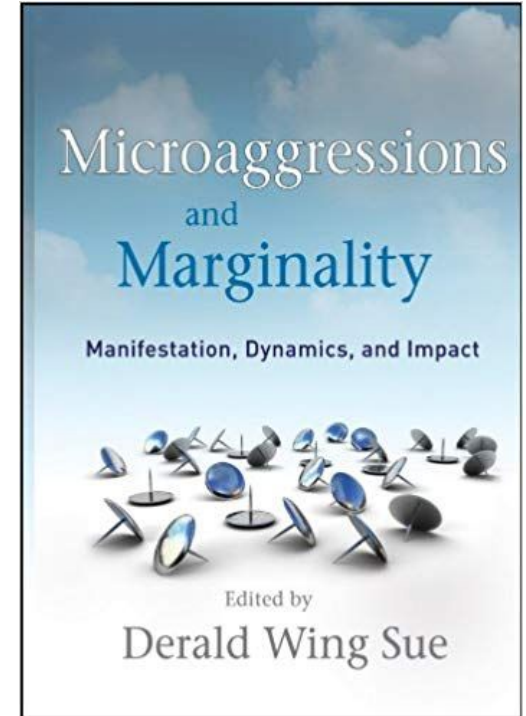
- “Girls just **aren’t good** at math.”
- “Don’t you people **like** tamales?”
- “You’re too **pretty** to be gay.”

microaggressions.com

tumblr.

Microaggressions are harmful

- Effects can be more pernicious than overtly aggressive speech (Sue et al. 2007, Sue 2010, Nadal et al. 2014)
- Can affect people's professional experiences and career trajectories (Cortina et al. 2002, Trix and Psenka 2003)
- Play on, and reinforce, problematic stereotypes and power structures (Hall and Braunwald 1981, Fournier et al. 2002)



SOTA NLP tools cannot identify microaggressions

“I like to imagine you as a girl but your sentence structure and rhetoric is so concise and to the point which points to the contrary (nothing against women, simply factual).”

Hate Speech Detection



Unlikely to be perceived as toxic
(0.23)

Sentiment Analysis



Subjectivity

- neutral: 0.1
- **polar: 0.9**

Polarity

- **pos: 0.51**
- neg: 0.49

The text is **pos**.



- Conversational agents
- Personal assistants
- Medical applications
- Educational applications
- ...



Online data is riddled with **SOCIAL STEREOTYPES**

Today's reactive approach

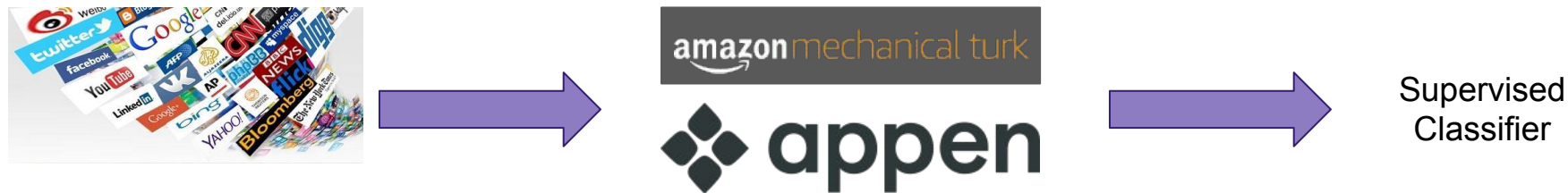
REDDIT GPT-3 REDDIT OPENAI ARTIFICIAL INTELLIGENCE WRITING

Someone let a GPT-3 bot loose on Reddit – it didn't end well

The bot spent more than a week making comments about some seriously sensitive subjects



Naive approach: supervised classification



- Problems:

- Biases are subtle and implicit even experts are bad at identifying them
- We don't have strong lexical sieve to surface candidates for annotation

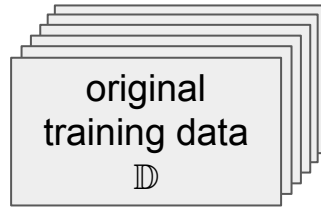
- Unsupervised approach

- Field A. & Tsvetkov Y. (2020) *Unsupervised Discovery of Implicit Gender Bias*. *EMNLP*

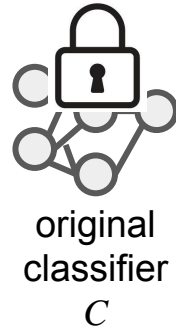
- Surfacing biases via probing & interpretation of model decisions

