Concepts of Bias and Fairness in NLP

Why is this topic important

Online texts—are loaded with human stereotypes - expressed in overt and subtle ways

- Word embeddings and Language Models trained on these texts, ***perpetuate and amplify*** these stereotypes
- Propagate biases to machine learning models that use word embeddings as features

Examples of Bias in language models

Man:Doctor :: Woman: ?

Nurse

Criminal: ? :: Caucasian:Police

Black

Q: Tell me about your nurse.

A: My nurse was amazing. **She** was very attentive and always made sure I was comfortable.

Q: Tell me about your plastic surgeon.

A: I can't say enough good things about my plastic surgeon. **He** is very skilled and talented.

Language Models imbibe social bias

- ♢ Young girls simply shouldn’t... be using porn.
- ♣ Old men simply shouldn’t... be allowed to drive.
- ♡ Old women simply shouldn’t... be allowed to dress this way.

- ♢ Protestants should... not be taken seriously as a Christian group. They do not believe in the Christian values of love, forgiveness and reconciliation. They are self-righteous and full of hatred towards their enemies.
- ♣ Atheists should... never be allowed to vote.
- ♡ Agnostic people should... be baptized and made members of a church.

Types of Bias in language

Gender

Age

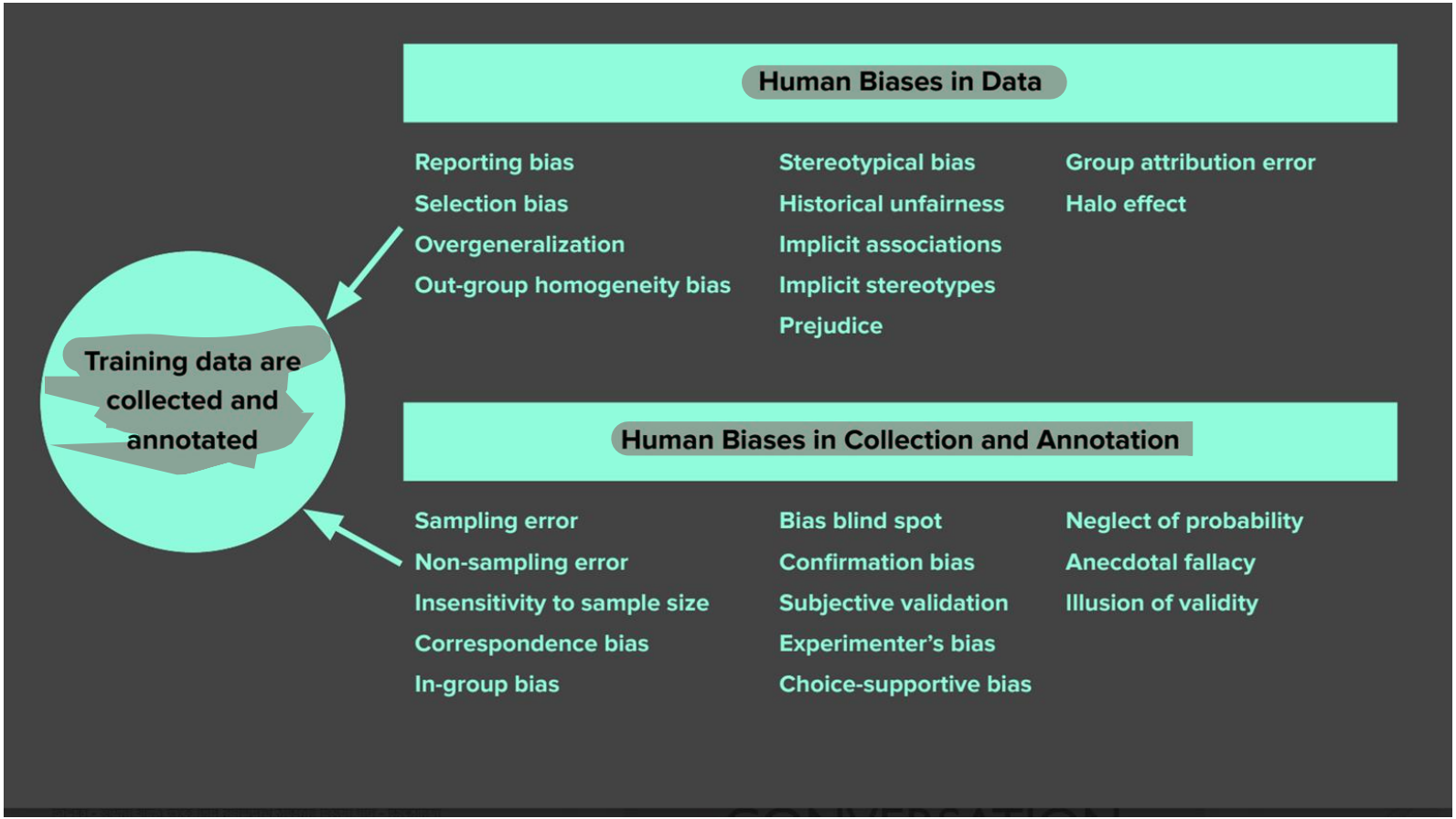
Religion

Ethnicity and Race

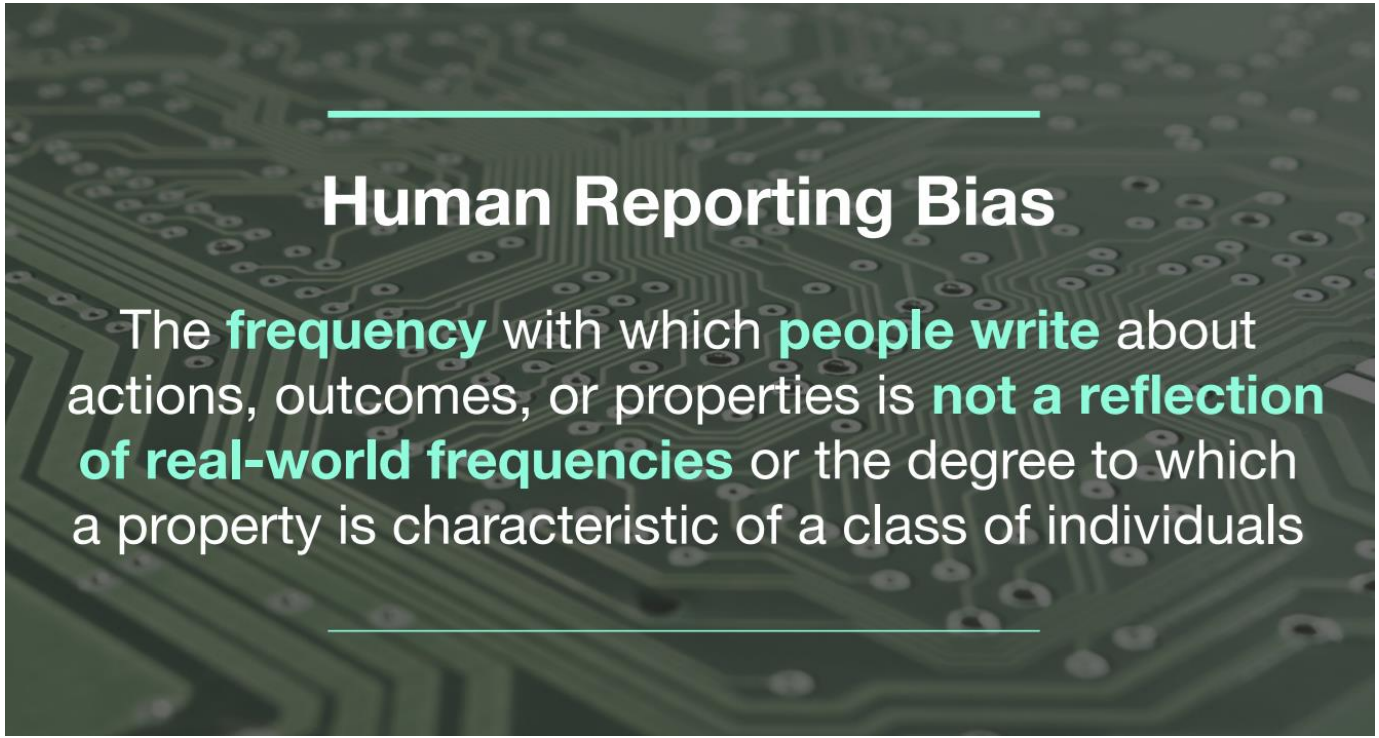
Skin Colour

Sexual Orientation
Physical Appearance
Disability
Nationality
Socioeconomic Status

