Case study on NLP: Identifying and Mitigating Unintended Demographic Bias in Machine Learning for NLP

Exploring Fairness in Machine Learning

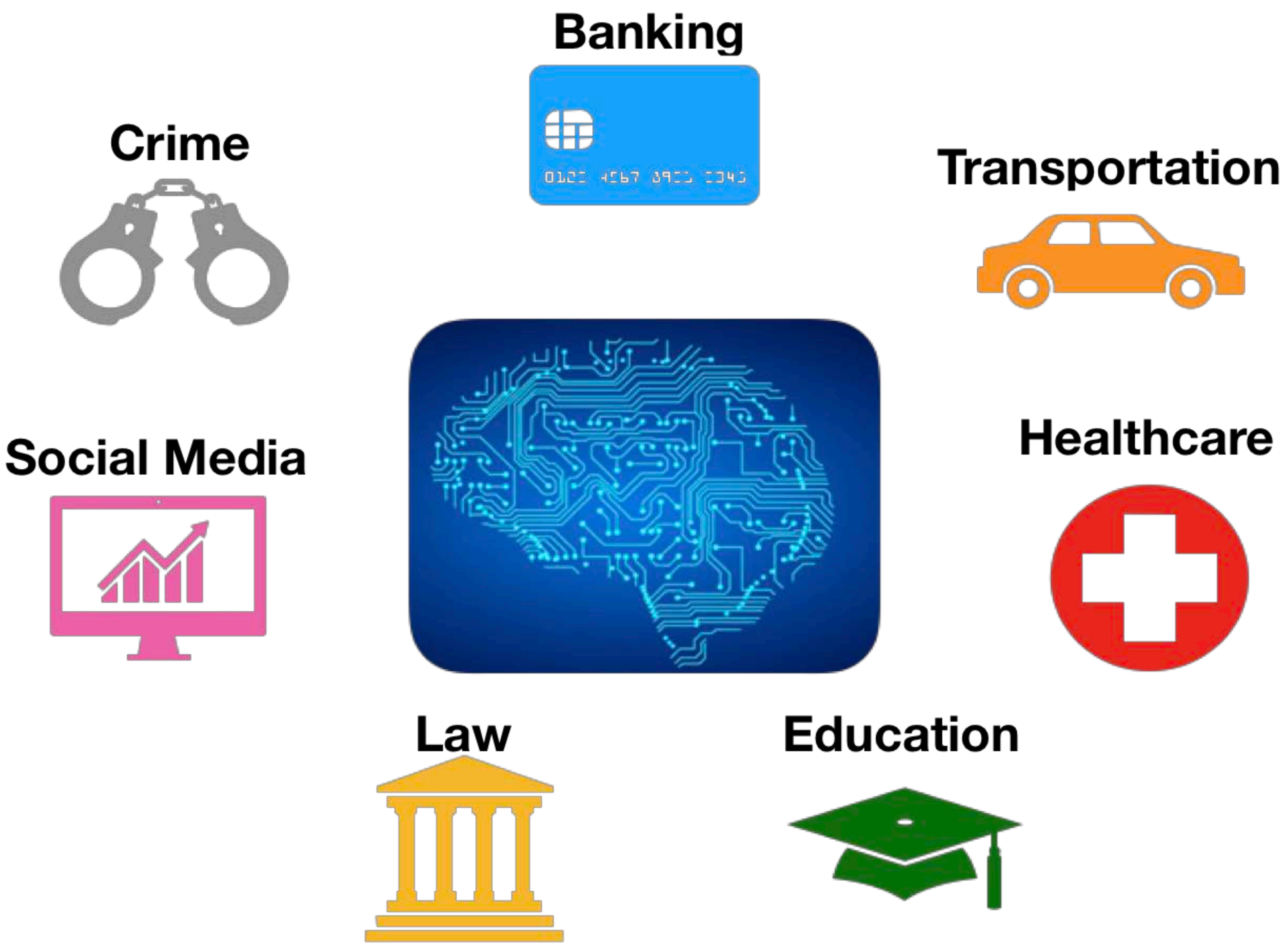
Researched by:
Christopher Sweeney
Researcher, MIT

