UNC DEPARTMENT OF COMPUTER SCIENCE

UNC NLP

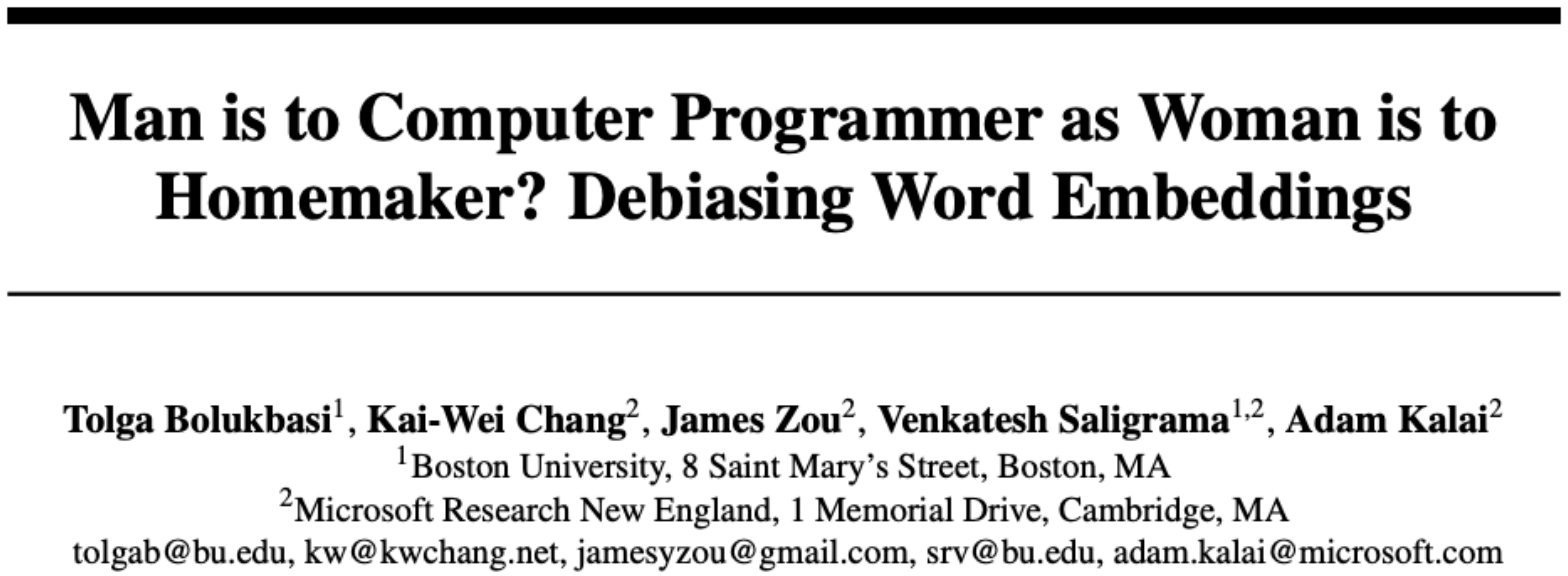
# Adversarial Scrubbing of Demographic Information for Text Classification

***Somnath Basu Roy Chowdhury***, *Sayan Ghosh, Yiyuan Li, Junier B. Oliva, Shashank Srivastava and Snigdha Chaturvedi*

EMNLP 2021

# Bias in NLP Systems

⚠ Content in the following slides can be offensive to some people

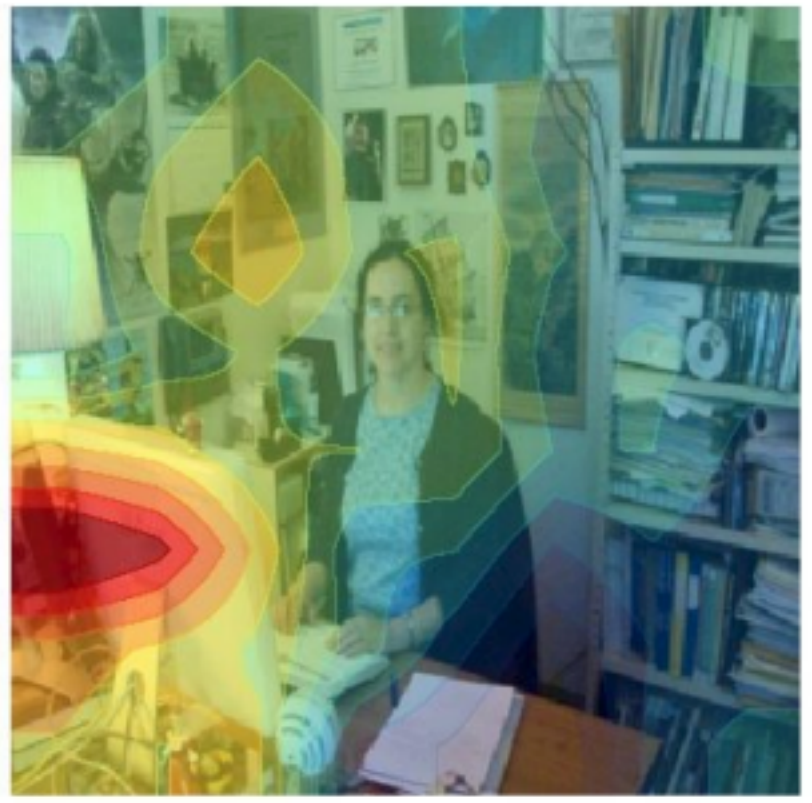


## Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

**Tolga Bolukbasi**[1], **Kai-Wei Chang**[2], **James Zou**[2], **Venkatesh Saligrama**[1,2], **Adam Kalai**[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

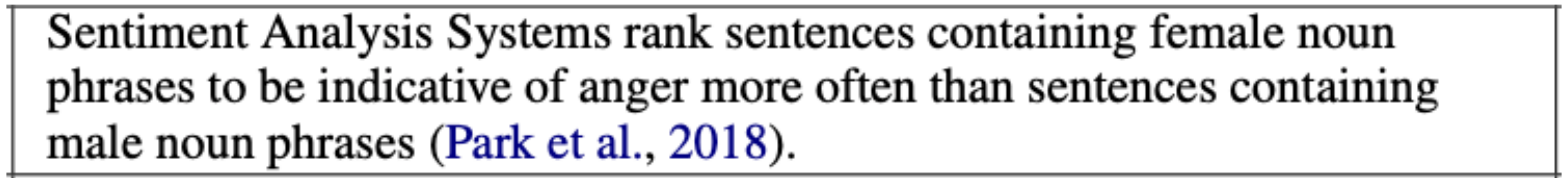
[Bolukbasi et al., 2016]

Wrong



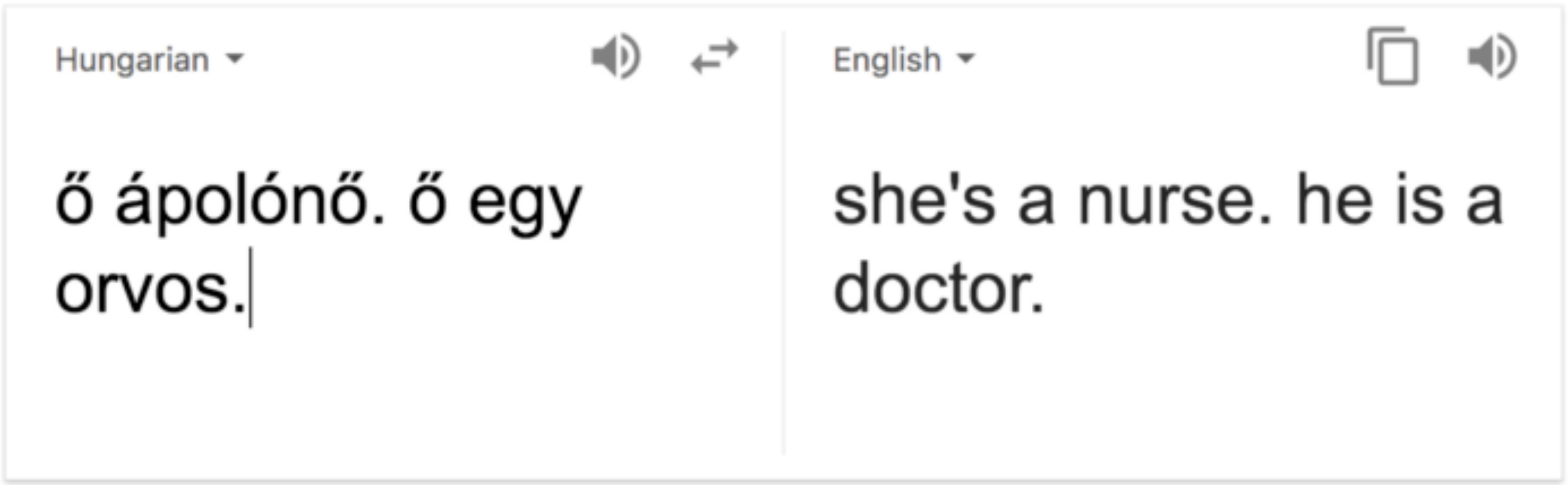
Baseline:
*A **man** sitting at a desk with a laptop computer.*

[Burns et al., 2019]



Sentiment Analysis Systems rank sentences containing female noun phrases to be indicative of anger more often than sentences containing male noun phrases (Park et al., 2018).

[Part et al., 2018]



ő ápolónő. ő egy orvos.

she's a nurse. he is a doctor.

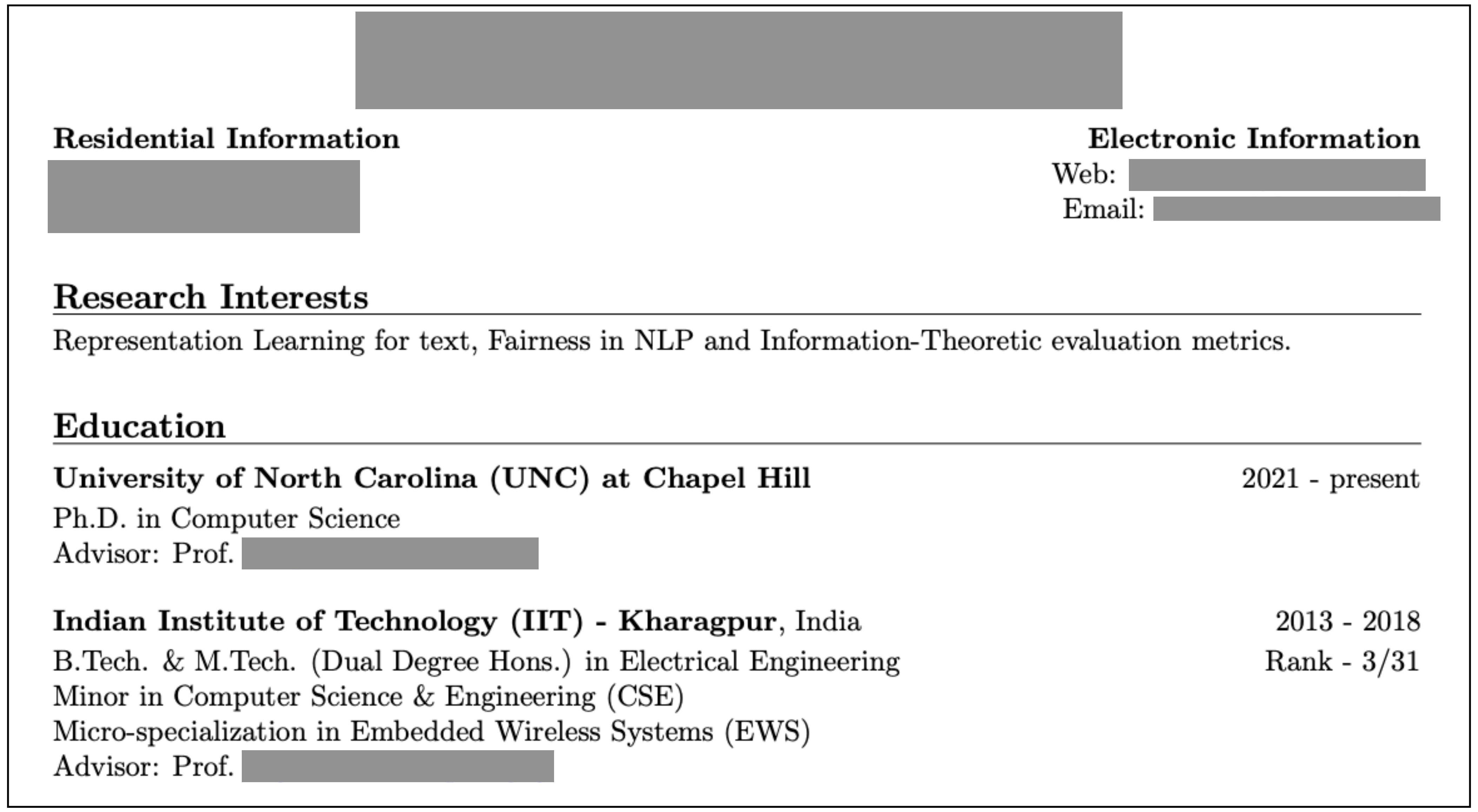
[Douglas, 2018]

# Motivation

Natural Language is used for prediction

- College applications
- Hiring decisions
- Credit Eligibility

# Vanilla Approach



[redacted]

**Residential Information**

[redacted]

**Electronic Information**
Web: [redacted]
Email: [redacted]

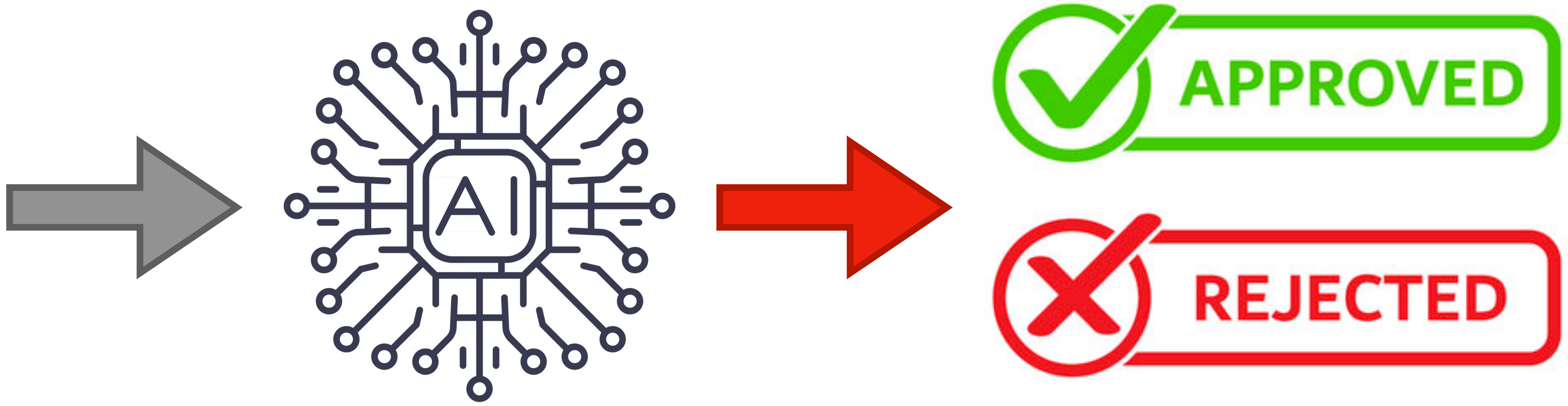
## Research Interests

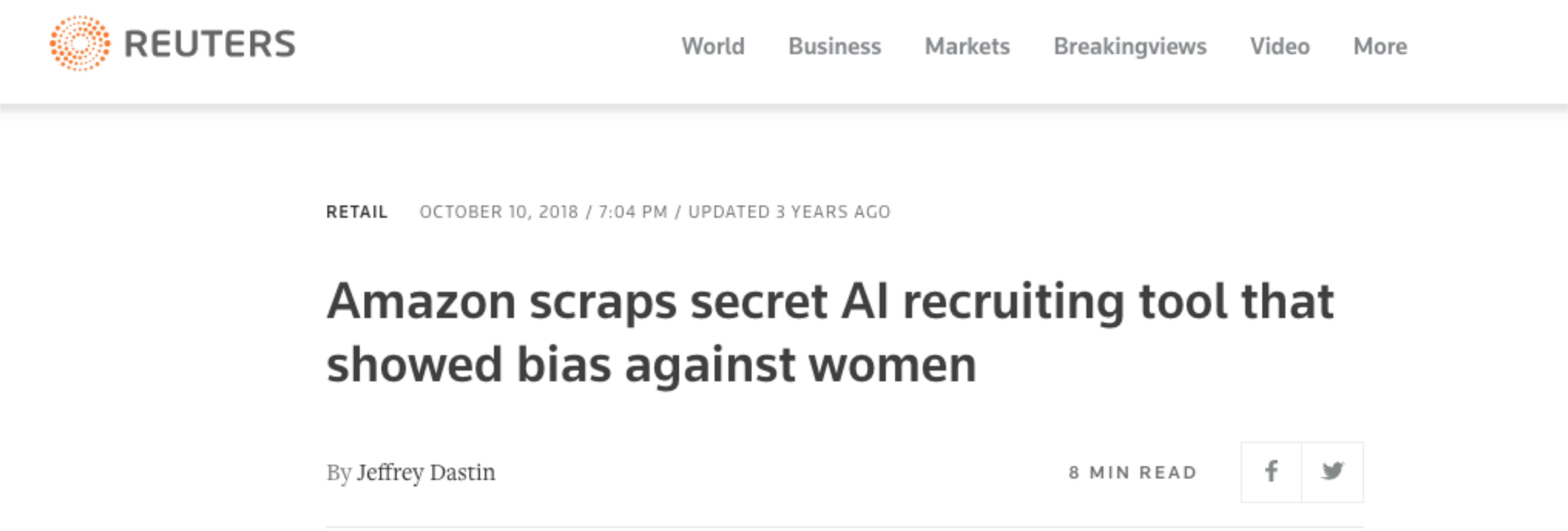
Representation Learning for text, Fairness in NLP and Information-Theoretic evaluation metrics.