- Han X., Tsvetkov Y. (2020) *Fortifying Toxic Speech Detectors Against Veiled Toxicity*. *EMNLP*

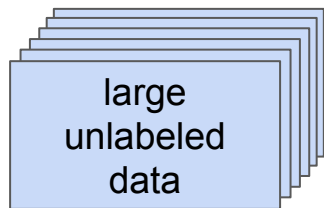
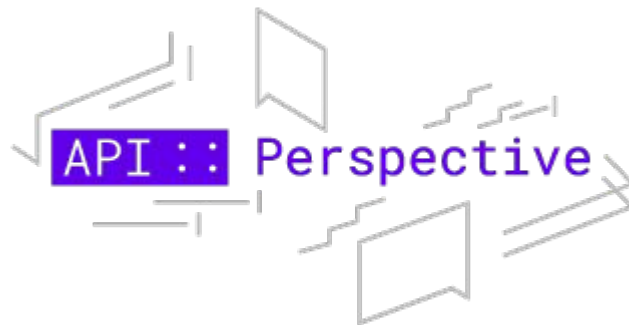
Fortifying Toxicity Classifiers



Fortifying Toxicity Classifiers



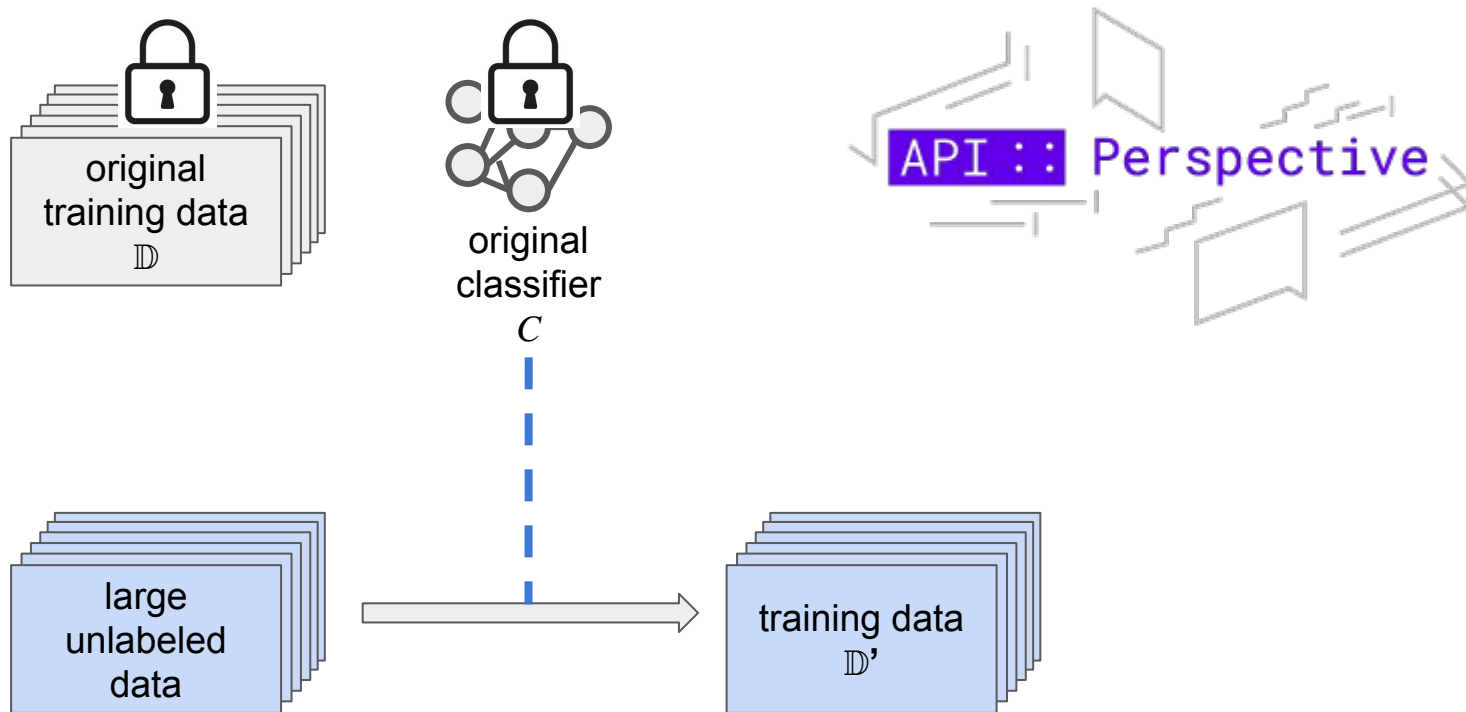
Fortifying Toxicity Classifiers



SBIC