Researched by:
Maryam Najafian
Research Scientist, MIT

Presented by:
Audace Nakeshimana
Researcher, MIT

MIT D-Lab USAID FROM THE AMERICAN PEOPLE CITE

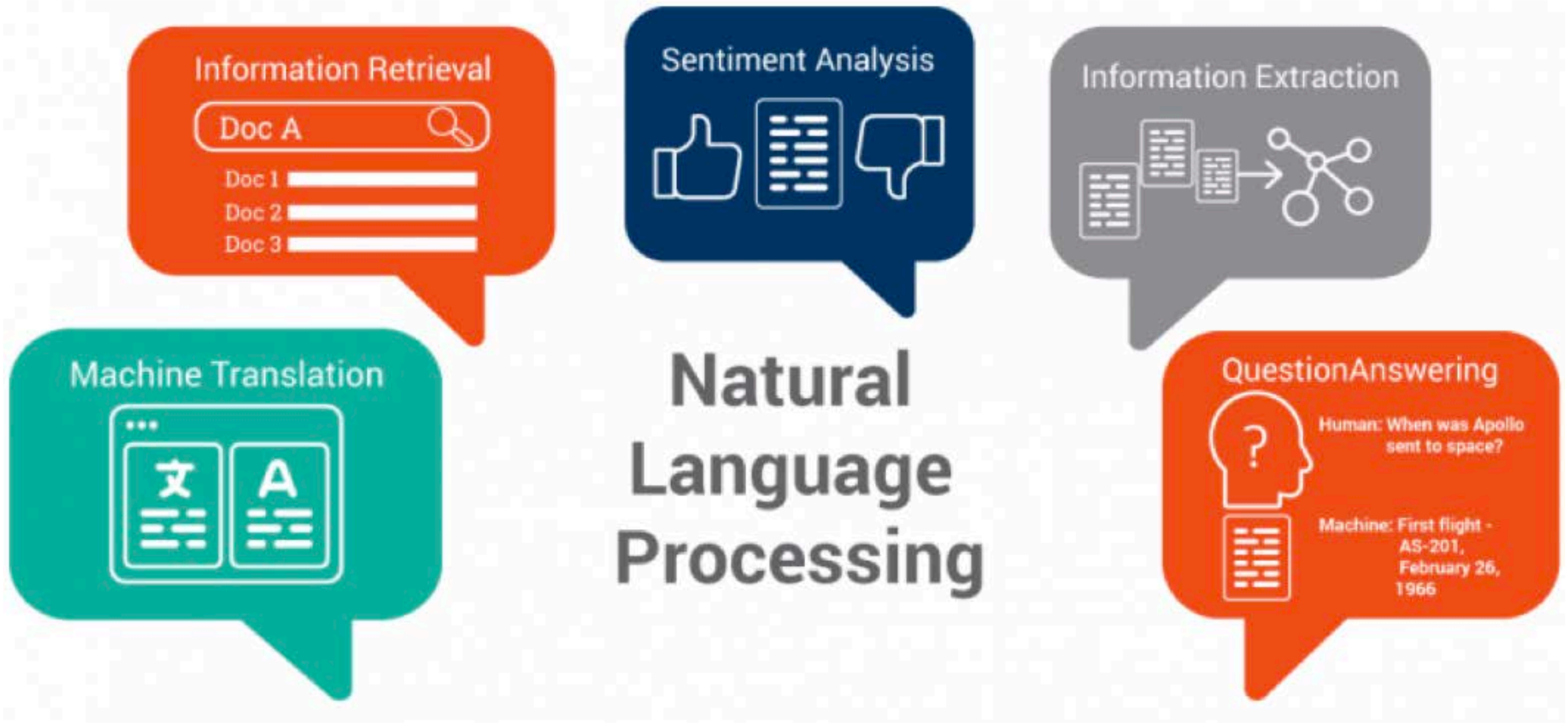
AI’s Power to Impact Society



Source: Sweeney & Najafian

Why Natural Language Processing?

- NLP is used in multiple domains (education, employment, social media, marketing).
- Many sources of unintended demographic bias in NLP pipeline.
- Data is widely available.



Source: Sweeney & Najafian

What is Unintended Demographic Bias

- Unintended: The bias is an adverse side effect, not deliberately learned
- Demographic: The bias is some form of inequality between demographic groups
- Bias: Artifact of the NLP pipeline that causes unfairness

Types of Unintended Demographic Bias

- Sentiment Bias: Artifact of the ML pipeline that causes unfairness in sentiment analysis algorithms
- Toxicity Bias: Artifact of the ML pipeline that causes unfairness in toxicity predictions algorithms

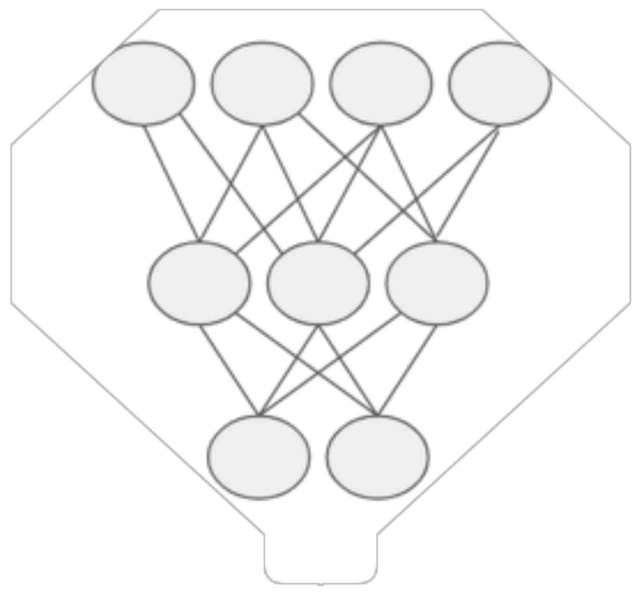
Types of Unintended Demographic Bias

Unfair toxicity classification example

Toxicity Classifier | **Unfair Decisions**

Sentence 1:

"I am American" → → Non-Toxic

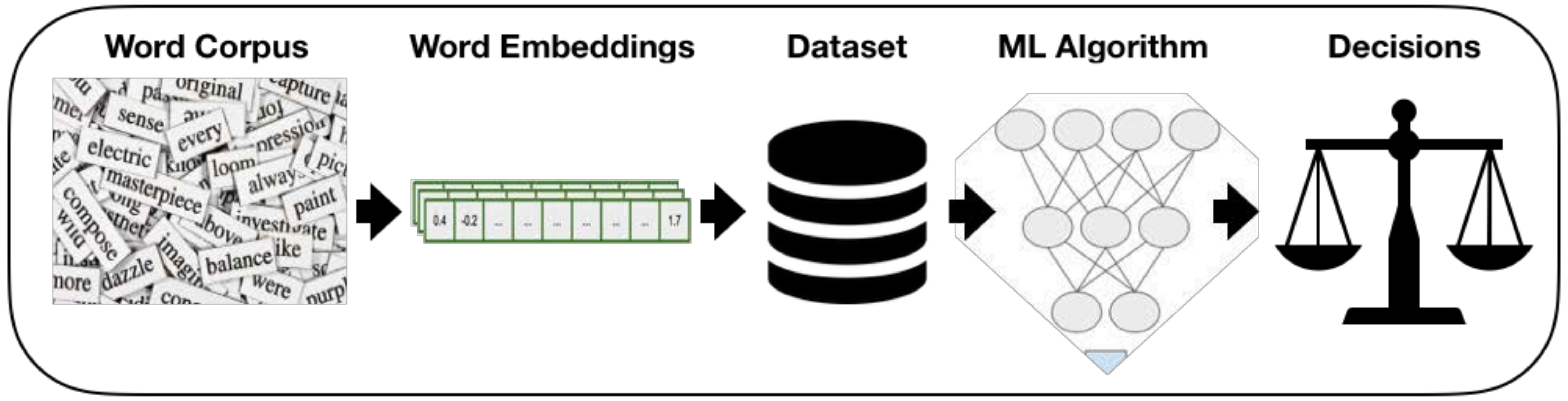


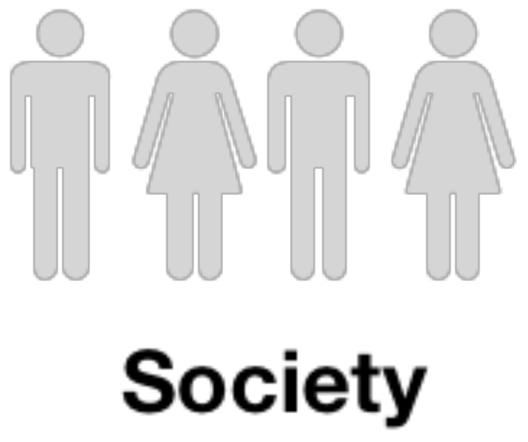
Sentence 2:

"I am Mexican" → → Toxic

Source: Sweeney & Najafian

Unintended Demographic Bias vs Unfairness





Unintended bias

Possible Unfairness (Discrimination)

Source: Sweeney & Najafian

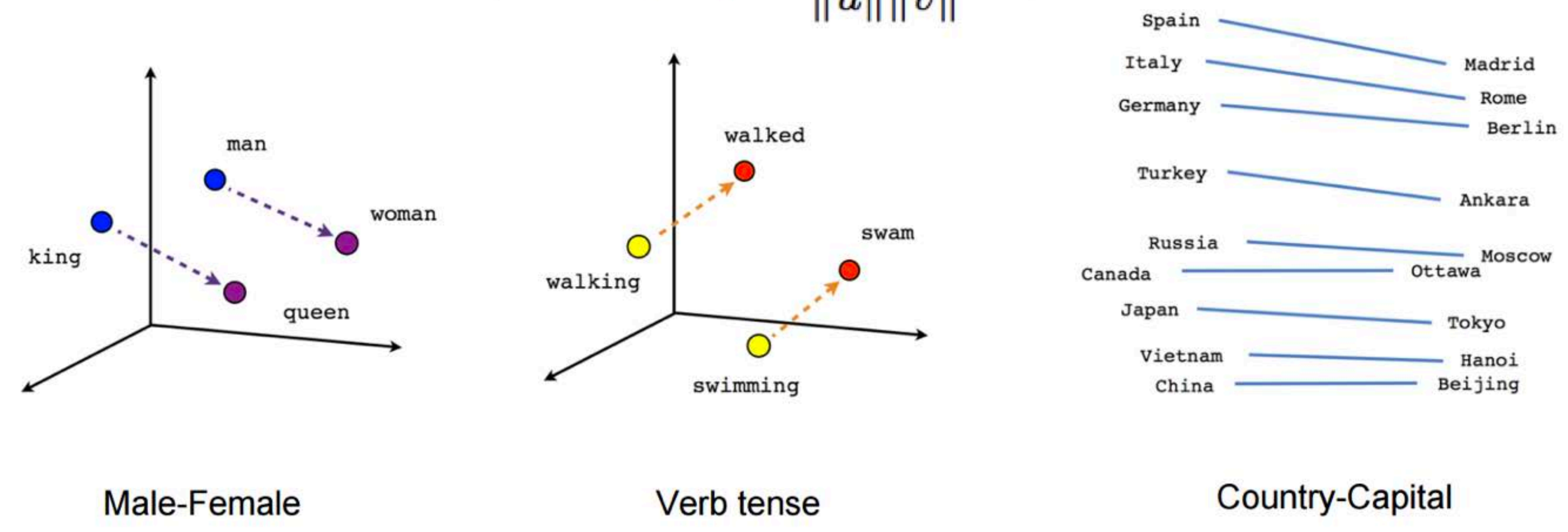
Research Summary

- Measuring Unintended Demographic Bias in word embeddings
- Using adversarial learning to mitigate word embedding bias
- PCA and Kernel methods to mitigate unintended bias
- Regression terms to mitigate unintended bias
- Evaluate methods against state-of-the-art bias mitigation methods on real NLP systems