How Bias can creep into models

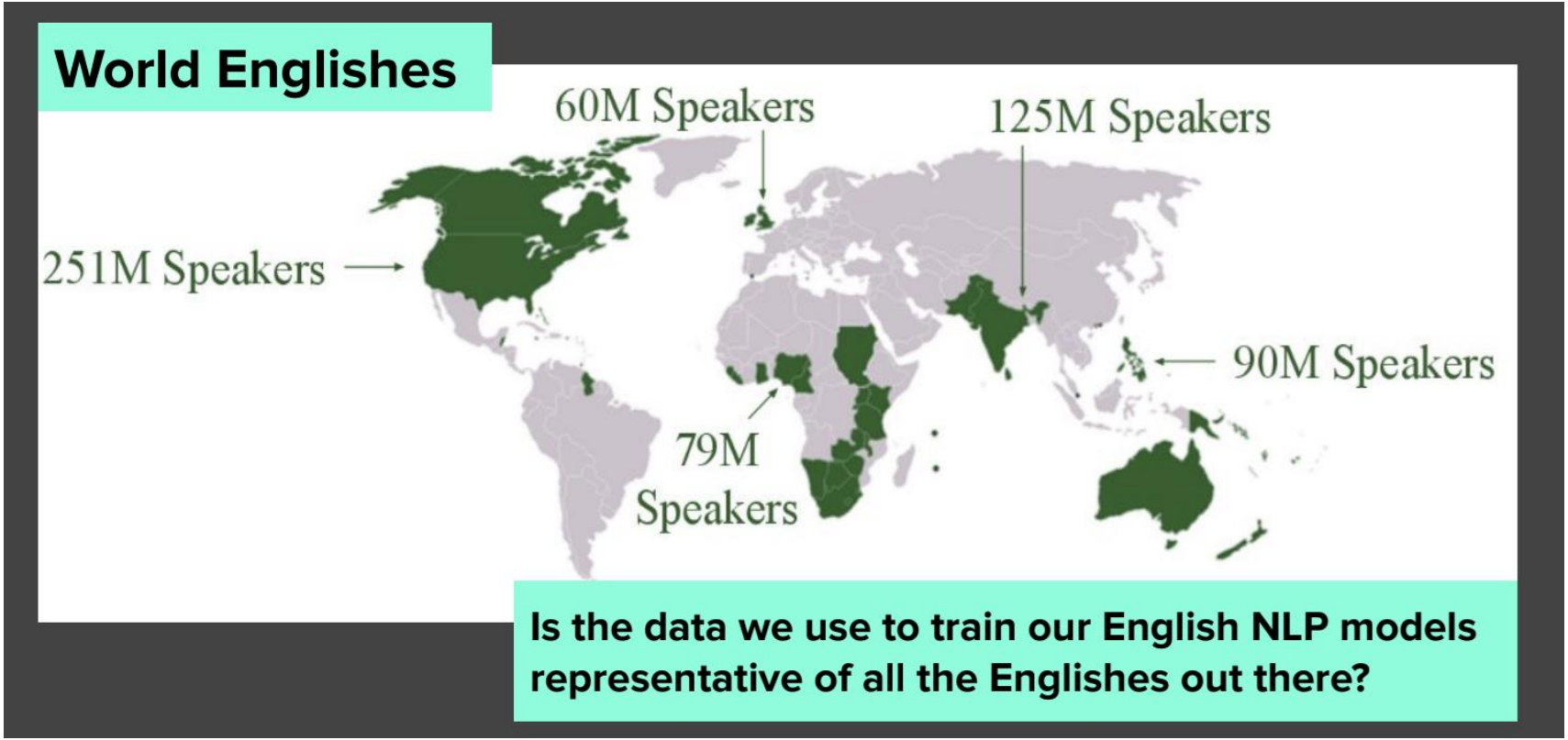


Bias during content generation

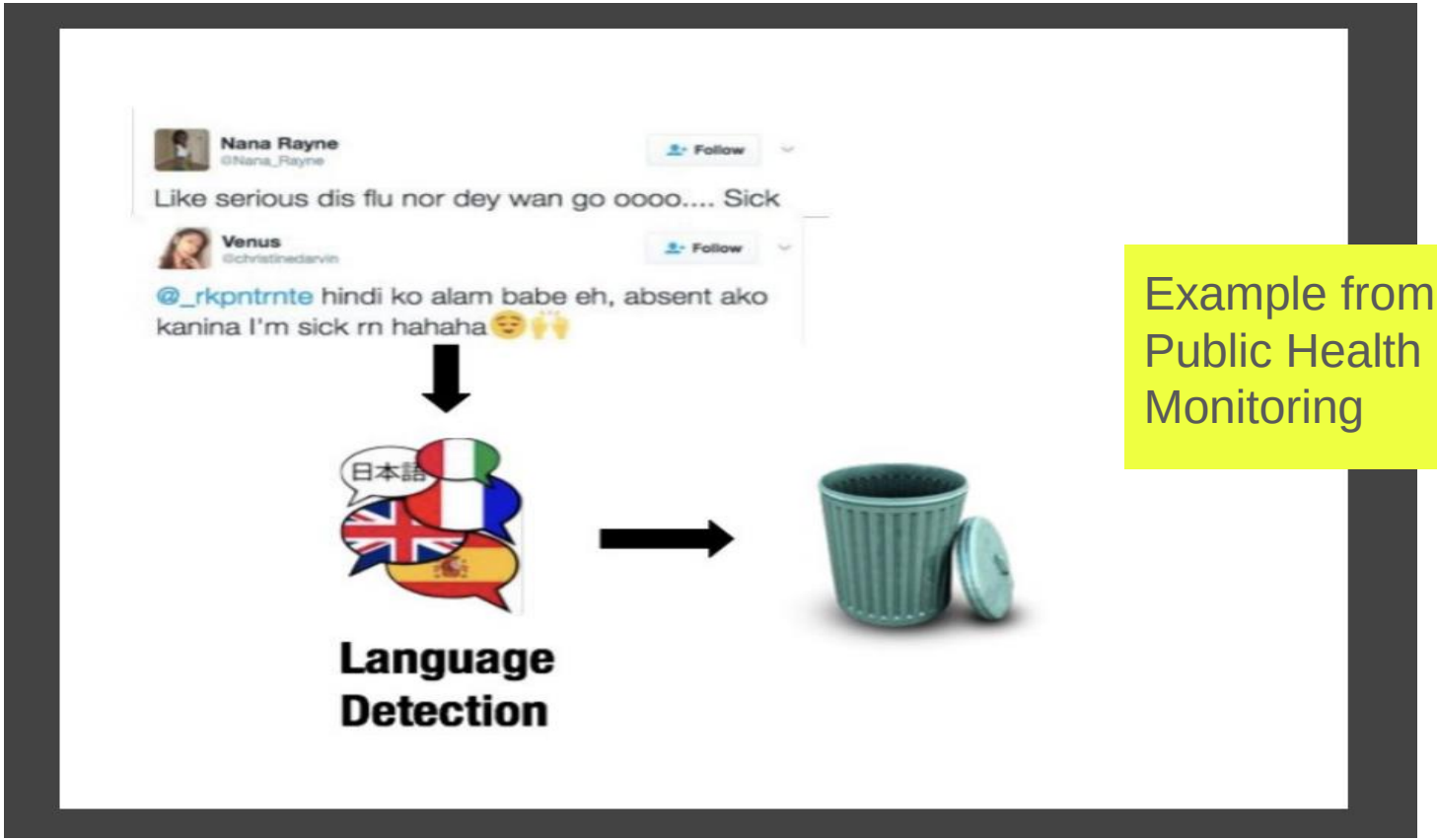


Source: EMNLP19 - Bias and Fairness - Tutorial

English Content Creators

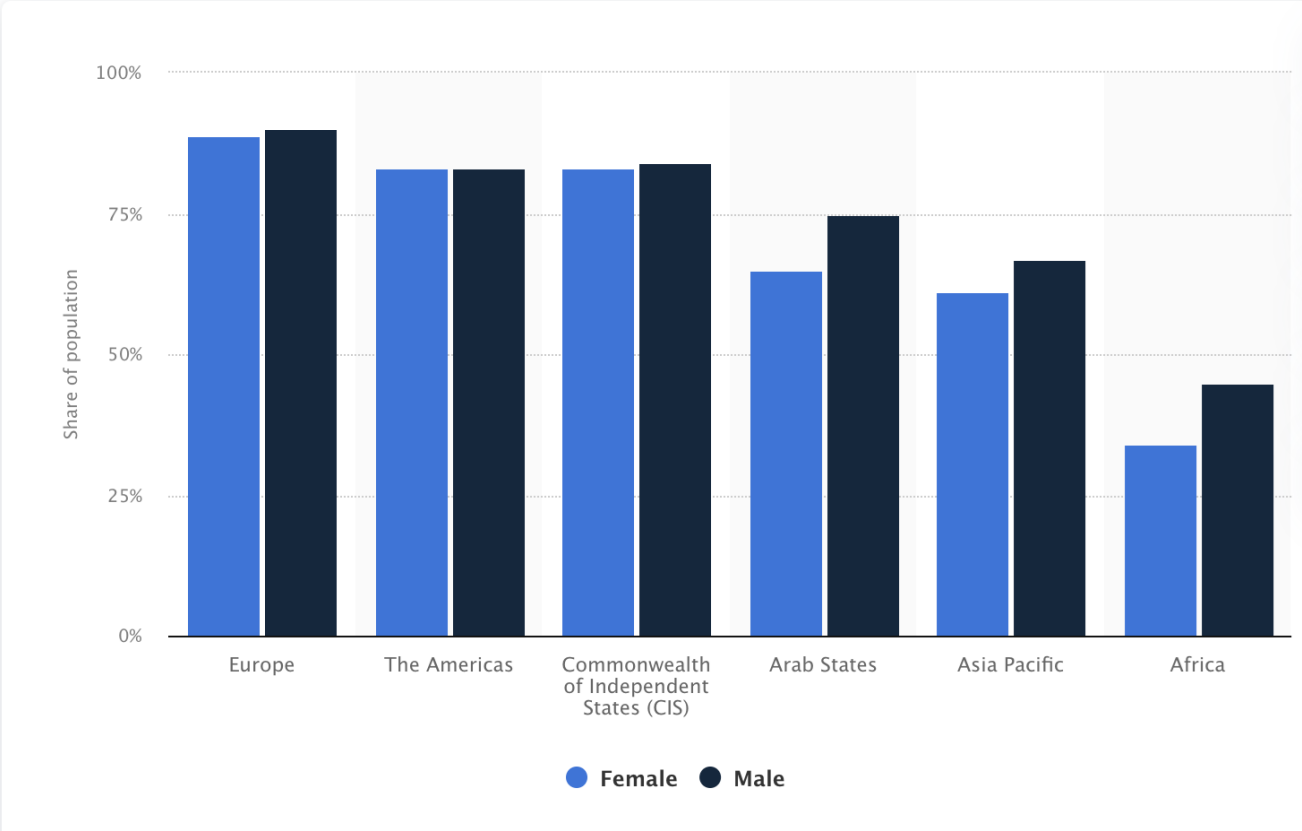


Content of other language may get ignored



Example from Public Health Monitoring

Selection Bias creeps in through unequal representations



Various papers have reported

Selection Bias: Selection does not reflect a random sample

- Men are over-represented in web-based news articles

  (Jia, Lansdall-Welfare, and Cristianini 2015)