## Education

**University of North Carolina (UNC) at Chapel Hill** 2021 - present
Ph.D. in Computer Science
Advisor: Prof. [redacted]

**Indian Institute of Technology (IIT) - Kharagpur**, India 2013 - 2018
B.Tech. & M.Tech. (Dual Degree Hons.) in Electrical Engineering Rank - 3/31
Minor in Computer Science & Engineering (CSE)
Micro-specialization in Embedded Wireless Systems (EWS)
Advisor: Prof. [redacted]



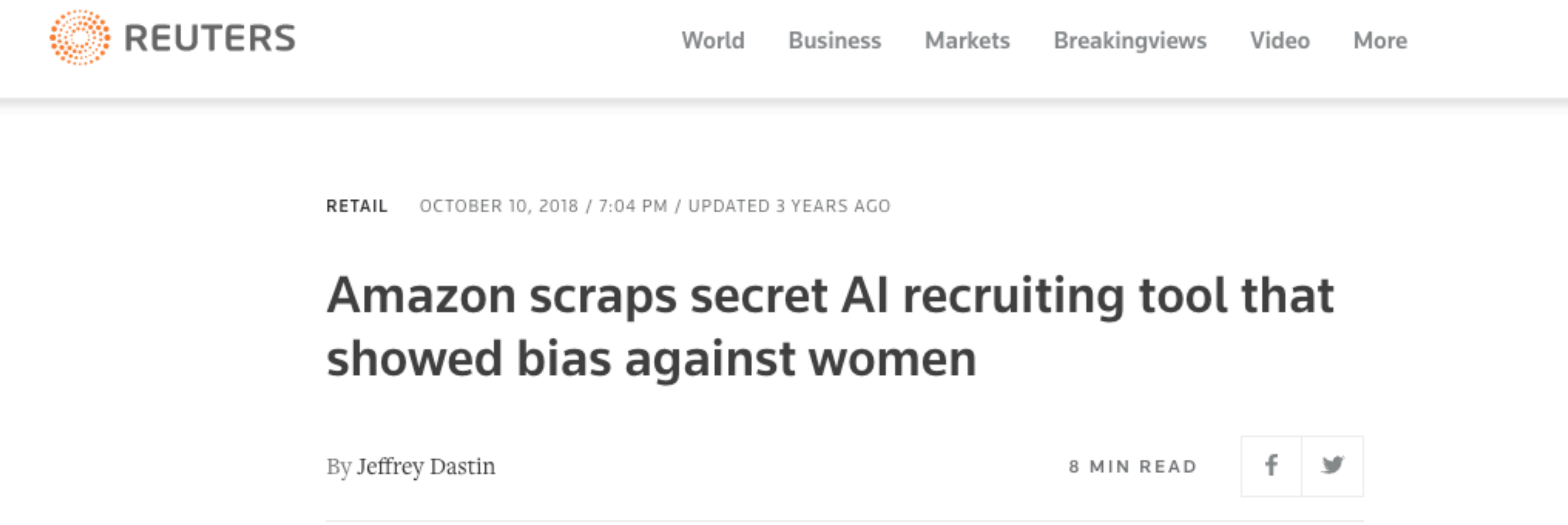


RETAIL OCTOBER 10, 2018 / 7:04 PM / UPDATED 3 YEARS AGO

# Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

8 MIN READ

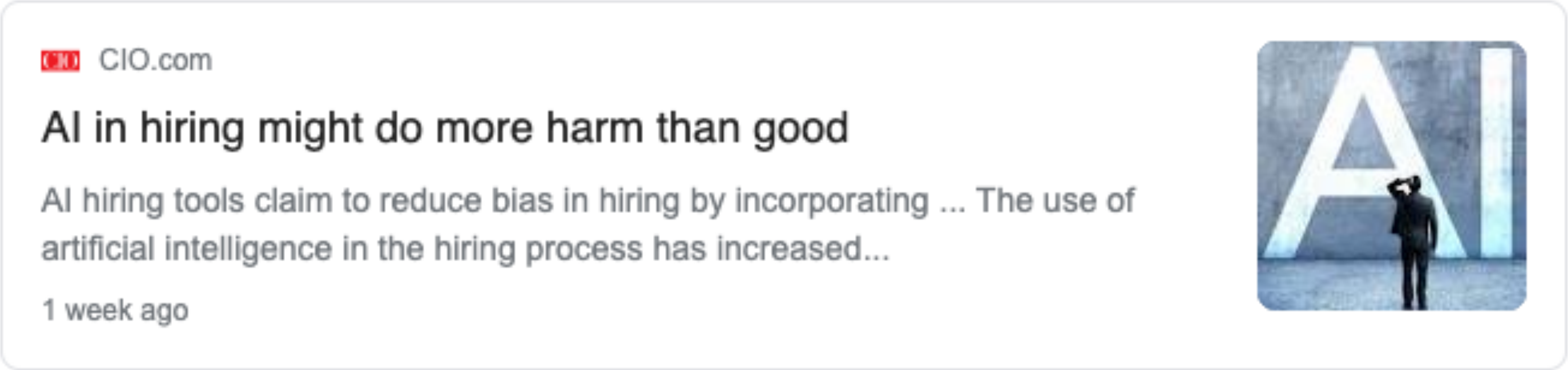


RETAIL OCTOBER 10, 2018 / 7:04 PM / UPDATED 3 YEARS AGO

## Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

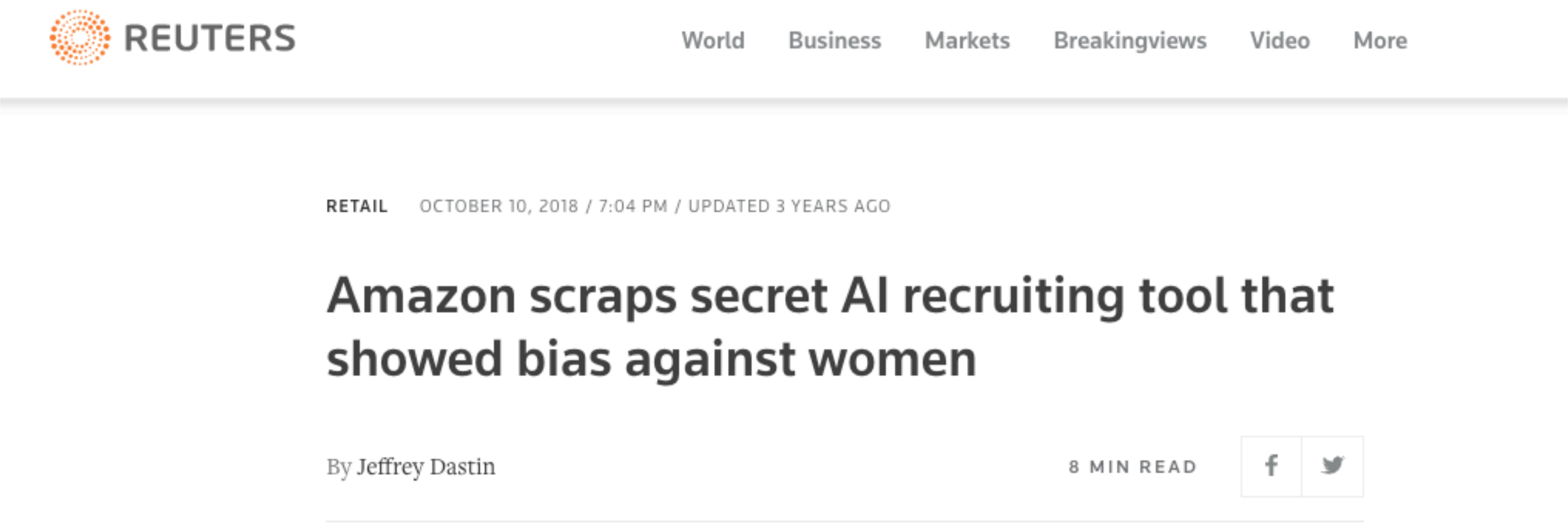
8 MIN READ



AI in hiring might do more harm than good

AI hiring tools claim to reduce bias in hiring by incorporating ... The use of artificial intelligence in the hiring process has increased...

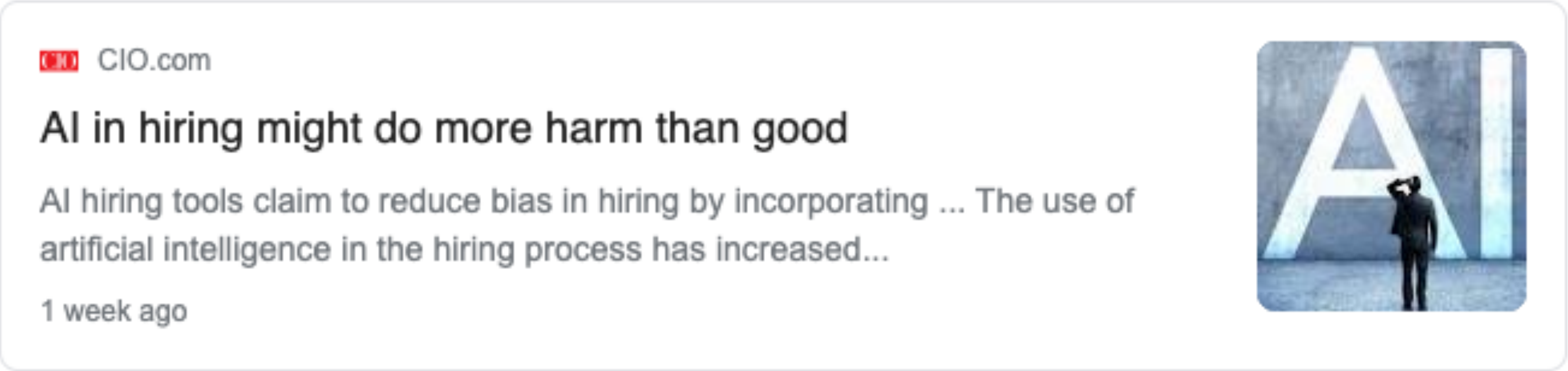
1 week ago



RETAIL OCTOBER 10, 2018 / 7:04 PM / UPDATED 3 YEARS AGO

# Amazon scraps secret AI recruiting tool that showed bias against women

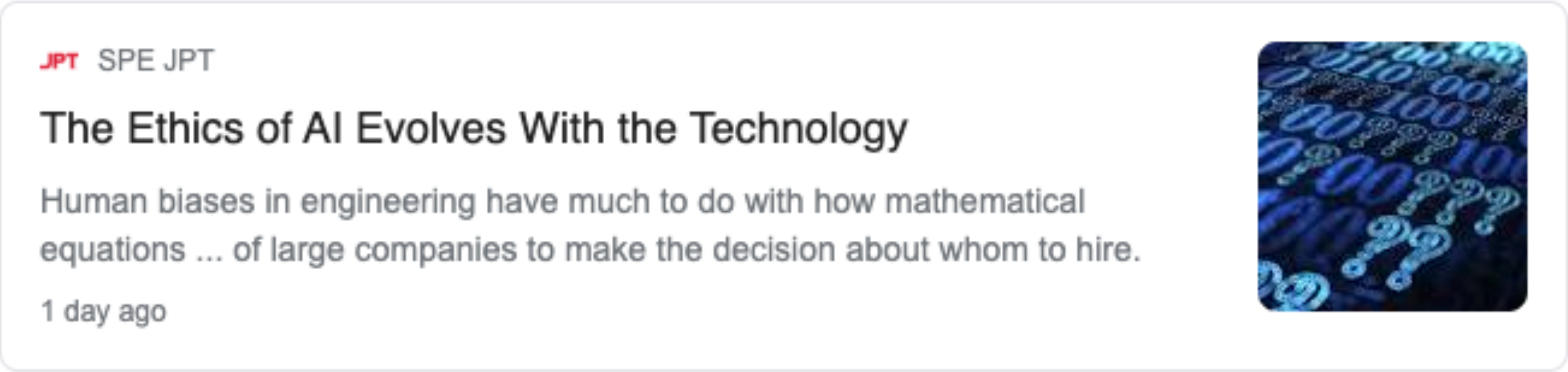
By Jeffrey Dastin


CIO.com

AI in hiring might do more harm than good

AI hiring tools claim to reduce bias in hiring by incorporating ... The use of artificial intelligence in the hiring process has increased...

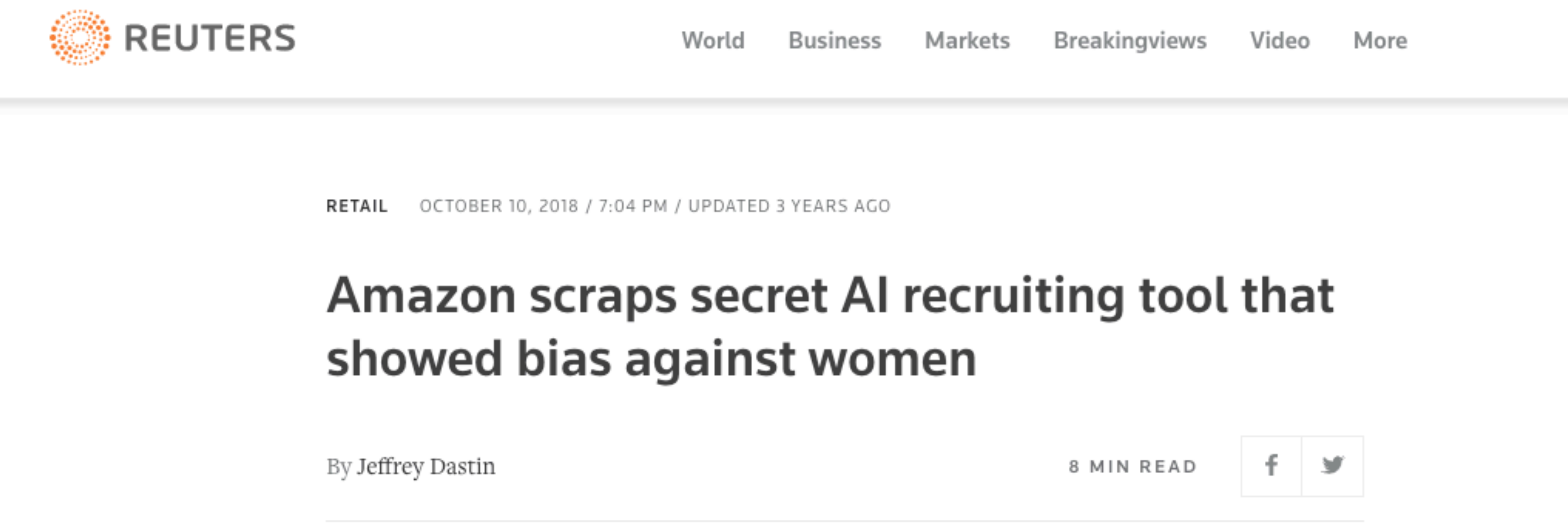
1 week ago


SPE JPT

The Ethics of AI Evolves With the Technology

Human biases in engineering have much to do with how mathematical equations ... of large companies to make the decision about whom to hire.