Part 1:
Measuring Word Embedding Bias

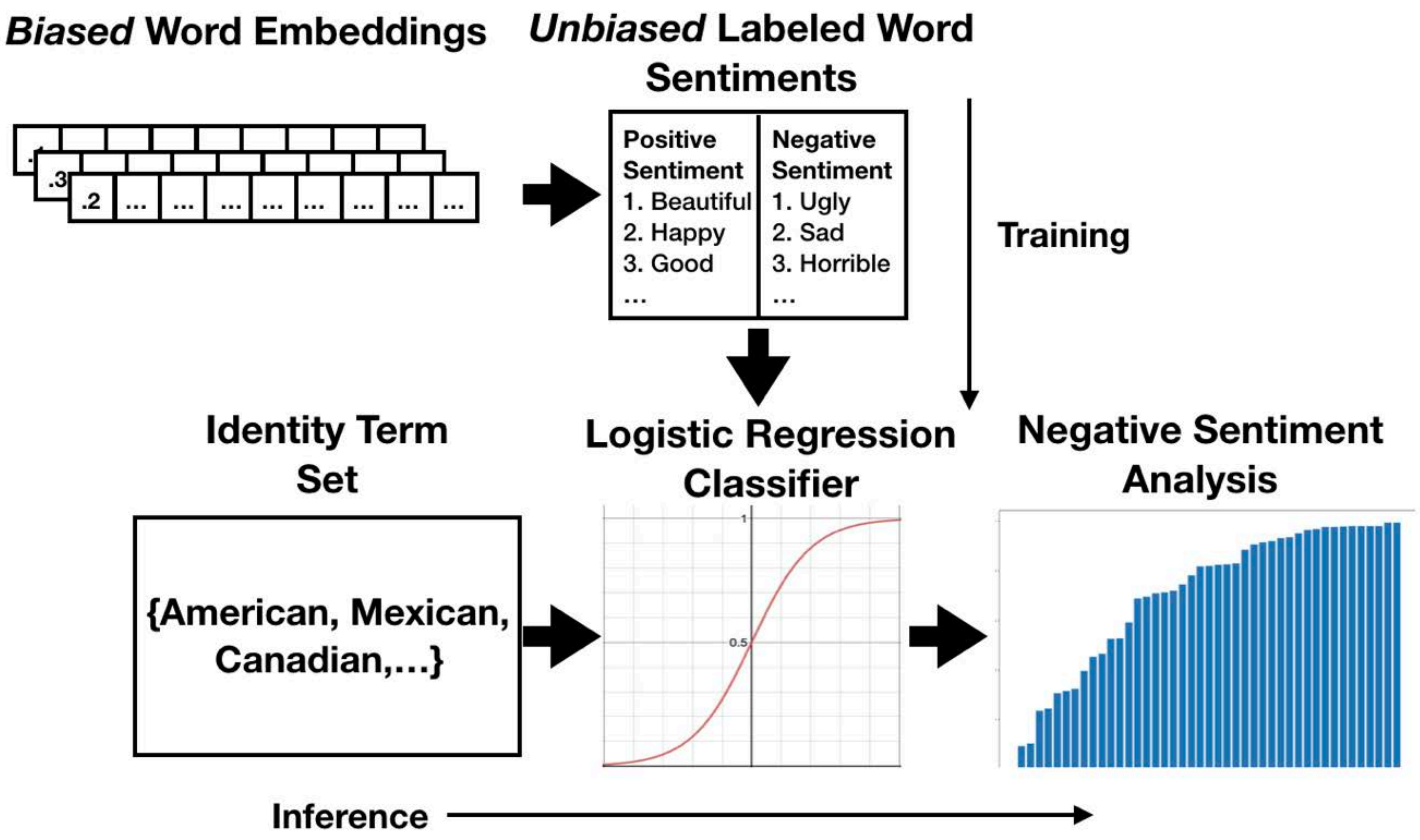
Measuring Word Embedding Bias

$$\cos(u, v) = \frac{u \cdot v}{\|u\| \|v\|}.$$



Man -> Woman as Computer Scientist -> Homemaker (Bolukbasi. '16)

How to Measure Sentiment Bias in Word Embeddings?

