

# Hands-on Tutorial: Measuring Unintended Bias in Text Classification Models with Real Data

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, Lucy Vasserman

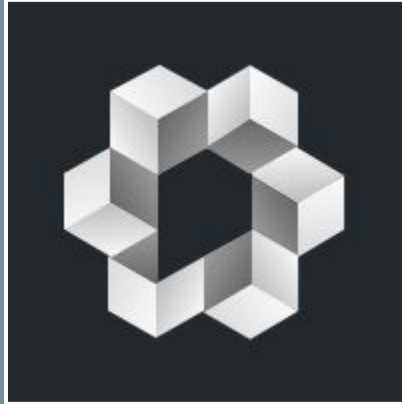


# Outline

- Who we are
- The Problem
- Civil Comments Data
- Training a deep learning model
- Metrics for ML fairness
- Hands-on tutorial

# Conversation-AI

ML to help improve  
online conversations  
at scale



# Perspective API

ML Models to score perceived impact of a comment on a conversation



Likely to be perceived as toxic (0.99) [Learn more](#)

[SEEM WRONG?](#)

You're a stupid idiot!

## False "toxic" positives

A naively trained model 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

# The Challenge

**Goal:** Train an ML model to detect toxicity in online comments.

## Questions:

- How can we determine if there is bias in such a model?
- Where does the bias come from?
- What can we do to mitigate bias?

# Civil Comments Dataset

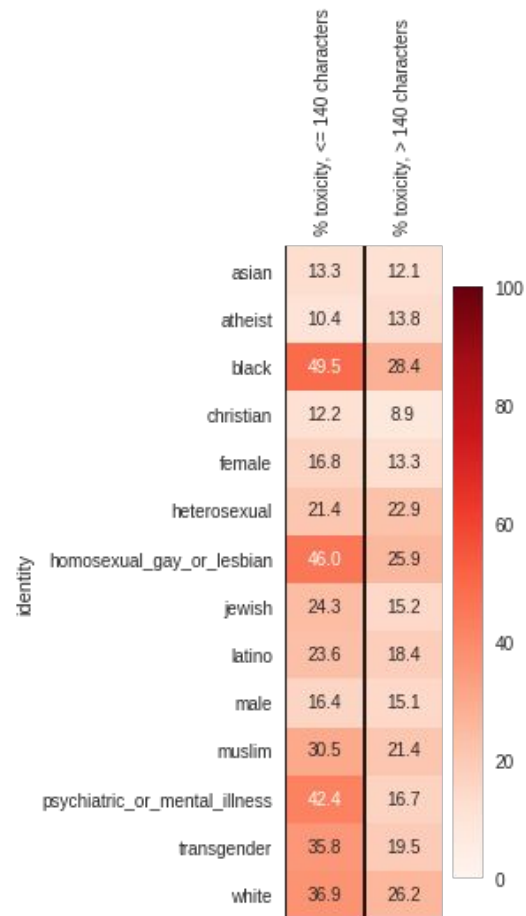
- ~2 million comments released by Civil Comments platform
- Collected via civil comments plugin on comment sections of online publications
- Annotated by Conversation AI for:
  - Toxicity, Obscenity, Sexually Explicit, Threats, Insults, Identity Based Attack
  - Subset (360k) annotated for various identity-related categories

**Disclaimer:** This dataset contains real comments from real users. Some of these comments may be very offensive.

# Assumptions

Dataset is reliable:

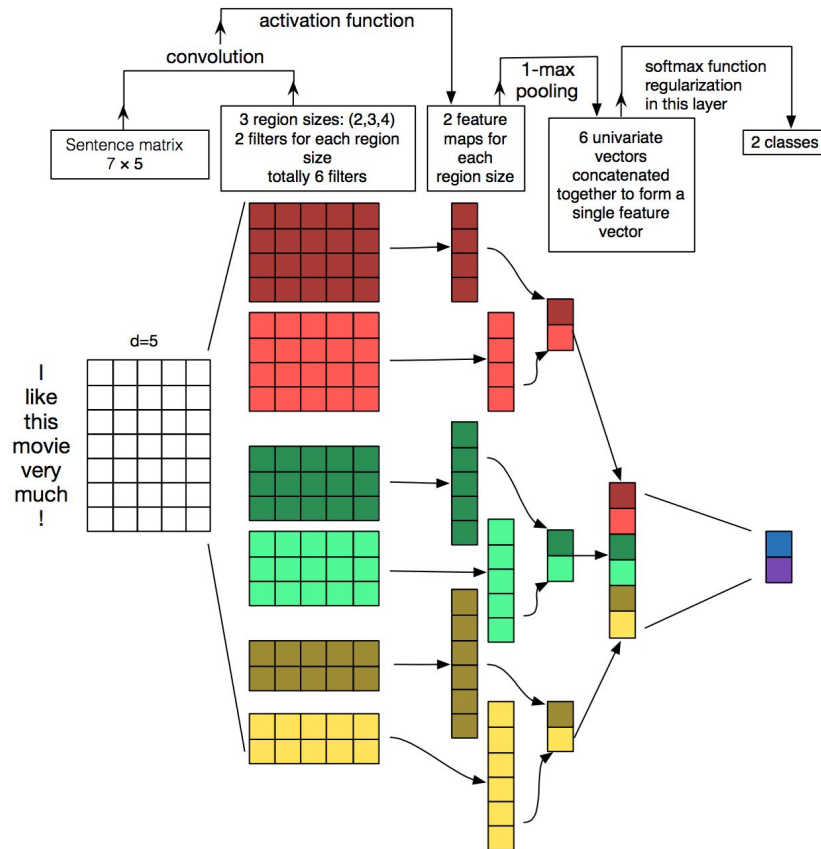
- Similar distribution as application
- Ignores annotator bias
- No causal analysis





# Deep Learning Model

- CNN architecture
- Pretrained GloVe Embeddings
- Keras Implementation

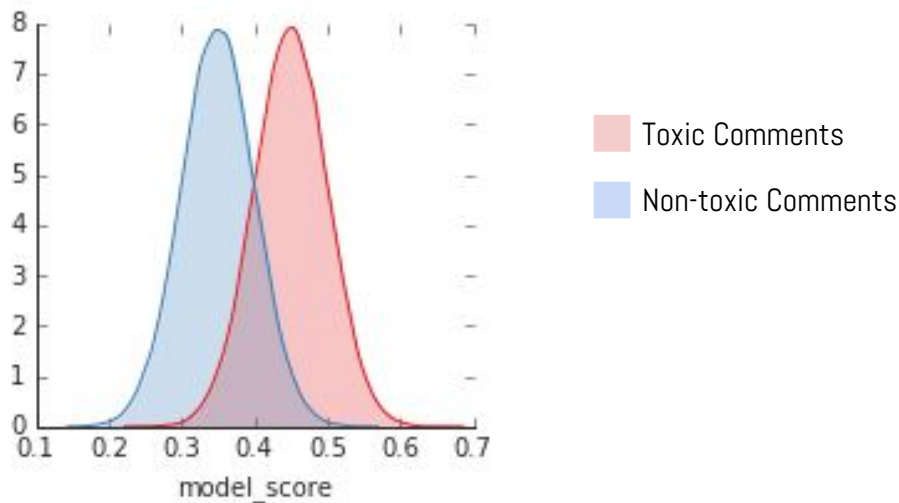


Source: Zhang, Y., & Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820.

# Measuring Model Performance

*How good is the model at distinguishing good from bad examples? (ROC-AUC)*

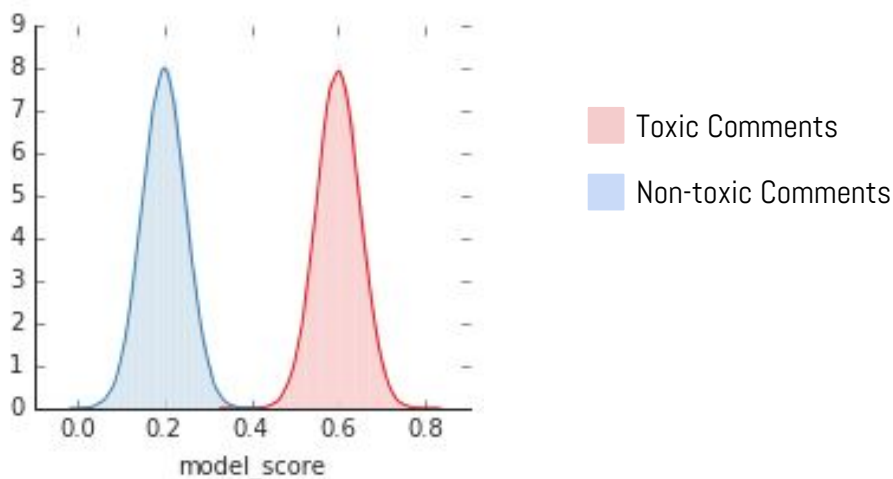
AUC (for a given test set) = Given two randomly chosen examples, one in-class (e.g. one is toxic and the other is not), AUC is the probability that the model will give the in-class example the higher score.



