

Ethical Issues in Online Abuse Detection

Svetlana Kiritchenko

Digital Technologies Research Centre



Abusive Language Online

Definition (for this presentation): any language that could offend, demean, or marginalize another person

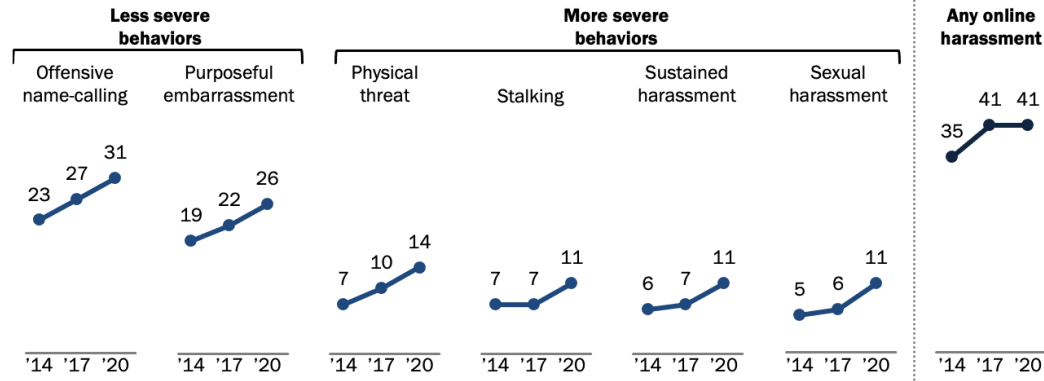
- covers the full range of inappropriate content from profanities and obscene expressions to threats and severe insults



Online Abuse

41% of Americans have experienced online harassment

% of U.S. adults who say they have personally experienced the following behaviors online



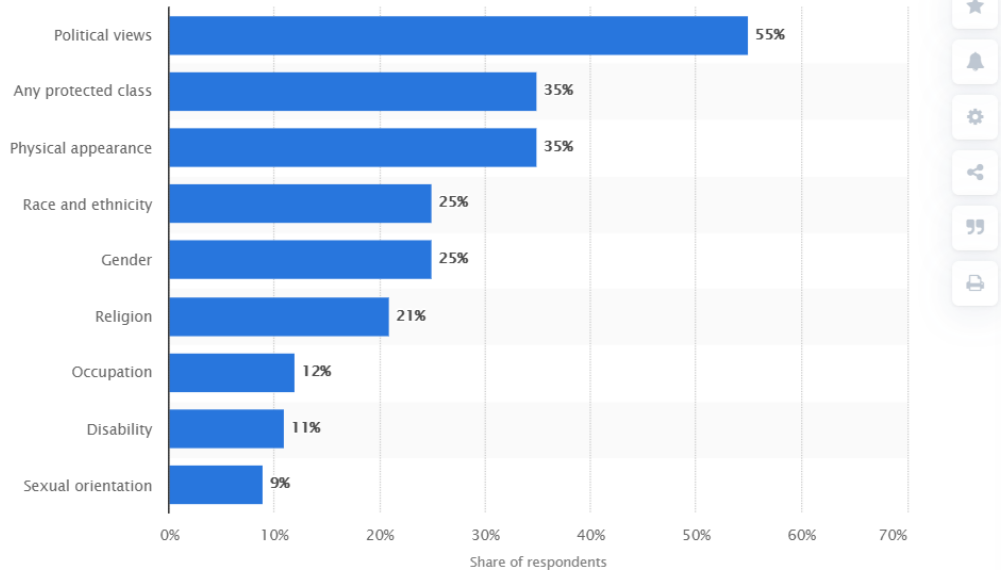
Note: Those who did not give an answer are not shown.
Source: Survey of U.S. adults conducted Sept. 8-13, 2020.
"The State of Online Harassment"

PEW RESEARCH CENTER

<https://www.pewresearch.org/internet/2021/01/13/personal-experiences-with-online-harassment/>

Online Abuse

Reasons for online harassment in US (2020)



© Statista 2021

<https://www.statista.com/statistics/971847/us-internet-online-harassment-reasons/>

Online Platforms Policies

[Help Center](#) > [Twitter Rules and policies](#) > [Hateful conduct policy](#)



Hateful conduct policy

Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories.