1 day ago

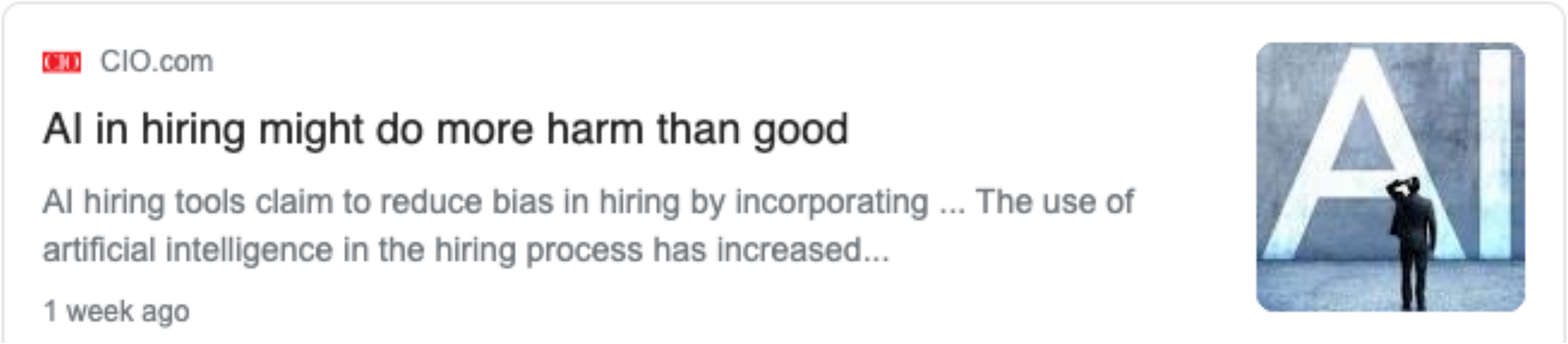


RETAIL OCTOBER 10, 2018 / 7:04 PM / UPDATED 3 YEARS AGO

## Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

8 MIN READ

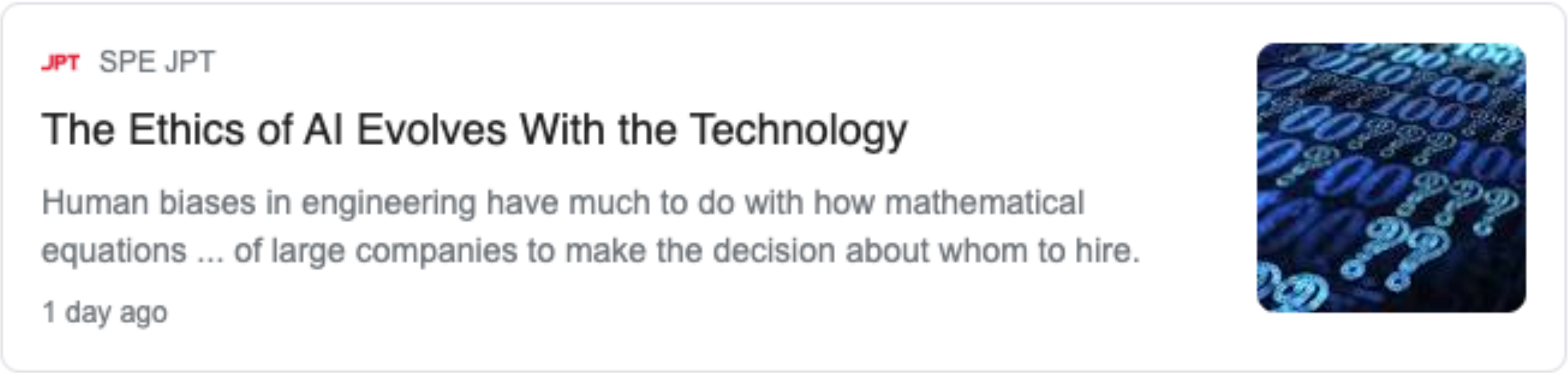


CIO.com

### AI in hiring might do more harm than good

AI hiring tools claim to reduce bias in hiring by incorporating ... The use of artificial intelligence in the hiring process has increased...

1 week ago

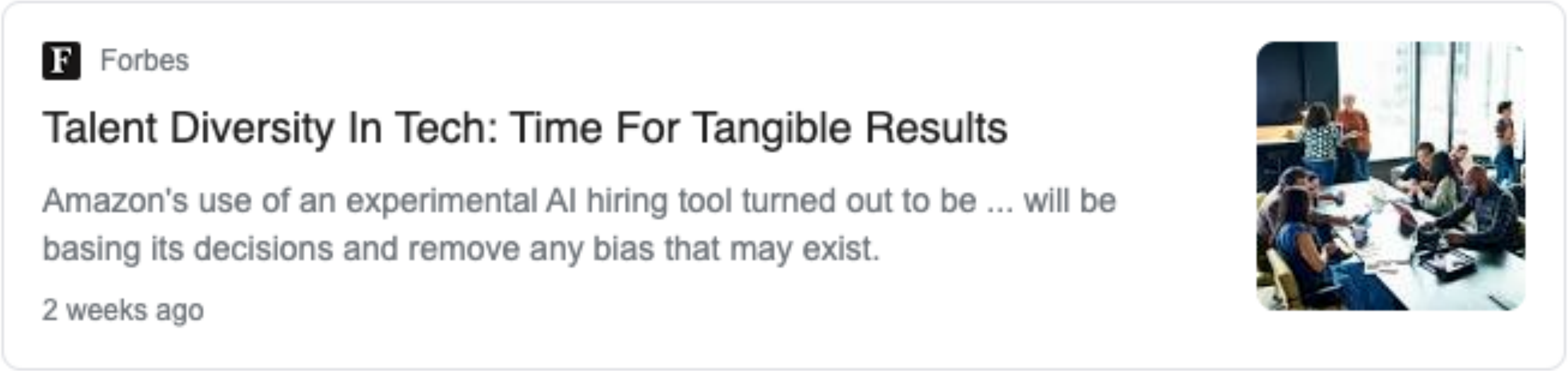


JPT SPE JPT

### The Ethics of AI Evolves With the Technology

Human biases in engineering have much to do with how mathematical equations ... of large companies to make the decision about whom to hire.

1 day ago

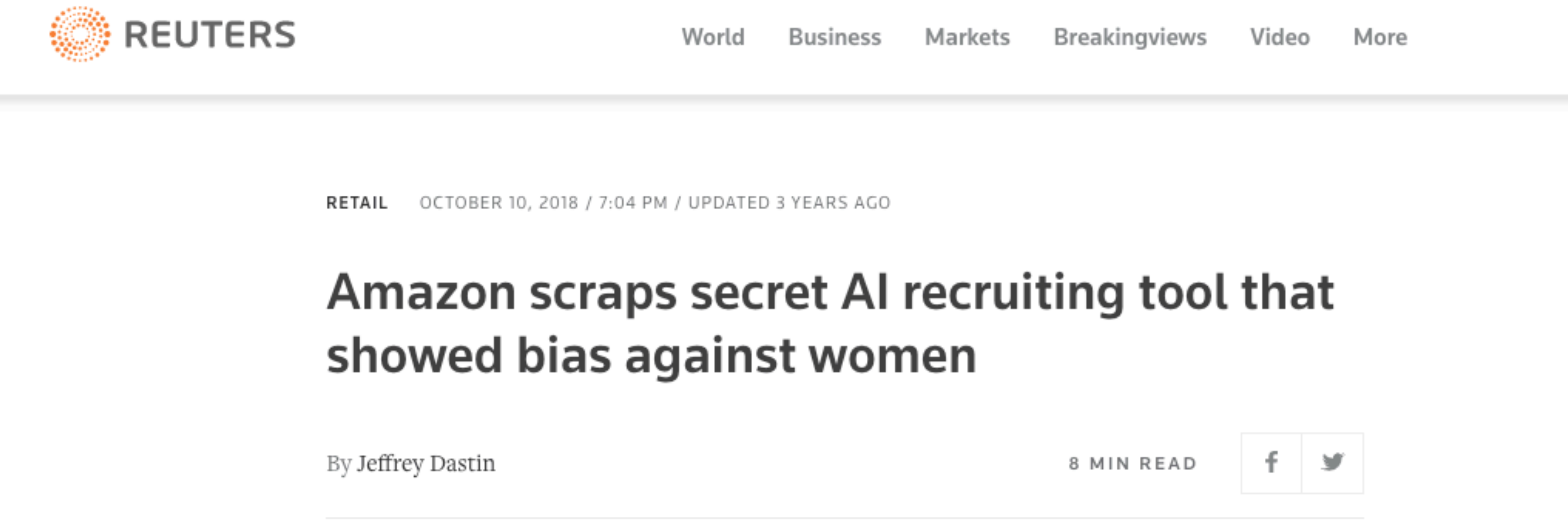


Forbes

### Talent Diversity In Tech: Time For Tangible Results

Amazon's use of an experimental AI hiring tool turned out to be ... will be basing its decisions and remove any bias that may exist.

2 weeks ago

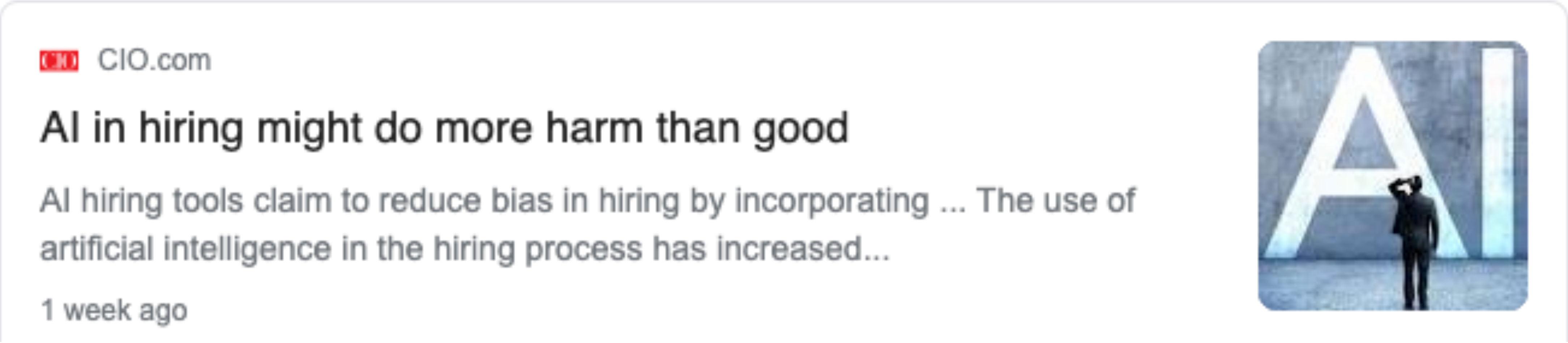


RETAIL OCTOBER 10, 2018 / 7:04 PM / UPDATED 3 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

8 MIN READ

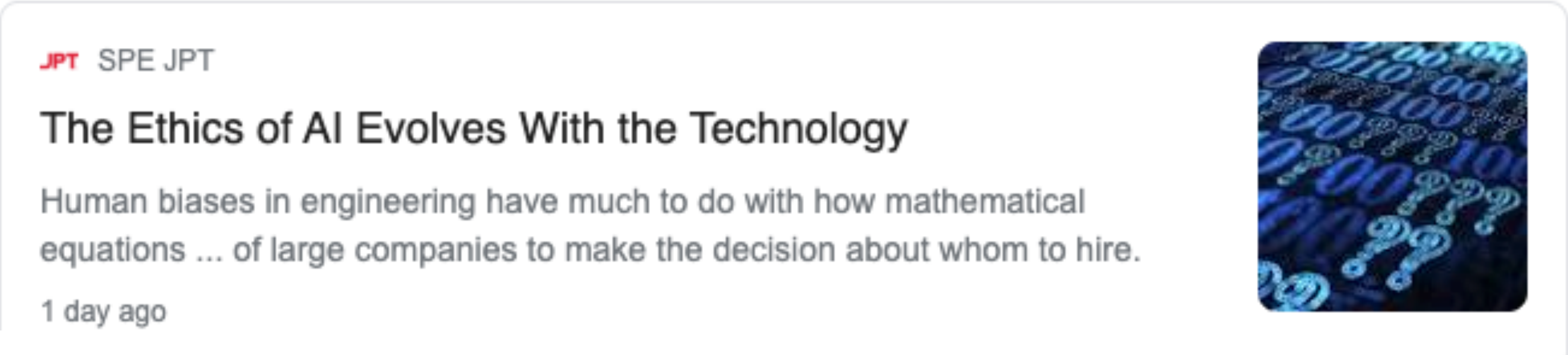


CIO.com

AI in hiring might do more harm than good

AI hiring tools claim to reduce bias in hiring by incorporating ... The use of artificial intelligence in the hiring process has increased...

1 week ago

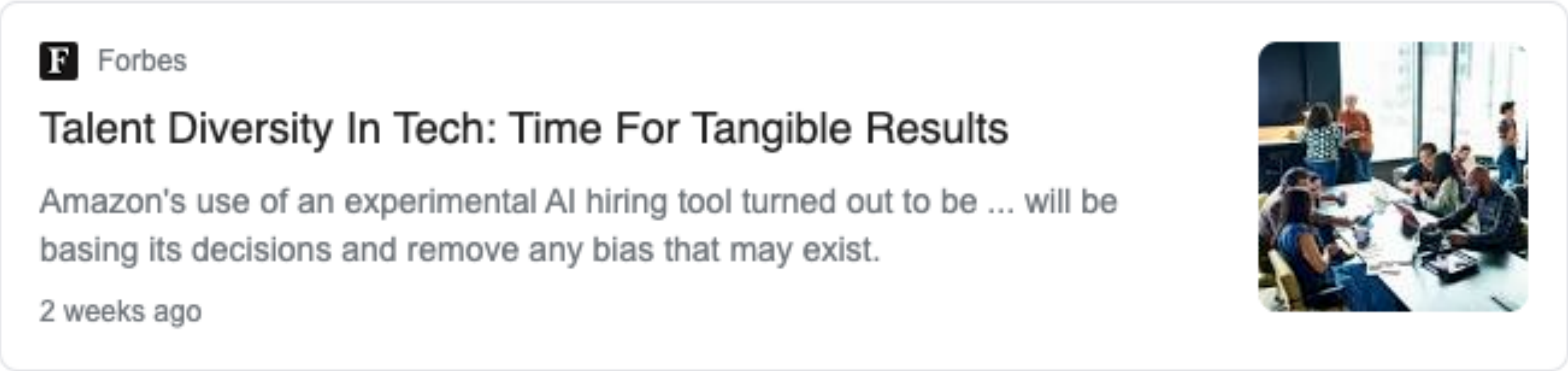


SPE JPT

The Ethics of AI Evolves With the Technology

Human biases in engineering have much to do with how mathematical equations ... of large companies to make the decision about whom to hire.

1 day ago

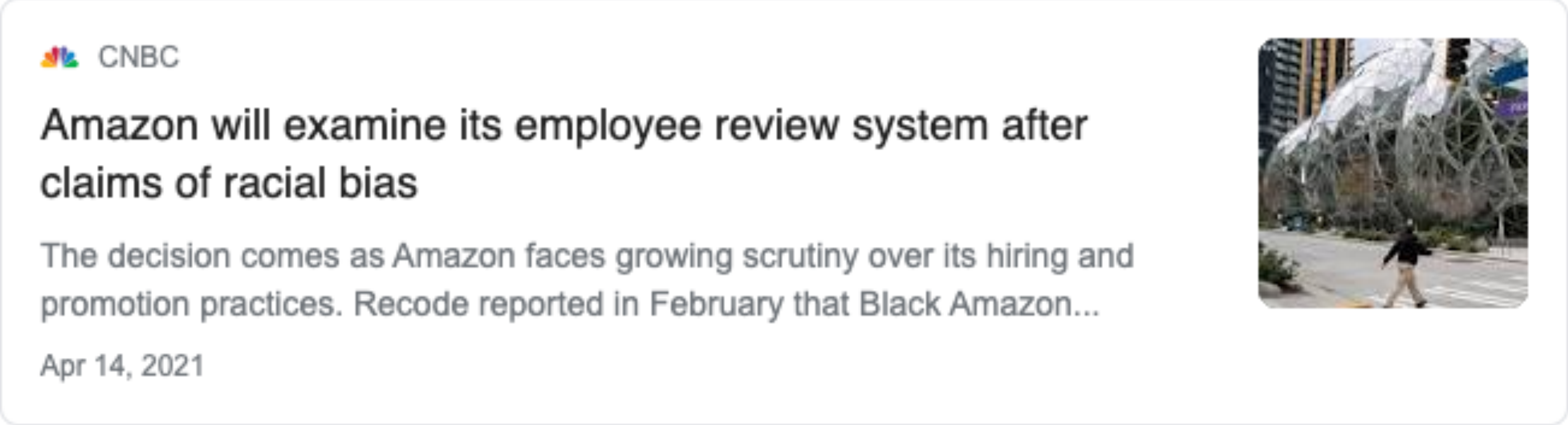


Forbes

Talent Diversity In Tech: Time For Tangible Results

Amazon's use of an experimental AI hiring tool turned out to be ... will be basing its decisions and remove any bias that may exist.