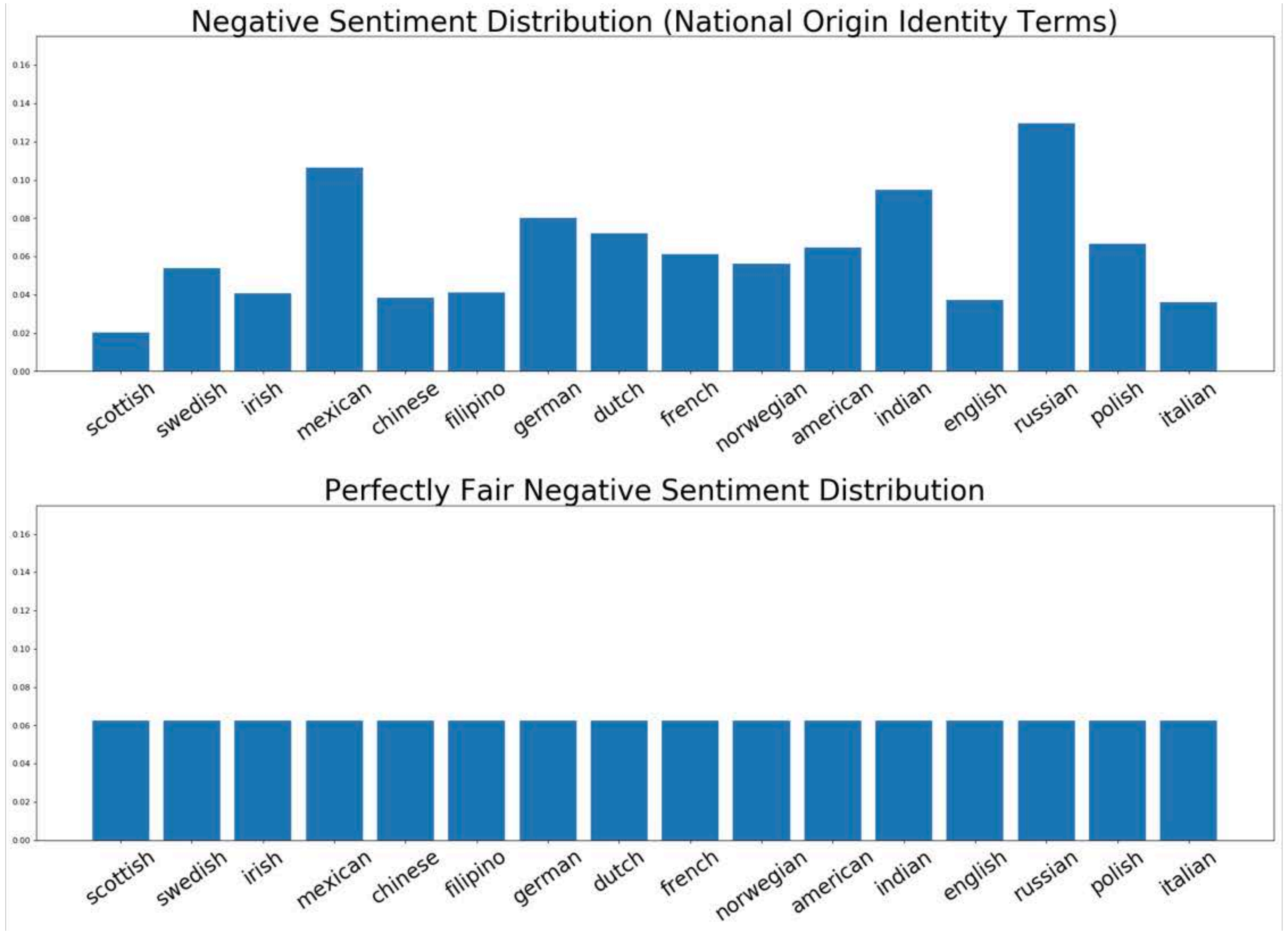


Source: Sweeney & Najafian

Relative Negative Sentiment Bias (RNSB)