# Measuring Model Performance

*How good is the model at distinguishing good from bad examples? (ROC-AUC)*

AUC (for a given test set) = Given two randomly chosen 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.

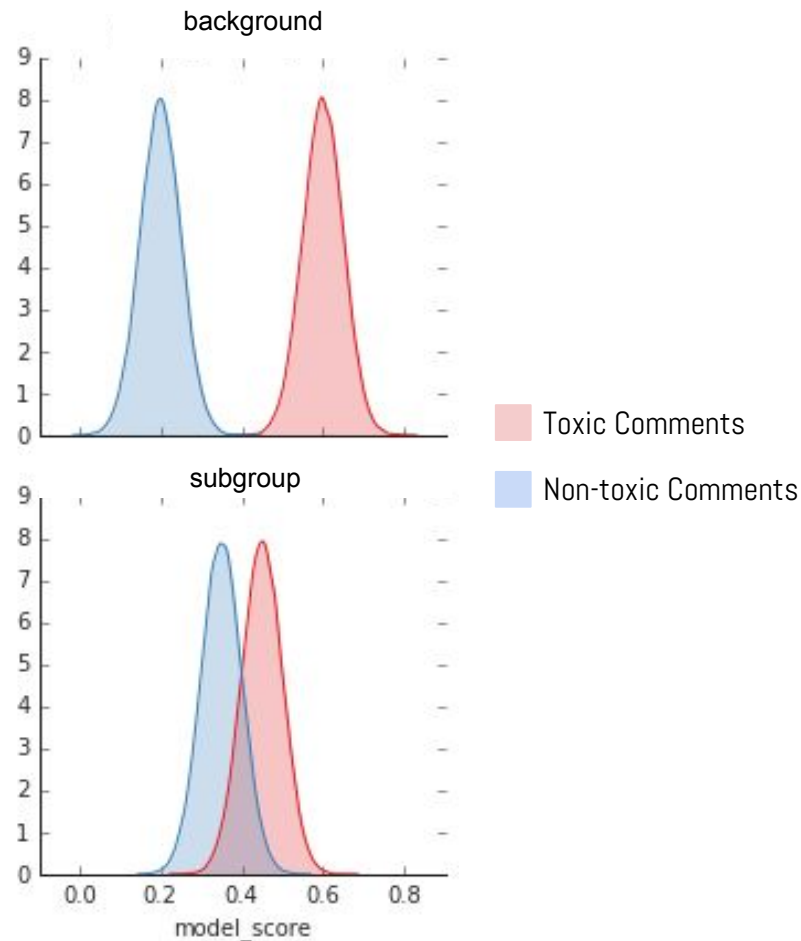


# Types of Bias

## *Low Subgroup Performance*

The model performs worse on subgroup comments than it does on comments overall.

**Metric:** Subgroup AUC



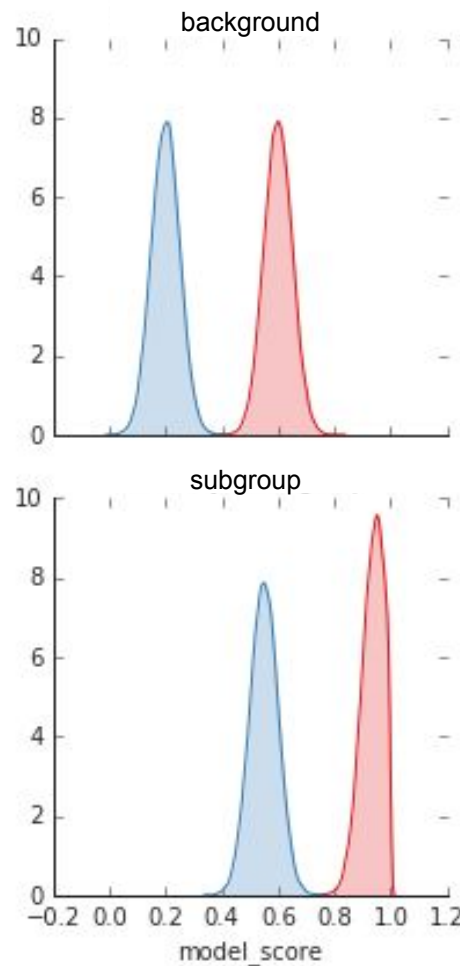
# Types of Bias

## *Subgroup Shift (Right)*

The model systematically scores comments from the subgroup higher.