2 weeks ago

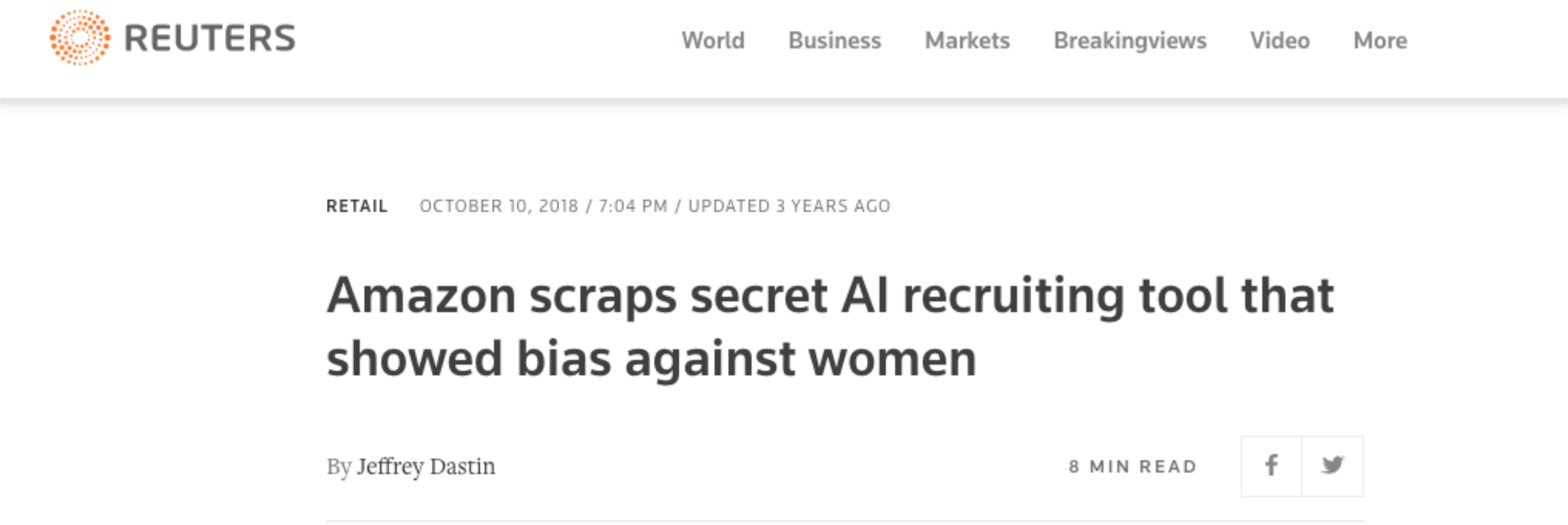


CNBC

Amazon will examine its employee review system after claims of racial bias

The decision comes as Amazon faces growing scrutiny over its hiring and promotion practices. Recode reported in February that Black Amazon...

Apr 14, 2021



REUTERS

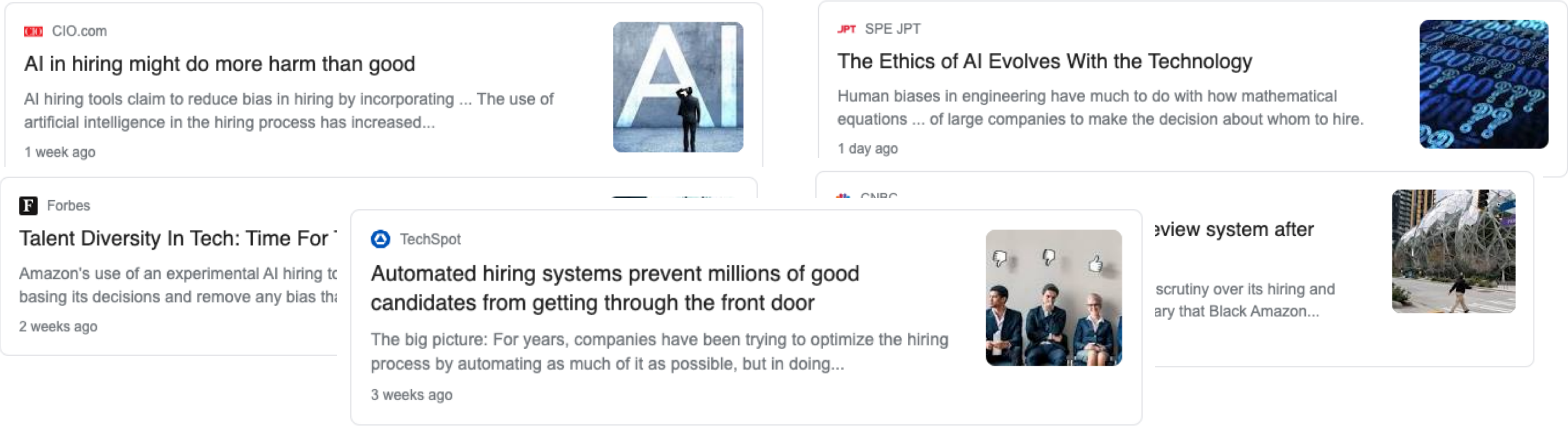
World Business Markets Breakingviews Video More

RETAIL OCTOBER 10, 2018 / 7:04 PM / UPDATED 3 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

8 MIN READ



CIO.com

AI in hiring might do more harm than good

AI hiring tools claim to reduce bias in hiring by incorporating ... The use of artificial intelligence in the hiring process has increased...

1 week ago

SPE JPT

The Ethics of AI Evolves With the Technology

Human biases in engineering have much to do with how mathematical equations ... of large companies to make the decision about whom to hire.

1 day ago

Forbes

Talent Diversity In Tech: Time For

Amazon's use of an experimental AI hiring tc
basing its decisions and remove any bias thi

2 weeks ago

TechSpot

Automated hiring systems prevent millions of good candidates from getting through the front door

The big picture: For years, companies have been trying to optimize the hiring process by automating as much of it as possible, but in doing...

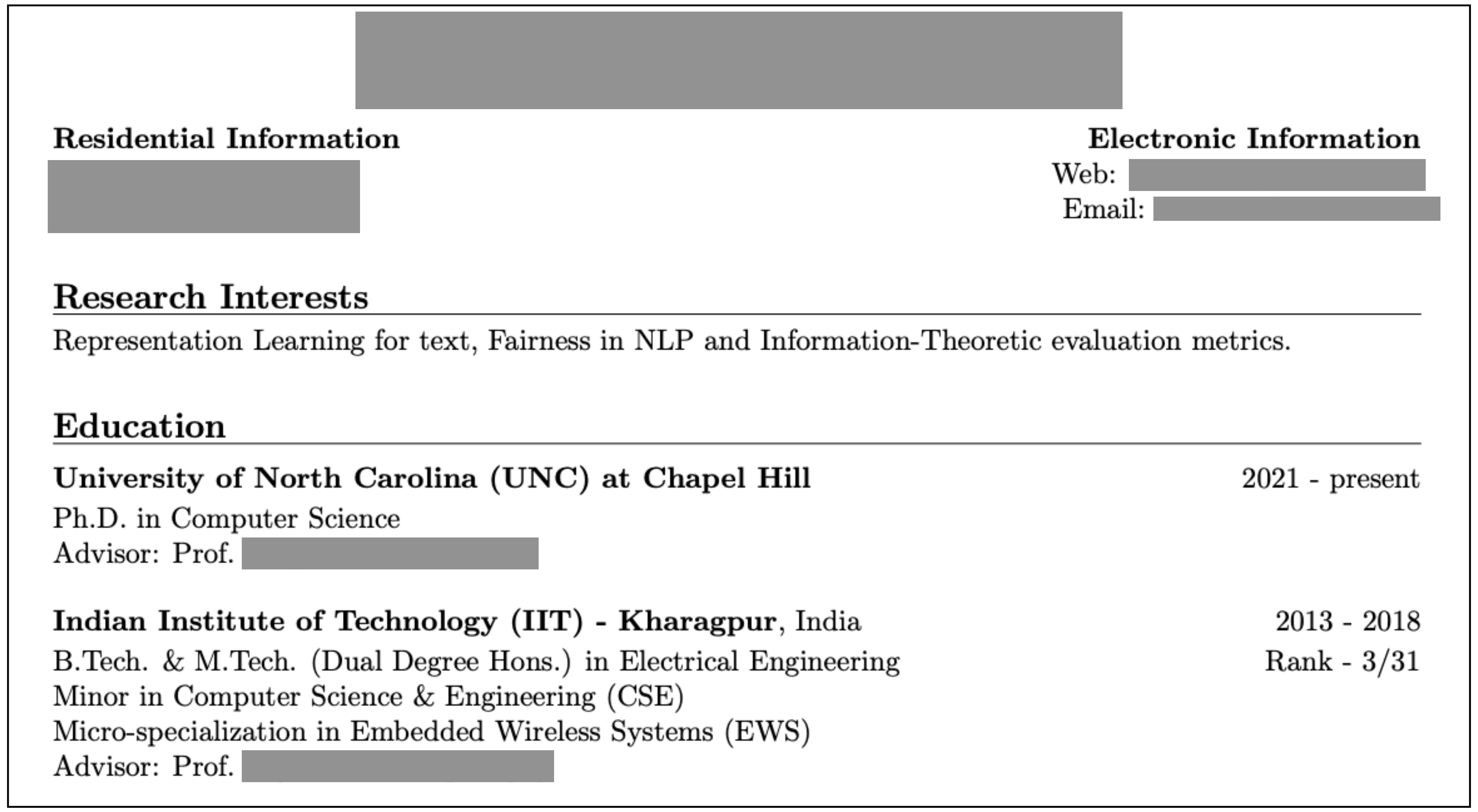
3 weeks ago

CNBC

eview system after

scrutiny over its hiring and
ary that Black Amazon...

# Text Classification



[redacted]

**Residential Information**

[redacted]

**Electronic Information**
Web: [redacted]
Email: [redacted]

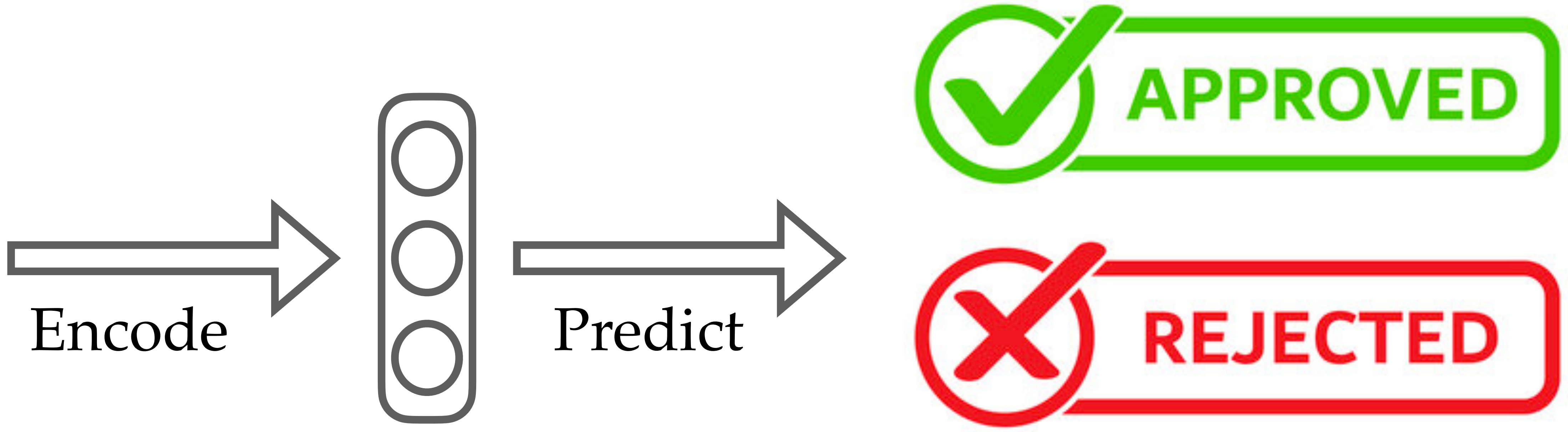
**Research Interests**

Representation Learning for text, Fairness in NLP and Information-Theoretic evaluation metrics.

**Education**

**University of North Carolina (UNC) at Chapel Hill** 2021 - present
Ph.D. in Computer Science
Advisor: Prof. [redacted]

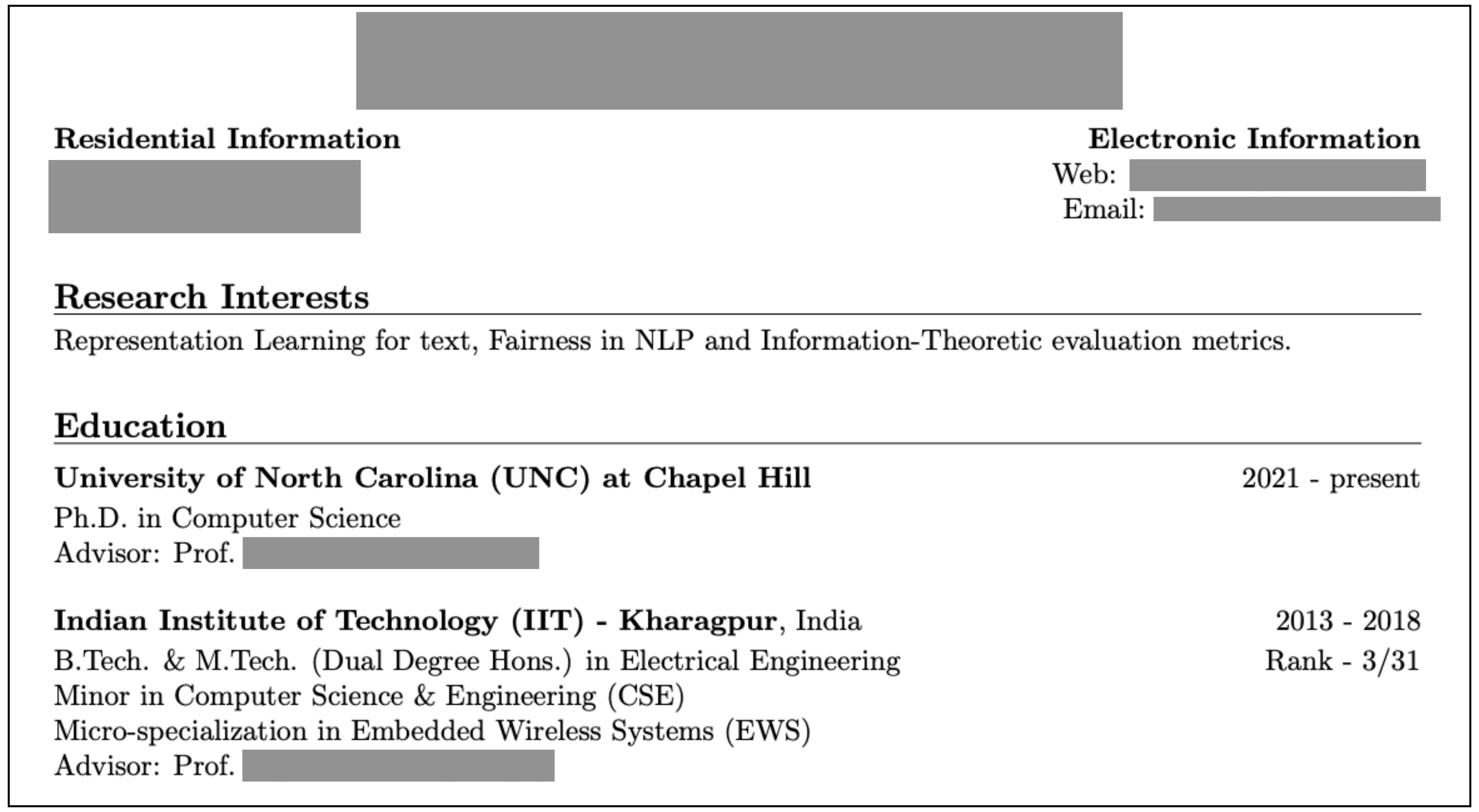
**Indian Institute of Technology (IIT) - Kharagpur**, India 2013 - 2018
B.Tech. & M.Tech. (Dual Degree Hons.) in Electrical Engineering Rank - 3/31
Minor in Computer Science & Engineering (CSE)
Micro-specialization in Embedded Wireless Systems (EWS)
Advisor: Prof. [redacted]



# Text Classification

- Natural Language is highly indicative of demographic attributes (gender/age/race)
- Models can encode such information *even without* having direct access to them

# Text Classification



[redacted]

**Residential Information**

[redacted]

**Electronic Information**

Web: [redacted]

Email: [redacted]

**Research Interests**

Representation Learning for text, Fairness in NLP and Information-Theoretic evaluation metrics.