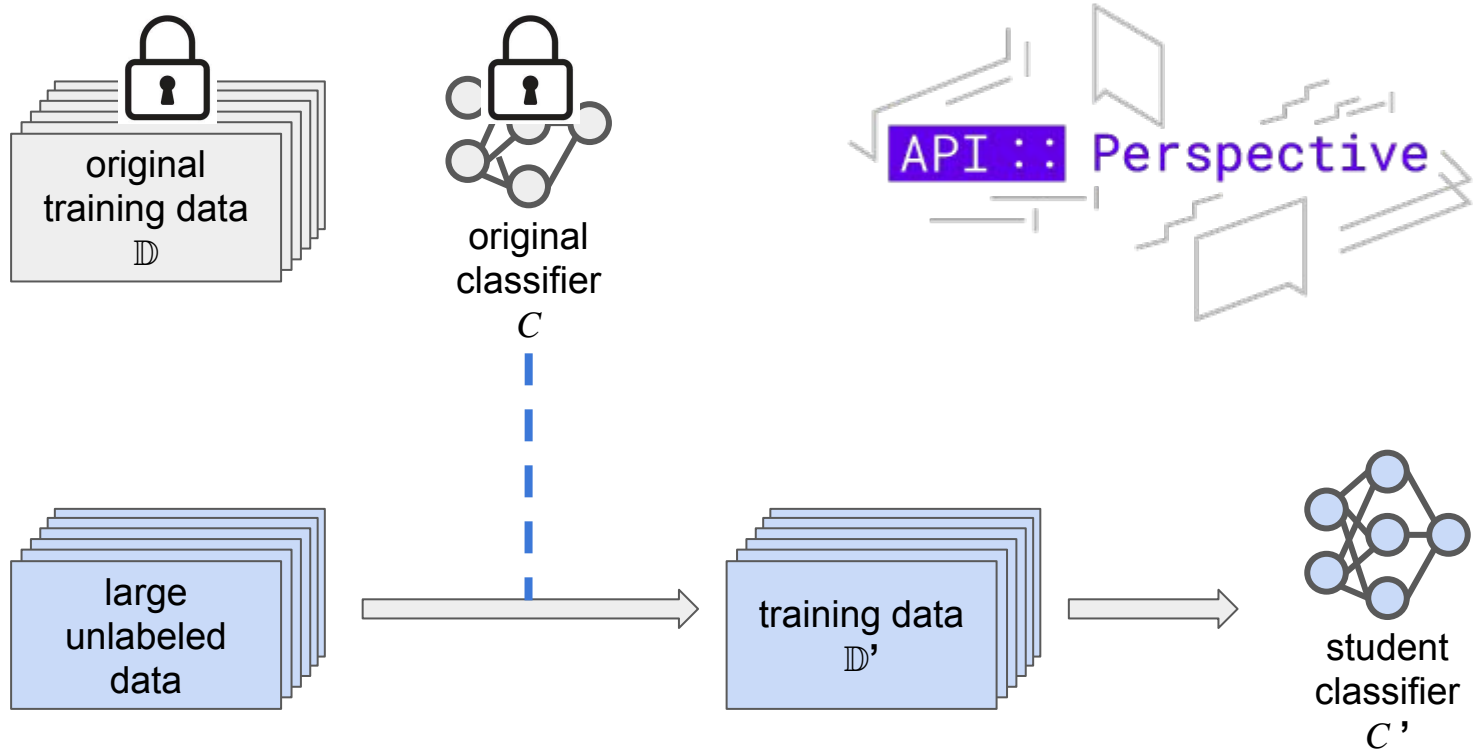
Fortifying Toxicity Classifiers



SBIC



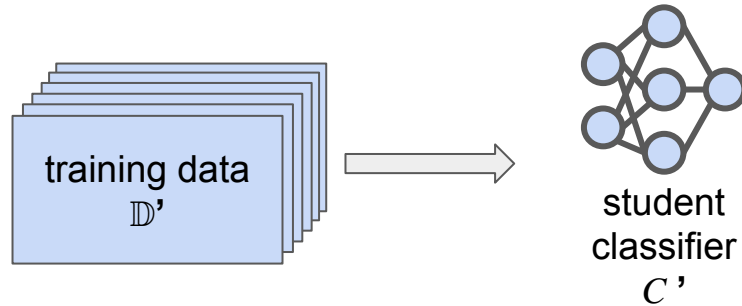
Fortifying Toxicity Classifiers



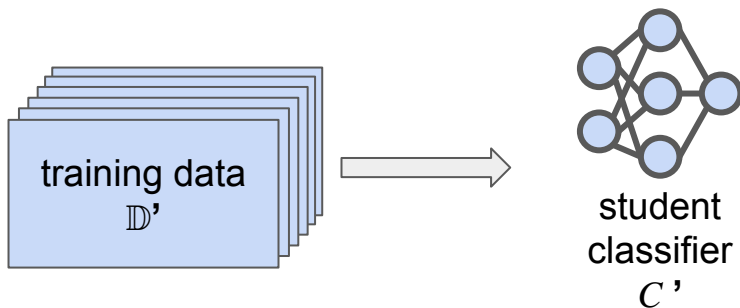
SBIC



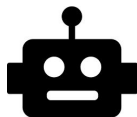
Fortifying Toxicity Classifiers



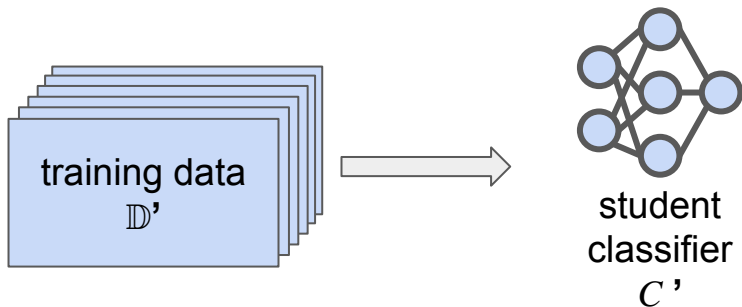
Fortifying Toxicity Classifiers



