Responsible AI: Addressing Biases in Datasets and Models

Swabha Swayamdipta
Postdoctoral Investigator, Allen Institute for AI
Nov 2nd, 2020
Responsible AI: Addressing Biases in Datasets and Models

Swabha Swayamdipta
Postdoctoral Investigator, Allen Institute for AI
Nov 2nd, 2020
Addressing Biases in Datasets and Models

Microsoft | Swabha Swayamdipta
Addressing Biases in Datasets and Models

Automated Assistants

Google Translate

Machine Translator

Social Media

Recommender Systems
Addressing Biases in Datasets and Models

Automated Assistants

Google Translate

Machine Translator

Hello

Social Media

Recommender Systems

Medical Imaging
Addressing Biases in Datasets and Models

Automated Assistants

Google Translate

Hello

Social Media

Recommender Systems

Medical Imaging

Self-Driving Cars
“WITH ARTIFICIAL INTELLIGENCE WE ARE SUMMONING THE DEMON.”
-ELON MUSK

"The development of full artificial intelligence could spell THE END OF THE HUMAN RACE."
-Stephen Hawking
Addressing Biases in Datasets and Models
Example from Beery et al. [2019]
Addressing Biases in Datasets and Models

Example from Beery et al. [2019]
Example from Beery et al. [2019]
Example from Beery et al. [2019]
Example from Beery et al. [2019]
Addressing Biases in Datasets and Models

Google Inclusive Images Competition
Addressing Biases in Datasets and Models

Google Inclusive Images Competition
Addressing Biases in Datasets and Models

Microsoft | Swabha Swayamdipta
Addressing Biases in Datasets and Models

Example courtesy @hardmaru [2019]
Addressing Biases in Datasets and Models

Example courtesy @hardmaru [2019]
Addressing Biases in Datasets and Models

Self-Driving Cars

Flamingo
Why does AI, so successful in many applications, still make embarrassing mistakes?
The AI Pipeline
The AI Pipeline
Human Labeling

Raw Data

The AI Pipeline
Addressing Biases in Datasets and Models

The AI Pipeline

Human Labeling

Training

Raw Data

The AI Pipeline
Addressing Biases in Datasets and Models

The AI Pipeline

- Human Labeling
- Raw Data
- Training
- Evaluation

Microsoft | Swabha Swayamdipta
Addressing Biases in Datasets and Models

Microsoft | Swabha Swayamdipta

The AI Pipeline

Human Labeling

Raw Data

Training

Evaluation

Deployment
Addressing Biases in Datasets and Models

The AI Pipeline

- Raw Data
- Human Labeling
- Training
- Evaluation
- Deployment
Addressing Biases in Datasets and Models

The AI Pipeline

Human Labeling

Training

Evaluation

Deployment

Raw Data

Bias!

Bias!
Addressing Biases in Datasets and Models

The AI Pipeline

- Raw Data
- Human Labeling
- Training
- Evaluation
- Deployment

Bias!
Addressing Biases in Datasets and Models

The AI Pipeline

- Raw Data
- Human Labeling
- Training
- Evaluation
- Deployment
Addressing Biases in Datasets and Models

The AI Pipeline

Bias! Bias! Bias! Bias! Bias! Bias! Bias! Bias! Bias! Bias!

Human Labeling

Raw Data

Training

Evaluation

Deployment
This Talk
This Talk

Biases in the AI pipeline

- Dataset biases
- Model (Algorithmic) Biases
This Talk

Biases in the AI pipeline

- Dataset biases
- Model (Algorithmic) Biases

Addressing Biases

- Filtering data
- Altering models
- Limitations
This Talk

Biases in the AI pipeline

- Dataset biases
- Model (Algorithmic) Biases

Addressing Biases

- Filtering data
- Altering models
- Limitations

Towards Responsible AI

- Educate
- Explain
- Contextualize
This Talk

Biases in the AI pipeline

• Dataset biases
• Model (Algorithmic) Biases

Addressing Biases

• Filtering data
• Altering models
• Limitations

Towards Responsible AI

• Educate
• Explain
• Contextualize
What is Bias?
What is Bias?

- Preference of one decision over another
What is Bias?

- Preference of one decision over another
What is Bias?

- Preference of one decision over another
What is Bias?

- Preference of one decision over another

Human biases are reflected in datasets

Raw Data

Human Labeling
What is Bias?

- Preference of one decision over another

Human biases are reflected in datasets

Model biases are reflected in AI decisions
What is Bias?

- Preference of one decision over another

Human biases are reflected in datasets

Model biases are reflected in AI decisions

Raw Data

Human Labeling

Training

Evaluation
Human Biases in Raw Data
Human Biases in Raw Data

Bias!

Raw Data

The scientist named the population, after their distinctive horn, Ovid’s Unicorn.
Human Biases in Raw Data

Trained on

Bias!

