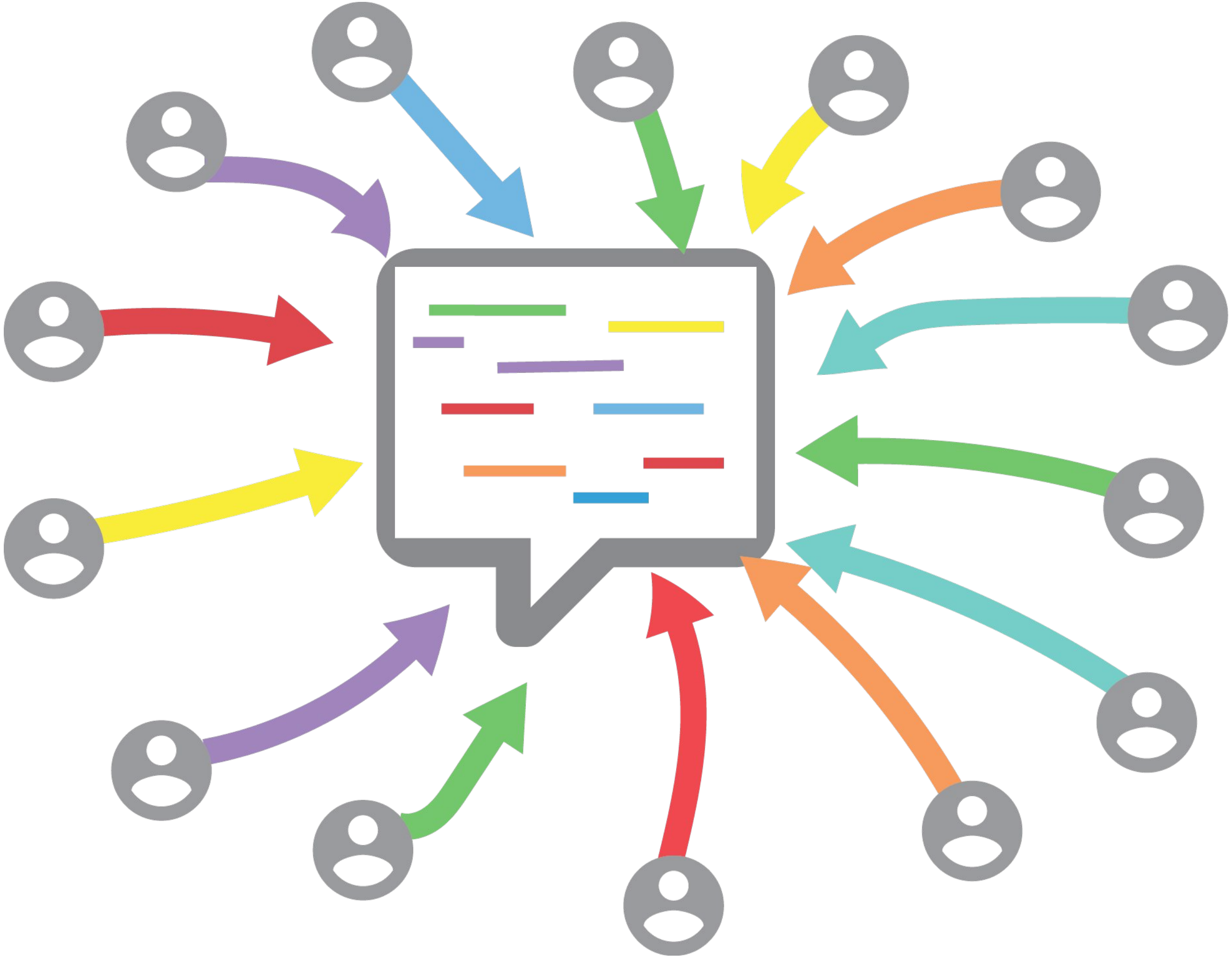
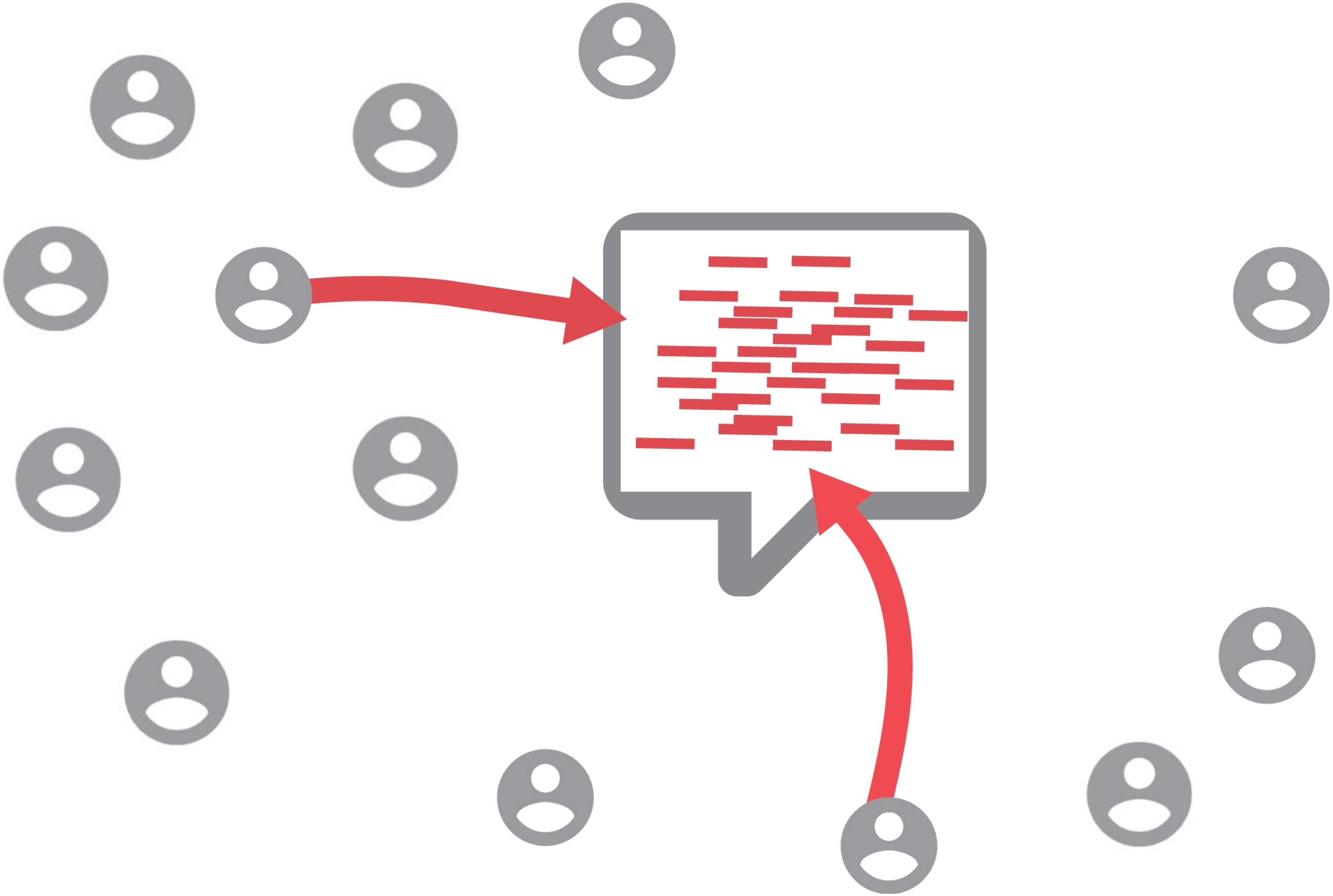