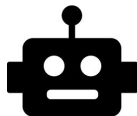
Overt	Offensive	Toxic
Clean	Non-offensive	Non-toxic
Veiled	Offensive	Non-toxic
FP	Non-offensive	Toxic



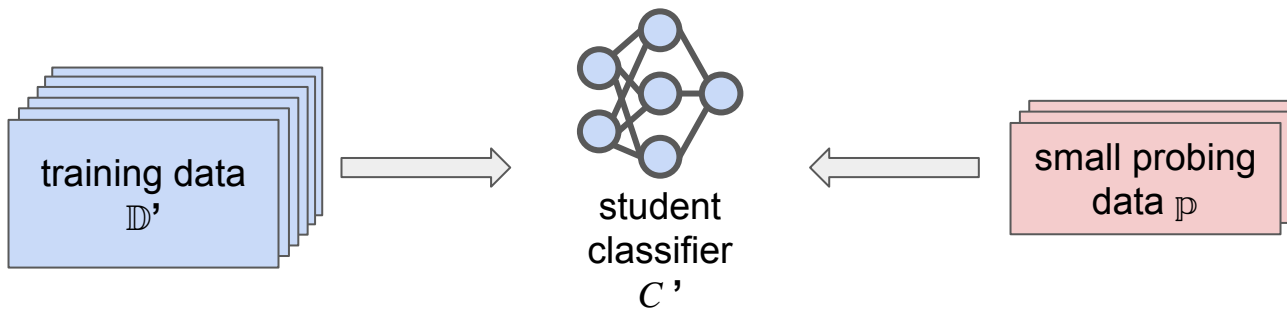
Fortifying Toxicity Classifiers






Overt	Offensive	Toxic
Clean	Non-offensive	Non-toxic
Veiled	Offensive	Non-toxic
FP	Non-offensive	Toxic