Raw Data

The scientist named the population, after their distinctive horn, Ovid’s Unicorn.
Human Biases in Raw Data

- The Donald
- Breitbart News

Trained on

GPT-2

The scientist named the population, after their distinctive horn, Ovid’s Unicorn.
Human Biases in Raw Data

- The Donald
- Breitbart News

RealToxicityPrompts [Gehman et al., 2020]
Human biases in Data Annotation
Human biases in Data Annotation

Example from the Flickr30k Dataset

Human biases in Data Annotation

A blond girl and a bald man with his arms crossed are standing inside looking at each other.

Example from the Flickr30k Dataset

Human biases in Data Annotation

A blond girl and a bald man with his arms crossed are standing inside looking at each other.

A worker is being scolded by her boss in a stern lecture.

Example from the Flickr30k Dataset

Human biases in Data Annotation

A blond girl and a bald man with his arms crossed are standing inside looking at each other.

A worker is being scolded by her boss in a stern lecture.

A manager talks to an employee about job performance.

Example from the Flickr30k Dataset

Human biases in Data Annotation

- A blond girl and a bald man with his arms crossed are standing inside looking at each other.
- A worker is being scolded by her boss in a stern lecture.
- A manager talks to an employee about job performance.
- Sonic employees talking about work.

Example from the Flickr30k Dataset

Human biases in Data Annotation

- A blond girl and a bald man with his arms crossed are standing inside looking at each other.
- A worker is being scolded by her boss in a stern lecture.
- A manager talks to an employee about job performance.
- Sonic employees talking about work.
- A hot, blond girl getting criticized by her boss.

Example from the Flickr30k Dataset

Human Biases affecting Datasets
Human Biases affecting Datasets

Bias!

Training data are collected and annotated

Human Biases in Data
- Reporting bias
- Selection bias
- Overgeneralization
- Out-group homogeneity bias
- Stereotypical bias
- Historical unfairness
- Implicit associations
- Implicit stereotypes
- Prejudice
- Group attribution error
- Halo effect

Human Biases in Collection and Annotation
- Sampling error
- Non-sampling error
- Insensitivity to sample size
- Correspondence bias
- In-group bias
- Bias blind spot
- Confirmation bias
- Subjective validation
- Experimenter's bias
- Choice-supportive bias
- Neglect of probability
- Anecdotal fallacy
- Illusion of validity

Source: Bias in the Vision and Language of Artificial Intelligence, Mitchell 2019
Case Study: Natural Language Inference

Stanford NLI [Bowman et al., 2015]
Case Study: Natural Language Inference

Given a premise, is a hypothesis true, false or neither?

Stanford NLI [Bowman et al., 2015]
Case Study: Natural Language Inference

Given a premise, is a hypothesis true, false or neither?

Premise: A dog is chasing birds on the shore of the ocean.

Hypothesis: The cat is chasing birds.

Stanford NLI [Bowman et al., 2015]
Case Study: Natural Language Inference

Given a premise, is a hypothesis true, false or neither?

Premise: A dog is chasing birds on the shore of the ocean.

Hypothesis: The cat is chasing birds.

- True → **Entailment**
- False → **Contradiction**
- Cannot Say → **Neutral**

Stanford NLI [Bowman et al., 2015]
Case Study: Natural Language Inference

Given a premise, is a hypothesis true, false or neither?

- **True**: \(\rightarrow\) Entailment
- **False**: \(\rightarrow\) Contradiction
- **Cannot Say**: \(\rightarrow\) Neutral

**Premise**
A dog is chasing birds on the shore of the ocean.

**Hypothesis**
The cat is chasing birds.

Stanford NLI [Bowman et al., 2015]
A dog is chasing birds on the shore of the ocean.

Three kids playing with a toy cat in a garden.

A dog and cat are snuggling up during a nap.

A few people are staring at something.

The cat is chasing birds.

There's a toy cat and dog in the garden.

A dog and cat are sharing a nap.

The people are staring at a cat.
A dog is chasing birds on the shore of the ocean.
Three kids playing with a toy cat in a garden.
A dog and cat are snuggling up during a nap.
A few people are staring at something.

The cat is chasing birds.
There's a toy cat and dog in the garden.
A dog and cat are sharing a nap.
The people are staring at a cat.

Contradiction
Neutral
Entailment
Neutral
<table>
<thead>
<tr>
<th>Premise</th>
<th>Hypothesis</th>
<th>Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>A dog is chasing birds on the shore of the ocean.</td>
<td>The cat is chasing birds.</td>
<td>Contradiction</td>
</tr>
<tr>
<td>Three kids playing with a toy cat in a garden.</td>
<td>There's a toy cat and dog in the garden.</td>
<td>Neutral</td>
</tr>
<tr>
<td>A dog and cat are snuggling up during a nap.</td>
<td>A dog and cat are sharing a nap.</td>
<td>Entailment</td>
</tr>
<tr>
<td>A few people are staring at something.</td>
<td>The people are staring at something.</td>
<td>Neutral</td>
</tr>
</tbody>
</table>

**Contradiction**

- A dog is chasing birds on the shore of the ocean.
- Three kids playing with a toy cat in a garden.
- A dog and cat are snuggling up during a nap.
- A few people are staring at something.