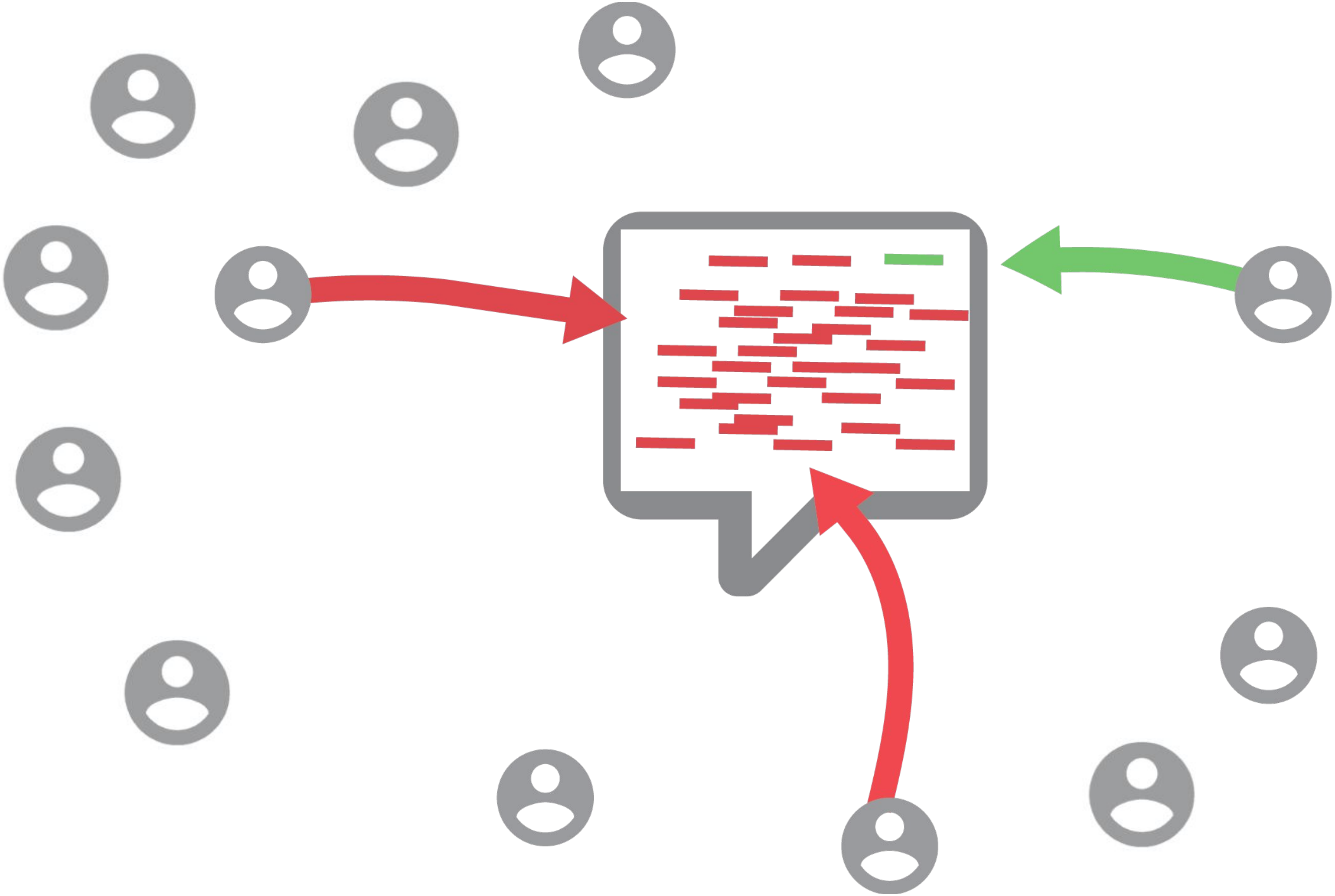
Perspective API