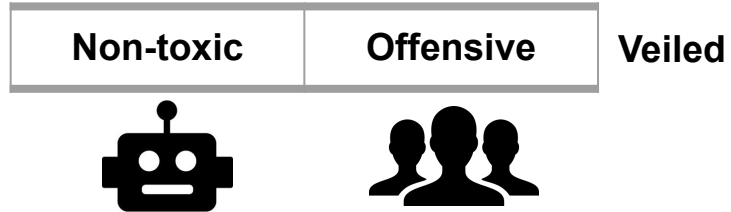


Fortifying Toxicity Classifiers

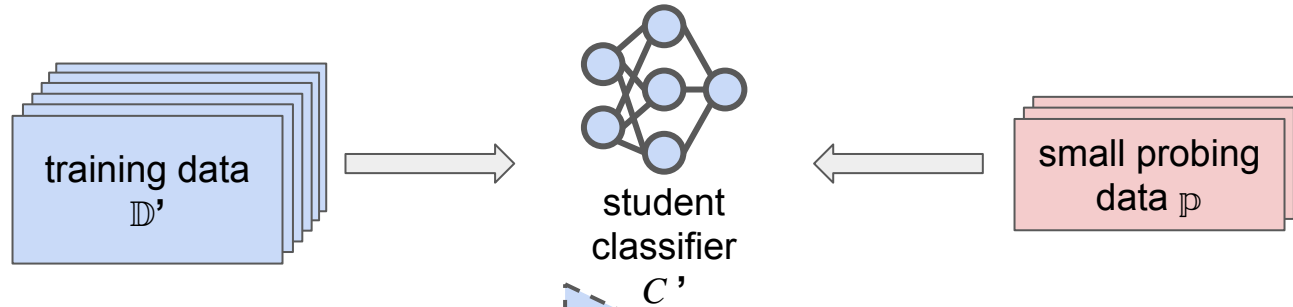


Overt	Offensive	Toxic
Clean	Non-offensive	Non-toxic
Veiled	Offensive	Non-toxic
FP	Non-offensive	Toxic

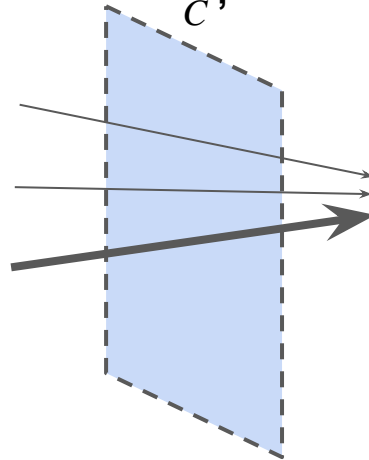
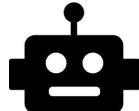






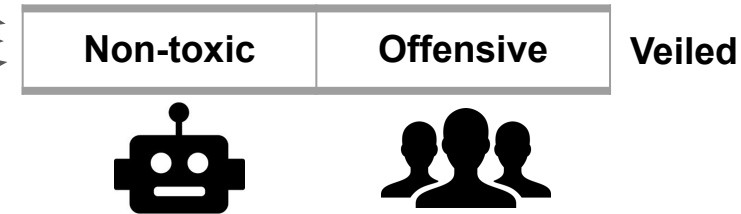
Fortifying Toxicity Classifiers



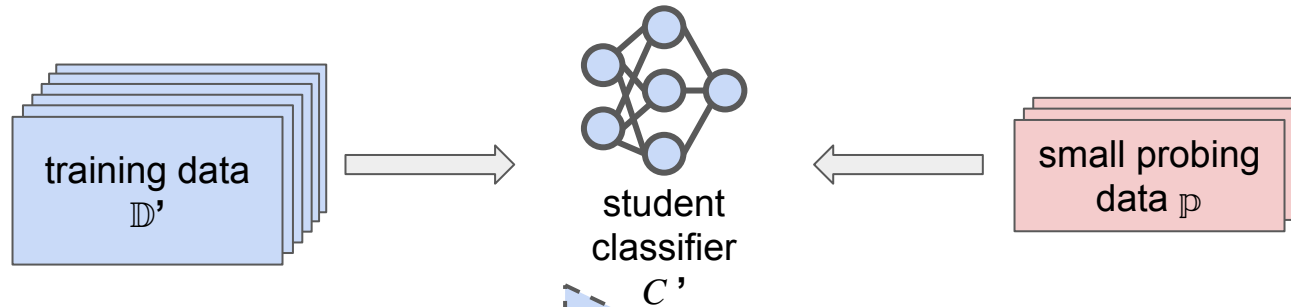
Overt	Offensive	Toxic
Clean	Non-offensive	Non-toxic
Veiled	Offensive	Non-toxic
FP	Non-offensive	Toxic