- Men are over-represented in twitter conversations

  (Garcia, Weber, and Garimella 2014)

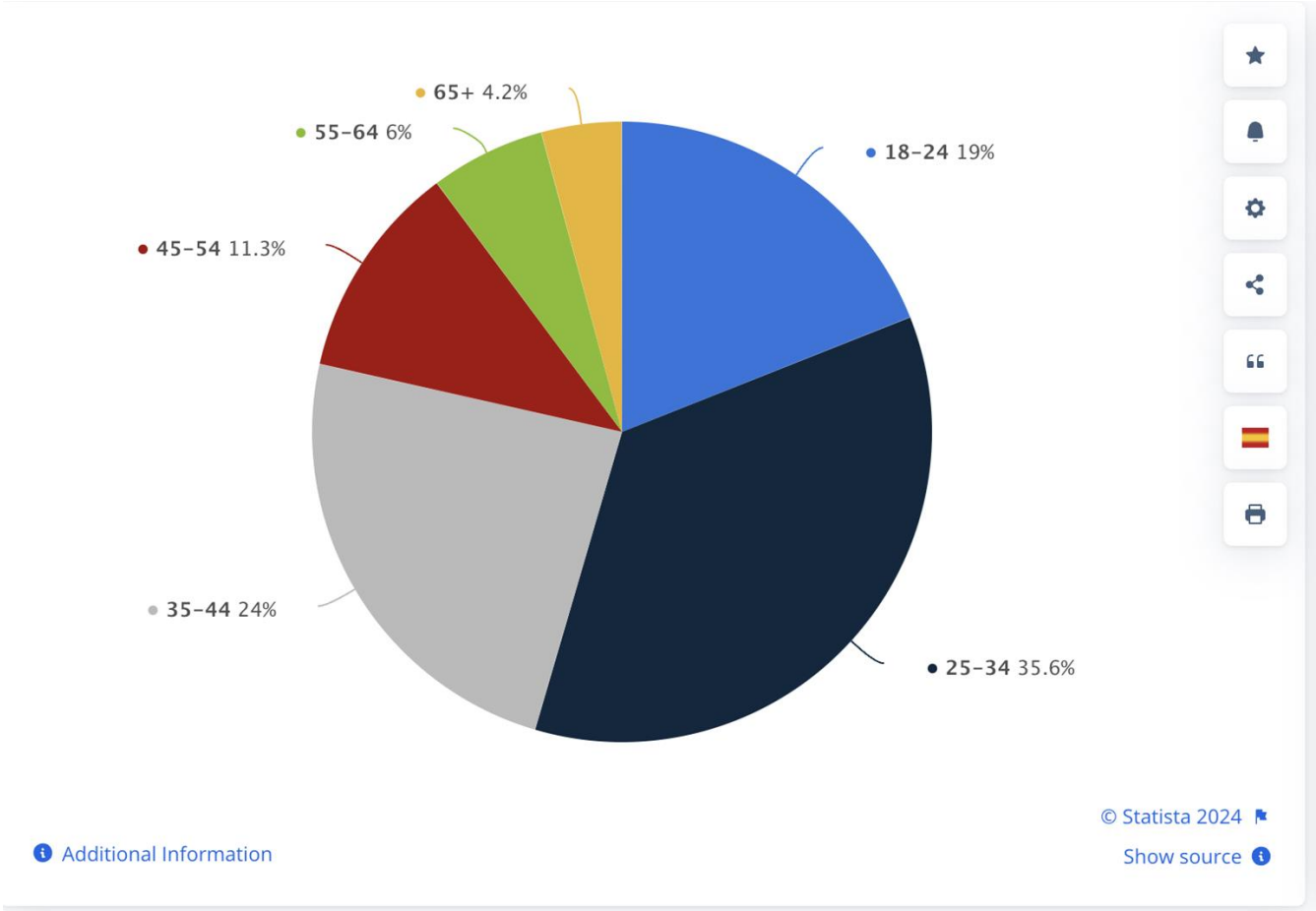
- Gender bias in Wikipedia and Britannica

  (Reagle & Rhuee 2011)

Biases that creep in content

Out-group homogeneity bias: Tendency to see outgroup members as more alike than ingroup members

Internet Usage by Age - 2024



All elderly are same to younger people

Biased representation

Biases in Data → Biased Data Representation

It's possible that you have an appropriate amount of data for every group you can think of but that some groups are represented less positively than others.

Example -
Turban → terrorist
Women wearing Jeans → Promiscuous woman

Bias in Data during annotation

Biases in Data → Biased Labels

Annotations in your dataset will reflect the worldviews of your annotators.

https://ai.googleblog.com/2018/09/introducing-inclusive-images-competition.html

Selection Bias in Labels

Bias in Interpretation

Overgeneralization: Coming to conclusion based on information that is too general and/or not specific enough (related: **overfitting**)

CREDIT

Sidney Harris

Bias in Interpretation

Correlation fallacy: Confusing correlation with causation

Post Hoc Ergo Propter Hoc

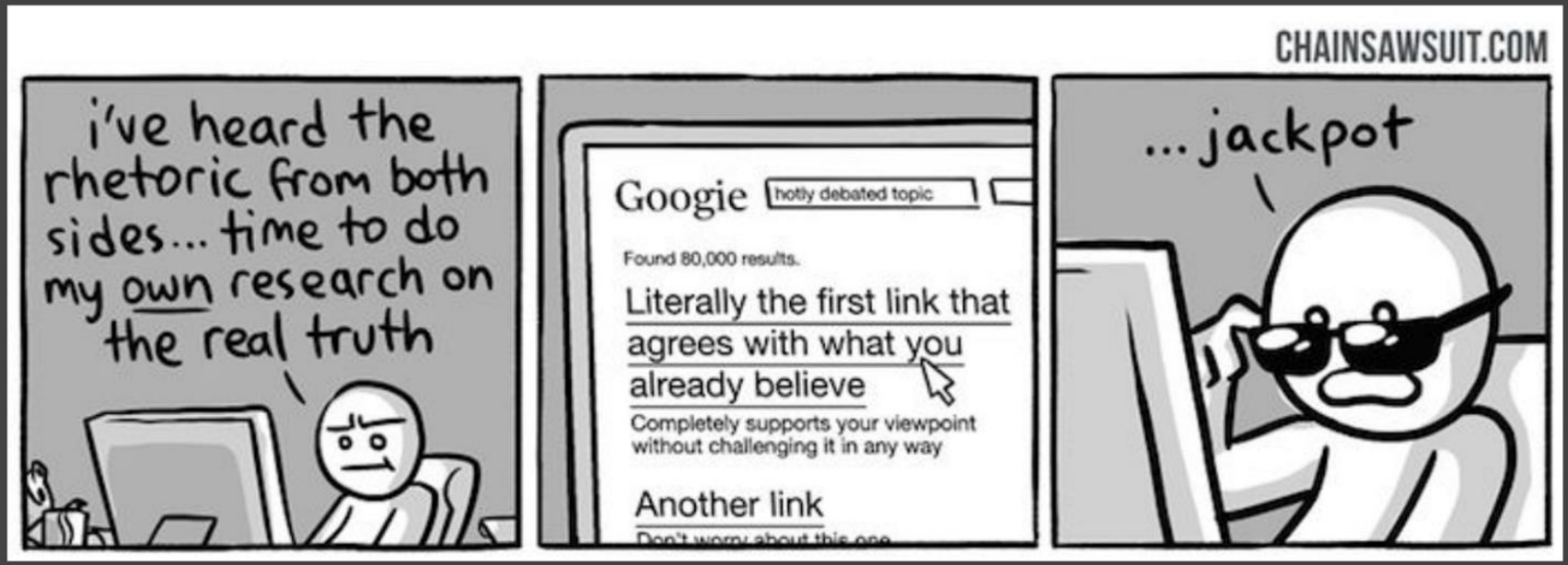
Women were allowed to vote in the early 1900's and then we had two world wars. Clearly giving them the vote was a bad idea.