P =

U =

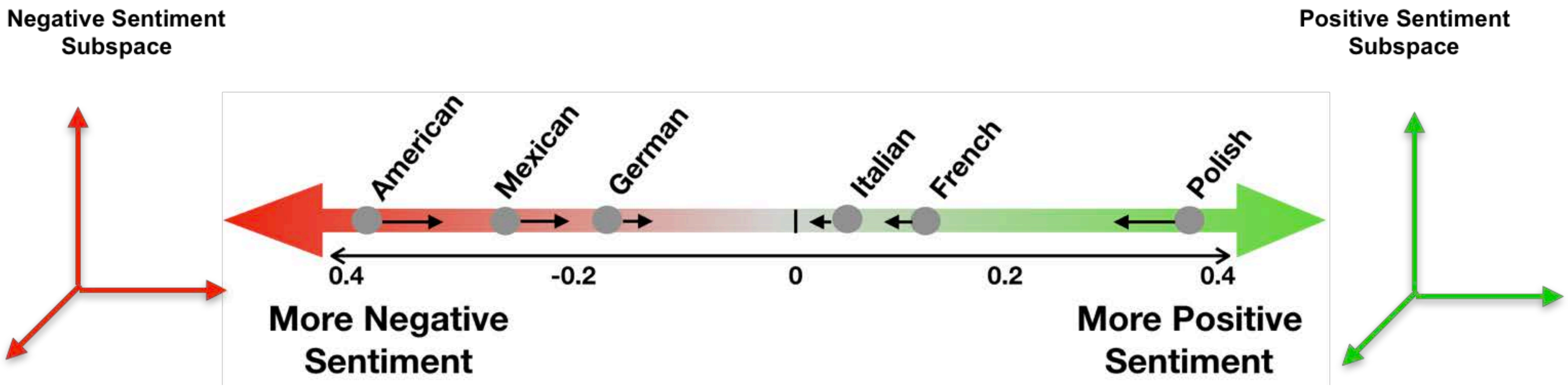


$$D_{KL}(P||U)$$

Source: Sweeney & Najafian

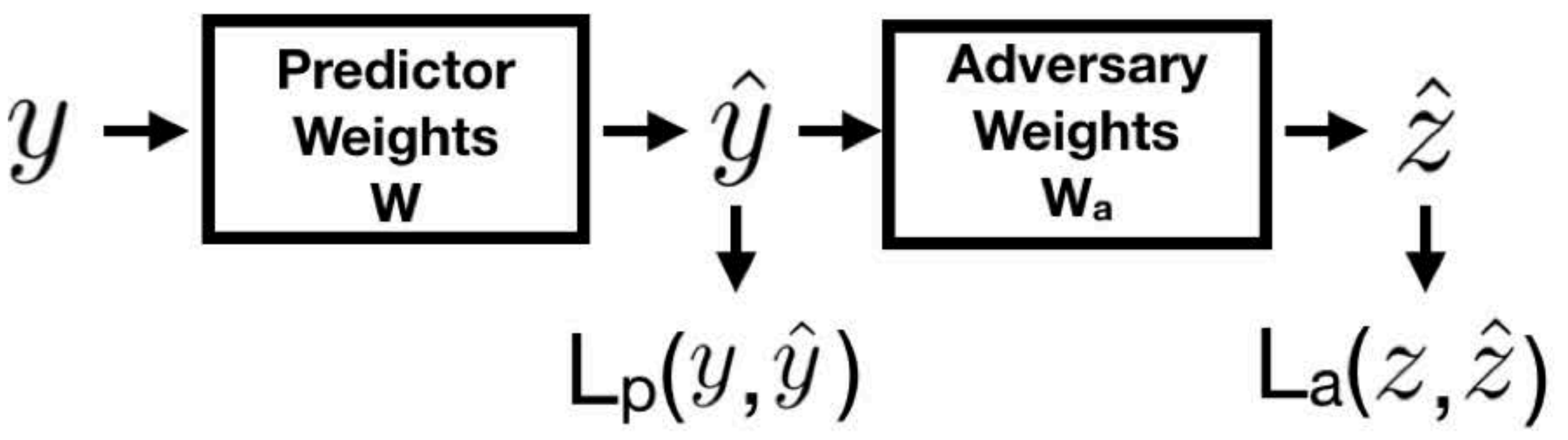
Part 2:
Mitigating Word Embedding Bias

Using Adversarial Learning to Debias Word Embeddings



Source: Sweeney & Najafian

Using Adversarial Learning to Debias Word Embeddings

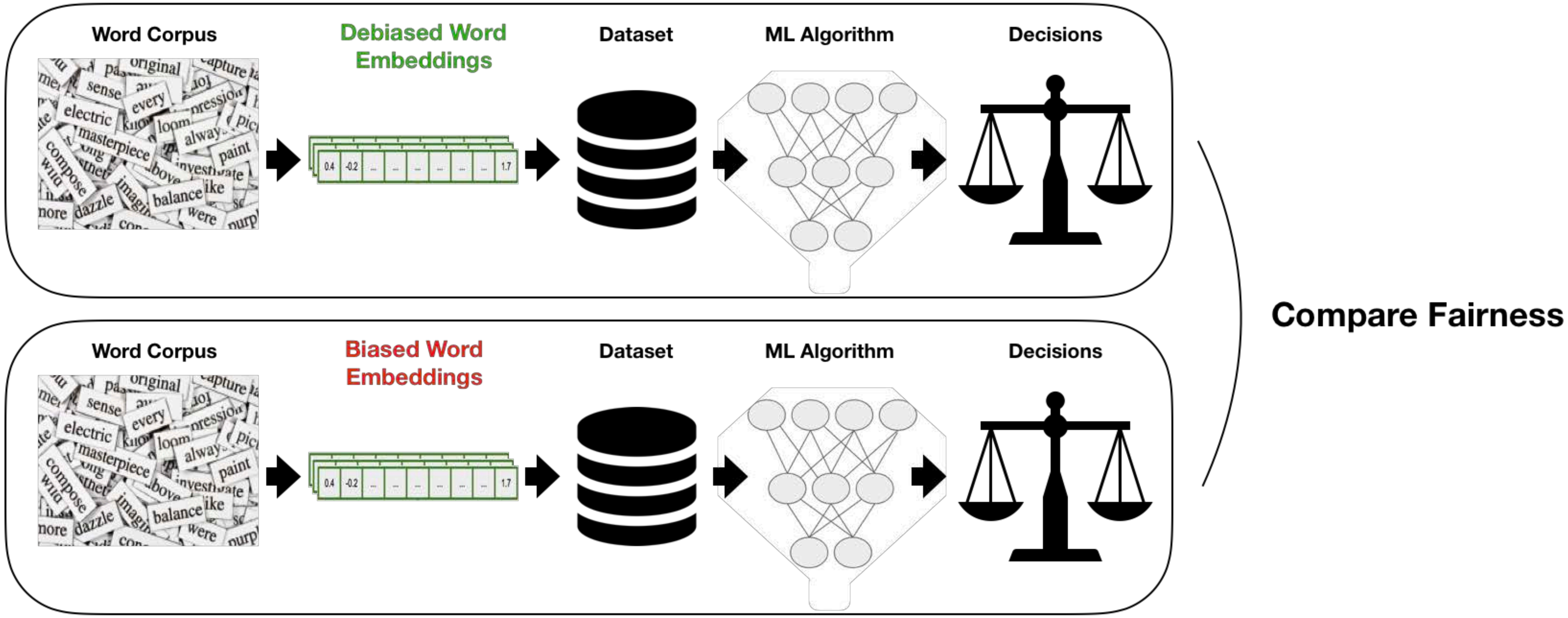


$L_a(z, \hat{z})$ Controls an adversary's ability to predict the projection of protected attributes onto the *directional sentiment vector.*

$L_p(y, \hat{y})$ Controls word vector distortion

Source: Sweeney & Najafian

Testing in Real World NLP Systems



Source: Sweeney & Najafian

How to Measure Fairness in a Downstream Classifier

Template Dataset

Template	#sent.
Sentences with emotion words:	
1. <Person> feels <emotional state word>.	1,200
2. The situation makes <person> feel <emotional state word>.	1,200
3. I made <person> feel <emotional state word>.	1,200
4. <Person> made me feel <emotional state word>.	1,200
5. <Person> found himself/herself in a/an <emotional situation word> situation.	1,200
6. <Person> told us all about the recent <emotional situation word> events.	1,200
7. The conversation with <person> was <emotional situation word>.	1,200

African American		European American	
Female	Male	Female	Male
Ebony	Alonzo	Amanda	Adam
Jasmine	Alphonse	Betsy	Alan
Lakisha	Darnell	Courtney	Andrew
Latisha	Jamel	Ellen	Frank
Latoya	Jerome	Heather	Harry
Nichelle	Lamar	Katie	Jack
Shaniqua	Leroy	Kristin	Josh
Shereen	Malik	Melanie	Justin
Tanisha	Terrence	Nancy	Roger
Tia	Torrance	Stephanie	Ryan

Kiritchenko, S., & Mohammad, S. M. (2018). Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.