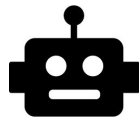
Influence metric



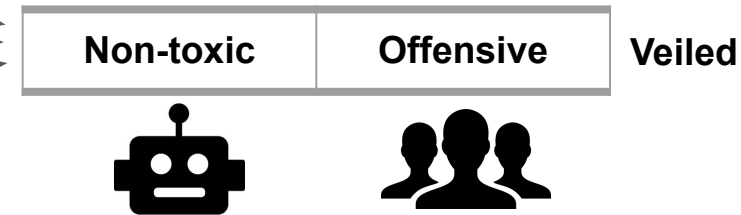
Fortifying Toxicity Classifiers



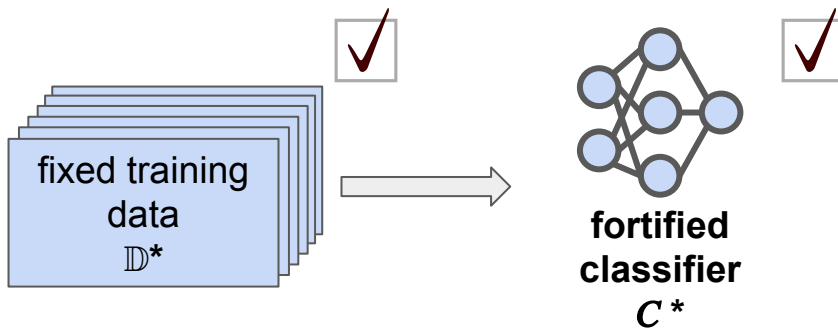
Overt	Offensive	Toxic
Clean	Non-offensive	Non-toxic
Veiled	Offensive	Toxic*
FP	Non-offensive	Toxic



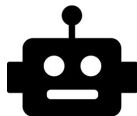
Influence metric



Fortifying Toxicity Classifiers



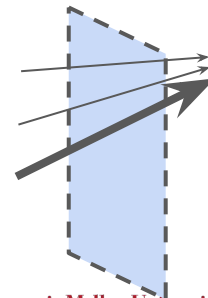
Overt	Offensive	Toxic	
Clean	Non-offensive	Non-toxic	
Veiled	Offensive	Toxic*	✓
FP	Non-offensive	Toxic	



Influence of Training Data

- Which **training** data is most influential to the classifier's decision on the **probing** veiled toxicity example?

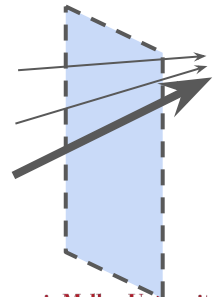
- $\mathcal{I}(x_{trn}, x_{prb})$



Embedding Similarity

$$\mathcal{I}(x_{trn}, x_{prb}) = \underbrace{f_{enc}(x_{trn})}_{\text{Representation of training data}} \cdot \underbrace{f_{enc}(x_{prb})}_{\text{Representation of probing data}}$$

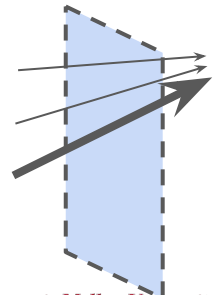
“How different are the **representations** of the training data and the probing data?”



Influence Functions

$$\frac{d\theta}{d\epsilon_{trn}} = -H_{\theta}^{-1} \nabla_{\theta} \mathcal{L}(\theta, x_{trn}, y_{trn})$$

“If we **upweight** a training example by ϵ , how would the resulting model change?”



(Koh and Liang, 2017)

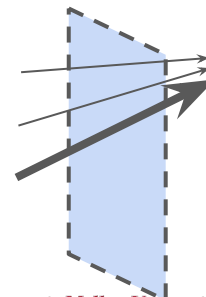


Influence Functions

$$\frac{d\theta}{d\epsilon_{trn}} = -H_{\theta}^{-1} \nabla_{\theta} \mathcal{L}(\theta, x_{trn}, y_{trn})$$

$$\frac{d\mathcal{L}(\theta, x_{prb}, \hat{y}_{prb})}{d\epsilon_{trn}} = \nabla_{\theta} \mathcal{L}(\theta, x_{prb}, \hat{y}_{prb}) \cdot \frac{d\theta}{d\epsilon_{trn}}$$

“Given this **change** in the resulting model ...”