Hateful imagery and display names: You may not use hateful images or symbols in your profile image or profile header. You also may not use your username, display name, or profile bio to engage in abusive behavior, such as targeted harassment or expressing hate towards a person, group, or protected category.


<https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy>

Content Moderation

online content



1.88B daily
active
Facebook users

 500M tweets
per day

human moderators



15,000 human
moderators at
Facebook

keep/delete




Icons by I Putu Kharismayadi, Bonegolem, Alec Dhuse (<https://thenounproject.com/>)

Automatic Content Moderation

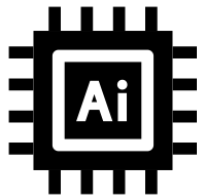
online content



1.88B daily
active
Facebook users

 500M tweets
per day

automatic moderation



keep/delete



Icons by I Putu Kharismayadi, SBTS, Alec Dhuse (<https://thenounproject.com/>)

ETHICAL ISSUES IN AUTOMATIC CONTENT MODERATION

Automatic Content Moderation Can Silence Marginalized Voices

NEWS

Facebook while black: Users call it getting 'Zucked,' say talking about racism is censored as hate speech

Jessica Guynn USA TODAY

Published 7:26 a.m. ET Apr. 24, 2019 | Updated 6:17 p.m. ET Jul. 9, 2020



Are black Facebook users censored from discussing racism online?

Carolyn Wysinger is a teacher and activist who says Facebook censors her from discussing racism online, sometimes locking her out of her account. USA TODAY

<https://www.usatoday.com/story/news/2019/04/24/facebook-while-black-zucked-users-say-they-get-blocked-racism-discussion/2859593002/>

Automatic Content Moderation Can Amplify Racial Bias

Vox

recode

The algorithms that detect hate speech online are biased against black people

A new study shows that leading AI models are 1.5 times more likely to flag tweets written by African Americans as “offensive” compared to other tweets.

By Shirin Ghaffary | Aug 15, 2019, 11:00am EDT



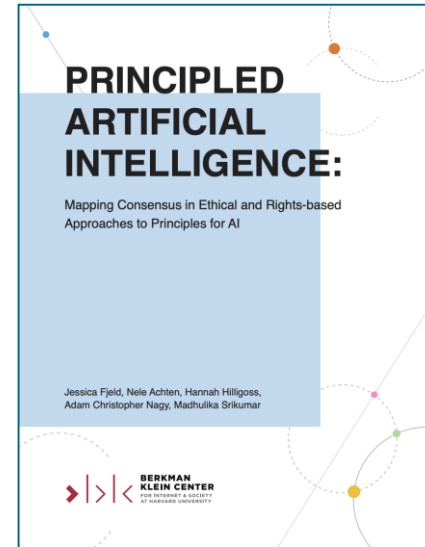
New studies show that AI trained to identify hate speech may actually end up amplifying racial bias. | Recode

<https://www.vox.com/recode/2019/8/15/20806384/social-media-hate-speech-bias-black-african-american-facebook-twitter>

Ethical and Human Rights Framework

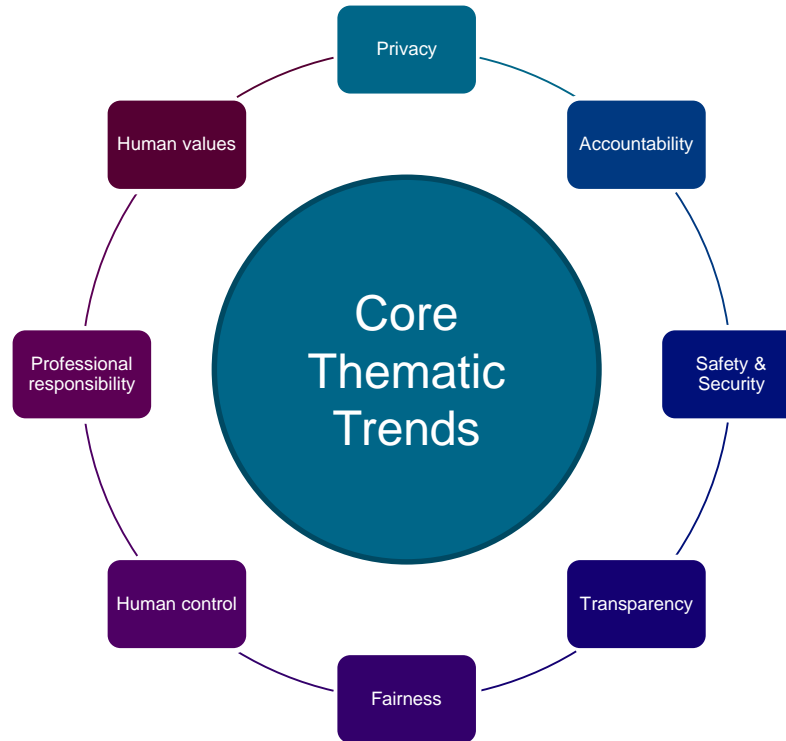
Covers:

- 36 prominent AI principles documents
- Organizations: governments, intergovernmental organizations, private sector, professional associations, advocacy groups, and multi-stakeholder initiatives
- Geographical areas: Latin America, East and South Asia, the Middle East, North America, and Europe

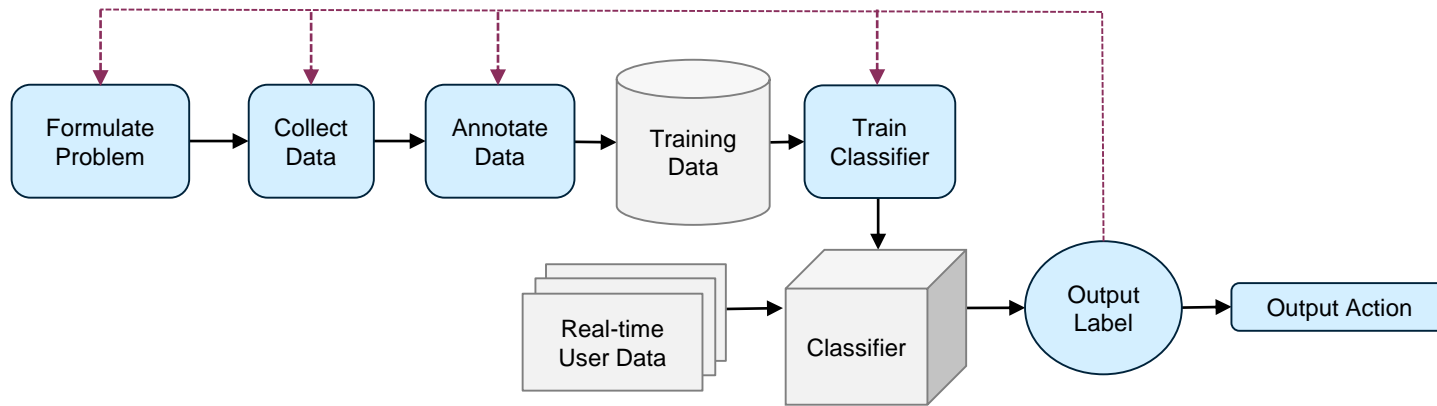


<https://cyber.harvard.edu/publication/2020/principled-ai>

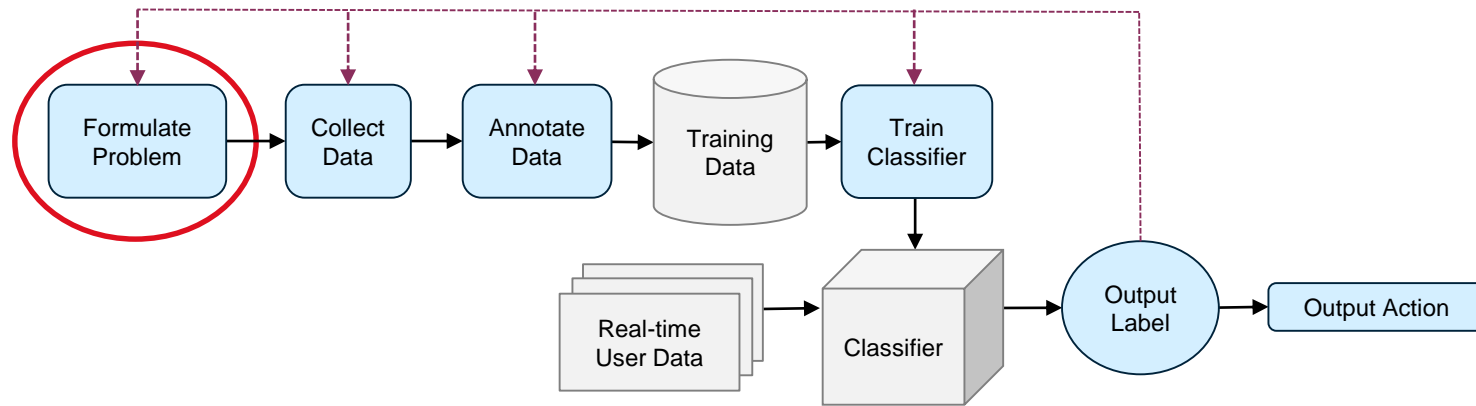
Ethical and Human Rights Framework



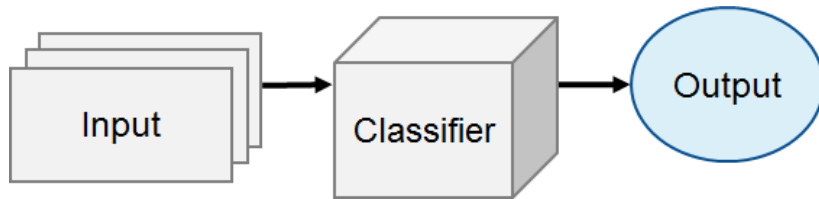
Machine Learning Pipeline



Formulate Problem



Formulate Problem



Input:

- Data source (Twitter, FB, etc.)
- Post, sentence, all user's content
- Context
- User info
- ...

Output:

- Categorical, numerical, sequential
- Number of categories
- Category definitions
- Target of abuse
- ...

Professional Responsibility

What is the right task? Consider long-term effects:

- Which groups will be affected by this system, and how?
- Are we solving real problems, or just the ones that are convenient to solve with the methods or data we have at hand?
- What are the possible future applications of such a system? Could it be used to silence political dissidents? Or marginalized groups discussing their own lived experiences?
- Who decides what constitutes offence or hate?



Icon by Margaret Hagan (<https://thenounproject.com/>)

Promotion of Human Values

Freedom of
speech



Respect for
equality and
dignity

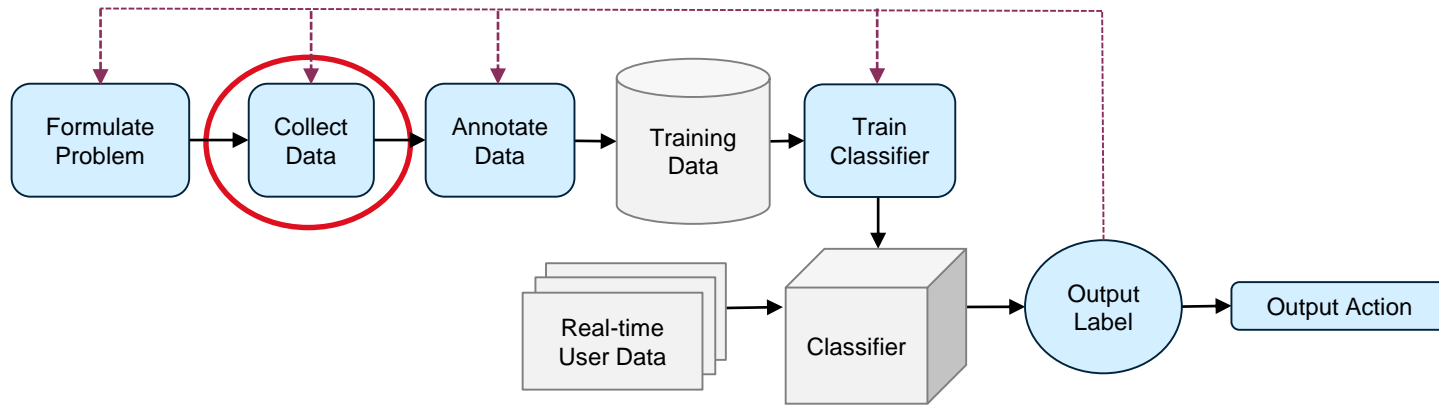
Online abuse can

- inflict significant psychological harm to its victims
- lead to physical violence
- silence the voices of minority groups and individuals through threats and offensive behavior

It is not only a **technical** problem, but also a **social** one.