**Metric:** BPSN AUC

(Background Positive Subgroup Negative)



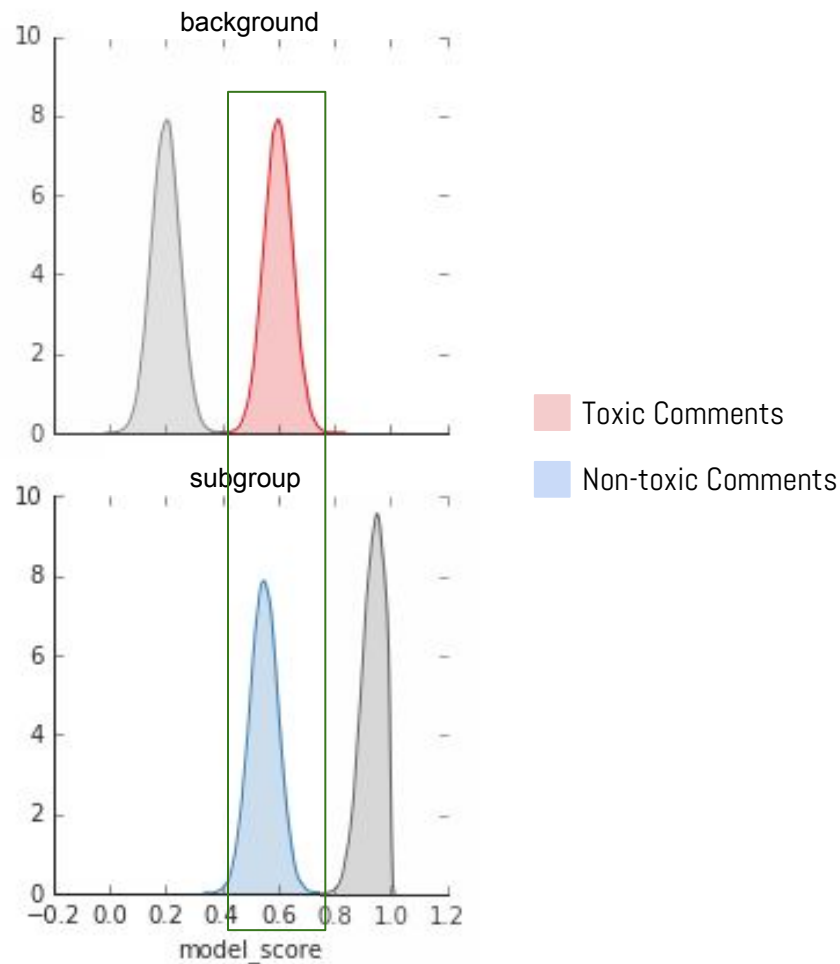
# Types of Bias

## *Subgroup Shift (Right)*

The model systematically scores comments from the subgroup higher.

**Metric:** BPSN AUC

(Background Positive Subgroup Negative)