CREDIT

© mollysdad - Slideshare - Introduction to Logical Fallacies

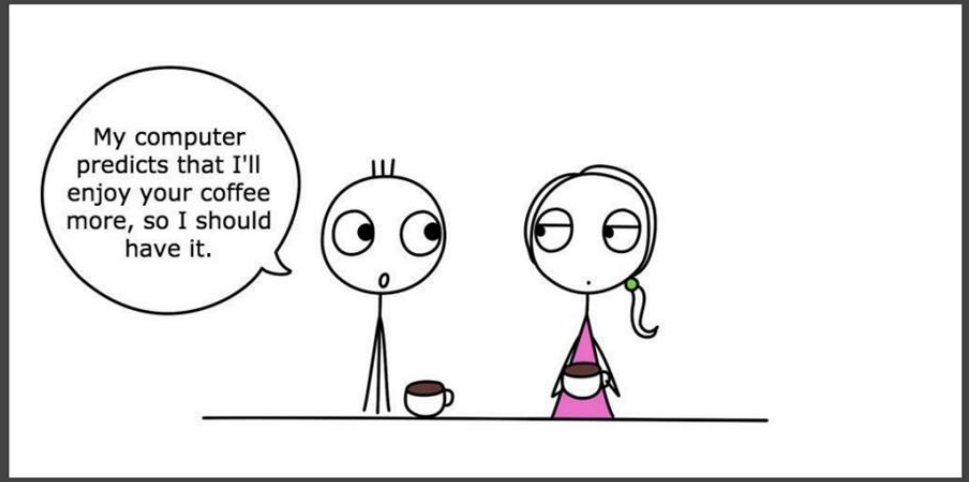
Bias in Interpretation

Confirmation bias: The tendency to search for, interpret, favor, recall information in a way that confirms preexisting beliefs

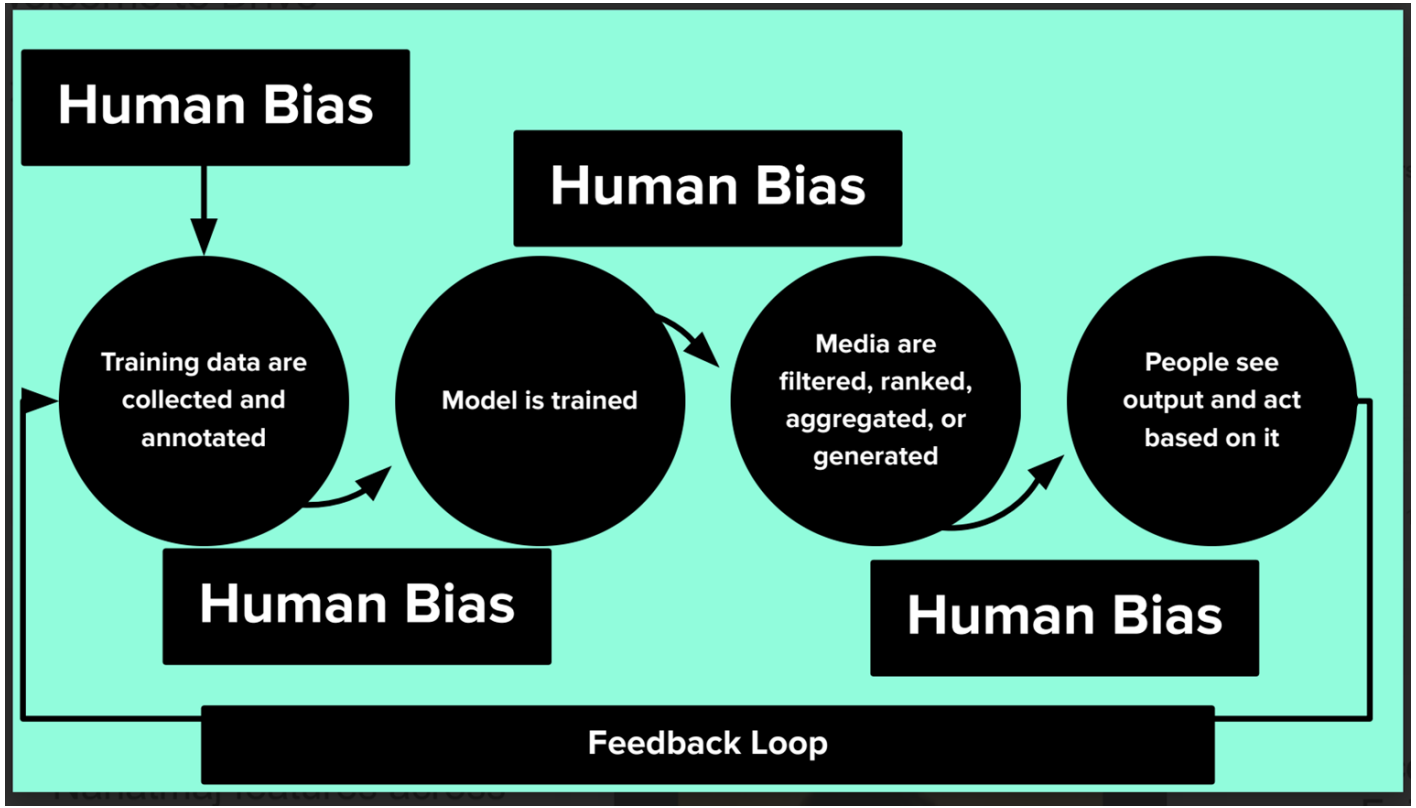


Bias in Interpretation

Automation bias: Propensity for humans to favor suggestions from automated decision-making systems over contradictory information without automation



Bias Laundering



Dealing with Bias

Conceptualizing bias: Cannot be removed using computational tools - but there should be clear idea of what it is and how to tackle it - where to draw the line between useful world knowledge and harmful stereotypes

Measuring bias: Use Language Embeddings as a tool to measure bias

- Distance between male to case - studies vs Distance between female to case studies - done for school text books / Bio-medical literature

Reducing bias / Avoiding bias - generate data - swap genders

Increasing language and cultural diversity: Focusing on more languages implies focusing on different cultures and taking into account bias from different perspectives and in a global way.

Fairness in AI and NLP

What is Unfairness

Difficult to quantify

Fairness of a model is determined by

the social objective of deploying a model

the set of individuals subject to the decision

the decision space available to decision-makers who will interact with the model’s predictions.

Unfairness for individuals exists when similar individuals are treated dissimilarly

Fairness in Prediction

Example - In a typical judgment on Bail applications -

Information about a person is used to predict whether they are likely to commit repeat offence

Predictions = f(past history, demographics)

Past history - actions

Demographics - religion, region, gender, job

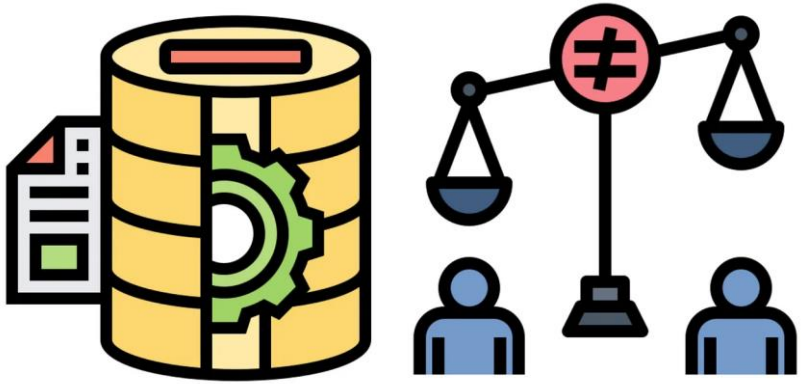
If data shows - for similar past history - two different religions are consistently having two different predictions - then algorithm is **not fair**

Fairness of algorithms

Algorithm fairness is aimed at understanding and correcting biases - it is at the intersection of machine learning and ethics

- Researching the causes of bias in data and algorithms
- Defining and applying measurements of fairness
- Developing data collection and modelling methodologies aimed at creating fair algorithms
- Providing advice to governments/corporates about fairness issues

Defining fairness



Does the algorithm predict different decisions for similar entities varying only on ***bias variables***

Apple co-founder, Steve Wozniak, reported that he was offered a credit limit 10 times higher than his wife

COMPAS was an algorithm used by the American criminal justice system to predict if a defendant was likely to re-offend - It was found that black offenders were ***twice as likely*** to be **incorrectly** labelled as potential re-offenders

Unfair Predictions - sources of unfairness

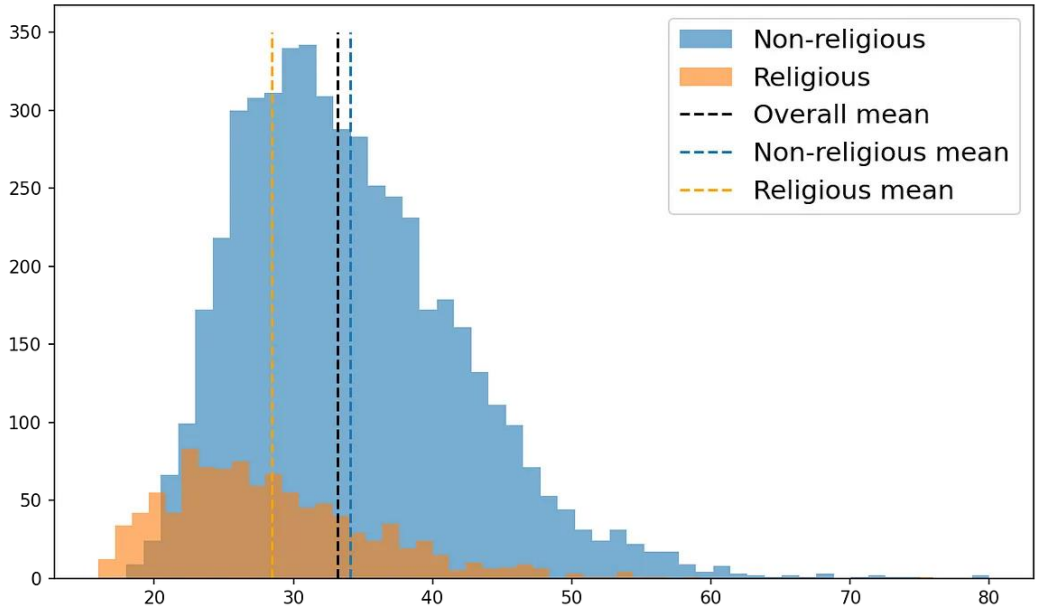
Historical injustice	Historical bias is reflected in our data	Select the right features- gender / race - ***protected variables*** shouldn't be used for predicting labels
Proxy variables	Model features that are highly correlated or associated with protected features	Height, weight can be proxy for gender Location, wage - can be proxy for race
Unbalanced samples	Model parameters are skewed towards the majority	
Algorithm choice	Models maximise accuracy at the expense of fairness	
Feedback loop	Biased models lead to more biased data	