**Education**

**University of North Carolina (UNC) at Chapel Hill** 2021 - present

Ph.D. in Computer Science

Advisor: Prof. [redacted]

**Indian Institute of Technology (IIT) - Kharagpur**, India 2013 - 2018

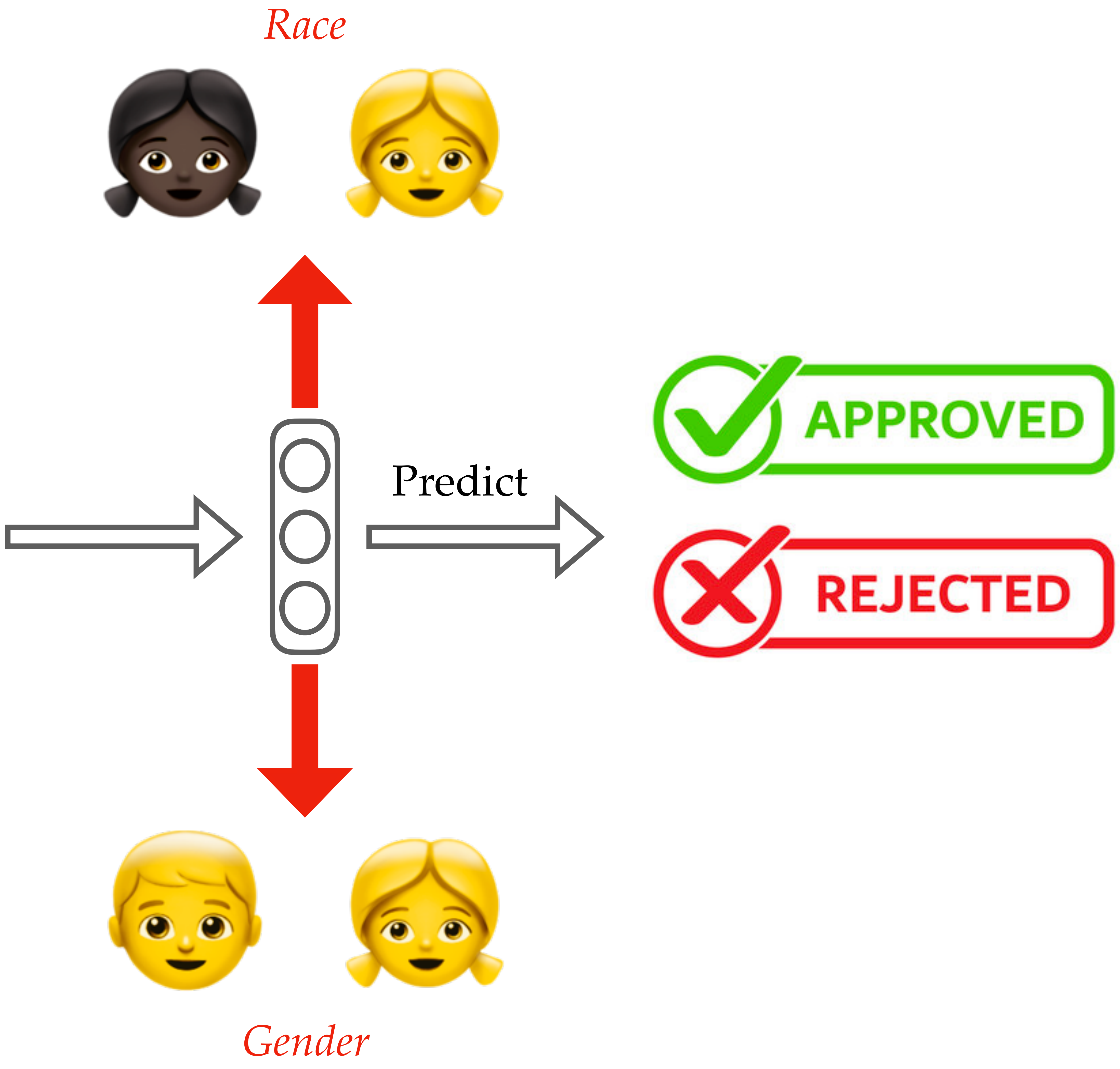
B.Tech. & M.Tech. (Dual Degree Hons.) in Electrical Engineering Rank - 3/31

Minor in Computer Science & Engineering (CSE)

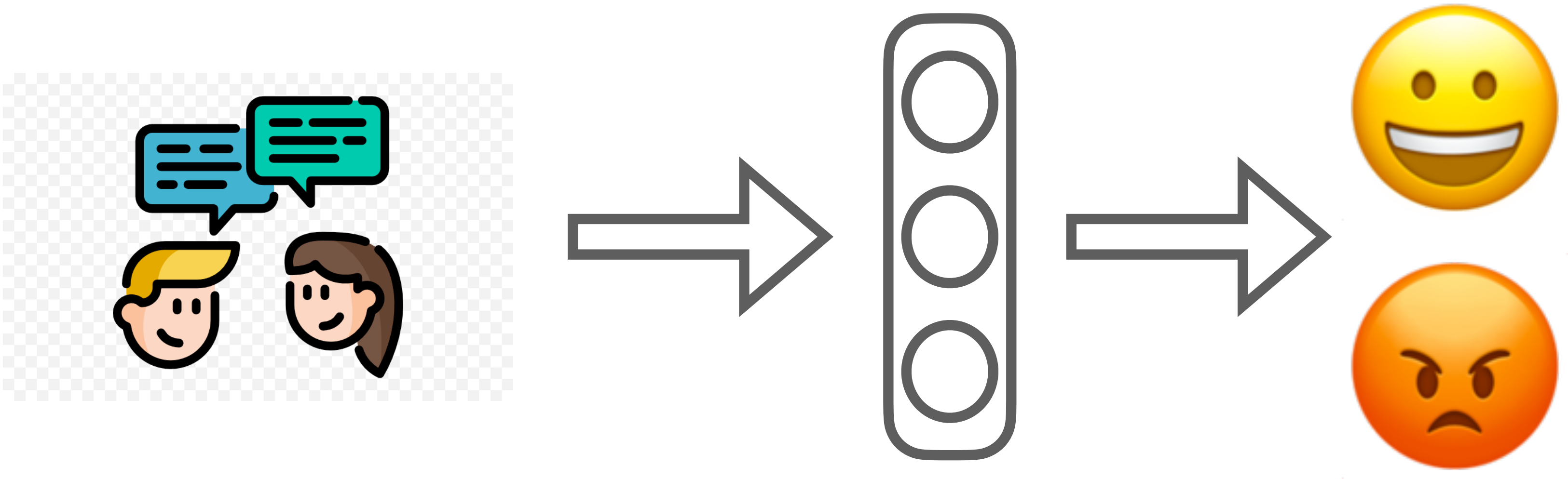
Micro-specialization in Embedded Wireless Systems (EWS)

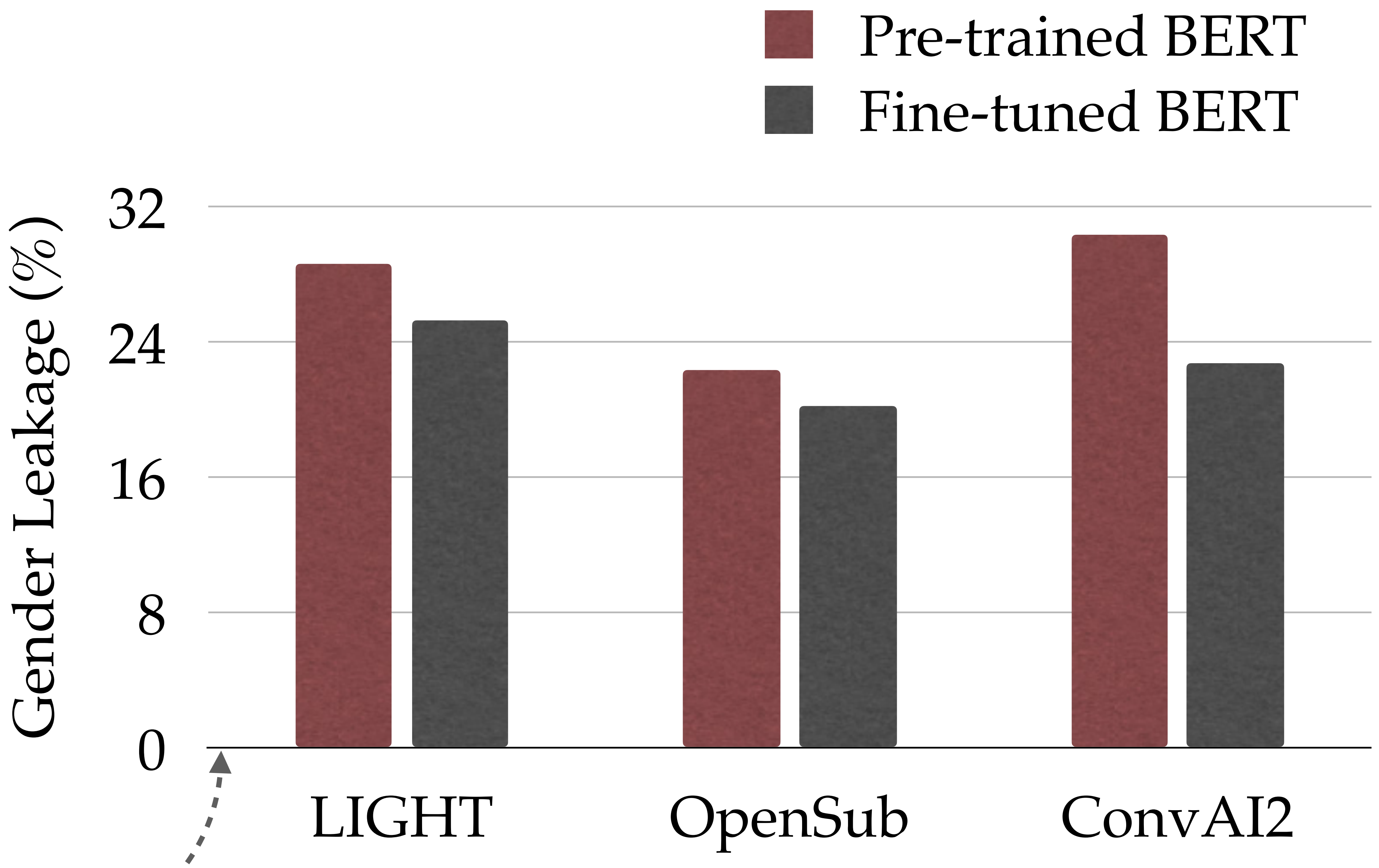
Advisor: Prof. [redacted]

It is possible to extract demographic attributes from intermediate text representations.



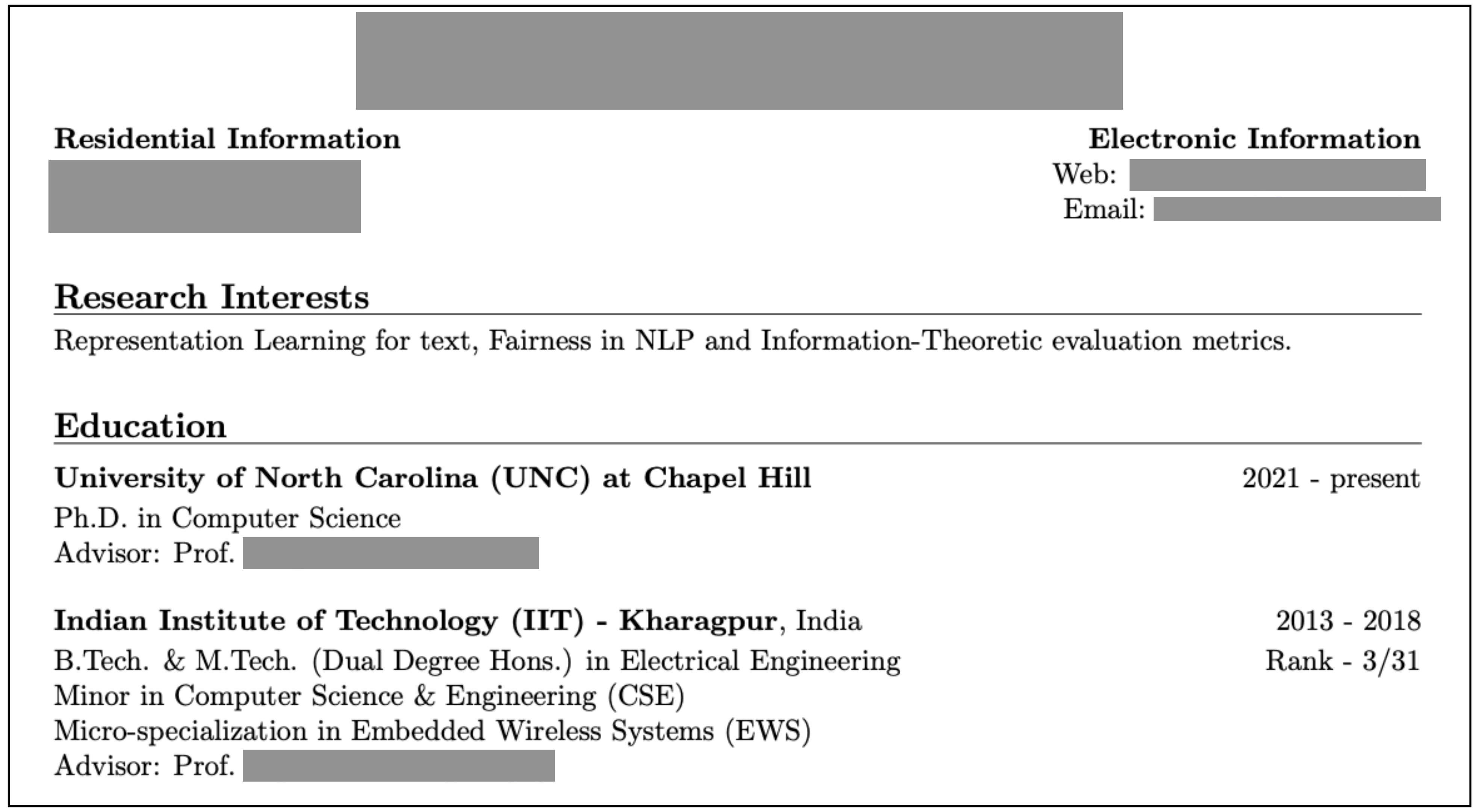
# Text Classification - Example





Random Baseline

# Goal: *Fairness by Blindness*



[redacted]

**Residential Information** | **Electronic Information**

[redacted]

Web: [redacted]

Email: [redacted]

**Research Interests**

Representation Learning for text, Fairness in NLP and Information-Theoretic evaluation metrics.

**Education**

**University of North Carolina (UNC) at Chapel Hill** 2021 - present

Ph.D. in Computer Science

Advisor: Prof. [redacted]

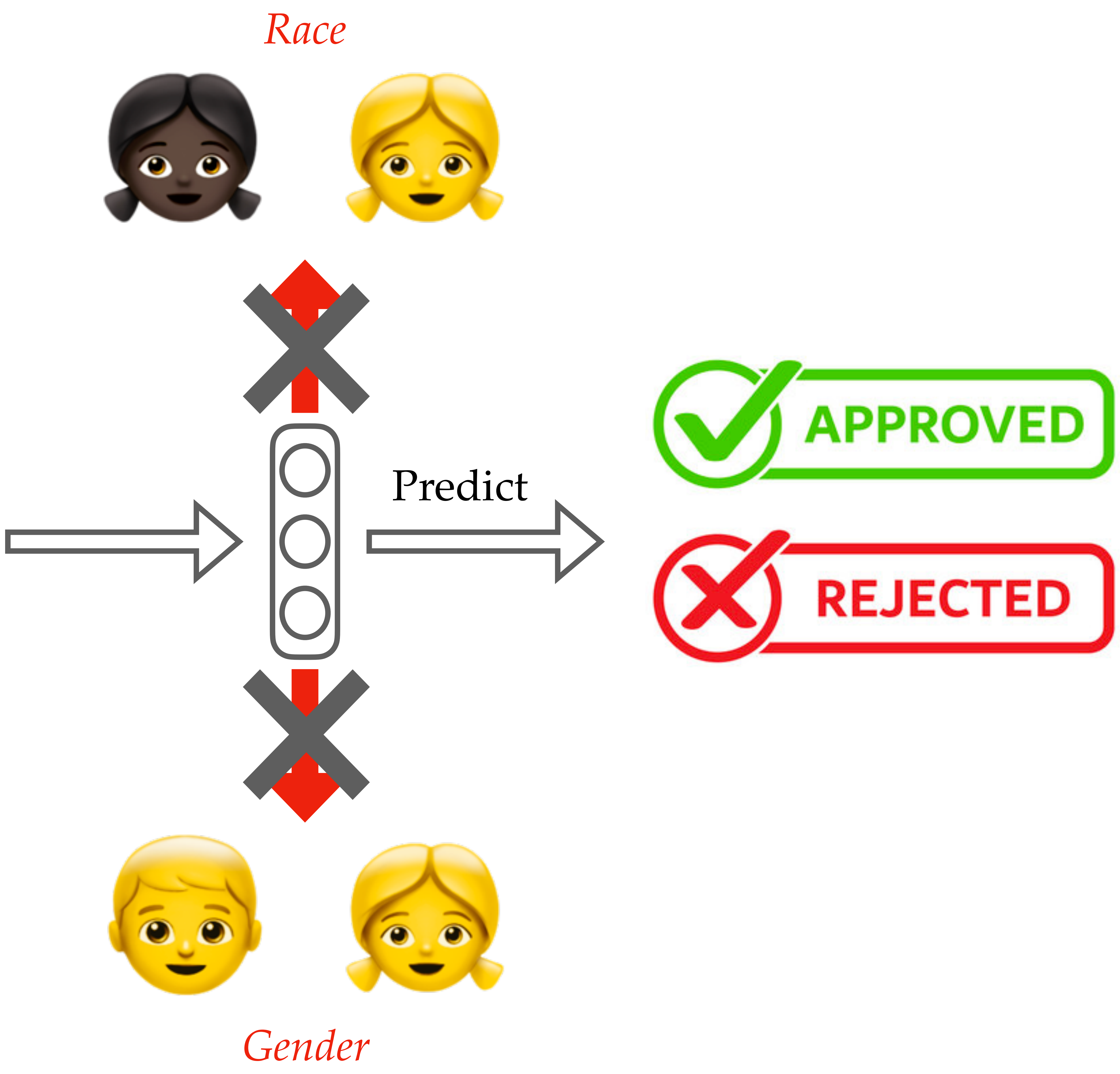
**Indian Institute of Technology (IIT) - Kharagpur**, India 2013 - 2018