How to Measure Fairness in a Downstream Classifier

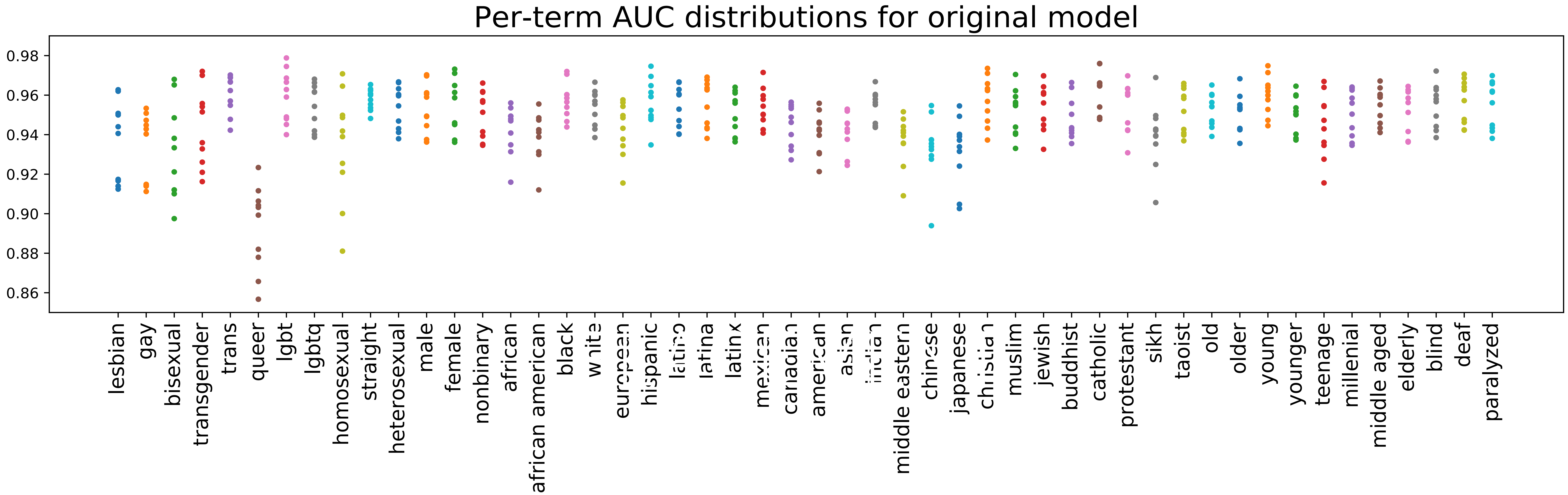
Template Dataset

Template Examples	Label
I am <IDENTITY>	Non-Toxic
I am a <IDENTITY> person, ask me anything	Non-Toxic
<IDENTITY> people are just like everyone else	Non-Toxic
I hate all <IDENTITY>	Toxic
I am a <IDENTITY> person and I hate your guts and think you suck	Toxic
<IDENTITY> people are gross and universally terrible	Toxic

Term	Toxic	Overall
atheist	0.09%	0.10%
queer	0.30%	0.06%
gay	3%	0.50%
transgender	0.04%	0.02%
lesbian	0.10%	0.04%
homosexual	0.80%	0.20%
feminist	0.05%	0.05%
black	0.70%	0.60%
white	0.90%	0.70%
heterosexual	0.02%	0.03%
islam	0.10%	0.08%
muslim	0.20%	0.10%
bisexual	0.01%	0.03%

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018, December). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 67-73). ACM.