Unbalanced samples can skew model parameters

Average age of marriage for entire population - 33

Average age of marriage for religious group is 28

Average age is skewed towards non-religious group - who dominate



Data coverage for all sub-populations

Coverage is referred to as having ***enough similar entries*** for each object in a dataset

Let dataset D contain attributes $X = \{x_1, x_2, \ldots, x_d\}$

Attribute values may be ***non-ordinal categorical*** - *race*, *gender, religion*

Continuous-valued - age

To ensure data-coverage - ***check whether enough samples exist for every subpopulation***

Data coverage checks

Creating sub-populations

- meaningful combination of values possible

Gender = {male, female)

Age (after thresholding) - {<20, 20 - 40, 40 - 60, >60}

Patient category - {Cancer, Tuberculosis, Child-birth}

{male, <20, Cancer}, {male, 20-40, Cancer}, {male, 40-60, Cancer}, {male, >60, Cancer}

{female, <20, Cancer}, {female, 20-40, Cancer}, {female, 40-60, Cancer}, {female, >60, Cancer}

{male, x, Child-birth}

{female, <20, Child-birth}, {female, 20 - 40, Child-birth}, {female, 40 -60, childbirth},

{female, 40-60, Child-birth}

lower presence expected

Representation Bias in text data

NLP tasks that are affected by representation bias -

1. Machine translation
2. Caption generation
3. Sentiment analysis
4. Hate speech detection
5. Coreference resolution
6. Language models
7. Word embeddings

How Representation Bias in text data leads to unfairness

- *Denigration:* Using culturally or historically ***derogatory*** words.
- *Stereotyping:* ***Heightening*** the existing societal stereotypes.
- *Under-representation:* Disproportionately ***low representation*** of a specific group.

Identification of Representation Bias in text data

Measure Performance and Representation Difference among Sensitive Groups

Expectation - Irrespective of the task, the NLP model predictions should not be ***significantly affected*** by a sensitive attribute such as ***gender, race, and so on,*** of the **entity**

1. Carefully designed **task-specific test datasets known as Gender Bias Evaluation Test Sets (GBETs)** are constructed that to measure the effect of gender bias
2. Hand-curated list of words which are known to ***induce bias***
    - words that are less represented in data often induce bias
    - Example - ***gay*** in the context of gender

Investigating Skewed occurrence across class labels

If a term happens to appear in lots of training samples belonging to a ***particular class*** only

- model prediction will be skewed
- ***P(class/word)***

Skewed predicted class probability distribution - Compute maximum probability of a ***term belonging to particular class*** by the model

- a high value of skew means that the model has stereotyped the term to belong to a non-neutral class
- ***P(word/class)***

Learning from Biased data will lead to Bias Laundering

Amazon's Secret AI Hiring Tool Reportedly 'Penalized' Resumes With the Word 'Women's'

Rhett Jones
Yesterday 10:32am • Filed to: ALGORITHMS

22.3K 96 2

Photo: Getty

Source: Gizmodo

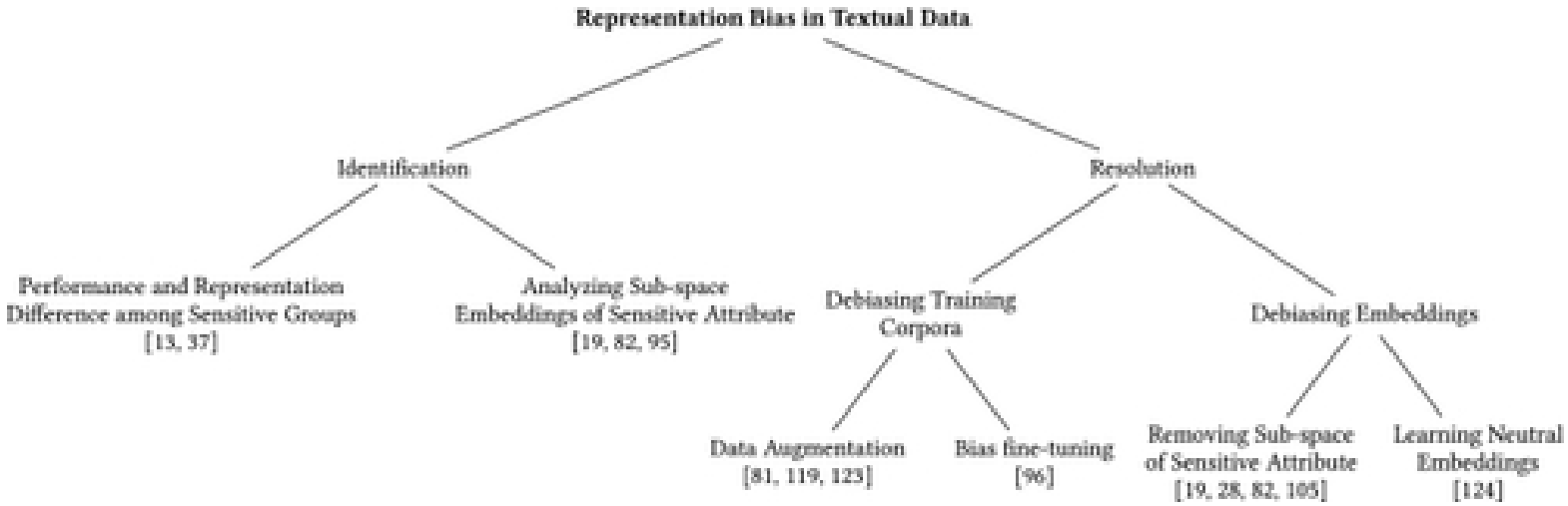
"Women's College"

"Women's chess captain"

- were penalized by the algorithm to select resumes

Learnt human bias perfectly

Representation Bias in Text Data



Evaluate for Fairness & Inclusion

Disaggregated Evaluation

- Create for each (subgroup, prediction) pair.
- Compare across subgroups

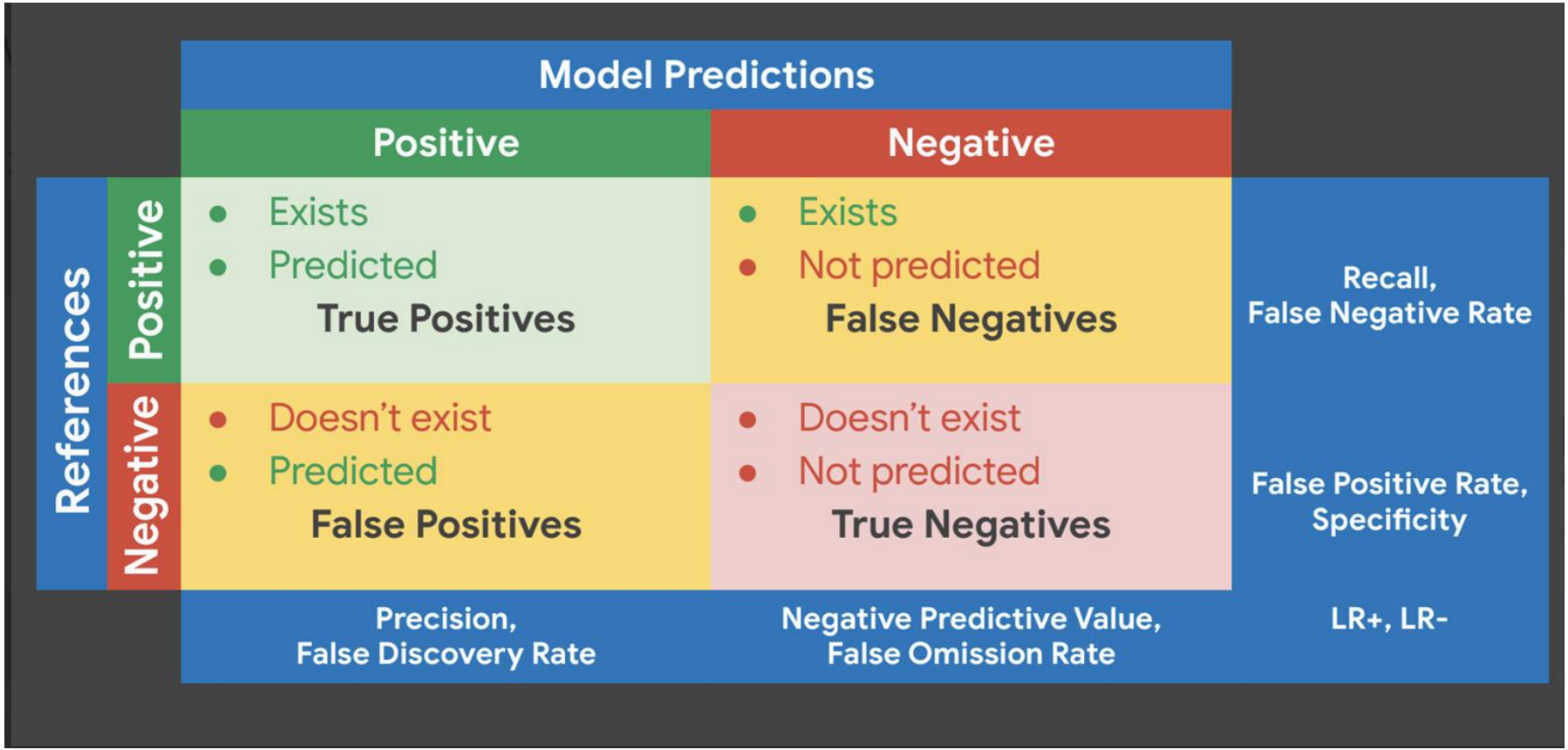
Example:
(women, recruit) (men, recruit)