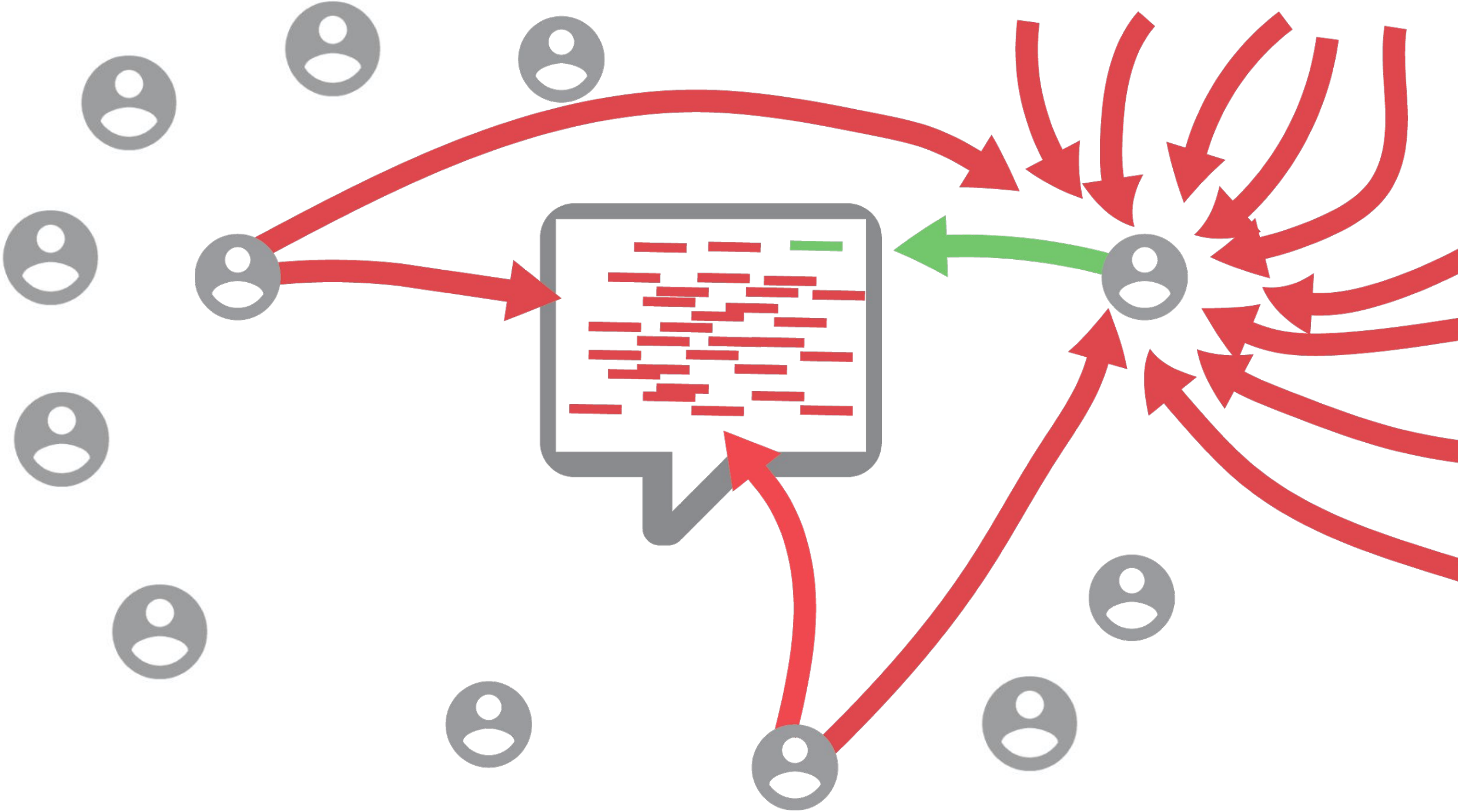
ML for good conversations at scale

mpellat@









CONVERSATION-AI

Research for good conversations at scale

conversationai.github.io

good conversations
How was I gonna do it?

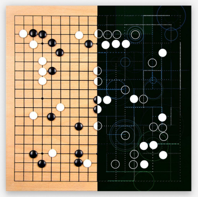
Three Kinds of Machine Learning

1. **Clustering problems (Unsupervised Learning)**

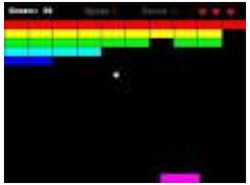
   Given: Metric (could be a context), cluster examples
   Word Embeddings (dimensionality reduction) ; image segmentation etc.

2. **Game playing problems (Reinforcement Learning)**

   Given: Way to score games, learn actions
   Need: computer can play the game quickly & gets a score.



win/lose



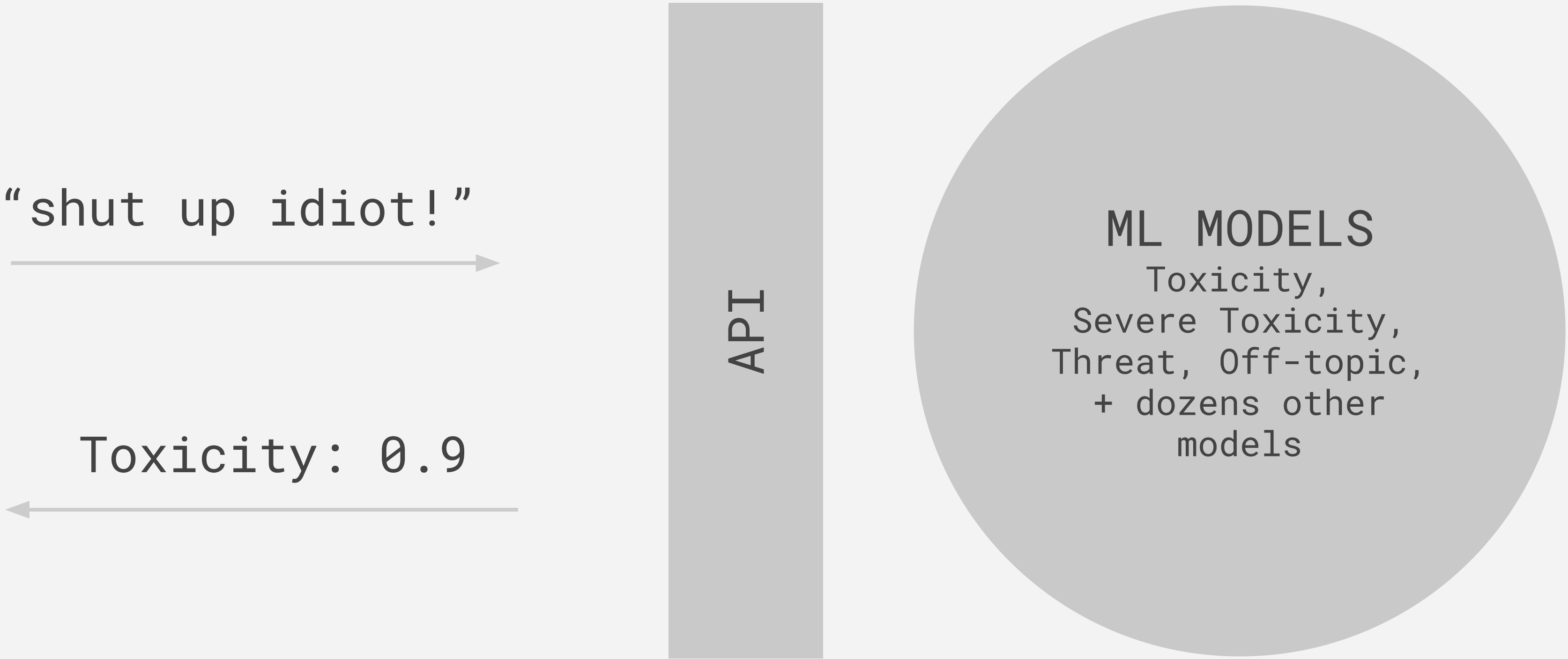
5 minutes

3. **Classification problems (Supervised Learning)**

   Given: Labelled Training data, learn how to label new examples

CLASSIFYING EMOTIONAL IMPACT

PERSPECTIVE API



“shut up idiot!”