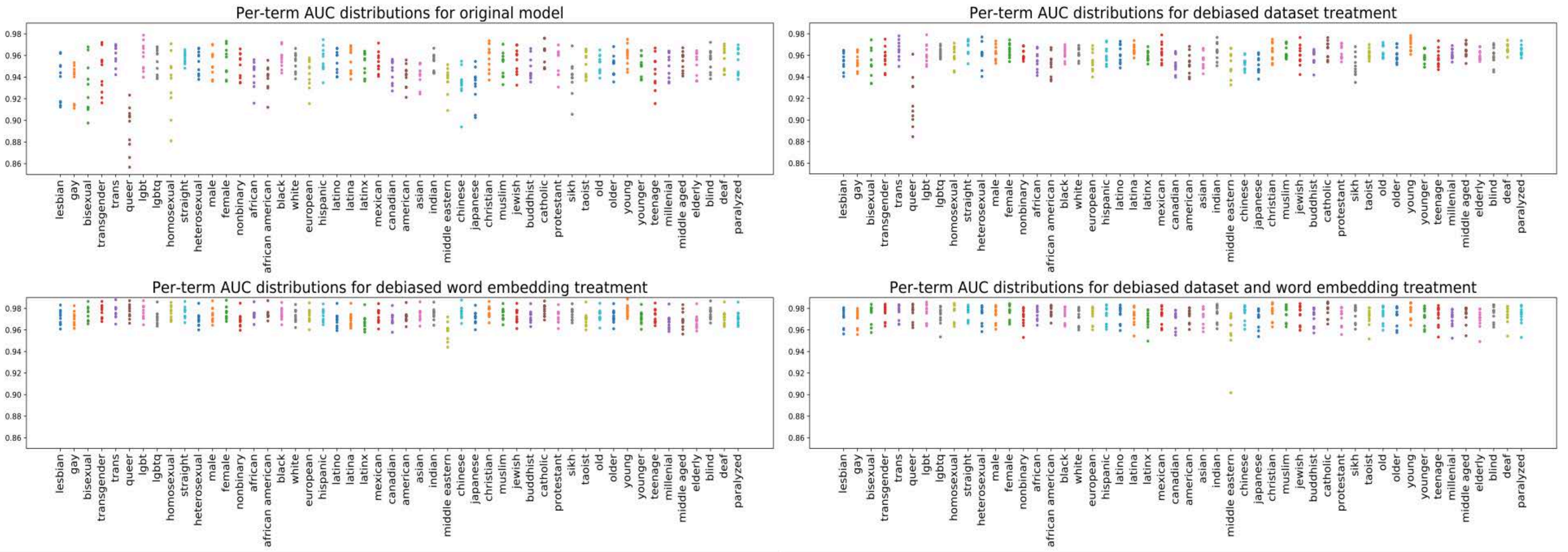
Results on a Real-World Toxicity Classifier



Source: Sweeney & Najafian

MITD-Lab USAID FROM THE AMERICAN PEOPLE CITE

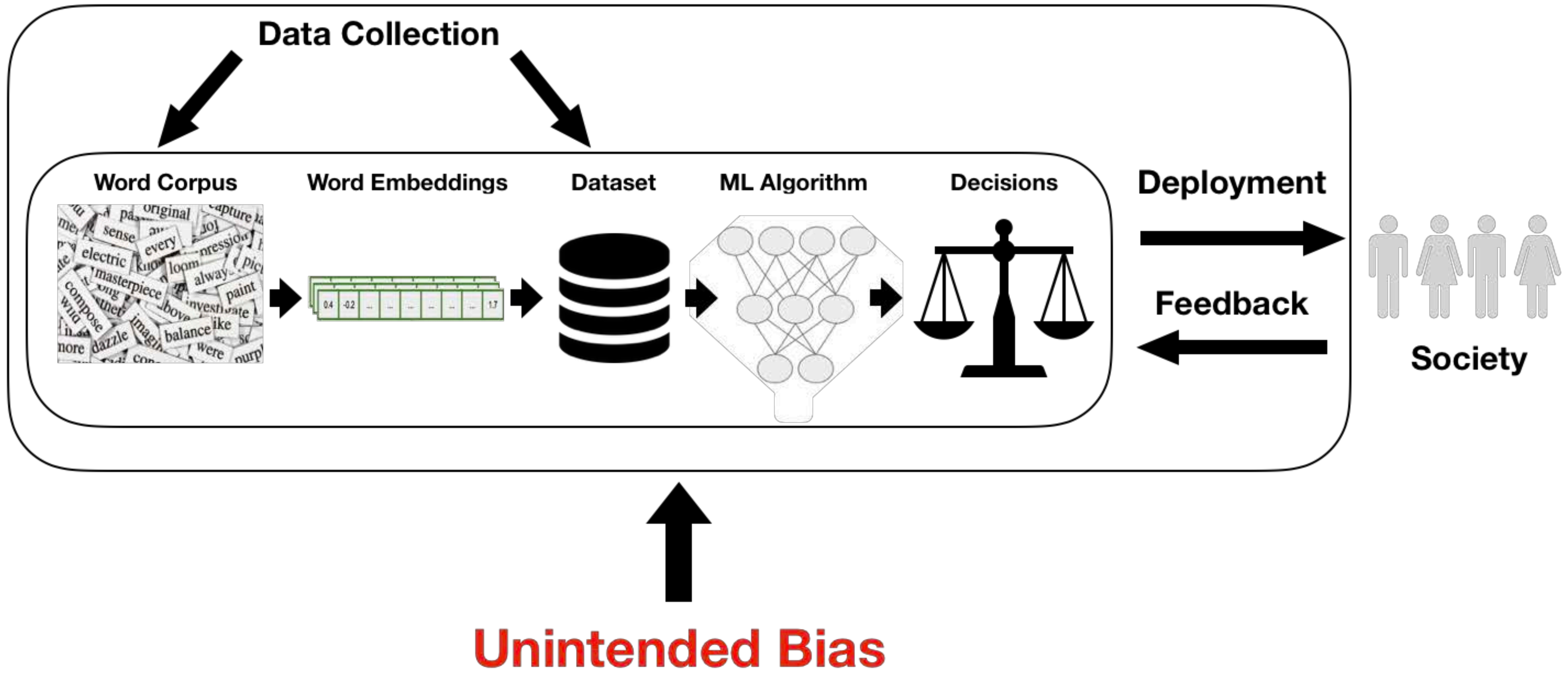
Comparisons to the State-of-the-Art Debiasing Techniques



Source: Sweeney & Najafian

Key Takeaways

- There is no silver bullet (various applications, various types of bias)
- Bias mitigation at all stages of the ML pipeline is essential
- Cannot all be solved in academia



Source: Sweeney & Najafian

Thank you

Audace Nakeshimana
Undergraduate Student and Researcher, MIT

audace@mit.edu