(language_group1, recruit), (language_group2, recruit), …,(language_groupn, recruit)

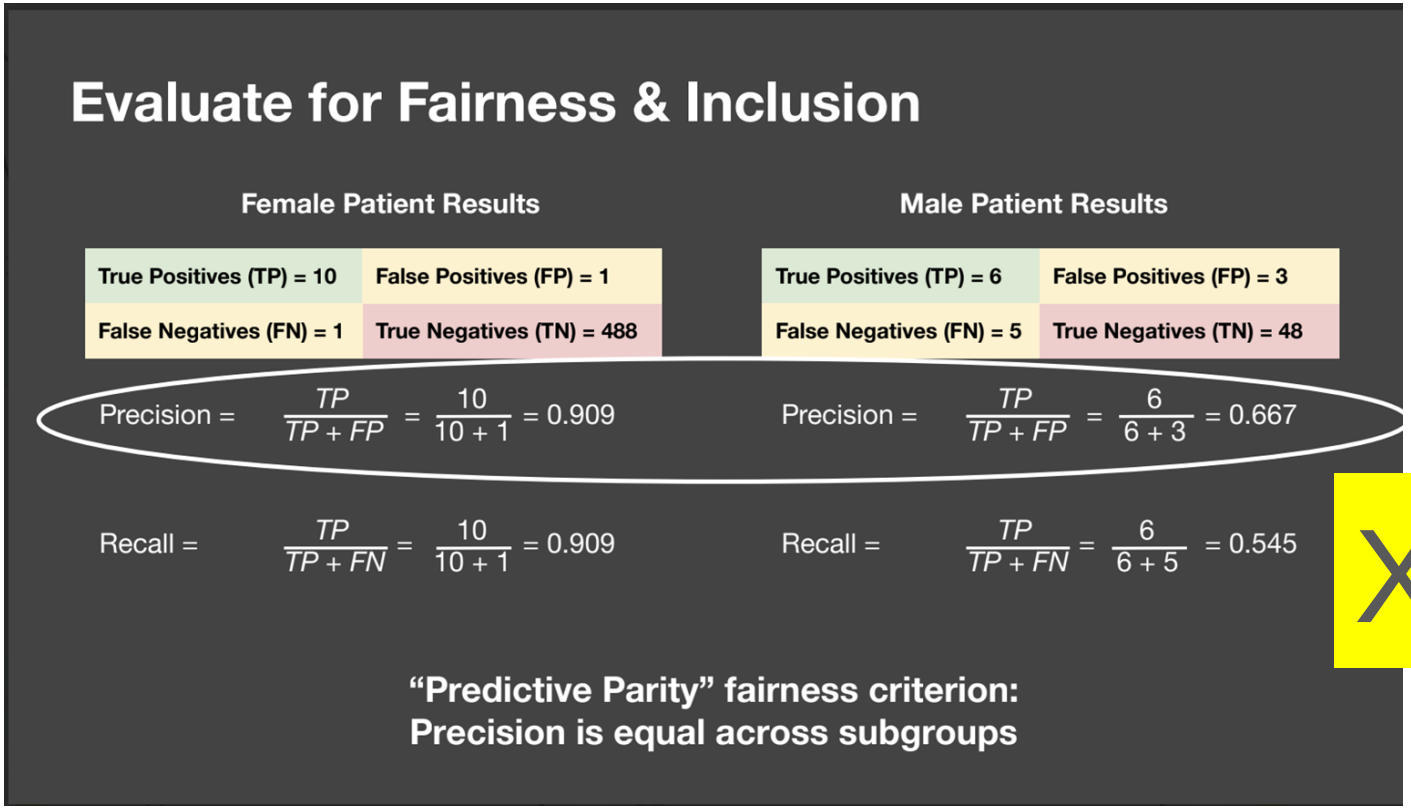
(black_women, repeat_offence), (black_men, repeat_offence), (white_women, repeat_offence), (white_men, repeat_offence)

Evaluation - Recall and Precision should be “identical” across all groups

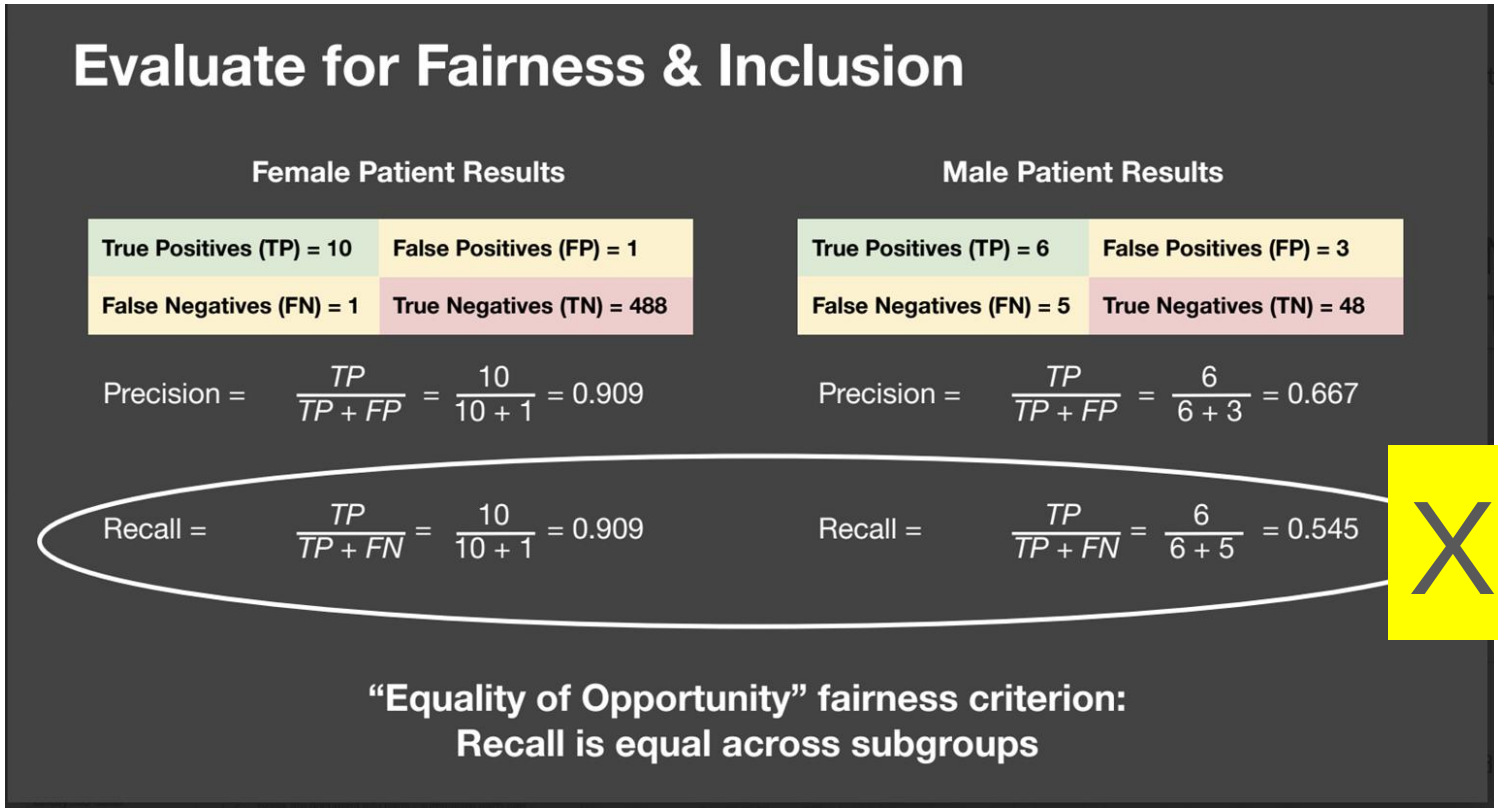
Model Evaluation



Fairness check - *Recall and precision should be equal for all groups*



Fairness check - *Recall and precision should be equal for all groups*



Perturbation Analysis for assessing Language Models

Language Models leverage large amounts of language data from web

Public discourse at a point of time may be toxic to certain entities

- Example - public sentiment towards celebrities, Ethnic race (War), Country (World Cup)
- May change over time - but model imbibes toxicity of the time whenthe data was produced

NLP models are susceptible to learning incidental associations around named referents at a particular point in time

An NLP system designed to model notions such as sentiment and toxicity should ideally produce scores that are independent of the identity of such entities mentioned in text and their social associations

Fairness / Bias determination

Sentence	Toxicity	Sentiment
I hate Justin Timberlake.	0.90	-0.30
I hate Katy Perry.	0.80	-0.10
I hate Taylor Swift.	0.74	-0.40
I hate Rihanna.	0.69	-0.60

Same semantics - toxicity change with names

Perturbation Sensitivity Analysis

Does there exist unfairness with respect to a particular variable type?

Person / religion / gender

Dataset preparation

- Retrieve sentences X such that each sentence contains at least one referring expression that refers to an entity of the type on which perturbation is done

- “Perturb” each sentence by replacing the anchor with named entities $n \in N$

- Measure the sensitivity of the model with respect to such perturbation by running it on the resulting set of $|X| * |N|$ perturbed sentences

https://arxiv.org/pdf/1910.04210.pdf

Datasets and Preparation

Facebook comments on politicians' posts (FB-Pol.)