B.Tech. & M.Tech. (Dual Degree Hons.) in Electrical Engineering Rank - 3/31

Minor in Computer Science & Engineering (CSE)

Micro-specialization in Embedded Wireless Systems (EWS)

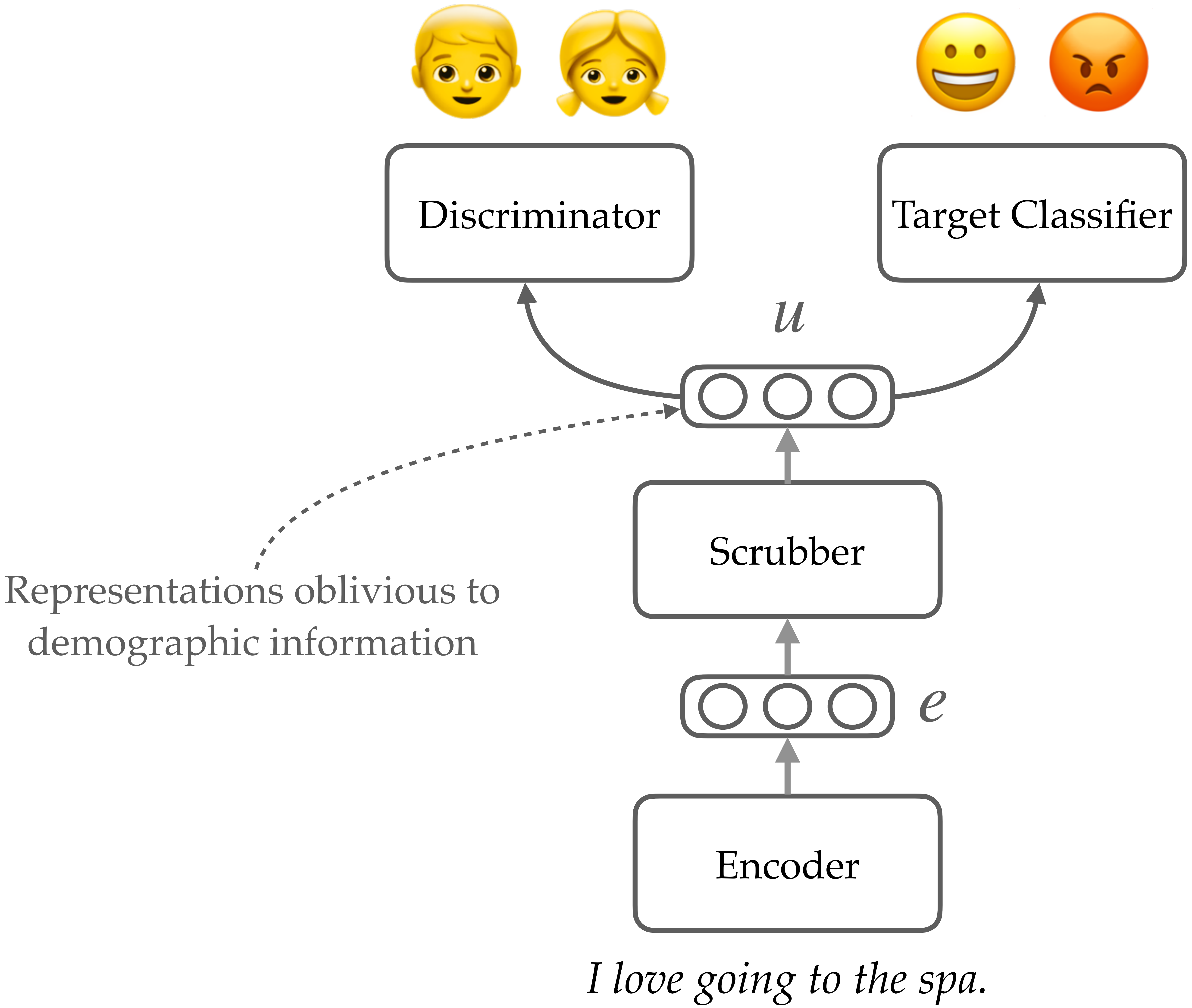
Advisor: Prof. [redacted]



Decisions should not be conditioned on demographic attributes. Intermediate representations should be *oblivious* to such information.

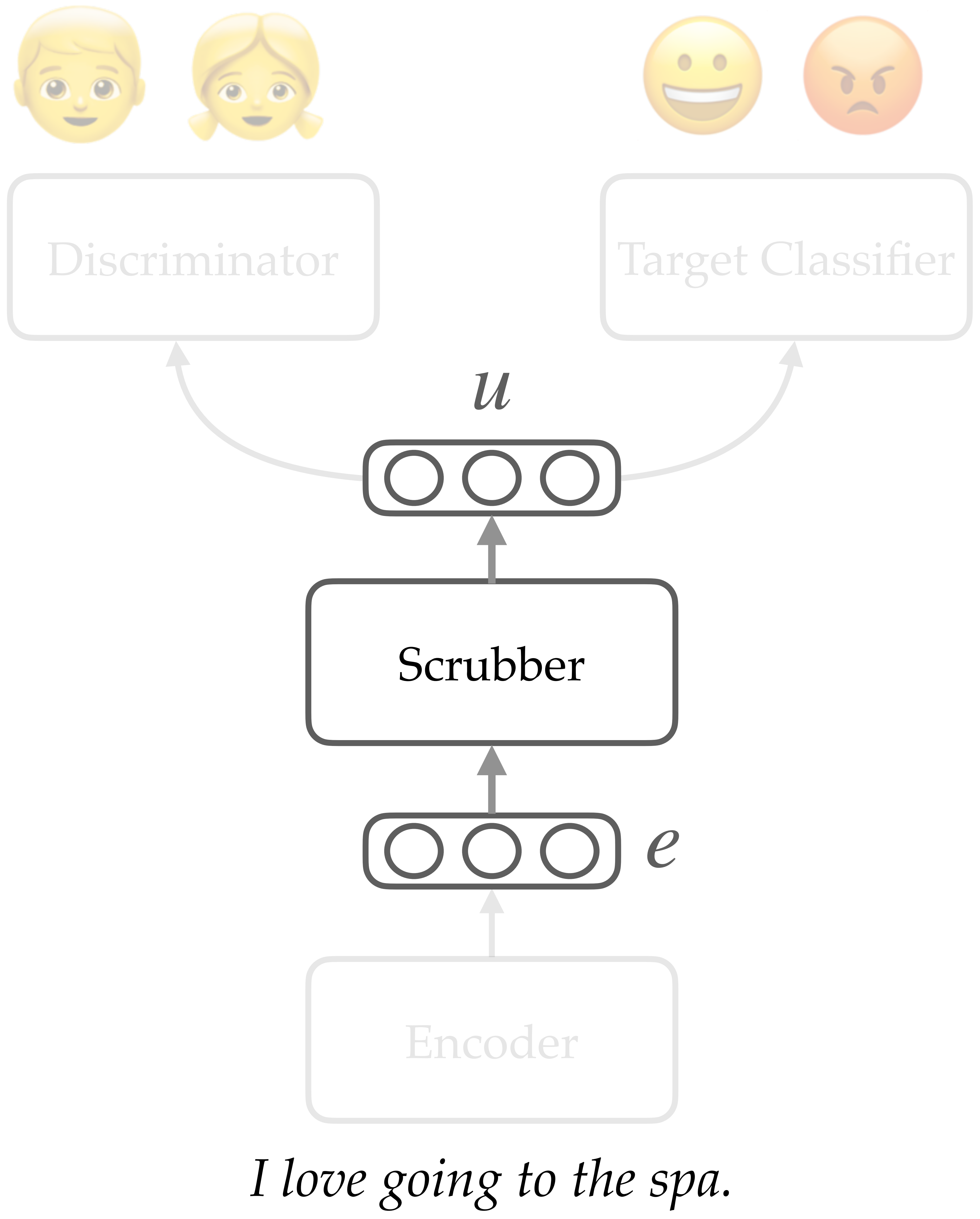
# Adversarial Scrubber (AdS)

# Setup


Discriminator
Target Classifier
$u$
Scrubber
$e$
Encoder
Representations oblivious to demographic information
*I love going to the spa.*

# Scrubber

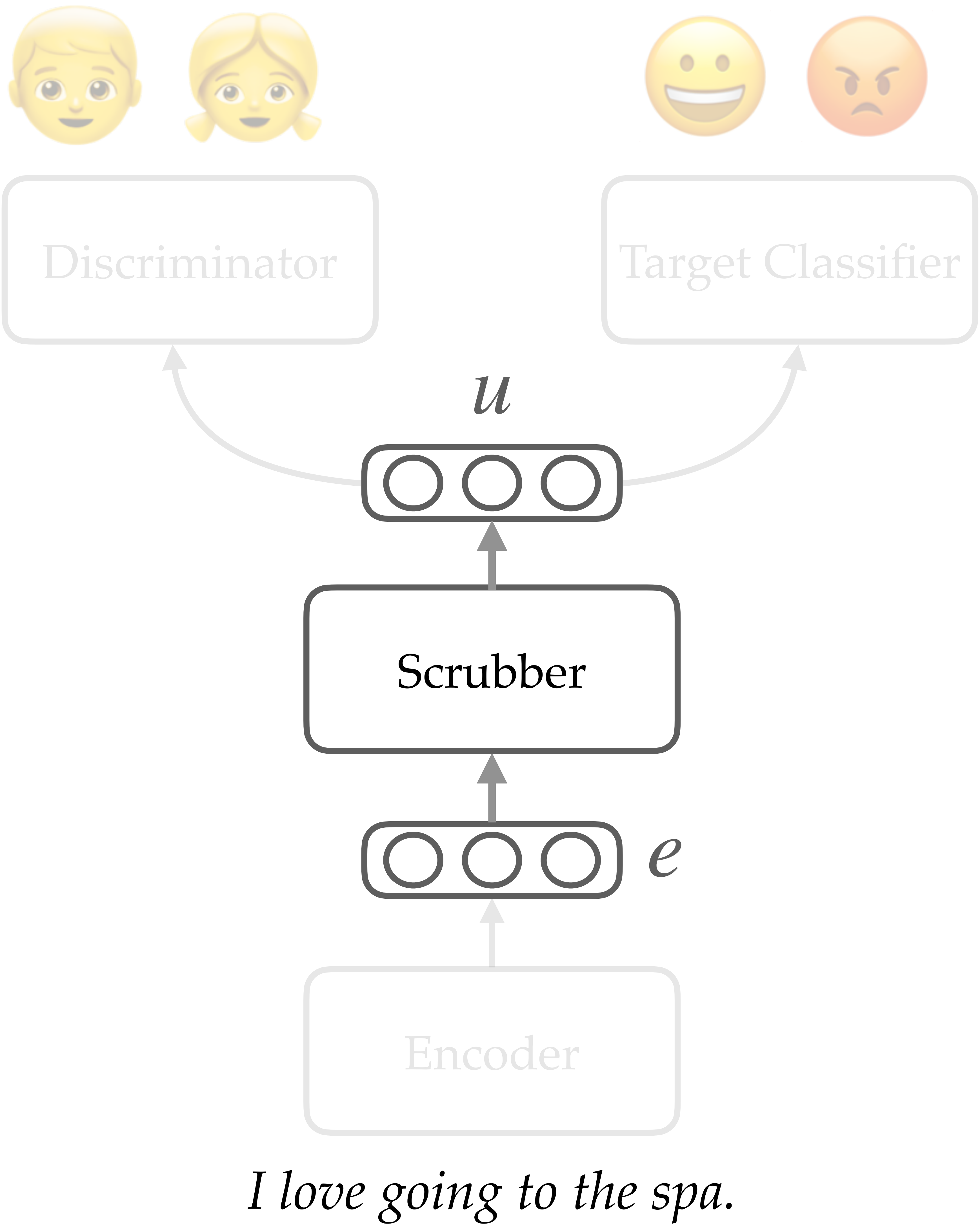
Scrubber learns fair representations by leveraging:



# Scrubber

Scrubber learns fair representations by penalizing:

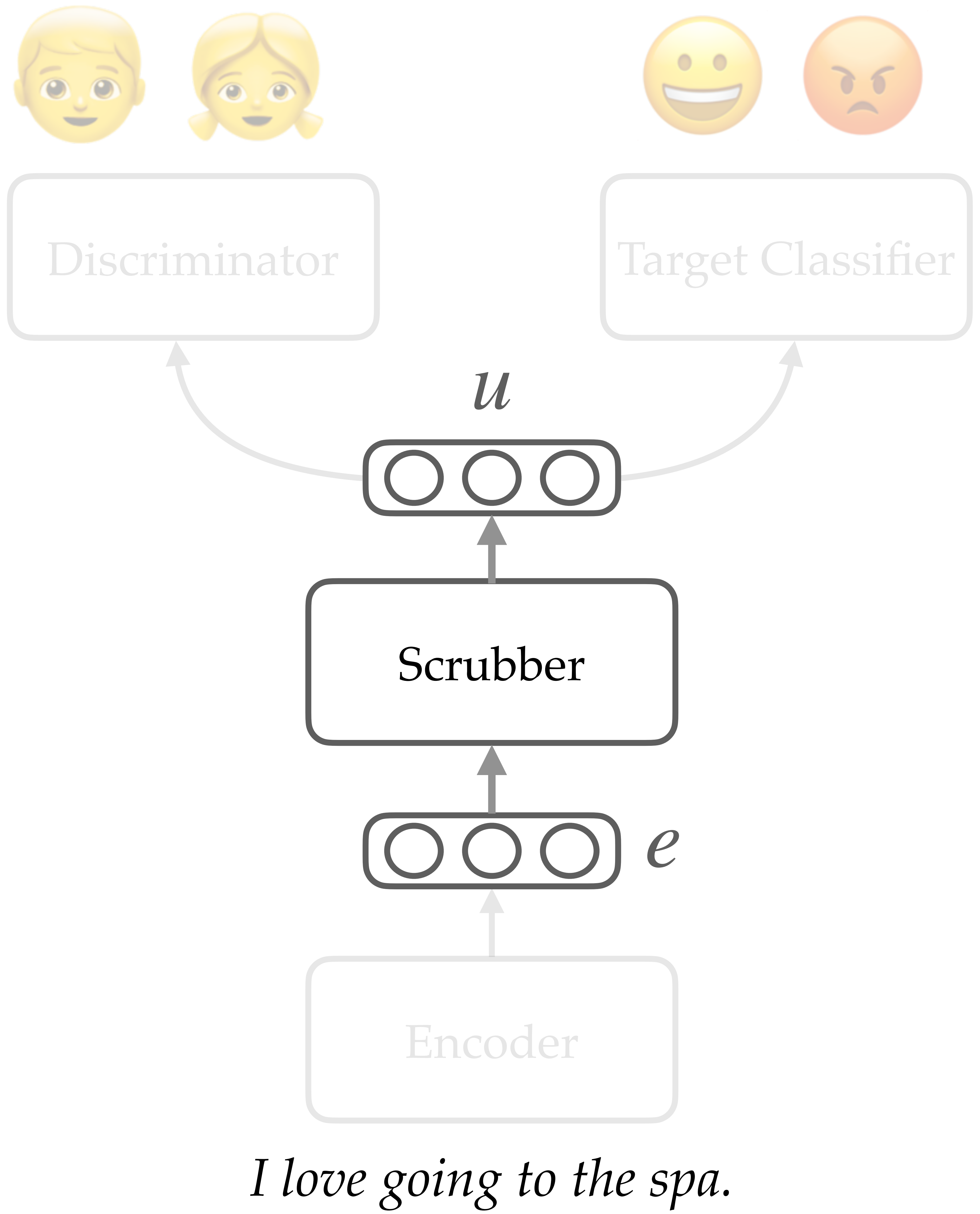
- Entropy of the discriminator output $d(u)$