Icon by iconcheese (<https://thenounproject.com/>)

Data Collection



Data Collection

Sampling strategies:

- Random sampling (results in between 0.1% and 3% of abusive content)
- Known abusive/profane words
- Words, phrases, and hashtags associated with abusive content
- Words describing target populations
- Users known for abusive behaviours
- ...



Icon by R Diepenheim (<https://thenounproject.com/>)

Fairness and Non-discrimination

Bias resulting from [skewed representation of vocabulary](#) (e.g., identity terms)

Term	Comment Length				
	20-59	60-179	180-539	540-1619	1620-4859
ALL	17%	12%	7%	5%	5%
gay	88%	77%	51%	30%	19%
queer	75%	83%	45%	56%	0%
homosexual	78%	72%	43%	16%	15%
black	50%	30%	12%	8%	4%
white	20%	24%	16%	12%	2%
wikipedia	39%	20%	14%	11%	7%
atheist	0%	20%	9%	6%	0%
lesbian	33%	50%	42%	21%	0%
feminist	0%	20%	25%	0%	0%
islam	50%	43%	12%	12%	0%
muslim	0%	25%	21%	12%	17%
race	20%	25%	12%	10%	6%
news	0%	1%	4%	3%	3%
daughter	0%	7%	0%	7%	0%

Due to over-representation of certain identity terms in abusive comments, the model associates the terms with the abusive class.

(Dixon et al., 2018)

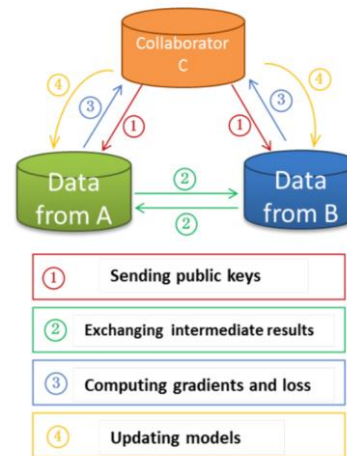
Privacy

Data collection and distribution for research:

- Public data without explicit consent from users
- May infer personal information
- "Right to be forgotten"

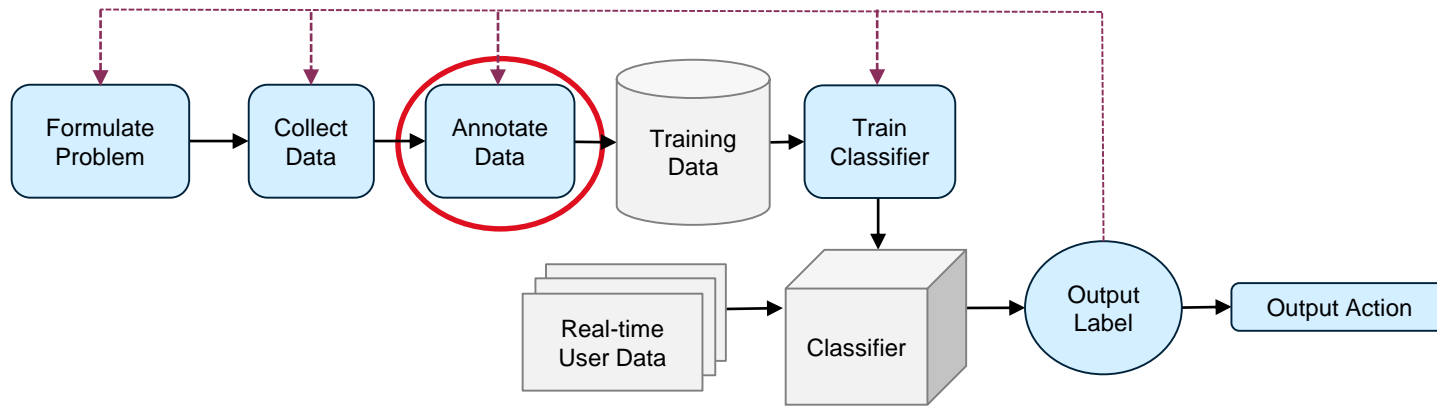
Privacy-preserving in commercial settings:

- Federated learning (models trained on decentralized servers, holding the user's data on the user's device and sharing only the model's parameters)