Toxicity: 0.9

API

ML MODELS

Toxicity,
Severe Toxicity,
Threat, Off-topic,
+ dozens other
models

perspectiveapi.com

SUCCESS METRICS

PARTICIPATION

Measured by the diversity of participants and overall engagement.

QUALITY

Measured by engagement and value of discussion experience.

EMPATHY

Measured by participants understanding of each other and change decisions.

VALUES

COMMUNITY
Tools for the community, by the community.

TOPIC-NEUTRALITY
It’s about how you discuss, not what you discuss

TRANSPARENCY
Open processes create open discussions

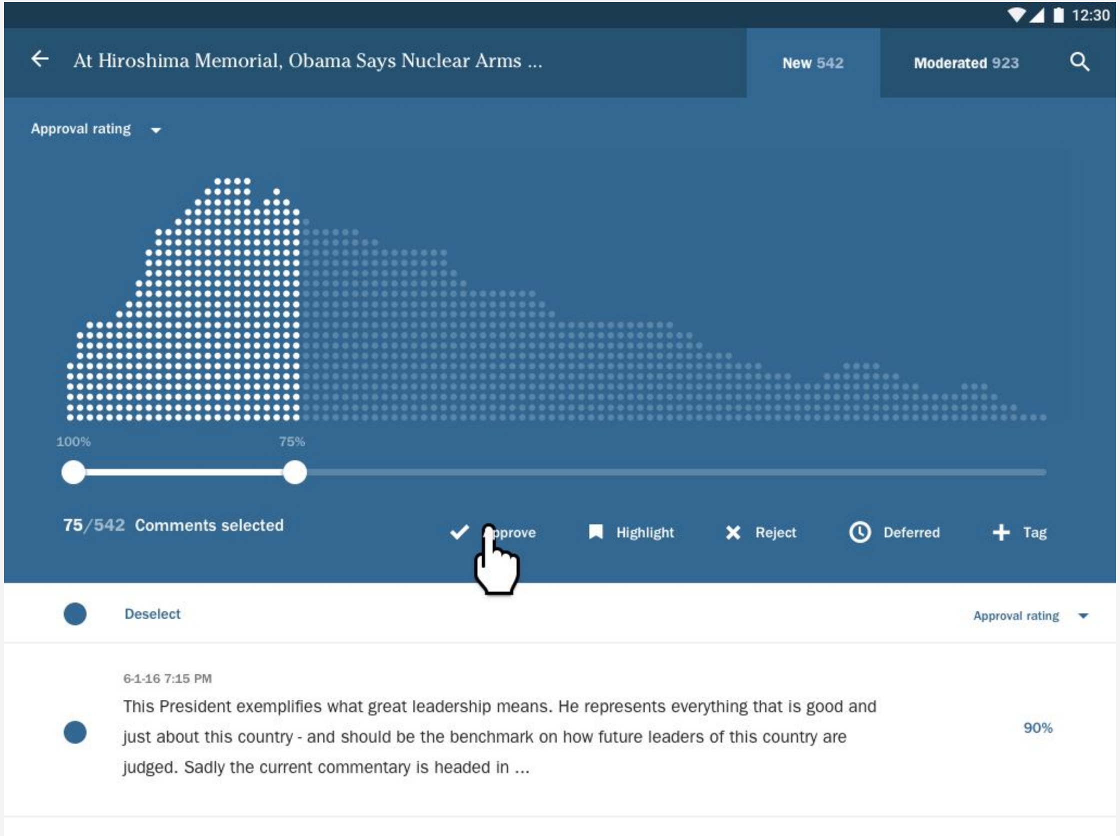
INCLUSIVITY
Diversity in participants and opinions make discussions better.

PRIVACY
It’s about what you say, whoever you are

HOW IS PERSPECTIVE API USED?

MODERATION

Help humans community managers review comments faster.



6-1-16 7:15 PM

This President exemplifies what great leadership means. He represents everything that is good and just about this country - and should be the benchmark on how future leaders of this country are judged. Sadly the current commentary is headed in ...