# Scrubber

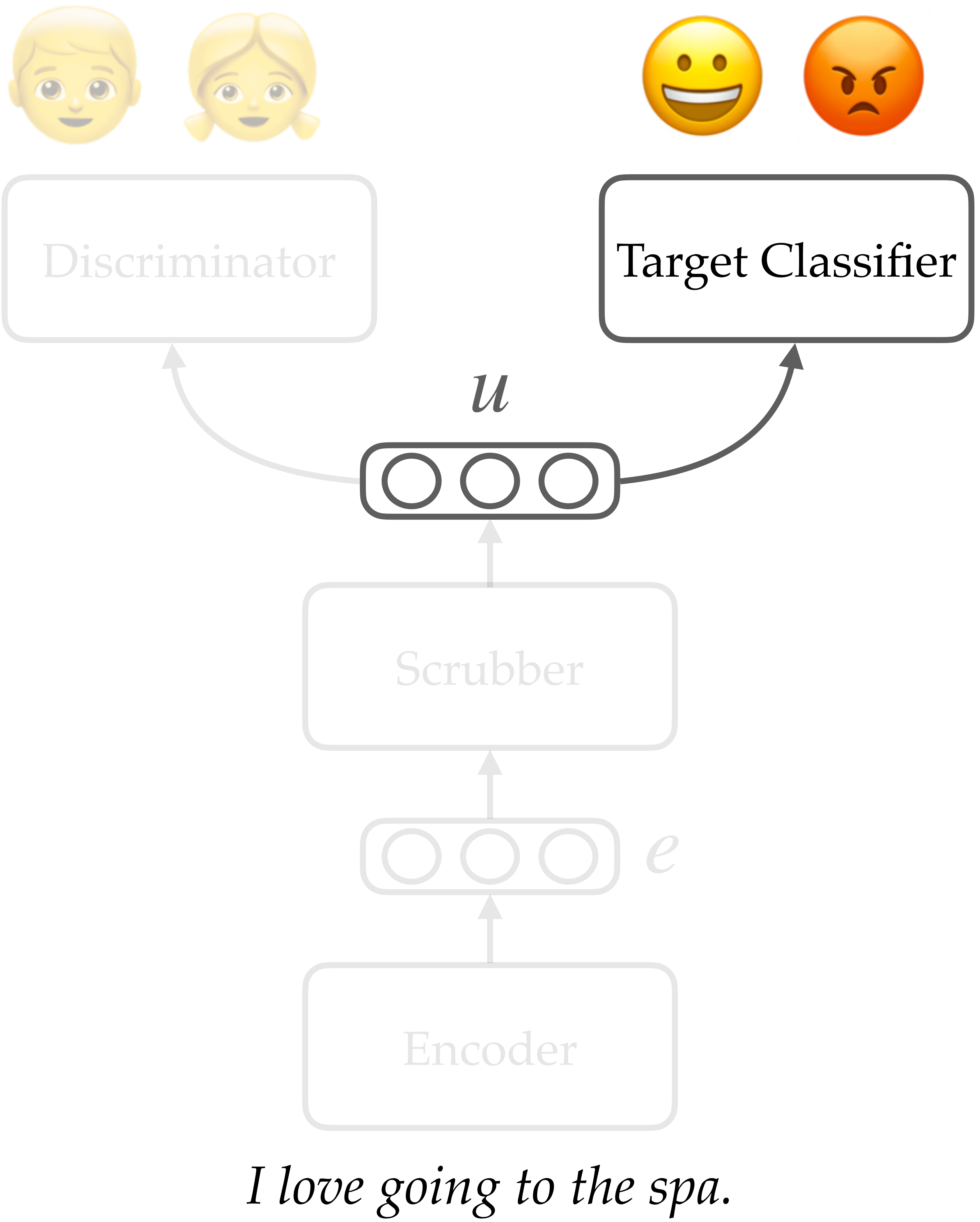
Scrubber learns fair representations by penalizing:

- Entropy of the discriminator output $d(u)$
- $\delta$-loss: penalises probability assigned to the correct target logit



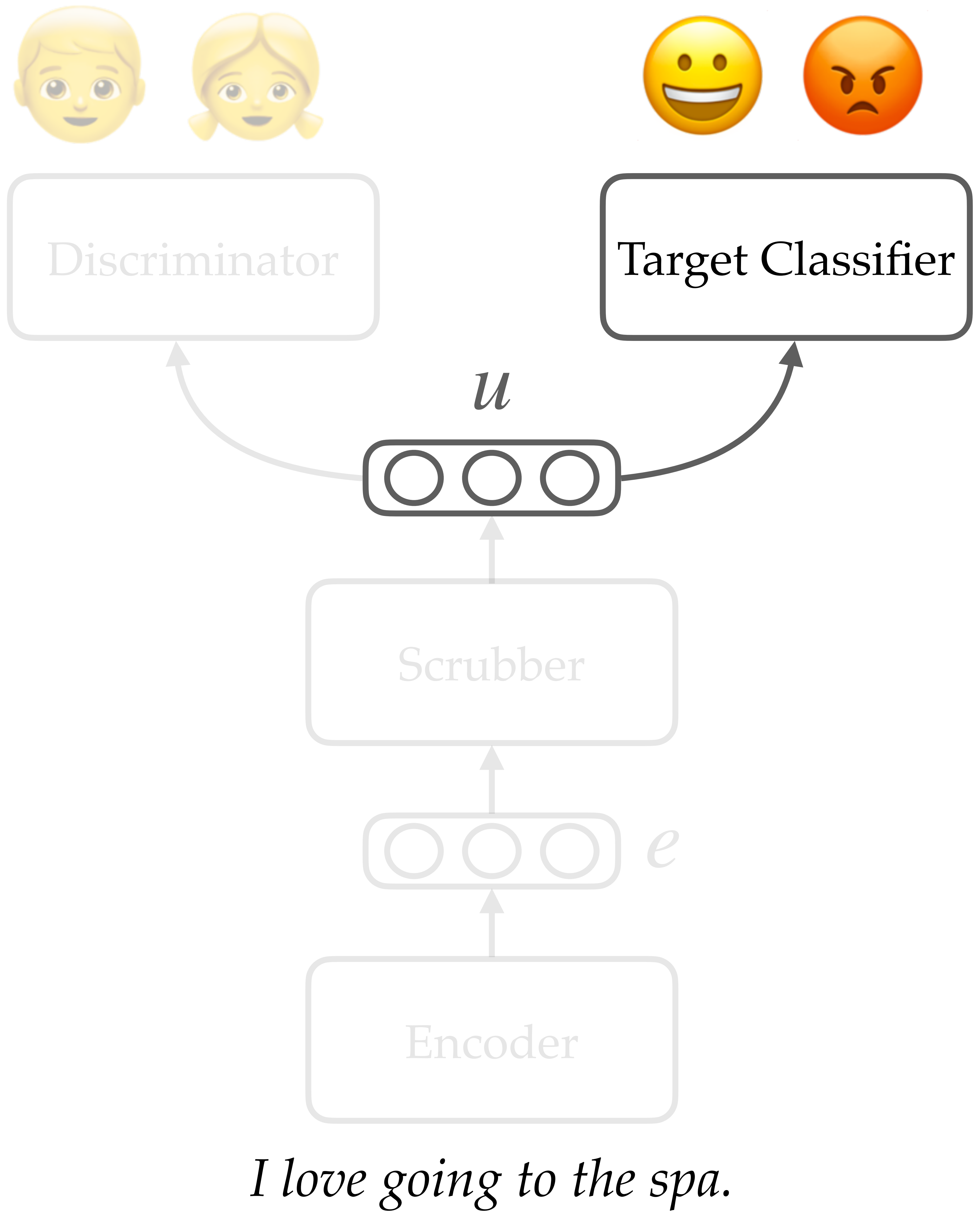
# Target Classifier

- Predicts target label from scrubbed representations ($u$)



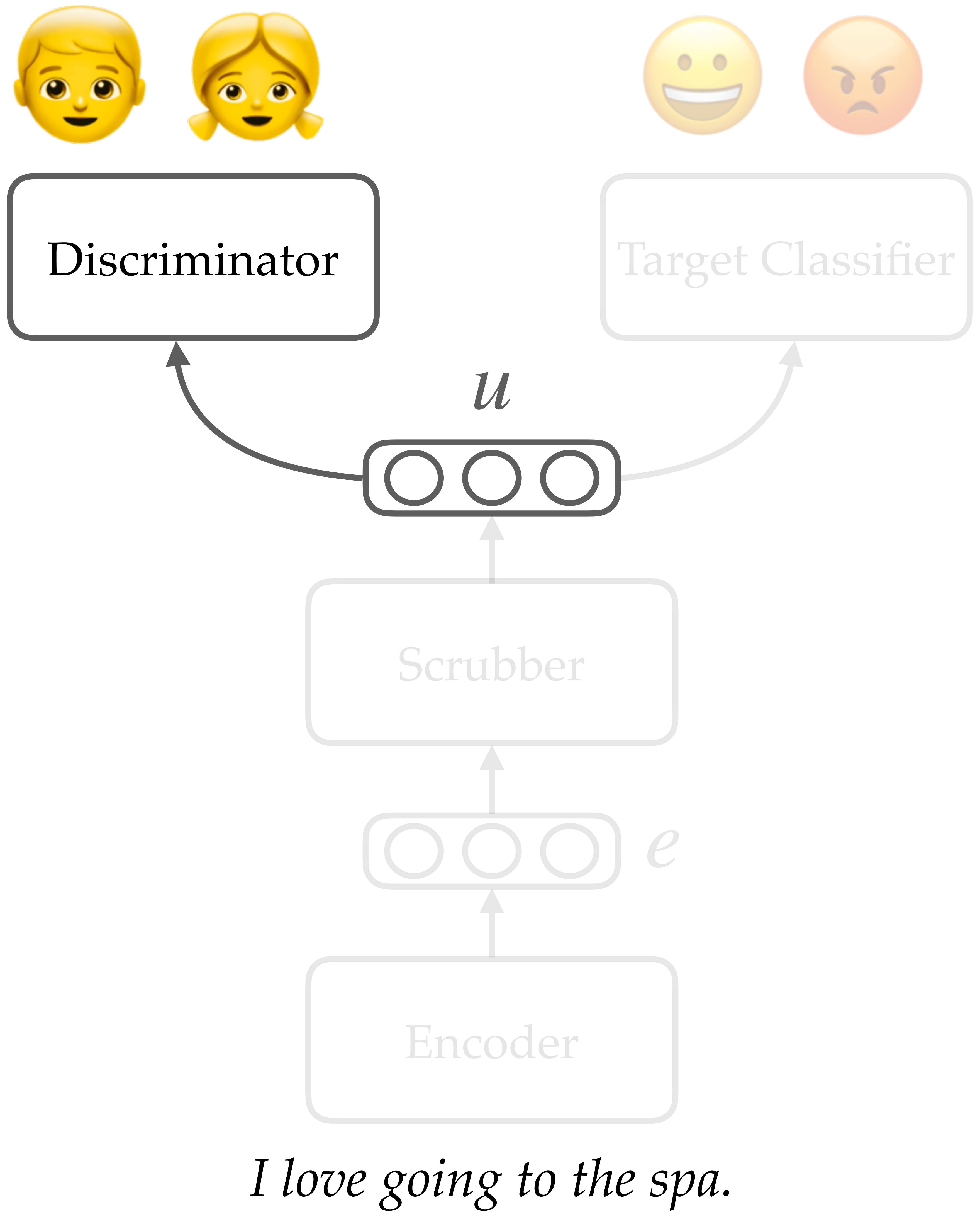
# Target Classifier

- Predicts target label from scrubbed representations ($u$)
- Optimizes cross-entropy loss



# Bias Discriminator

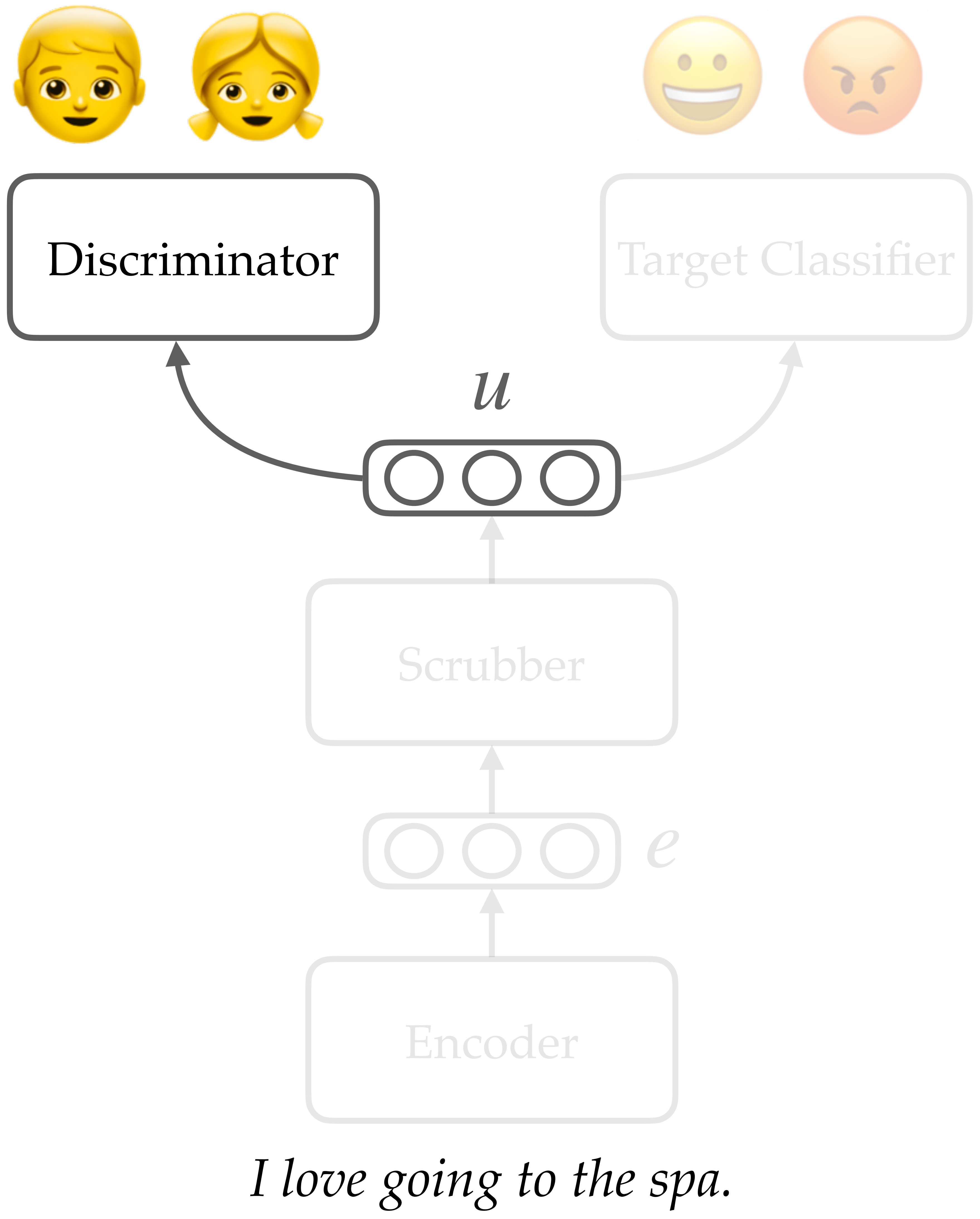
- Predicts protected variable from scrubbed representations ($u$)



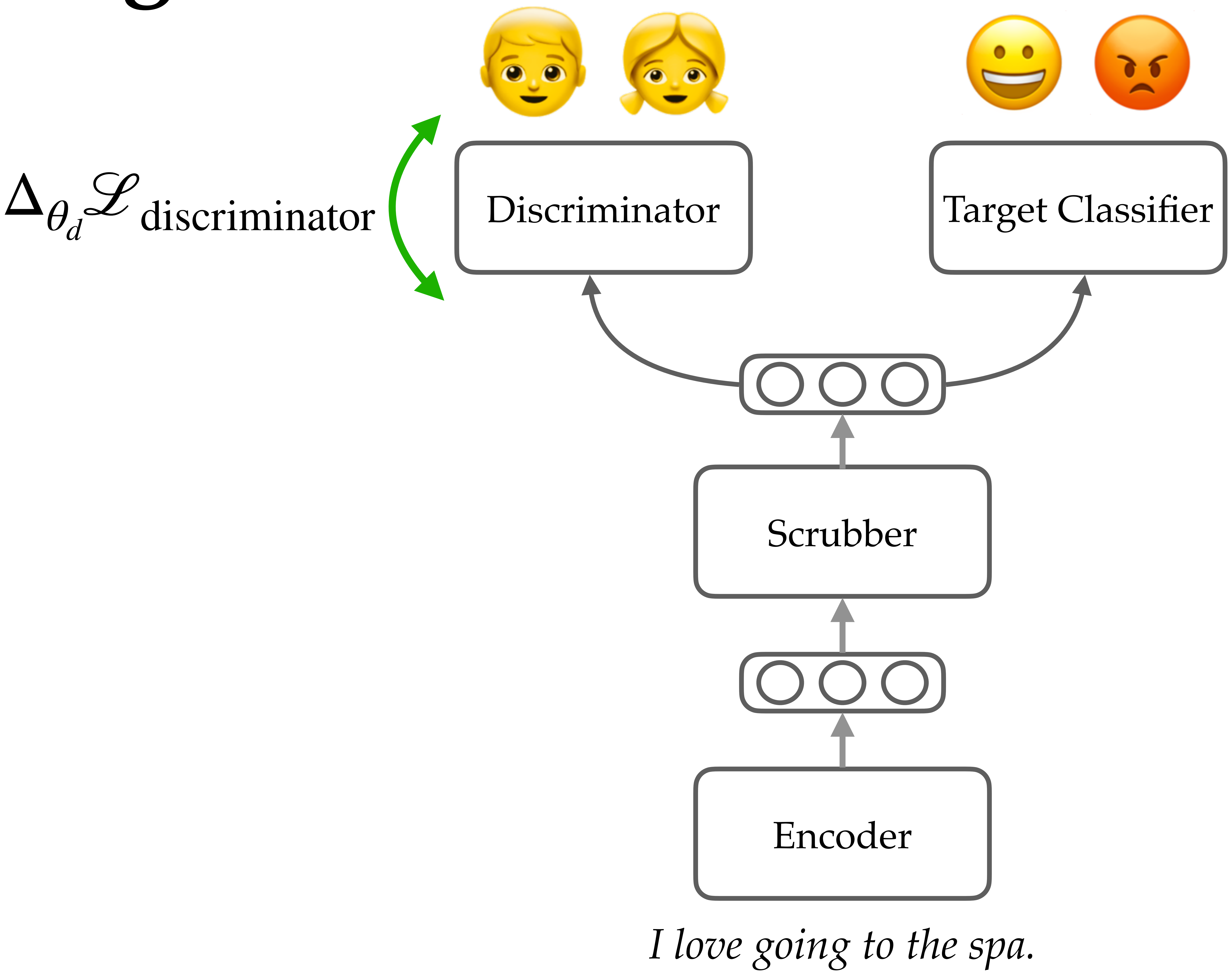
# Bias Discriminator

- Predicts protected variable from scrubbed representations ($u$)