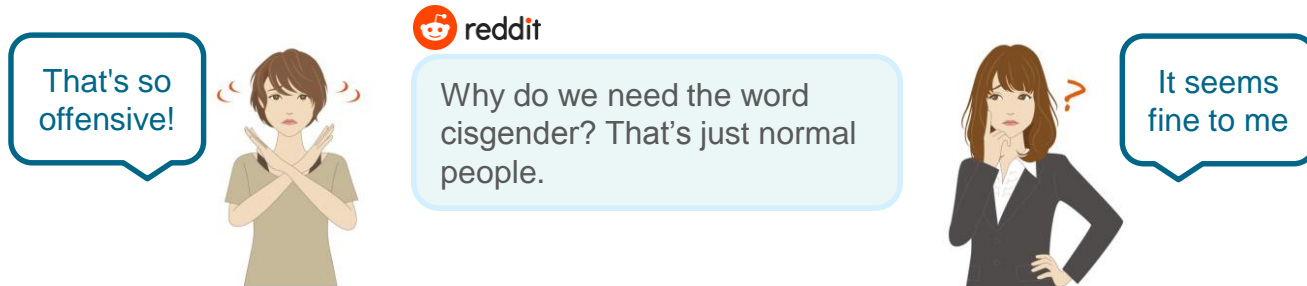
(Konecny et al., 2016; Yang et al., 2019)

Data Annotation



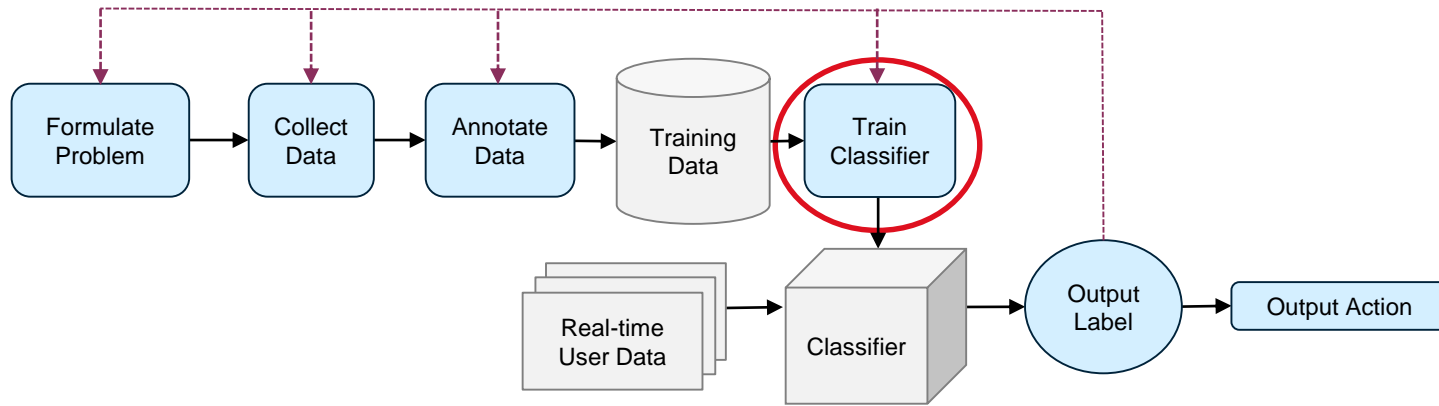
Bias in Manual Annotations

- Human judgement is subjective
- Complexities and ambiguities in abusive label definitions make the annotation task even harder
- Gender, age, education, first language affect moral judgement
- Insensitivity or unawareness of dialect can lead to biased annotations