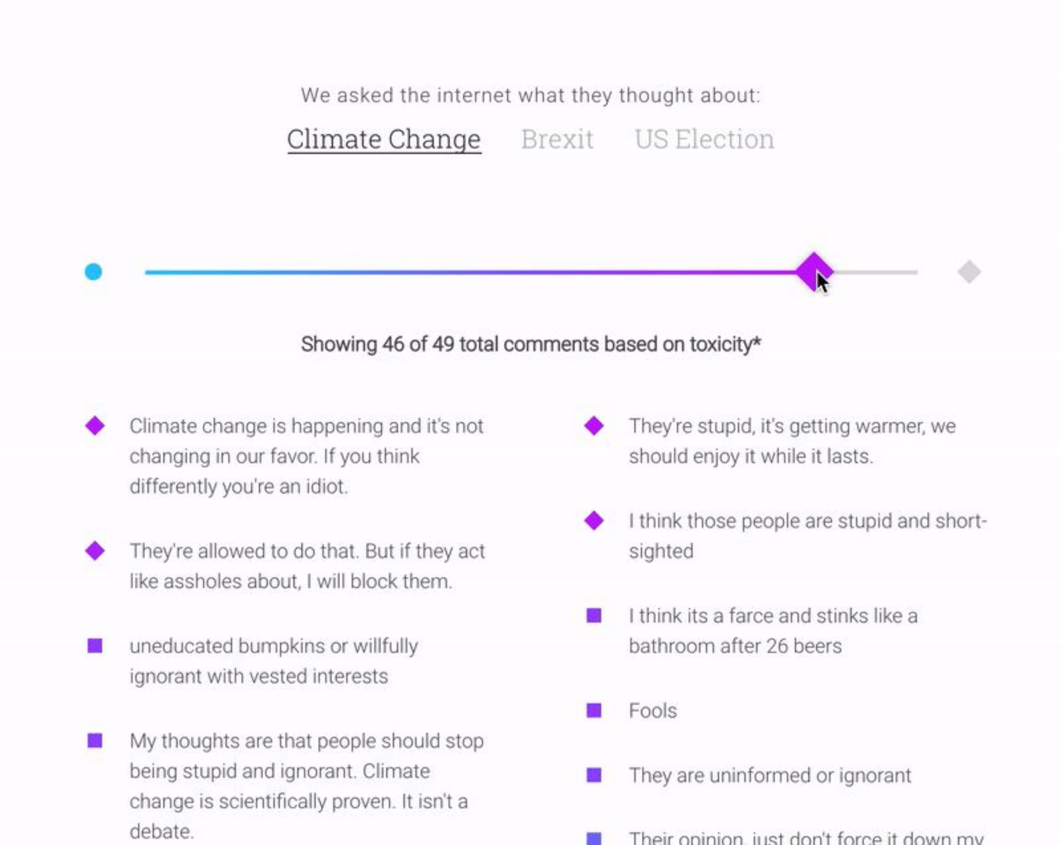
90%

AUTHORSHIP

Help people understand the impact of what they are writing

READERSHIP

Help people find/understand the comments they are interested in



We asked the internet what they thought about:

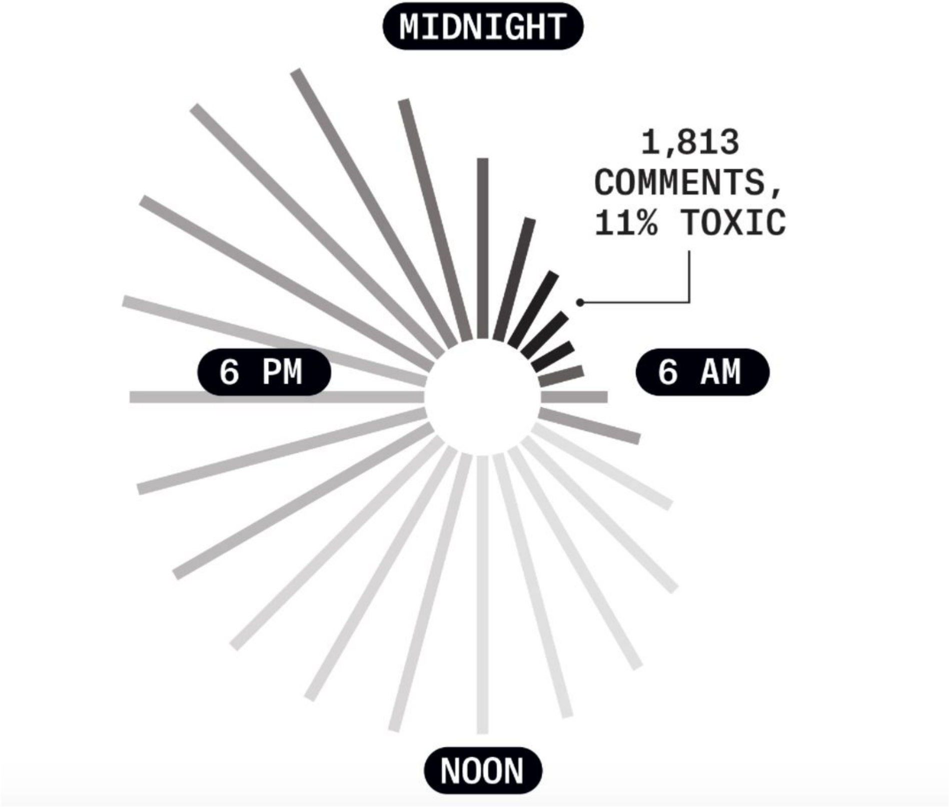
Climate Change   Brexit   US Election

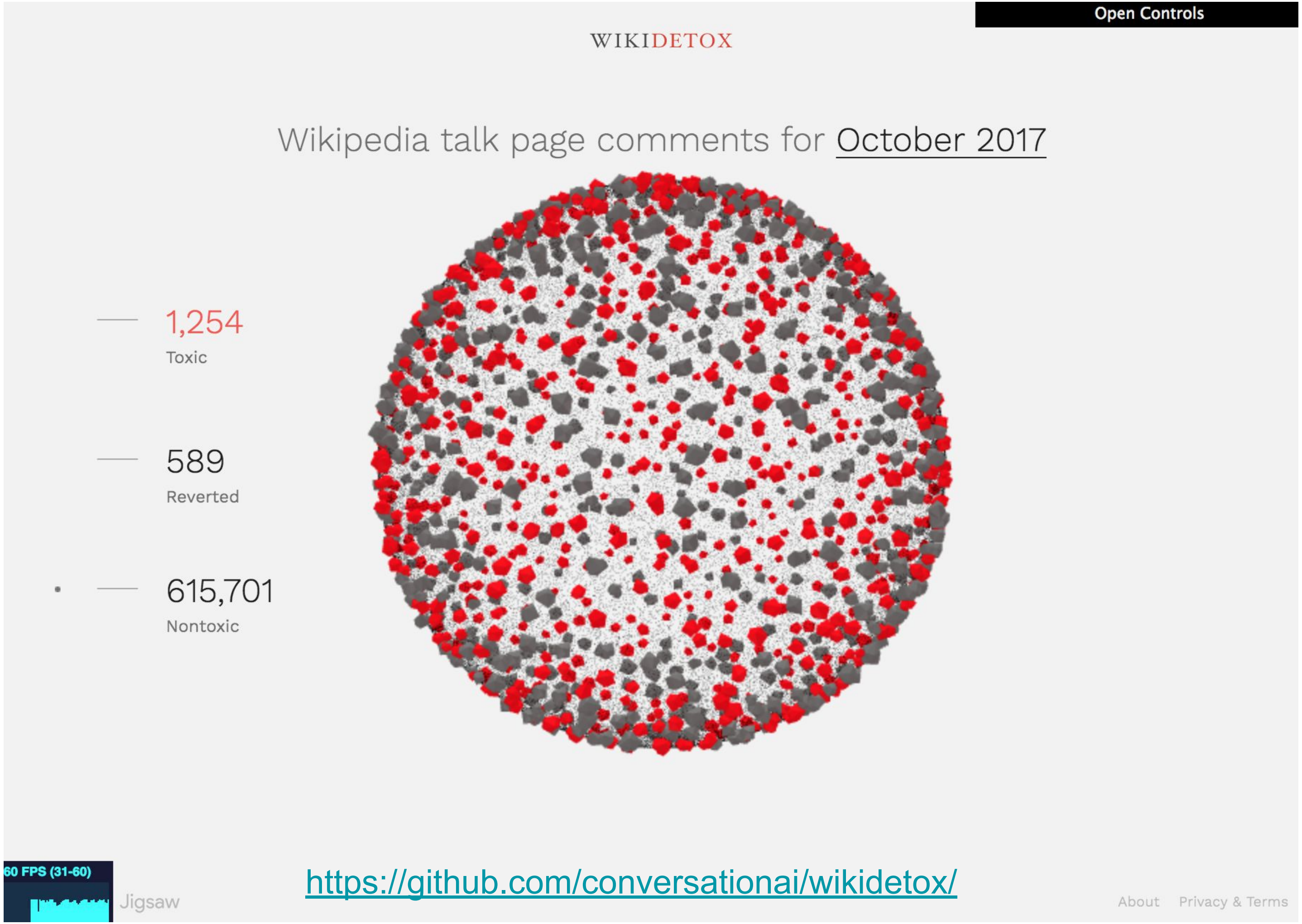
Showing 46 of 49 total comments based on toxicity*