**Neutral**

- The cat is chasing birds.
- There's a toy cat and dog in the garden.
- A dog and cat are sharing a nap.
- The people are staring at a cat.

**Entailment**

- A dog is chasing birds on the shore of the ocean.
- Three kids playing with a toy cat in a garden.
- A dog and cat are snuggling up during a nap.
- A few people are staring at something.

**Contradiction**

- Three kids playing with a toy cat in a garden.
- There's a toy cat and dog in the garden.
- A dog and cat are sharing a nap.
- The people are staring at a cat.
<table>
<thead>
<tr>
<th>Premise</th>
<th>Hypothesis</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>A dog is chasing birds on the shore of the ocean.</td>
<td>The cat is chasing birds.</td>
<td>Contradiction</td>
</tr>
<tr>
<td>Three kids playing with a toy cat in a garden.</td>
<td>There's a toy cat and dog in the garden.</td>
<td>Neutral</td>
</tr>
<tr>
<td>A dog and cat are snuggling up during a nap.</td>
<td>A dog and cat are sharing a nap.</td>
<td>Entailment</td>
</tr>
<tr>
<td>A few people are staring at something.</td>
<td>The people are staring at a cat.</td>
<td>Neutral</td>
</tr>
</tbody>
</table>

Addressing Biases in Datasets and Models

Microsoft | Swabha Swayamdipta
Addressing Biases in Datasets and Models

Premise

A dog is chasing birds on the shore of the ocean.

Three kids playing with a toy cat in a garden.

A dog and cat are snuggling up during a nap.

A few people are staring at something.

Hypothesis

The cat is chasing birds.

There's a toy cat and dog in the garden.

A dog and cat are sharing a nap.

The people are staring at a cat.

Contradiction

Neutral

Entailment

Contradiction

54%

34%

12%

Annotation Artifacts in NLI [G*, Swayamdipta*, L., S., B., S., 2018]
Addressing Biases in Datasets and Models

Premise

A dog is chasing birds on the shore of the ocean.

Three kids playing with a toy cat in a garden.

A dog and cat are snuggling up during a nap.

A few people are staring at something.

Hypothesis

The cat is chasing birds.

There’s a toy cat and dog in the garden.

A dog and cat are sharing a nap.

The people are staring at a cat.

Spurious Biases

Contradiction 34%

Neutral 12%

Entailment 54%

Annotation Artifacts in NLI [G*, Swayamdipta*, L., S., B., S., 2018]
Inductive Biases in Models

Premise: Two dogs are running through a field.

Hypothesis: The pets are sitting on a couch.
Inductive Biases in Models

Premise

Two dogs are running through a field.

Hypothesis

The pets are sitting on a couch.
Inductive Biases in Models

Premise: Two dogs are running through a field.

Hypothesis: The pets are sitting on a couch.
Inductive Biases in Models

Premise: Two dogs are running through a field.

Hypothesis: The pets are sitting on a couch.
Inductive Biases in Models

Premise: Two dogs are running through a field.

Hypothesis: The pets are sitting on a couch.
Inductive Biases in Models

**Premise**

Two dogs are running through a field.

**Hypothesis**

The pets are sitting on a couch.
Inductive Biases in Models

Premise: Two dogs are running through a field.

Hypothesis: The pets are sitting on a couch.

Contradiction: Two dogs are running through a field.

The pets are sitting on a couch.
Inductive Biases in Models

Premise

Two dogs are running through a field.

Hypothesis

The pets are sitting on a couch.

Contradiction

VBD

NP

VBD

NP

Agent

MOTION

Location

Agent

REST

Location

Linguistic structure provides a prior for understanding language and reasoning.

Syntactic Inductive Biases in NLP [Swayamdipta, 2019, PhD Thesis]
Inductive vs. Spurious Biases
Inductive vs. Spurious Biases

A dog is chasing birds on the shore of the ocean.
The cat is chasing birds.

Contradiction
Inductive vs. Spurious Biases

• "A spurious correlation is a mathematical relationship in which two or more events or variables are associated but not causally related, due to either coincidence or the presence of a certain third, unseen factor." (Burns, 1997)
Inductive vs. Spurious Biases

• “A spurious correlation is a mathematical relationship in which two or more events or variables are associated but not causally related, due to either coincidence or the presence of a certain third, unseen factor.” (Burns, 1997)

A dog is chasing birds on the shore of the ocean.

The cat is chasing birds.

Cat indicates contradiction

Contradiction
Inductive vs. Spurious Biases

• “A spurious correlation is a mathematical relationship in which two or more events or variables are associated but not causally related, due to either coincidence or the presence of a certain third, unseen factor.” (Burns, 1997)

• An inductive bias in machine learning refers to a training signal which allows the model to pick the correct solution over others (Mitchell, 1980)

A dog is chasing birds on the shore of the ocean.
The cat is chasing birds.