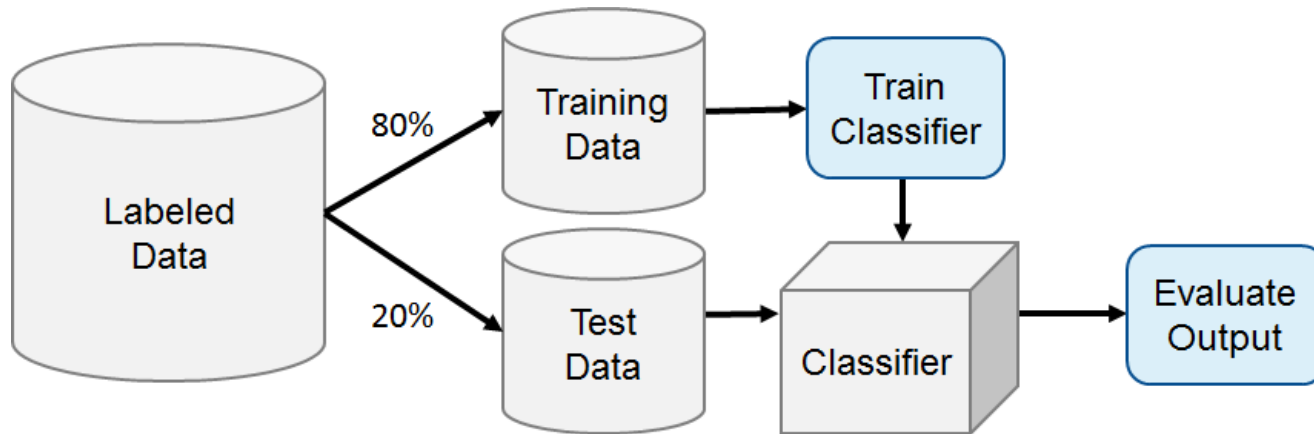


(Tversky & Kahneman, 1974; Breitfeller et al., 2019; Wilhelm & Joeckel, 2019; Sap et al., 2019; Al Kuwatly et al., 2020)

Training Classifier



Classifier Performance Evaluation



Training and test data come from [the same distribution](#)

Safety

Does the system perform as intended?

One safety risk comes from the **mismatch between training and test environments**

- Real-world data is ever-changing and rarely matches the data on which the model has been trained
- New topics and new types of abuse emerge (e.g., anti-Asian COVID-related hate speech)
- System performance can drop substantially on new data

Macro-averaged F1-scores for a classifier trained and tested on different abusive datasets

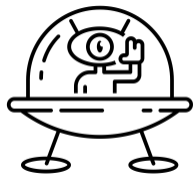
Training dataset	Test dataset		
	Wikipedia Toxic	East-Asian Prejudice	COVID-Hate
Wikipedia Toxic	0.82	0.27	0.69
East-Asian Prejudice	-	0.84	0.75

Security

Is the system vulnerable to malicious attacks by unauthorized third parties?

Abuse detection systems are vulnerable against:

- Adversarial insertion of typos, innocuous words (e.g., *love*), specific rare words
- Change of word boundaries
- Vowel substitution and duplication



Martians are disgusting and should be killed



94.88% likely
to be toxic.

MartiansAreDisgustingAndShouldBeKilled love



9.84% likely
to be toxic.

(Hosseini et al., 2017; Gröndahl et al., 2018; Kalin et al., 2020; Kurita et al., 2020; icon by Symbolon)

Transparency and Explainability

- Make **the process of creating an automatic system** understandable by different stakeholders
 - document all design decisions (datasheets for datasets, model cards, fact sheets)
 - audit the system
- Explain **the underlying dynamics of opaque algorithms**, such as deep neural networks
 - to improve a trained model
 - to inform feature engineering
 - to direct future data collection
- Explain **automatic outputs to end users**
 - to inform human decision making
 - to build trust



Icon by Wichai Wi (<https://thenounproject.com/>)

Transparency and Explainability

Types of explanations:

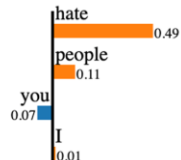
- Linear models with interpretable features (e.g., words)
- Confidence estimates (probability of the output being correct)
- Saliency maps (LIME, SHAP)

Prediction probabilities



non-offensive

offensive



Text with highlighted words

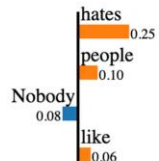
I **hate** people like you.