- Climate change is happening and it's not changing in our favor. If you think differently you're an idiot.
- They're allowed to do that. But if they act like assholes about, I will block them.
- uneducated bumpkins or willfully ignorant with vested interests
- My thoughts are that people should stop being stupid and ignorant. Climate change is scientifically proven. It isn't a debate.
- They're stupid, it's getting warmer, we should enjoy it while it lasts.
- I think those people are stupid and short-sighted
- I think its a farce and stinks like a bathroom after 26 beers
- Fools
- They are uninformed or ignorant
- Their opinion, just don't force it down my

Visualize trends

Examples of visualizations created by Wired magazine using Perspective





Wikipedia talk page comments for October 2017

1,254
Toxic

589
Reverted

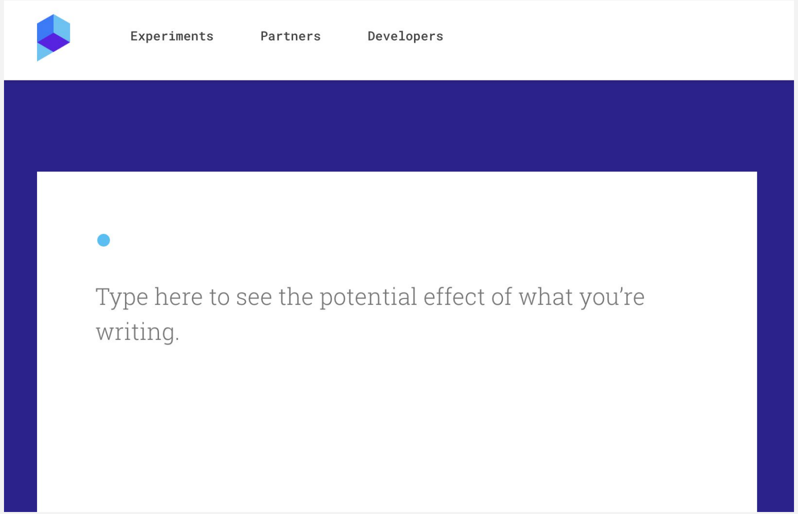
615,701
Nontoxic

https://github.com/conversationai/wikidetox/

DID IT WORK?

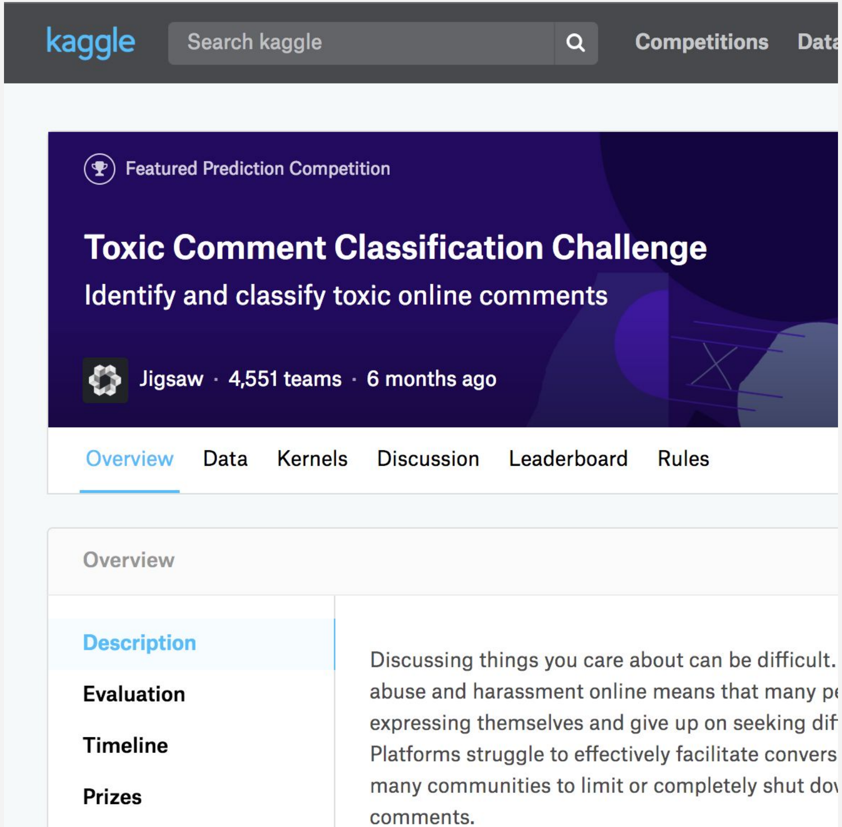
SOME LEARNINGS

STAY TRANSPARENT



Type here to see the potential effect of what you're writing.