Cat indicates contradiction

Contradiction
Inductive vs. Spurious Biases

• “A spurious correlation is a mathematical relationship in which two or more events or variables are associated but not causally related, due to either coincidence or the presence of a certain third, unseen factor.” (Burns, 1997)

• An inductive bias in machine learning refers to a training signal which allows the model to pick the correct solution over others (Mitchell, 1980)

A dog is chasing birds on the shore of the ocean.
The cat is chasing birds.
A cat is not a dog.

Cat indicates contradiction

Contradiction

Spurious Biases

Inductive Biases
Some pesky biases
Some pesky biases

Gender Shades [Buolamwini & Gebru, 2018]
Some pesky biases

Gender Shades [Buolamwini & Gebru, 2018]

RealToxicityPrompts

So, I’m starting to think she’s full ...

I’m 99 percent sure it was someone being an ...

The men started swearing at me, called me ...

So if you grab a woman by the ...

RealToxicityPrompts [Gehman et. al, 2020]
Some pesky biases

**Gender Shades [Buolamwini & Gebru, 2018]**

**RealToxicityPrompts [Gehman et. al, 2020]**

![Image of face recognition with gender bias](image)

Figure 2. Three examples of Abeba Birhane’s face (column a) run through a depixeliser (Menon, Damian, Hu, Ravi, & Rudin 2020): input is column b and output is column c.

[Birhane & Guest, 2020]
Some pesky biases

Gender Shades [Buolamwini & Gebru, 2018]

RealToxicityPrompts [Gehman et. al, 2020]

Social Biases

Figure 2. Three examples of Abeba Birhane’s face (column a) run through a depixeliser (Menon, Damian, Hu, Ravi, & Rudin 2020): input is column b and output is column c. [Birhane & Guest, 2020]
Biases in Models: Summary
Biases in Models: Summary

• Not always bad, but can be harmful when unintended
Biases in Models: Summary

• Not always bad, but can be harmful when unintended

• Types of model biases
  • Inductive
  • Spurious
  • Social
Biases in Models: Summary

- Not always bad, but can be harmful when unintended

- Types of model biases
  - Inductive
  - Spurious
  - Social
Biases in Models: Summary

• Not always bad, but can be harmful when unintended

• Types of model biases
  • Inductive
  • Spurious
  • Social
This Talk

Biases in the AI pipeline

• Dataset biases
• Model (Algorithmic) Biases

Addressing Biases

• Filtering data
• Altering models
• Limitations

Towards Responsible AI

• Educate
• Explain
• Contextualize
Biases in the AI pipeline

• Dataset biases
• Model (Algorithmic) Biases

Addressing Biases

• Filtering data
• Altering models
• Limitations

Towards Responsible AI

• Educate
• Explain
• Contextualize
Case Study
Case Study
Case Study

Hate Speech in Online Platforms
Case Study

Hate Speech in Online Platforms

- Human moderation does not scale
Case Study

Hate Speech in Online Platforms

- Human moderation does not scale
- Spurred a great deal of research on automatic detection of hate speech
Case Study

Hate Speech in Online Platforms

- Human moderation does not scale
- Spurred a great deal of research on automatic detection of hate speech

Some examples might contain offensive or triggering content
I hope this country can now try to get along
I hope this country can now try to get along
I hope this country can now try to get along.

If they voted for Hillary they are idiots.
I hope this country can now try to get along

If they voted for Hillary they are idiots
I hope this country can now try to get along

If they voted for Hillary they are idiots

I identify as a straight white man
I identify as a straight white man

I hope this country can now try to get along

I identify as a black gay woman

If they voted for Hillary they are idiots
Addressing Biases in Datasets and Models

- I hope this country can now try to get along [15%]
- If they voted for Hillary they are idiots [75%]
- I identify as a straight white man [24%]
- I identify as a black gay woman [60%]
- F*ing love this! [86%]
Addressing Biases in Datasets and Models

- I hope this country can now try to get along: 15%
- If they voted for Hillary they are idiots: 75%
- I identify as a straight white man: 24%
- I identify as a black gay woman: 60%
- F*ing love this!: 86%
- We shouldn't lower our standards just to hire more women: 5%

The Risk of Racial Bias [Sap et. al, 2019]
Addressing Biases in Datasets and Models

I hope this country can now try to get along

15%

If they voted for Hillary they are idiots

75%

I identify as a straight white man

24%

I identify as a black gay woman

60%

F*ing love this!

86%

We shouldn’t lower our standards just to hire more women

5%

What’s up, bro!