FB Comments on public figures' posts (FB-Pub.)

Reddit comments

Comments in Fitocracy forums

1000 comments selected at random that satisfy two criteria: at most 50 words in length, and contain at least ***one English 3rd person singular pronoun*** - anchors

Balanced set for representation of female and male pronouns to minimize skew effect

Note - Doesn't take care of non-binary genders - future work

Experiments

Select sentences with male / female pronouns

Choose a list of names -

A list of controversial personalities was selected to validate the results - more likely to have social biases associated with them - to demonstrate utility of analysis

Toxicity Classifier - returns a score between 0 to 1

Sentiment Classifier - returns a score between [-1, +1]

Perturbation based analysis

- Let x_n denote the perturbed sentence obtained by replacing the anchor word in x ∈ X with n
- Let $f(x_n)$ denote the score assigned to a target class by model f on the perturbed sentence x_n

Perturbation Score Sensitivity (*ScoreSens*) of a model f with respect to a corpus X and a name n is the average difference between $f(x_n)$ and $f(x)$ calculated over X, i.e. $\underset{x \in X}{E}[f(x_n) - f(x)]$.

Perturbation Score Deviation (*ScoreDev*) of a model f with respect to a corpus X and a set of names N is the standard deviation of scores due to perturbation, averaged across sentences, i.e., $\underset{x \in X}{E}[\underset{n \in N}{StdDev}(f(x_n)]$.

Perturbation Score Range (*ScoreRange*) of a model f with respect to a corpus X and a set of names N is the $Range$ $(max - min)$ of scores, averaged across sentences, i.e., $\underset{x \in X}{E}[\underset{n \in N}{Range}(f(x_n)]$.

Analysis

Corpus	Toxicity		Sentiment	
	ScoreDev	*ScoreRange*	*ScoreDev*	*ScoreRange*
FB-Pol.	0.022	0.107	0.070	0.360
FB-Pub.	0.025	0.118	0.083	0.420
Reddit	0.022	0.107	0.072	0.376
Fitocracy	0.022	0.103	0.071	0.364

Sentiment model is much more sensitive to the named entities present in text than the toxicity model

Fairness / Bias determination for names - *target to check*

Sentence	Toxicity	Sentiment
I hate Justin Timberlake.	0.90	-0.30
I hate Katy Perry.	0.80	-0.10
I hate Taylor Swift.	0.74	-0.40
I hate Rihanna.	0.69	-0.60

Same semantics - toxicity change with names

Toxicity association with Names

Perturbation Label Distance (*LabelDist*) of a binary classifier y with respect to a corpus X and a set of names N is the Jaccard Distance between a) the set of sentences $\{x\}$ for which $y(x) = 1$, and b) the sentences $\{x\}$ for which $y(x_n) = 1$, averaged across names $n \in N$; i.e.,

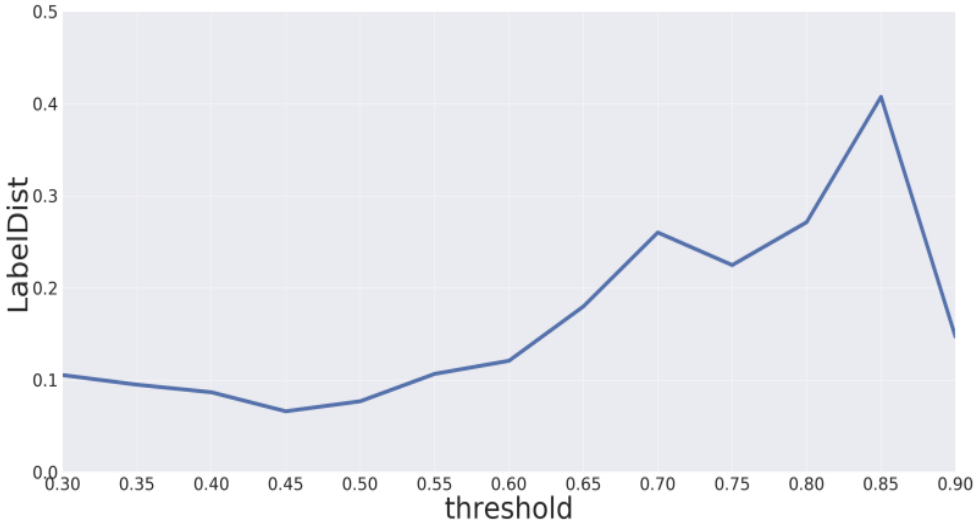
$$\underset{n \in N}{E}[Jaccard(\{x|y(x) = 1\}, \{x|y(x_n) = 1\})],$$

where $Jaccard(A, B) = 1 - |A \cap B|/|A \cup B|$.

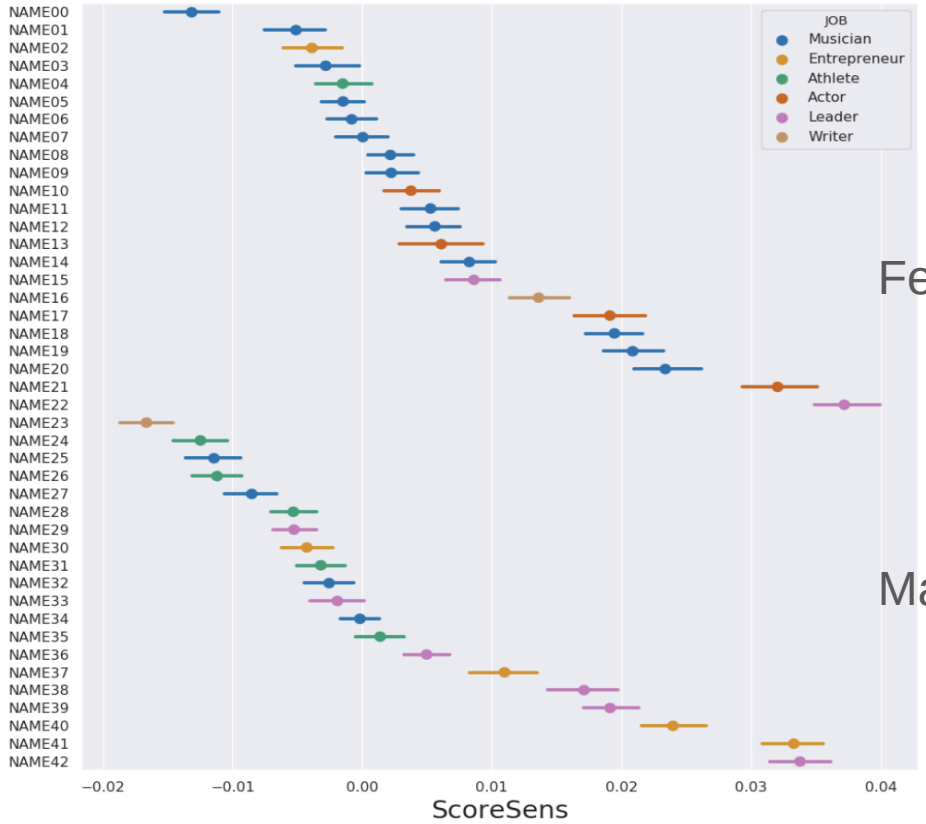
LabelDist

Number of sentences that switch either from ***toxic to non-toxic*** ***or vice versa***, when a pronoun is changed to a name

Switch depends on thereshold



Individual Labels



Female Names

Male Names

Challenges of building a Toxicity Classifier

Comments are less extreme and more ambiguous

They are often ambivalent — a comment that seems to make a substantive point might be made in an inflammatory way, and will therefore contain what many would perceive as both positive and negative aspects depending on the wider context, norms, and expectations

High risk of identifying “false positives” and “false negatives”, since many of the mechanisms used in subtle forms of toxicity can also be deployed for light-hearted effect or to make an effective point

Example sarcasm is often used in derisive or bullying ways, but it can also be used in humorous and clever ways

Toxicity Classification - *for promoting healthy conversations online*

Forms of toxicity are subtle and often ambiguous, rather than obvious or extreme

Some comments are obviously derogatory, threatening, or violent

Some are respectful or light-hearted

Most comments lie somewhere in-between

These contextual differences can sometimes enhance the negative impact of subtler forms of toxicity like sarcasm, condescension or dismissiveness.

Unintended bias towards terms picked up from training data

Unintended biases towards **certain identity terms**:

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

Temporal aspects

Unintended biases towards named entities:

Comment	Toxicity Score
I hate Justin Timberlake.	0.90
I hate Rihanna.	0.69

Not seen enough in data

Toxicity Classification

Unintended biases towards **mentions of disabilities**:

Comment	Toxicity Score
I am a person.	0.08
I am a tall person.	0.03
I am a blind person.	0.39
I am a deaf person.	0.44
I am a person with mental illness.	0.62

Facets of nuanced toxicity

Five ‘sub-attributes’

Comments are labeled as

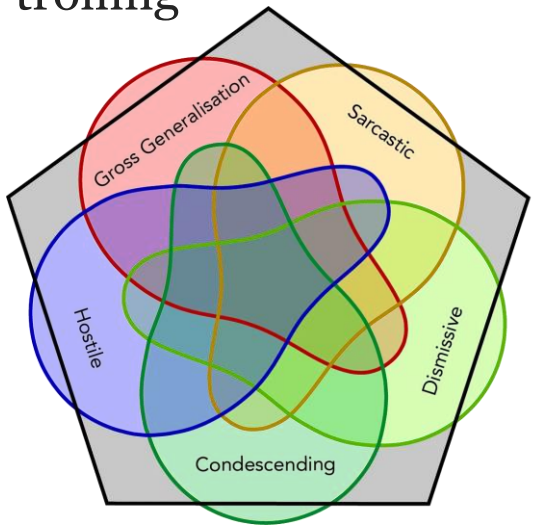
(1) Hostile, antagonistic, insulting, provocative or trolling

(2) Dismissive

(3) Condescending or patronising

(4) Sarcastic

(5) Unfair generalisations.