Prediction probabilities



non-offensive

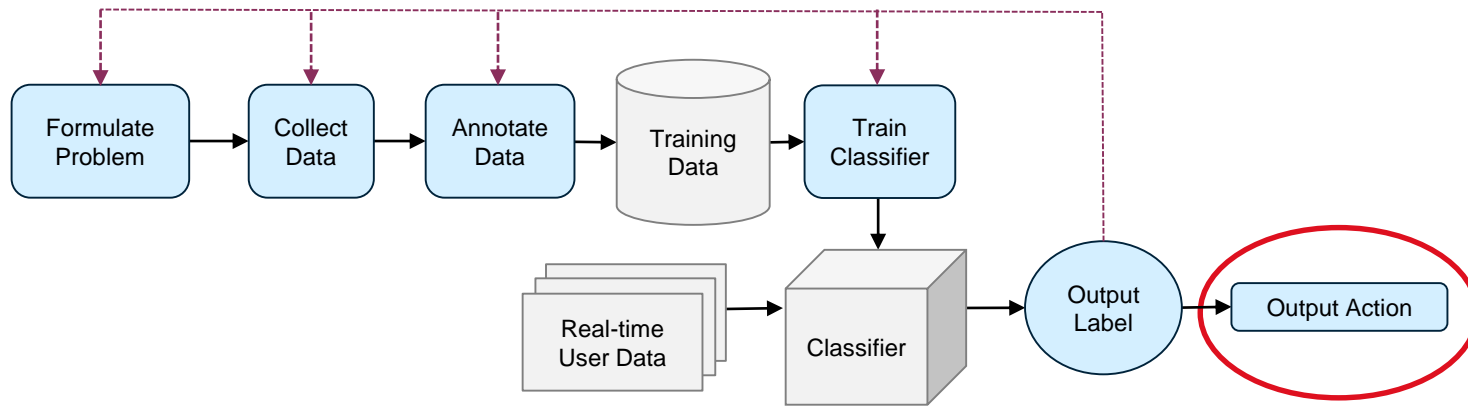
offensive



Text with highlighted words

Nobody **hates** people like you.

Output Action



Human Control of Technology

Users must be able to appeal automated decisions and request human review or even opt out of automated decisions entirely.

Two scenarios for human review:

- Pre-moderation: an automatic system flags potentially problematic content for human review before posting
- Post-moderation: fully automatic moderation followed by human review if requested by users



Icon by Gan Khoon Lay (<https://thenounproject.com/>)

Accountability

Organizations (e.g., social media corporations) that develop and deploy AI systems should be accountable for the systems' outcomes and impacts on the social and natural world:

- better transparency on the processes and results of content moderation
- meaningful opportunities for users to appeal any content removal
- justification for any content removal decisions
- Internal and external algorithmic audit

Moving Forward

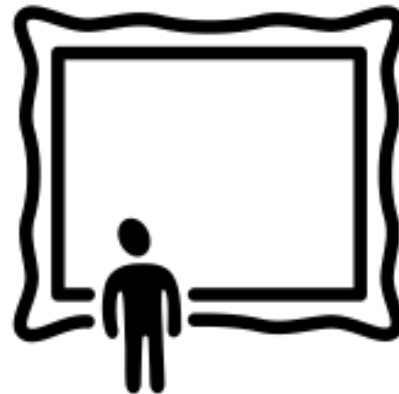
Some areas of focus:

- Reimagining the task:
 - Moving away from coarse-grained definitions of abuse
 - Moving towards flexible, right-respecting moderation (e.g., quarantining, nudging)
 - Extending the task beyond detection (e.g., counter-narrative, public education)
- Going beyond text and including multi-media inputs
- Advancing explainability
- Grounding research in work from other disciplines
- Engaging affected communities

Conclusions

Looking at the bigger picture:

- How has the problem been formulated, and by whom?
- Where is the data coming from, and is it representative?
- Who is annotating the data, and what are their implicit biases and beliefs?
- Is a binary label of "abusive" or "not abusive" truly sufficient?
- How can the decision be explained?
- How can the decision be appealed?



Further Readings

Journal of Artificial Intelligence Research (2021)

Submitted 12/20; published 06/21

Confronting Abusive Language Online: A Survey from the Ethical and Human Rights Perspective

Svetlana Kiritchenko

Isar Nejadgholi

Kathleen C. Fraser

National Research Council Canada

1200 Montreal Rd., Ottawa, ON, Canada

SVETLANA.KIRITCHENKO@NRC-CNRC.GC.CA

ISAR.NEJADGHOLI@NRC-CNRC.GC.CA

KATHLEEN.FRASER@NRC-CNRC.GC.CA

Preprint available at ArXiv: <https://arxiv.org/abs/2012.12305>

AND about 200 articles referenced in the paper



Isar Nejadgholi



Kathleen C. Fraser



Esma Balkir