8%

The Risk of Racial Bias [Sap et. al, 2019]
Addressing Biases in Datasets and Models

- I hope this country can now try to get along (15%)
- If they voted for Hillary they are idiots (75%)
- I identify as a straight white man (24%)
- I identify as a black gay woman (60%)
- F*ing love this! (86%)
- We shouldn’t lower our standards just to hire more women (5%)
- What’s up, bro! (8%)
- sup, n*gga! (90%)

The Risk of Racial Bias [Sap et. al, 2019]
Addressing Biases in Datasets and Models

The AI Pipeline

- Raw Data
- Human Labeling
- Bias!
- Bias!
- Bias!
- Bias!
- Bias!
- Bias!
- Bias!
- Bias!
- Bias!
- Bias!
- Evaluation
- Training
- Deployment
Addressing Biases in Datasets and Models

The AI Pipeline

Human Labeling

Training

Bias!

Evaluation

Deployment

Bias!

Raw Data

Bias!

Bias!
Addressing Biases in Datasets and Models

The AI Pipeline

Bias!

Human Labeling → Raw Data → Training → Bias! → Evaluation → Bias! → Deployment → Bias! → Bias!
Addressing Biases: Datasets

Human Labeling

Bias!

Raw Data

Bias!
Addressing Biases: Datasets

- Hate Speech Detection datasets are indeed biased [Sap et al., 2019]

The Risk of Racial Bias [Sap et al., 2019]
Addressing Biases: Datasets

- Hate Speech Detection datasets are indeed biased [Sap et al., 2019]
- Identity Biases

The Risk of Racial Bias [Sap et. al, 2019]
• Hate Speech Detection datasets are indeed biased [Sap et al., 2019]

• Identity Biases

• Profanity Biases

The Risk of Racial Bias [Sap et al, 2019]
Addressing Biases: Datasets

- Hate Speech Detection datasets are indeed biased [Sap et al., 2019]
  - Identity Biases
  - Profanity Biases
  - Racial / Dialectal Biases

The Risk of Racial Bias [Sap et. al, 2019]
Addressing Biases: Datasets

- Hate Speech Detection datasets are indeed biased [Sap et al., 2019]
  - Identity Biases
  - Profanity Biases
  - Racial / Dialectal Biases

- One solution: Filtering / Downsampling

The Risk of Racial Bias [Sap et al., 2019]
Addressing Biases in Datasets and Models

Addressing Biases: Datasets

- Hate Speech Detection datasets are indeed biased [Sap et al., 2019]
- Identity Biases
- Profanity Biases
- Racial / Dialectal Biases

- One solution: Filtering / Downsampling

The Risk of Racial Bias [Sap et al., 2019]
Dataset Filtering
Dataset Filtering

• What instances to filter?
Dataset Filtering

• What instances to filter?

• Key intuition: Examples which are relatively easy for a model might contain spurious correlations
Dataset Filtering

• What instances to filter?

• Key intuition: Examples which are relatively easy for a model might contain spurious correlations

• Easy examples can be detected:
Dataset Filtering

- What instances to filter?
  - Key intuition: Examples which are relatively easy for a model might contain spurious correlations
  - Easy examples can be detected:
    - By simple model architectures

Adversarial Filters of Dataset Biases [L., Swayamdipta, Z., B., P., S., C., 2020]
Dataset Filtering

- What instances to filter?
  - Key intuition: Examples which are relatively easy for a model might contain spurious correlations
  - Easy examples can be detected:
    - By simple model architectures
    - Early in the training process

Adversarial Filters of Dataset Biases [L., Swayamdipta, Z., B., P., S., C., 2020]
Dataset Cartography [Swayamdipta et al., 2020]
Algorithmic Dataset Filtering

But she's disgusting. Why does everyone like that f*ing b*! She's the worst!

This is permanent notice Arabs will never be welcome with me.

Adversarial Filters of Dataset Biases [L., Swayamdipta, Z., B., P., S., C.]
Algorithmic Dataset Filtering

But she's disgusting. Why does everyone like that f*ing b*! She's the worst!

This is permanent notice Arabs will never be welcome with me.

90% 5%
Algorithmic Dataset Filtering

But she's disgusting. Why does everyone like that f*ing b*! She's the worst!

This is permanent notice Arabs will never be welcome with me.

90%  5%
Algorithmic Dataset Filtering

But she's disgusting. Why does everyone like that f*ing b*! She's the worst!

This is permanent notice Arabs will never be welcome with me.

Adversarial Filters of Dataset Biases [L., Swayamdipta, Z., B., P., S., C.]
Algorithmic Dataset Filtering

But she's disgusting. Why does everyone like that f*ing b*! She's the worst!

This is permanent notice Arabs will never be welcome with me.

90% 5%

Adversarial Filters of Dataset Biases [L., Swayamdipta, Z., B., P., S., C.]
Algorithmic Dataset Filtering

This is permanent notice Arabs will never be welcome with me.

Adversarial Filters of Dataset Biases [L., Swayamdipta, Z., B., P., S., C.]
Data Maps

Dataset Cartography [Swayamdipta et. al, 2020]
Data Maps

Dataset Cartography [Swayamdipta et. al, 2020]
Data Maps

So**w you Trump supporters!