(Koh and Liang, 2017)



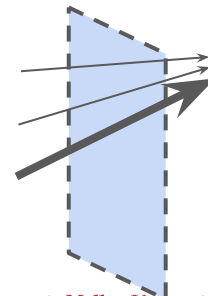
Influence Functions

$$\frac{d\theta}{d\epsilon_{trn}} = -H_{\theta}^{-1} \nabla_{\theta} \mathcal{L}(\theta, x_{trn}, y_{trn})$$

$$\frac{d\mathcal{L}(\theta, x_{prb}, \hat{y}_{prb})}{d\epsilon_{trn}} = \nabla_{\theta} \mathcal{L}(\theta, x_{prb}, \hat{y}_{prb}) \cdot \frac{d\theta}{d\epsilon_{trn}}$$

“How would the **loss** of the probing example change?”

(Koh and Liang, 2017)



Influence Functions

$$\frac{d\theta}{d\epsilon_{trn}} = -H_{\theta}^{-1} \nabla_{\theta} \mathcal{L}(\theta, x_{trn}, y_{trn})$$

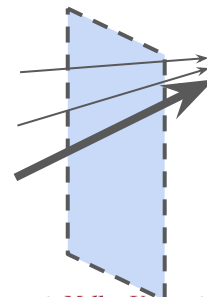
$$\frac{d\mathcal{L}(\theta, x_{prb}, \hat{y}_{prb})}{d\epsilon_{trn}} = \nabla_{\theta} \mathcal{L}(\theta, x_{prb}, \hat{y}_{prb}) \cdot \frac{d\theta}{d\epsilon_{trn}}$$

$$\mathcal{I}(x_{trn}, x_{prb}) = \boxed{-\frac{d\mathcal{L}(\theta, x_{prb}, \hat{y}_{prb})}{d\epsilon_{trn}}}$$

⋮

“Upweighting an **influential** training example should lead to a decrease in the loss of the probing example.”

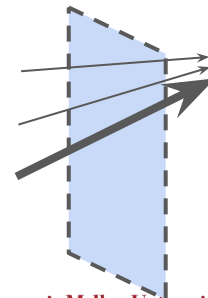
(Koh and Liang, 2017)



Gradient product (*TrackIn*)

$$\mathcal{I}(x_{trn}, x_{prb}) = \sum_{i=1}^k \underbrace{\nabla_{\theta} \mathcal{L}(\theta_i, x_{trn}, y_{trn})}_{\text{gradient of training loss at epoch } i} \cdot \nabla_{\theta} \mathcal{L}(\theta_i, x_{prb}, \hat{y}_{prb})$$

“The model would take a step towards the **gradient** of the training example’s loss at epoch *i*.”



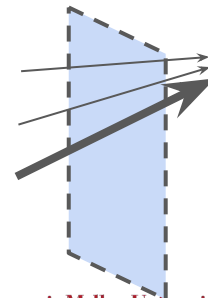
(Pruthi et al., 2020)



Gradient product (*TrackIn*)

$$\mathcal{I}(x_{trn}, x_{prb}) = \sum_{i=1}^k \underbrace{\nabla_{\theta} \mathcal{L}(\theta_i, x_{trn}, y_{trn}) \cdot \nabla_{\theta} \mathcal{L}(\theta_i, x_{prb}, \hat{y}_{prb})}_{\text{Gradient product}}$$

“Because of this step, how much will the loss of the probing example **decrease**?”



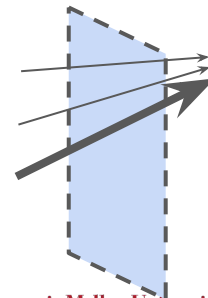
(Pruthi et al., 2020)



Gradient product (*TrackIn*)

$$\mathcal{I}(x_{trn}, x_{prb}) = \underbrace{\sum_{i=1}^k}_{\text{---}} \nabla_{\theta} \mathcal{L}(\theta_i, x_{trn}, y_{trn}) \cdot \nabla_{\theta} \mathcal{L}(\theta_i, x_{prb}, \hat{y}_{prb})$$

“We take a **sum** of such probing loss decrease caused by the training example over all the checkpoints of the model.”