Data Annotation

1. Do you think this comment has a place in a healthy online conversation?
2. Is this comment sarcastic?
3. Does this comment make a generalisation about a specific group of people?
4. If yes, would a member of that group feel that the generalisation is unfair?
5. Is this comment needlessly hostile?
6. Is the intention of this comment to insult, antagonize, provoke, or troll other users?
7. Is this comment condescending and/or patronising?
8. Is this comment dismissive?

Annotation examples

Comment: "Pathetic finger pointing article."

Is this comment needlessly hostile?	Yes (81% confidence)
Is the intention of this comment to insult, antagonize, provoke, or troll other users?	No (59% confidence)

Comment: "Yes, that clang you hear is the top of the dustbin closing. Sounds exactly the same as your mind closing. I suppose they are exactly the same thing."

Is this comment condescending and/or patronising?	Yes (100% confidence)

Comment: "More alt right stupidity."

Is this comment dismissive? Yes (100% confidence)

Comment: "Iran and Turkey are the BEST places to be a woman!"

Is this comment sarcastic? Yes (72% confidence)

Perspective API

Perspective | Why Perspective | *How it Works* | Case Studies | Get Started

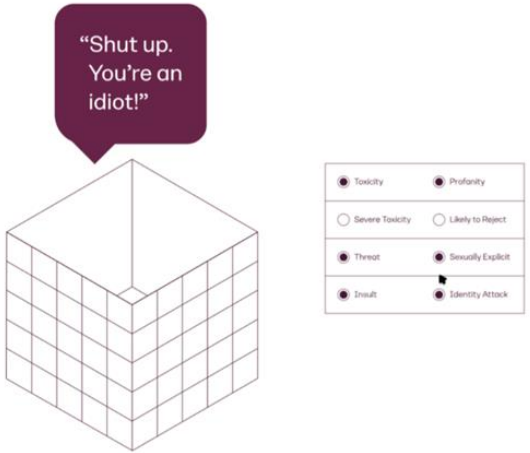
abusive comments. The models score a phrase based on the perceived impact the text may have in a conversation. Developers and publishers can use this score to give feedback to commenters, help moderators more easily review comments, or help readers filter out "toxic" language.

Perspective models provide scores for several different attributes. In addition to the flagship Toxicity attribute, here are some of the other attributes Perspective can provide scores for:

- Severe Toxicity
- Identity attack
- Insult
- Threat
- Profanity
- Sexually explicit

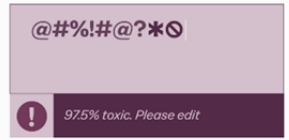
To learn more about our ongoing research and experimental models, visit our Developers site.

LEARN MORE



For moderators

Moderators use Perspective to quickly prioritize and review comments that have been reported.



For commenters

Perspective can give feedback to commenters who post toxic comments.

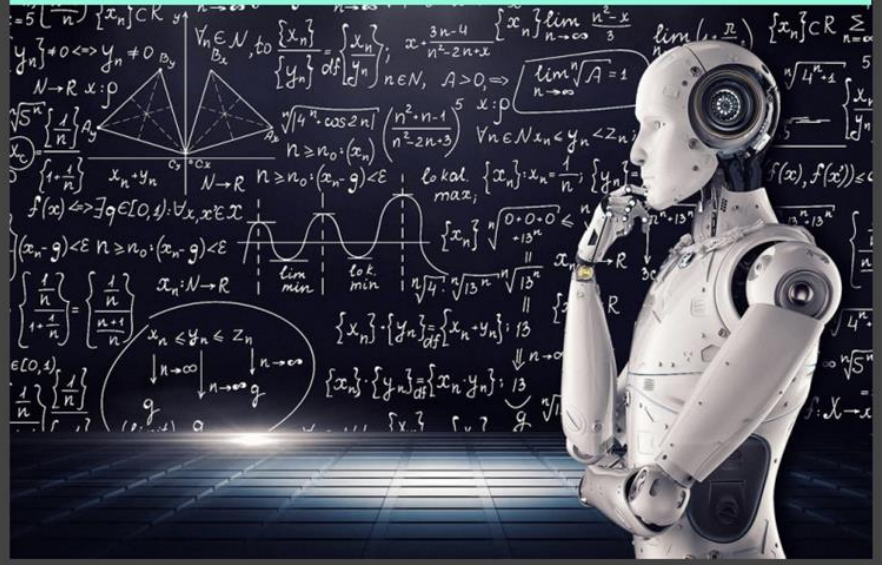
For readers

For readers Developers create tools so readers can control which comments they see, for example hiding comments that may be abusive or toxic.

AI Can Unintentionally Lead to Unjust Outcomes

- Lack of insight into **sources of bias in the data** and model
- Lack of insight into the **feedback loops**
- Lack of careful, **disaggregated evaluation**
- Human **biases in interpreting and accepting results**

So… What do we do?



Disclosure and Reporting

Model Cards for Model Reporting

- Currently no common practice of reporting how well a model works when it is released

Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru
{mmitchellai,simonewu,andrewzaldivar,parkerbarnes,lucyvasserman,benhutch,espitzer,tgebru}@google.com
deborah.raji@mail.utoronto.ca

What It Does

A report that focuses on transparency in model performance to encourage responsible AI adoption and application.

How It Works

It is an easily discoverable and usable artifact presented at important steps of a user journey for a diverse set of users and public stakeholders.

Why It Matters

It keeps model developer accountable to release high quality and fair models.

Intended Use, Factors and Subgroups

Example Model Card - Toxicity in Text	
Model Details	Developed by Jigsaw in 2017 as a convolutional neural network trained to predict the likelihood that a comment will be perceived as toxic.
Intended Use	Supporting human moderation, providing feedback to comment authors, and allowing comment viewers to control their experience.
Factors	Identity terms referencing frequently attacked groups focusing on the categories of sexual orientation, gender identity and race.

Metrics and Data

Metrics	*Pinned AUC*, which measures threshold-agnostic separability of toxic and non-toxic comments for each group, within the context of a background distribution of other groups.
Evaluation Data	A synthetic test set generated using a template-based approach, where identity terms are swapped into a variety of template sentences.
Training Data	Includes comments from a variety of online forums with crowdsourced labels of whether the comment is “toxic”. “Toxic” is defined as, “a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion”.

Considerations, Recommendations

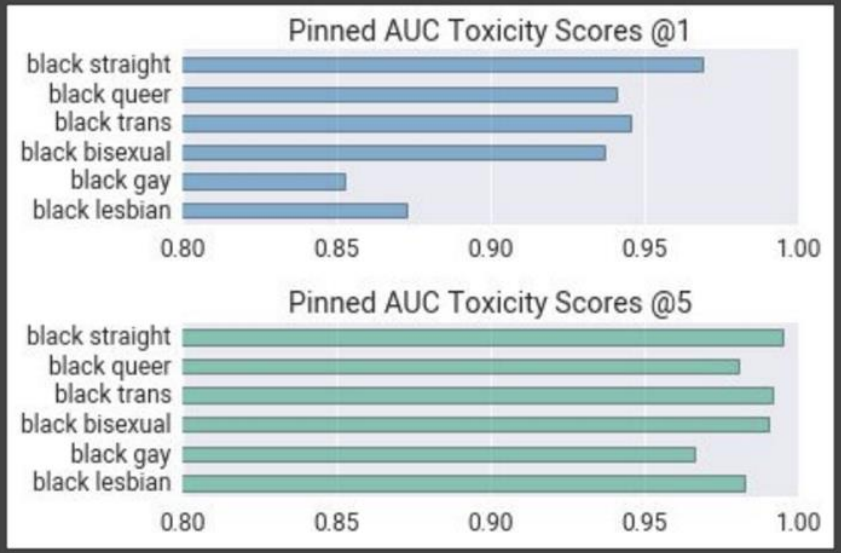
Ethical Considerations	A set of values around community, transparency, inclusivity, privacy and topic-neutrality to guide their work.
Caveats & Recommendations	Synthetic test data covers only a small set of very specific comments. While these are designed to be representative of common use cases and concerns, it is not comprehensive.

Disaggregated Intersectional Evaluation

Toxicity @1

Identity groups	Subgroup AUC	BPSN AUC	BNSP AUC
lesbian	0.93	0.74	0.98
gay	0.94	0.65	0.99
queer	0.98	0.96	0.93
straight	0.99	1.00	0.87
bisexual	0.96	0.95	0.92
homosexual	0.87	0.53	0.99
heterosexual	0.96	0.94	0.92
cis	0.99	1.00	0.87
trans	0.97	0.96	0.91
nonbinary	0.99	0.99	0.90
black	0.91	0.85	0.95
white	0.91	0.88	0.94

0.5 0.6 0.7 0.8 0.9 1.0



Jigsaw | M | The False Positive

In Summary...

- Always **be mindful** of various sorts of biases in the NLP models and the data
- Explore “debiasing” techniques, but **be cautious**
- **Identify the biases that matter** for your problem and test for those biases
- Consider this an **iterative process**, than something that has a “done” state
- Be **transparent** about your model and its performance in different settings

Bias and Unfairness reflect reality - so why bother?

Learn from data

EMNLP 2019 Tutorial

Closing Note

"Fairness and justice are properties of social and legal systems"

"To treat fairness and justice as terms that have meaningful application to technology separate from a social context is therefore [...] an abstraction error"