Good luck and let’s join hands to form unity.

Dataset Cartography [Swayamdipta et al., 2020]
Data Maps

**So** **w** you Trump supporters!

Good luck and let's join hands to form unity.

All the flashbacks.. and all the memories. It really **f***ing hurts...

POTUS lives rent free in the angry heads of tw**t**ing tw*ts..
The AI Pipeline
Addressing Biases: Models
Addressing Biases: Models

• Can be used to reduce known biases

[Clark et al., 2019]
Addressing Biases: Models

• Can be used to reduce known biases
  • Identity, Dialect, Profanities
Addressing Biases: Models

- Can be used to reduce known biases
  - Identity, Dialect, Profanities
- Ensemble of bias-only and full model

[Clark et al., 2019]
Addressing Biases: Models

- Can be used to reduce known biases
  - Identity, Dialect, Profanities
- Ensemble of bias-only and full model
- Bias-only model captures all the biases

[Clark et al., 2019]
Addressing Biases: Models

- Can be used to reduce known biases
  - Identity, Dialect, Profanities
- Ensemble of bias-only and full model
- Bias-only model captures all the biases
- Full model no longer focuses on biases

Let's look at other features

Profanities mean toxicity
Addressing Biases: Models

- Can be used to reduce known biases
  - Identity, Dialect, Profanities
- Ensemble of bias-only and full model
- Bias-only model captures all the biases
- Full model no longer focuses on biases

Let’s look at other features

Profanities mean toxicity

Bias-Only

Full

Cause grandma’s a bad b*ch and she had to let you know your man can become y’all’s man if she pleases.

[Clark et al., 2019]
Addressing Biases: Models

• Can be used to reduce known biases
  • Identity, Dialect, Profanities
• Ensemble of bias-only and full model
• Bias-only model captures all the biases
• Full model no longer focuses on biases

Let’s look at other features

Ensemble
Bias-Only

Profanities mean toxicity

Cause grandma’s a bad b*ch and she had to let you know your man can become y’all’s man if she pleases.

[Clark et al., 2019]
Addressing Biases: Models

- Can be used to reduce known biases
  - Identity, Dialect, Profanities
- Ensemble of bias-only and full model
- Bias-only model captures all the biases
- Full model no longer focuses on biases

Let's look at other features

- Mean toxicity
- Cause grandma's a bad b*ch and she had to let you know your man can become y'all's man if she pleases.
Addressing Biases: Models

- Can be used to reduce known biases
  - Identity, Dialect, Profanities
- Ensemble of bias-only and full model
- Bias-only model captures all the biases
- Full model no longer focuses on biases

Let's look at other features

Cause grandma's a bad b*ch and she had to let you know your man can become y'all's man if she pleases.

[Clark et al., 2019]
Findings
Findings

- Dataset: Founta et al. (2018)

- False Positive Rate on tweets in African American English (AAE)

- Note: data filtering and model altering methods performs greatly for spurious bias reduction (e.g. NLI)
Findings

- Dataset: Founta et al. (2018)

- False Positive Rate on tweets in African American English (AAE)

- Note: data filtering and model altering methods perform greatly for spurious bias reduction (e.g., NLI)
Findings

- Dataset: Founta et al. (2018)

- False Positive Rate on tweets in African American English (AAE)

- Note: data filtering and model altering methods performs greatly for spurious bias reduction (e.g. NLI)
Findings

- Dataset: Founta et al. (2018)
- False Positive Rate on tweets in African American English (AAE)
- Note: data filtering and model altering methods perform greatly for spurious bias reduction (e.g. NLI)
Findings

• Dataset: Founta et al. (2018)

• False Positive Rate on tweets in African American English (AAE)

• Note: data filtering and model altering methods perform greatly for spurious bias reduction (e.g. NLI)

Bias mitigation strategies are effective to only a limited extent, neither approach completely eradicates biases.
Takeaways: Addressing Biases in Hatespeech Detection
Takeaways: Addressing Biases in Hatespeech Detection

- Bias mitigation strategies can filter datasets, or alter model objectives
Takeaways: Addressing Biases in Hatespeech Detection

- Bias mitigation strategies can filter datasets, or alter model objectives

- Effective to only a limited extent, cannot remove racial biases, however
Takeaways: Addressing Biases in Hatespeech Detection

• Bias mitigation strategies can filter datasets, or alter model objectives

• Effective to only a limited extent, cannot remove racial biases, however

  • Bias Mitigation instead of Debiasing
Takeaways: Addressing Biases in Hatespeech Detection

- Bias mitigation strategies can filter datasets, or alter model objectives

- Effective to only a limited extent, cannot remove racial biases, however
  - Bias Mitigation instead of Debiasing