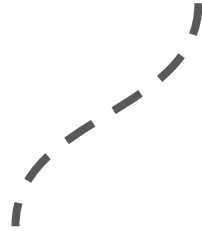


(Pruthi et al., 2020)

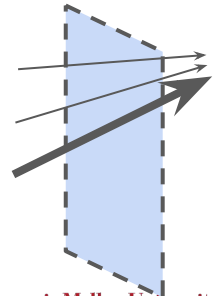


Training Loss

$$\mathcal{I}(x_{trn}) = \mathcal{L}(\theta, x_{trn}, y_{trn})$$



“A high training loss means the example is **hard** to learn.”



Dataset

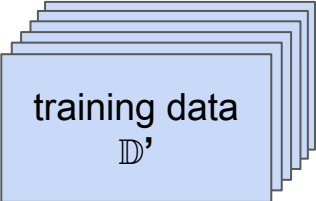
Social Bias Inference Corpus (SBIC)

- 45K social media posts, primarily from Reddit and Twitter.



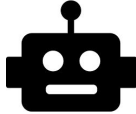


- Crowdsourced annotations of offensiveness, target group, etc.

Dataset



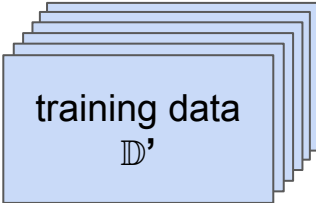
Overt	Offensive	Toxic
Clean	Non-offensive	Non-toxic
Veiled	Offensive	Non-toxic
FP	Non-offensive	Toxic



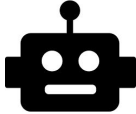
SBIC



Dataset



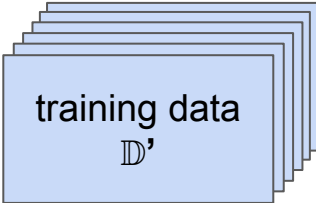
Overt	Offensive	Toxic	$tox > 0.8$
Clean	Non-offensive	Non-toxic	$tox \approx 0.17$
Veiled	Offensive	Non-toxic	$tox \approx 0.17$
FP	Non-offensive	Toxic	

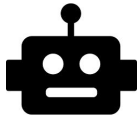
SBIC



Dataset



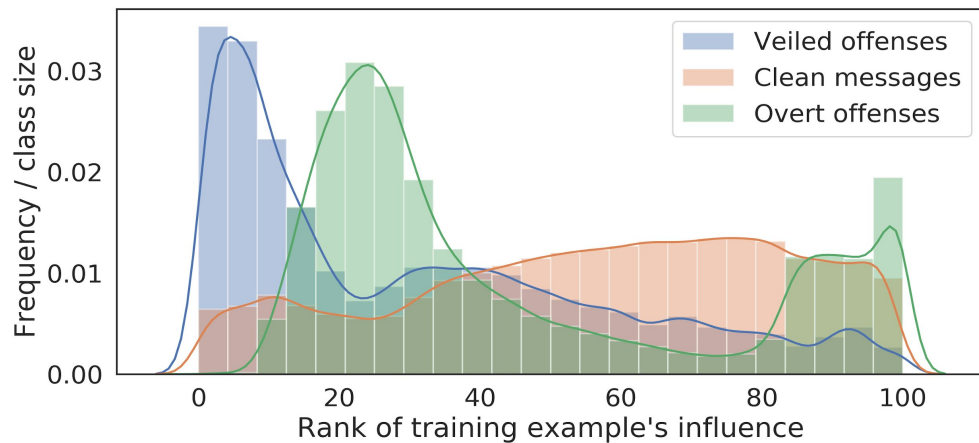
Overt	Offensive	Toxic	$tox > 0.8$	2K
Clean	Non-offensive	Non-toxic	$tox \approx 0.17$	8K
Veiled	Offensive	Non-toxic	$tox \approx 0.17$	2K + 100
FP	Non-offensive	Toxic		



SBIC



Unveiling disguised toxicity via probing & interpreting model decisions



		Veiled	Clean	Overt
Model	Operation			
<u>Original</u>		1.2	99.6	97.2
<u>Gradient product</u>	fix top 2000	37.5	97.6	98.0
	flip top 2000	51.1	87.6	99.5
<u>Gold</u>		76.0	94.8	98.2

Pointers

- Computational ethics readings, lectures
http://demo.clab.cs.cmu.edu/ethical_nlp/
- NeurIPS Keynote: Kate Crawford, The Trouble with Bias
<https://goo.gl/qgeMKO>



Thank you!

Computational Ethics Lab

- Examples of projects: <https://bit.ly/2Vw9aaA>

ytsvetko@cs.cmu.edu