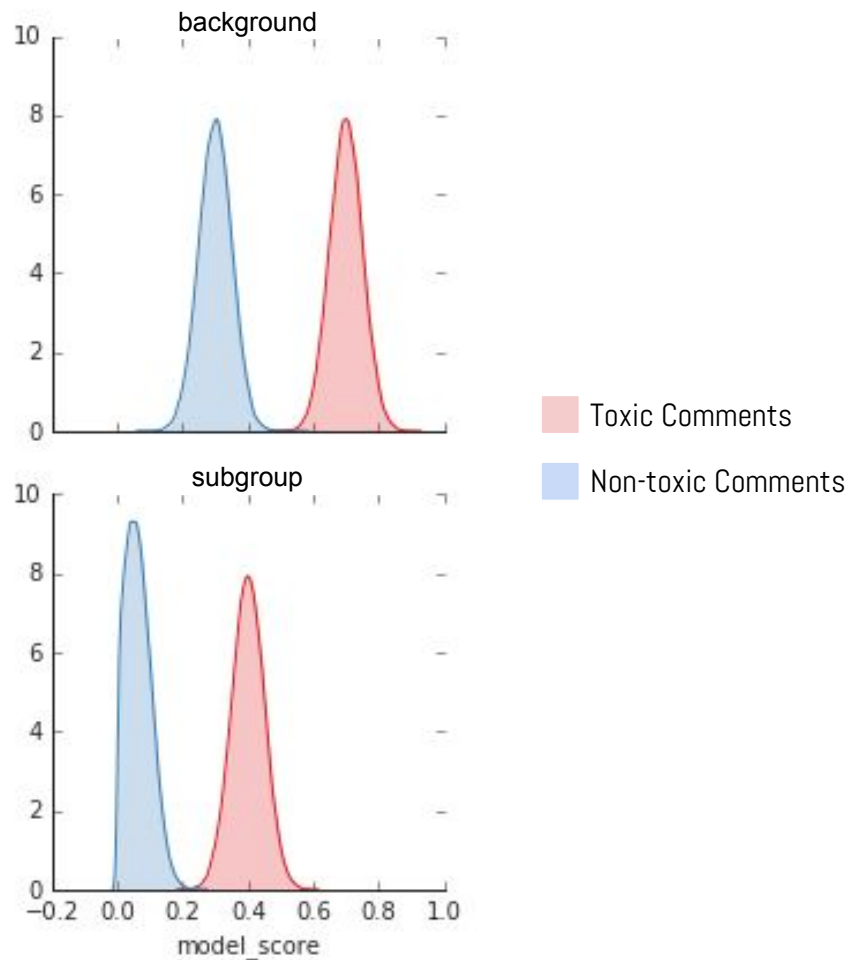
# Types of Bias

## *Subgroup Shift (Left)*

The model systematically scores comments from the subgroup lower.

**Metric:** BNSP AUC

(Background Negative Subgroup Positive)



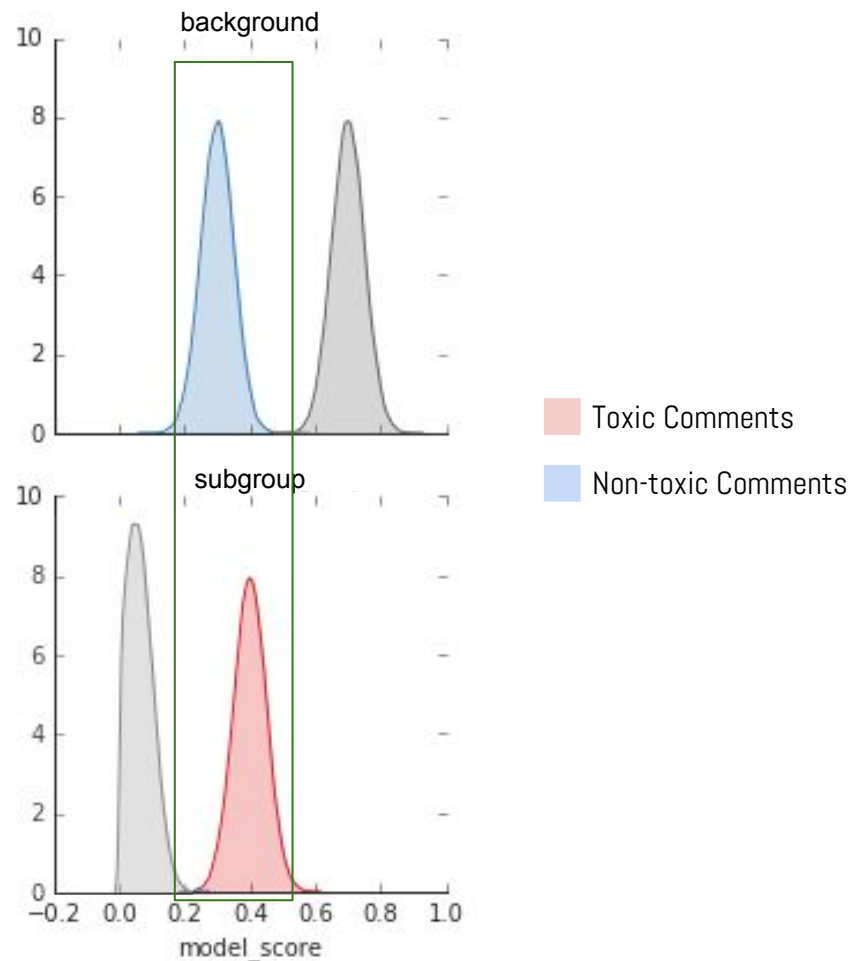
# Types of Bias

## *Subgroup Shift (Left)*

The model systematically scores comments from the subgroup lower.

**Metric:** BNSP AUC

(Background Negative Subgroup Positive)





# Converstion AI / Fairness Resources

- [Measuring and Mitigating Unintended Bias in Text Classification](#)
- [Conversation AI Research Post](#)
- [Unintended Bias Github Repository](#)
- [Unintended Bias Blog Posts](#)
- [Google Developers Blog](#)
- [Fairness Crash Course](#)

# Tutorial

# Tutorial Setup

1. Navigate to colaboratory at <https://bit.ly/2RnMKEf>
2. Click connect in the top right corner
3. SHIFT + ↵ to run a cell

Questions