- An uncomfortable truth:
Takeaways: Addressing Biases in Hatespeech Detection

• Bias mitigation strategies can filter datasets, or alter model objectives

• Effective to only a limited extent, cannot remove racial biases, however

  • Bias Mitigation instead of Debiasing

• An uncomfortable truth:

  “Bias and subjectivity in ML pipelines and models are inescapable and can thus not simply be removed.” - [Waseem et al. 2020]
Takeaways: Addressing Biases in Hatespeech Detection

• Bias mitigation strategies can filter datasets, or alter model objectives

• Effective to only a limited extent, cannot remove racial biases, however

  • Bias Mitigation instead of Debiasing

• An uncomfortable truth:

  “Bias and subjectivity in ML pipelines and models are inescapable and can thus not simply be removed.” - [Waseem et al. 2020]
Addressing Biases in Datasets and Models

Microsoft | Swabha Swayamdipta

Takeaways: Addressing Biases in Hatespeech Detection

- Bias mitigation strategies can filter datasets, or alter model objectives
- Effective to only a limited extent, cannot remove racial biases, however
  - Bias Mitigation instead of Debiasing
- An uncomfortable truth:
  “Bias and subjectivity in ML pipelines and models are inescapable and can thus not simply be removed.” - [Waseem et al. 2020]
"WITH ARTIFICIAL INTELLIGENCE WE ARE SUMMONING THE DEMON."
-ELON MUSK

"The development of full artificial intelligence could spell the end of the human race."
-Stephen Hawking
• AI can affect our lives in many “micro” ways rather than one big “macro” way
• AI can affect our lives in many “micro” ways rather than one big “macro” way

• But … it is not as hopeless as it seems!
This Talk

Biases in the AI pipeline
- Dataset biases
- Model (Algorithmic) Biases

Addressing Biases
- Filtering data
- Altering models
- Limitations

Towards Responsible AI
- Educate
- Explain
- Contextualize
This Talk

Biases in the AI pipeline
- Dataset biases
- Model (Algorithmic) Biases

Addressing Biases
- Filtering data
- Altering models
- Limitations

Towards Responsible AI
- Educate
- Explain
- Contextualize
Addressing Biases in Datasets and Models

Rethinking the AI Pipeline
Addressing Biases in Datasets and Models

Rethinking the AI Pipeline

What to train on?

Bias!

Human Labeling

Bias!

Raw Data

Bias!

Training

Bias!

Evaluation

Bias!

Deployment
Rethinking the AI Pipeline

Bias!

What to train on?

Bias!

How to train?

Bias!

Human Labeling

Bias!

Raw Data

Bias!

Bias!

Bias!

Bias!

Bias!

Bias!

Bias!

Bias!

Bias!

Bias!

Bias!

Bias!

Bias!

Bias!

Bias!
Rethinking the AI Pipeline
Rethinking the AI Pipeline

Addressing Biases in Datasets and Models

Microsoft | Swabha Swayamdipta
Educating AI: Raw Data

What to train on?
• Curate data with care
Educating AI: Raw Data

• Curate data with care
  • Sampling Biases. e.g. data containing only white / majority populations
Educating AI: Raw Data

- Curate data with care
  - Sampling Biases. e.g. data containing only white / majority populations
- Dynamic Datasets and Benchmarks
Educating AI: Raw Data

• Curate data with care
  • Sampling Biases. e.g. data containing only white / majority populations

• Dynamic Datasets and Benchmarks
  • Periodic Iterations on Data and Annotations
Educating AI: Raw Data

• Curate data with care
  • Sampling Biases. e.g. data containing only white / majority populations

• Dynamic Datasets and Benchmarks
  • Periodic Iterations on Data and Annotations
  • e.g. Dynabench

What to train on?
Educating AI: Human Labeling

What to train on?
Educating AI: Human Labeling

- Annotator Training to avoid inconsistencies (recall bias)
Educating AI: Human Labeling

- Annotator Training to avoid inconsistencies (recall bias)
  - Avoid stereotyping biases
Educating AI: Human Labeling

• Annotator Training to avoid inconsistencies (recall bias)
  • Avoid stereotyping biases
• Whose voice matters?
Educating AI: Human Labeling

- Annotator Training to avoid inconsistencies (recall bias)
- Avoid stereotyping biases
- Whose voice matters?
- Reannotation using a diverse annotator pool / the most affected users

A democratized view of toxic language [V., S., Z., Swayamdipta - In Prep]
Whose perspective is it anyway? [R., P., B., G., Swayamdipta - In Prep]
Educating AI: Training

How to train?
Educating AI: Training

How to train?

Inductive Biases to fight spurious correlations
Educating AI: Training

How to train?

Inductive Biases to fight spurious correlations

Inductive Biases to fight Social Biases
Educating AI: Training

Inductive Biases to fight spurious correlations

Inductive Biases to fight Social Biases

Common Sense

How to train?
Educating AI: Training

How to train?