OPEN DATA



Featured Prediction Competition

Toxic Comment Classification Challenge

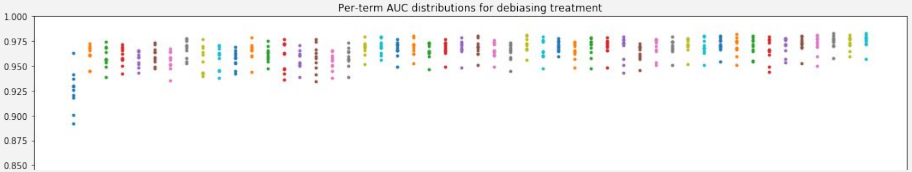
Identify and classify toxic online comments

Jigsaw · 4,551 teams · 6 months ago

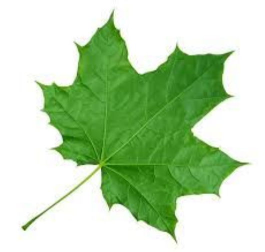
Overview

Discussing things you care about can be difficult. abuse and harassment online means that many pe expressing themselves and give up on seeking dif Platforms struggle to effectively facilitate convers many communities to limit or completely shut dov comments.

MEASURE BIAS



Unintended Bias



False "toxic" positives

A naively trained model on will have some strong unintended biases illustrated by these false-positive examples...

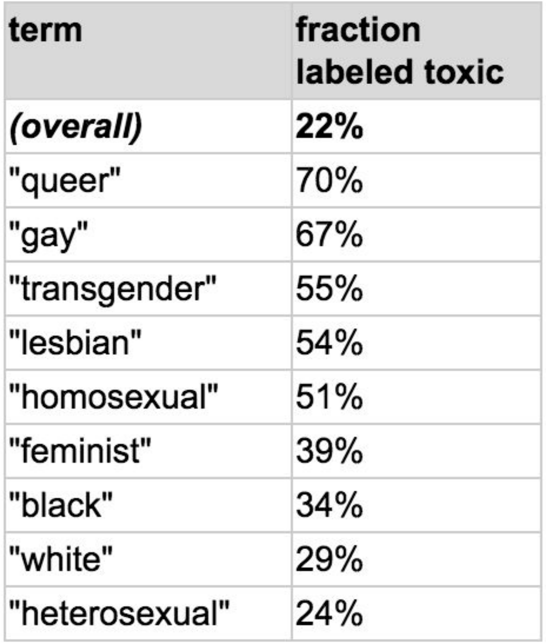
Comment	Toxicity score
The Gay and Lesbian Film Festival starts today.	0.82
Being transgender is independent of sexual orientation.	0.52
A Muslim is someone who follows or practices Islam.	0.46

How did this happen?

ML over-generalizes due to:

- Insufficient data
- The 'real' distribution is skewed

The model is not able to distinguish toxic from non-toxic uses of many identity words (and some others too, e.g. donkey)



term	fraction labeled toxic
(overall)	**22%**
"queer"	70%
"gay"	67%
"transgender"	55%
"lesbian"	54%
"homosexual"	51%
"feminist"	39%
"black"	34%
"white"	29%
"heterosexual"	24%

Unintended Model Bias vs Unfairness

- **Model: Unintended Bias** (A subset of examples has an unintended score distribution)
  Application: Unfairness (Unfair impact on people)

- Unintended bias can easily lead to unfair applications.

- **Every application of ML needs to consider the potential impact of unintended bias on the application's impact on society (fairness, inclusivity, etc).**
  - Unintended bias can lead to behaviour that increases, or decreases, the prevalence of mentions of an identity group (or it may have not effect); e.g. human pre-moderation, post-moderation, and batch moderation respectively.

How to measure Unintended Bias?

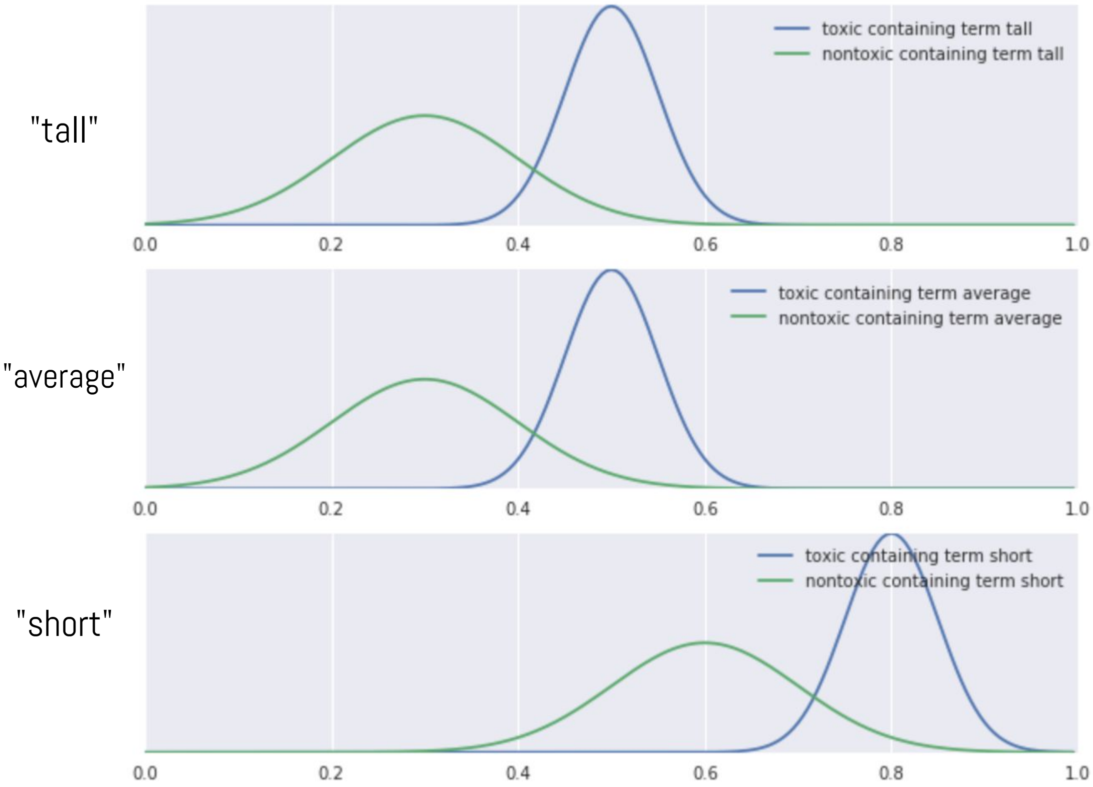
How good is the model at distinguishing good from bad examples? (ROC-AUC)
AUC (for a given test set) = Given two examples, one in-class (e.g. one is toxic and the other is not), AUC is the probability that the model will give the in-class example the higher score.