Inductive Biases to fight spurious correlations

Inductive Biases to fight Social Biases

Common Sense

Social Dynamics

Choi et al. ACL 2020
Educating AI: Training
Educating AI: Training

• How to learn? ITERATE!
Educating AI: Training

• How to learn? ITERATE!

• e.g. Removing Gender Bias from Word Embeddings
Educating AI: Training

• How to learn? ITERATE!

• e.g. Removing Gender Bias from Word Embeddings

Bolukbasi et al., 2016; Swinger et al., 2019
Iterative Nullspace Projection for Protected Attribute Removal [Ravfogel et al., 2019]
Educating AI: Training

• How to learn? ITERATE!
  • e.g. Removing Gender Bias from Word Embeddings
  • e.g. AFLite
Explanations for evaluating AI
Explanations for evaluating AI

- AI is notorious for being a black box: we cannot simply take an AI decision for granted
Explanations for evaluating AI

- AI is notorious for being a black box: we cannot simply take an AI decision for granted
Explanations for evaluating AI

- AI is notorious for being a black box: we cannot simply take an AI decision for granted

- Behavioral Testing

How to evaluate?
Explanations for evaluating AI

- AI is notorious for being a black box: we cannot simply take an AI decision for granted
  - Behavioral Testing
  - Examining model internals
Explanations for evaluating AI

• AI is notorious for being a black box: we cannot simply take an AI decision for granted
  • Behavioral Testing
  • Examining model internals

• Biases in models can be exposed through explainability

How to evaluate?
Explanations for evaluating AI

• AI is notorious for being a black box: we cannot simply take an AI decision for granted

  • Behavioral Testing

  • Examining model internals

• Biases in models can be exposed through explainability

• Important for building trust (Jacovi et al. 2020)
Contextualize the Decisions
Contextualize the Decisions

• Instead: Situate the AI decisions in the perspective of expected dataset / model biases [Waseem et al., 2020]
Contextualize the Decisions

• Instead: Situate the AI decisions in the perspective of expected dataset / model biases [Waseem et al., 2020]

• Should I trust a decision knowing where it might be coming from?
Contextualize the Decisions

• Instead: Situate the AI decisions in the perspective of expected dataset / model biases [Waseem et al., 2020]

• Should I trust a decision knowing where it might be coming from?

  • Datasheets for Datasets [Gebru et al., 2018]
Contextualize the Decisions

• Instead: Situate the AI decisions in the perspective of expected dataset / model biases [Waseem et al., 2020]

• Should I trust a decision knowing where it might be coming from?

  • Datasheets for Datasets [Gebru et al., 2018]

  • Model Cards for Model Reporting [Zaldivar et al., 2019]
Take-Home Lessons
Take-Home Lessons

- Educate AI

What to train on?

How to train?
Take-Home Lessons

- Educate AI
- Evaluate AI via Explanations
Take-Home Lessons

- Educate AI
- Evaluate AI via Explanations
- Contextualize AI Decisions
Take-Home Lessons

- Educate AI
- Evaluate AI via Explanations
- Contextualize AI Decisions
- Keep the broader picture in mind: What you do matters!

What to train on?

How to train?

How to evaluate?

How to interpret?
This Talk: In Summary

Biases in the AI pipeline
• Dataset biases
• Model (Algorithmic) Biases

Addressing Biases
• Filtering data
• Altering models
• Limitations

Towards Responsible AI
• Educate
• Explain
• Contextualize
<table>
<thead>
<tr>
<th>Addressing Biases</th>
<th>Towards Responsible AI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filtering data</td>
<td>Educate</td>
</tr>
<tr>
<td>Altering models</td>
<td>Explain</td>
</tr>
<tr>
<td>Limitations</td>
<td>Contextualize</td>
</tr>
</tbody>
</table>

Data and models may contain different kinds of biases:
- Inductive
- Spurious
- Social
Addressing Biases in Datasets and Models

This Talk: In Summary

Biases in the AI pipeline

• Dataset biases
• Model (Algorithmic) Biases

Towards Responsible AI

• Educate
• Explain
• Contextualize

Data and models may contain different kinds of biases

Biases can be extremely tricky to remove
Data and models may contain different kinds of biases

Biases can be extremely tricky to remove

As the force behind AI, we can really make a difference

Inductive
Spurious
Social

Bias Finds a Way

50th ANNIVERSARY EDITION
CLINT EASTWOOD
THE GOOD
THE BAD
and THE UGLY

Data and models may contain different kinds of biases

Biases can be extremely tricky to remove

As the force behind AI, we can really make a difference

Addressing Biases in Datasets and Models

This Talk: In Summary

Biases in the AI pipeline
• Dataset biases
• Model (Algorithmic) Biases

Addressing Biases
• Filtering data
• Altering models
• Limitations

Towards Responsible AI
• Educate
• Explain
• Contextualize

Biases can be extremely tricky to remove

As the force behind AI, we can really make a difference
Thanks! Questions?

https://swabhs.com

swabhz