*Pinned AUC (for a given term, **t**, in a test set) =*
*AUC(all N examples with **t** & N representative examples from the test set)*

Pinned AUC < AUC if the model gives unusually high (or low) scores to examples containing the term **t**. PinnedAUCΔ = if AUC > PinnedAUC then (AUC - PinnedAUC) else 0.

Unintended bias for identity terms = ∑ PinnedAUCΔ(**t**, **s**), for each identity term **t** in a balanced test set **s** (e.g. a synthetic test set based on templates with identity terms)

Pinned AUC



$$PinnedAUC(t) = AUC(D_t + sample(D))$$

for identity term ***t*** and full dataset ***D***

	AUC	Pinned AUC
Tall	0.93	0.84
Average	0.93	0.84
Short	0.93	**0.79**
Combined	0.79	N/A

Mitigating unintended bias: re-balance the dataset

Where to get non-toxic examples about terms that are most frequently in toxic comments?

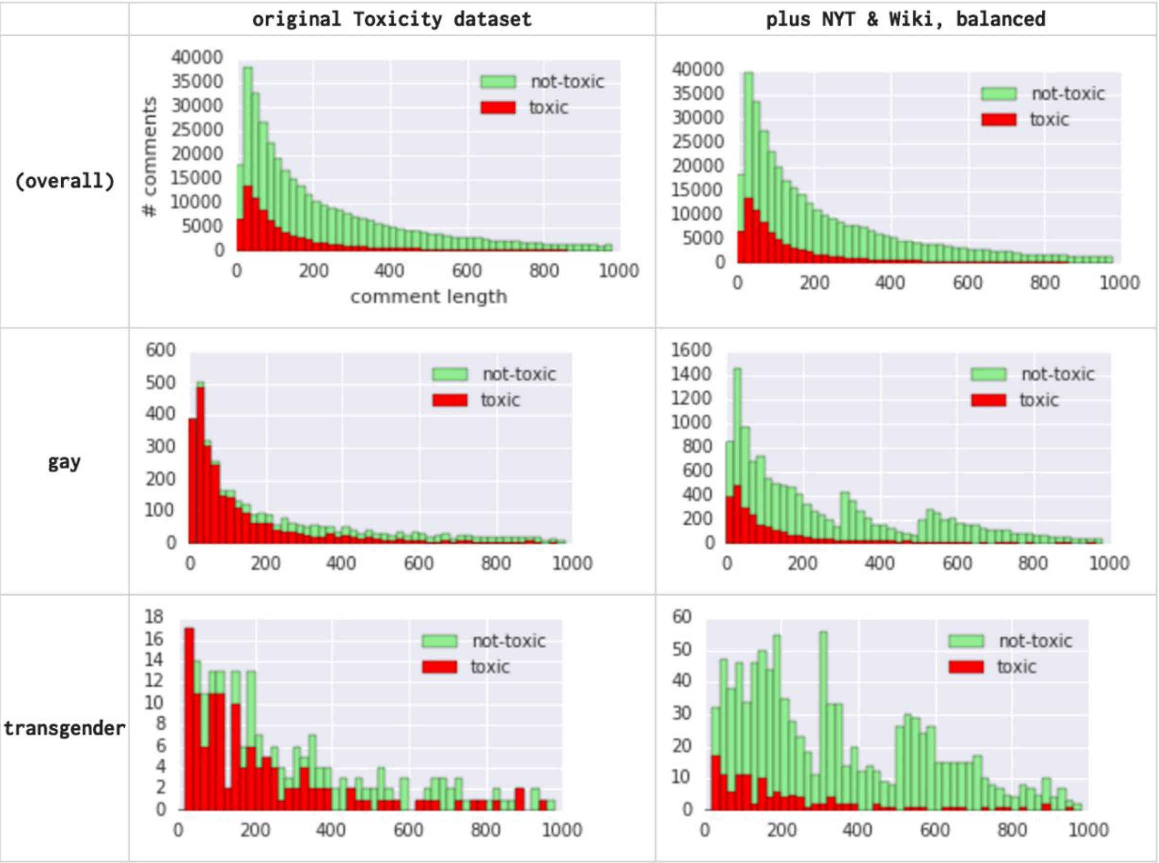
- Wikipedia Article Pages! (or other reviewed sources; reviewed comments, articles, etc)
- Re-balance the examples for each term (by length, this is important)

Potential issues:

- Text in article pages is not the same as text in comments, will this work?
- Will you have enough examples?

Mitigating unintended bias? re-balance the data

term	**fraction labeled toxic**
(overall)	**22%**
"queer"	70%
"gay"	67%
"transgender"	55%
"lesbian"	54%
"homosexual"	51%
"feminist"	39%
"black"	34%
"white"	29%
"heterosexual"	24%



False positives - some improvement

Comment	Old	New
The Gay and Lesbian Film Festival starts today.	0.82	0.01
Being transgender is independent of sexual orientation.	0.52	0.05
A Muslim is someone who follows or practices Islam.	0.46	0.13

Overall AUC for old and new classifiers within noise of retraining.

Many open questions

- Where to get a balances test set of identity terms?
- Should we be doing a squared error calculation?

Adversarial examples from public demos help a lot too.

But this does not make a 'perfect' model - that does not exist, a lot more hard work is needed here, and this will be a challenge for a long time.

https://github.com/conversationai/unintended-ml-bias-analysis
(built on Wikipedia, includes ML models, and mitigation methods)

THANKS!

Marie Pellat

PerspectiveAPI.com

DEMO LINKS

- Authorship + Slider
- NYT moderator
- Wikipedia (unsorted) vs (sorted)
- Wikiviz
- Disqus Toxicity Filter
- Coral Project
- Wonder Chrome Extension
- Kaggle (public ML competitions on toxicity)

TEAM LINKS

- Perspective API (public version has only limited set of models)
- API documentation
- Team research page
- Blog