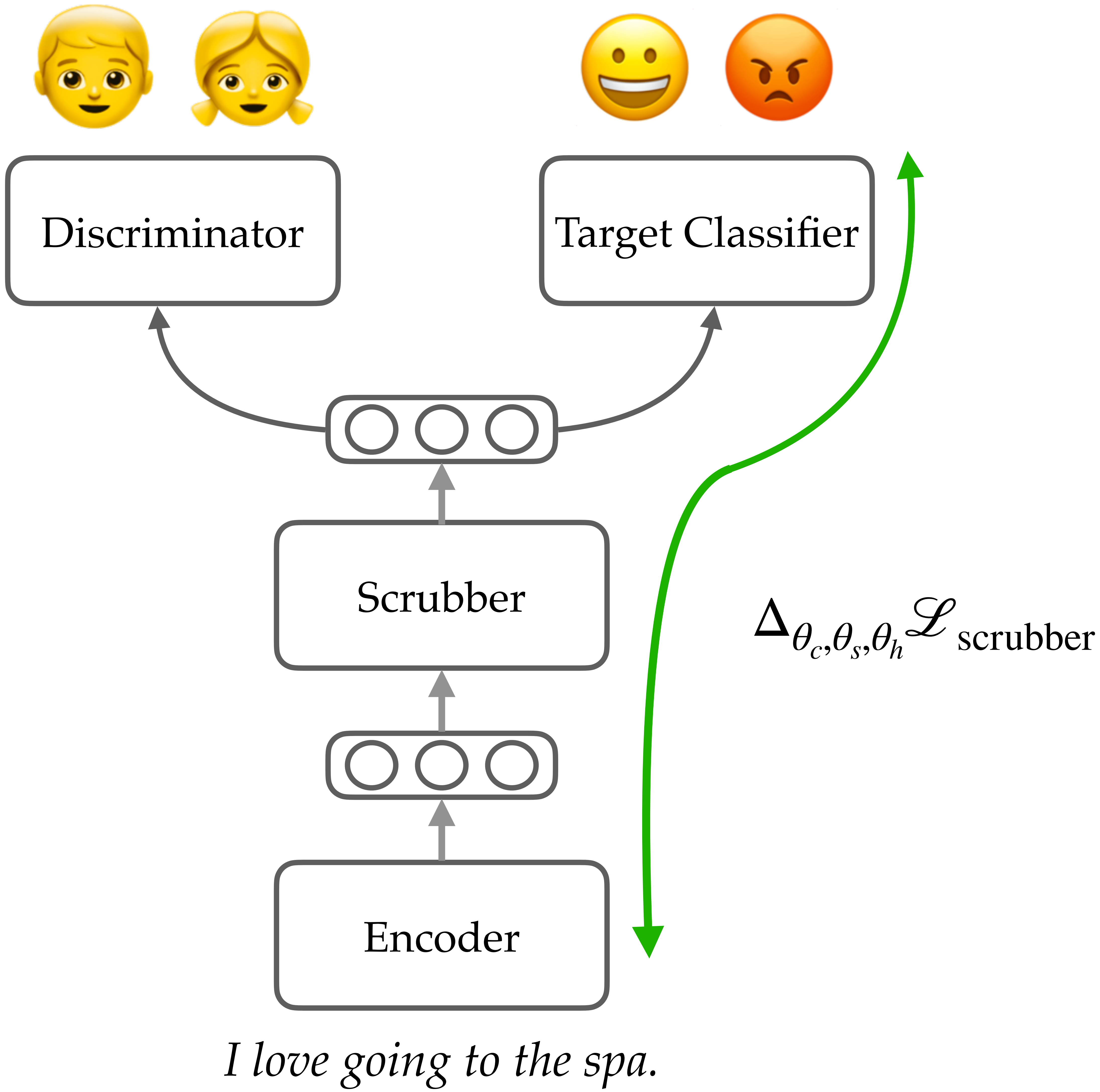
- Optimizes cross-entropy loss



# Training

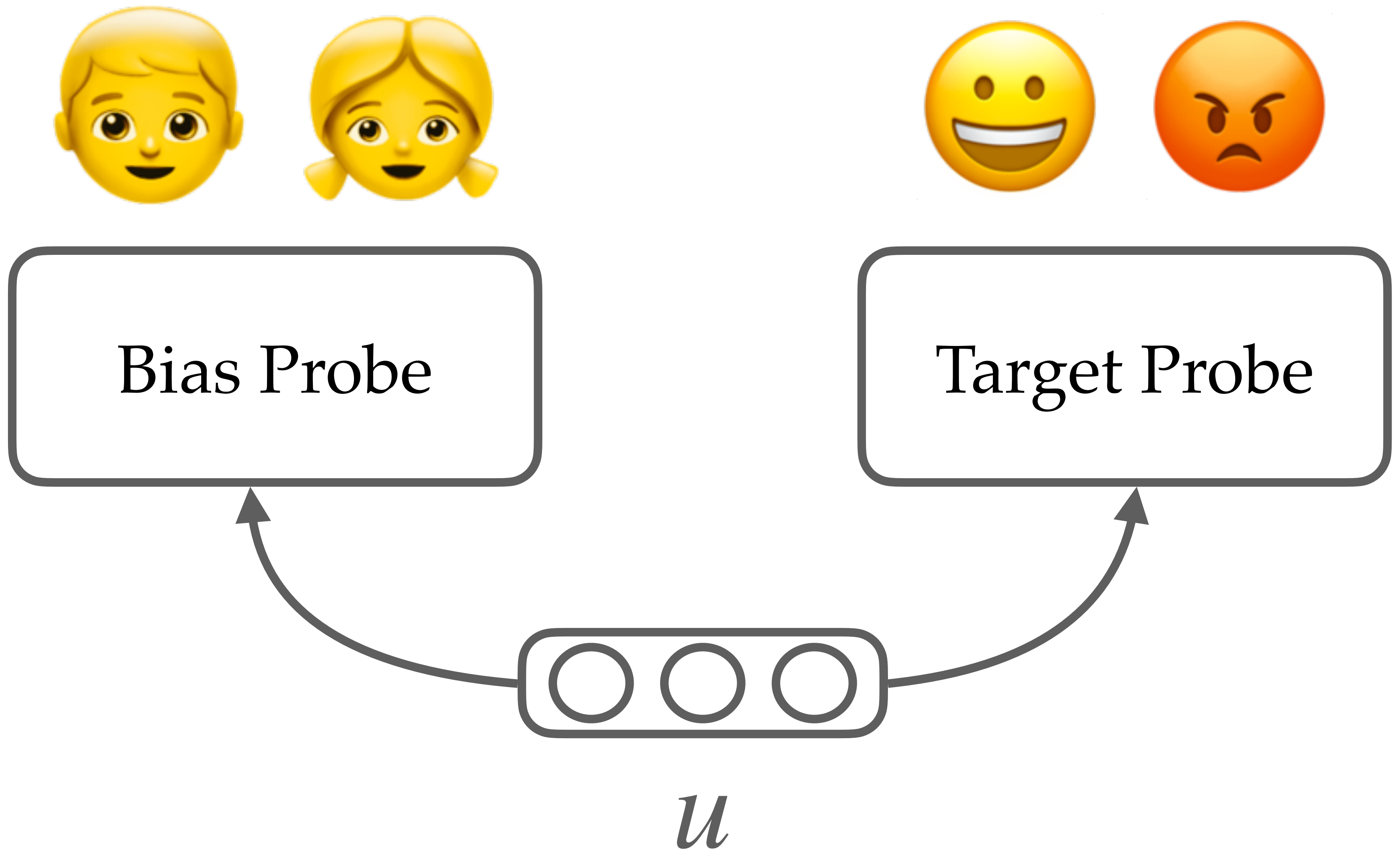

$\Delta_{\theta_d}\mathscr{L}_{\text{discriminator}}$
Discriminator
Target Classifier
Scrubber
Encoder
*I love going to the spa.*

# Training

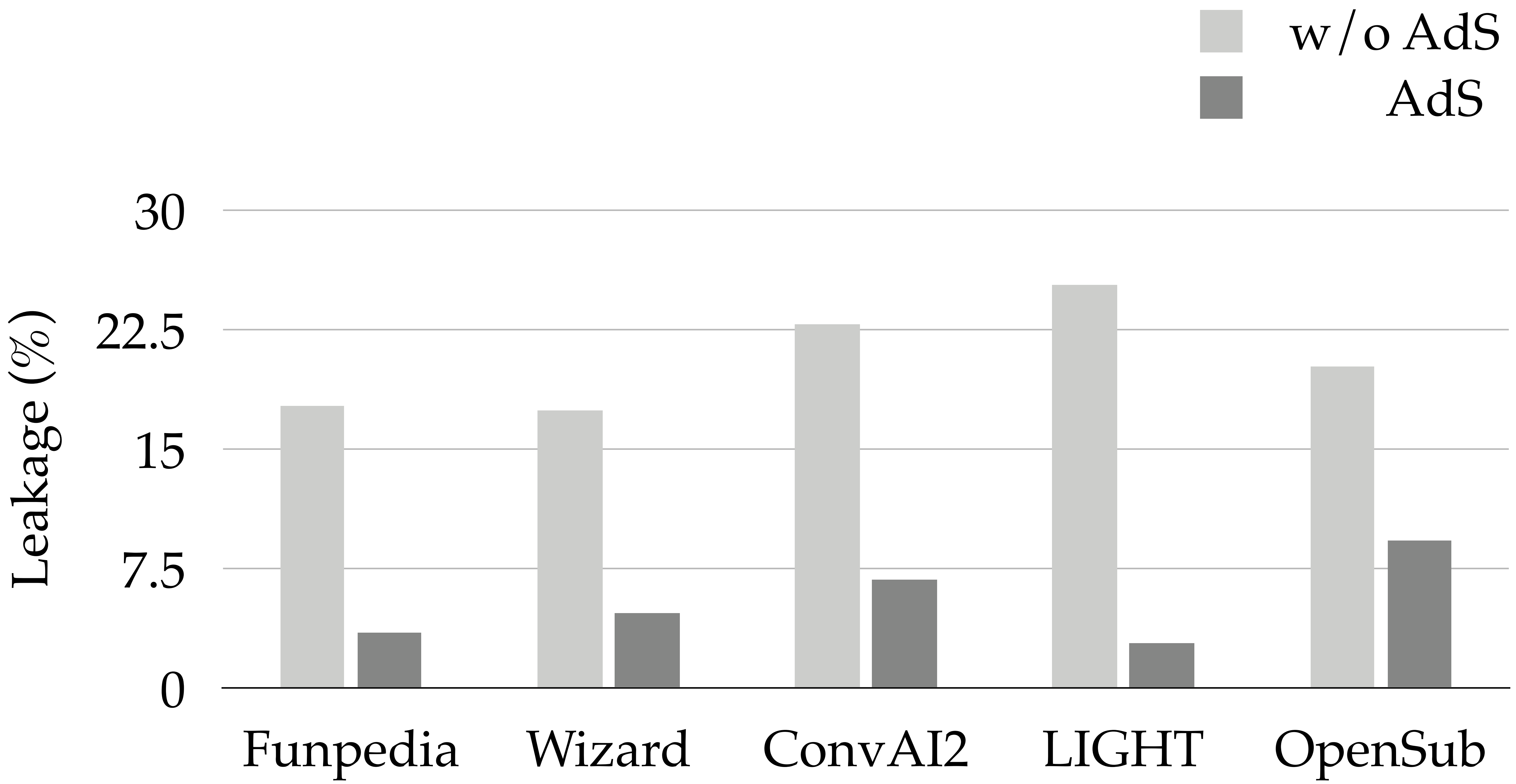

Discriminator
Target Classifier
Scrubber
Encoder
$\Delta_{\theta_c,\theta_s,\theta_h}\mathscr{L}_{\text{scrubber}}$
*I love going to the spa.*

# Probing

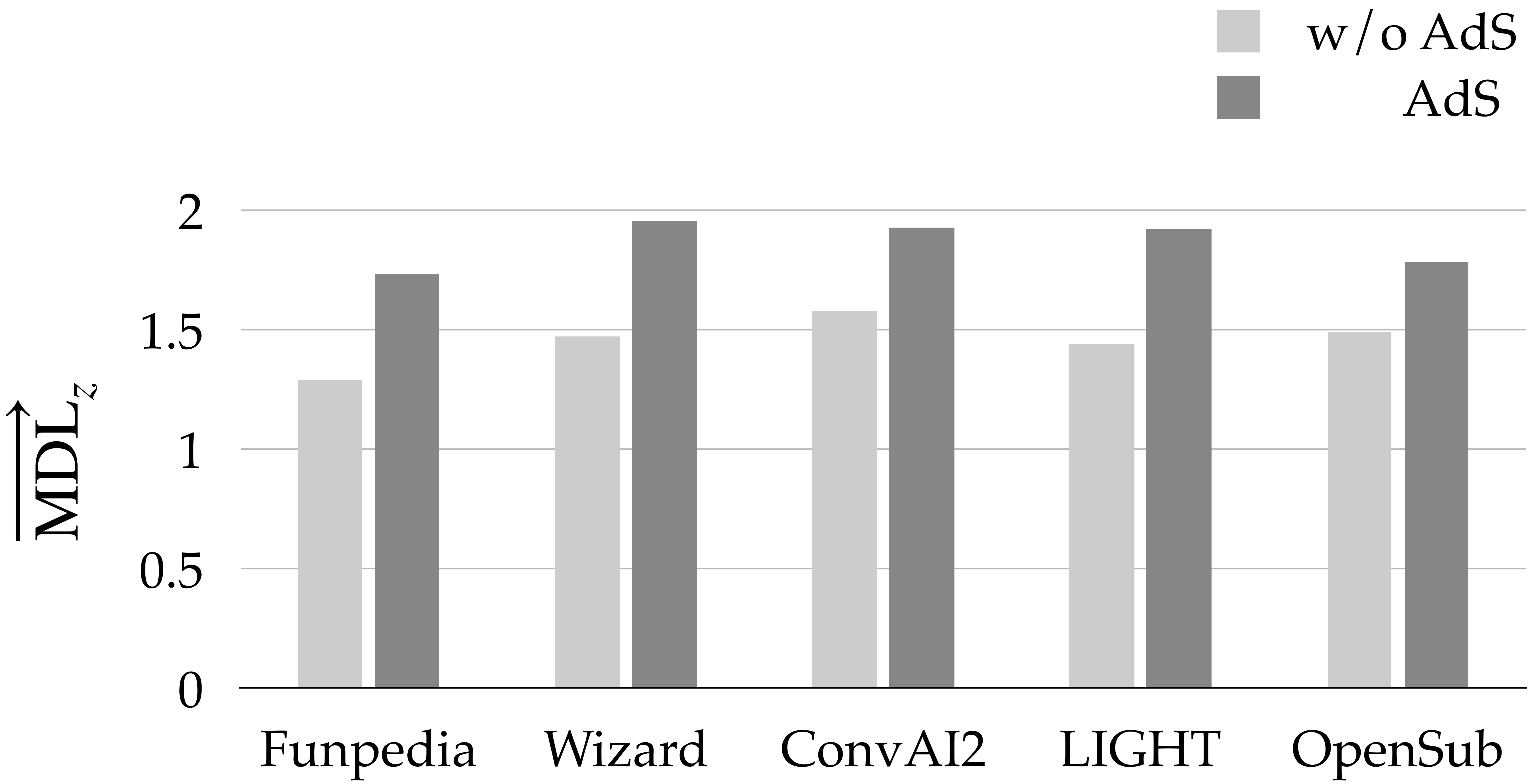
- 8 datasets from domains: *dialogue*, *tweet* and *biography* classification
- 1-layer MLP Classifier
- Metrics evaluated:
  - Accuracy
  - Minimum Description Length (MDL)
  - Target task accuracy



# Results - Dialogue datasets

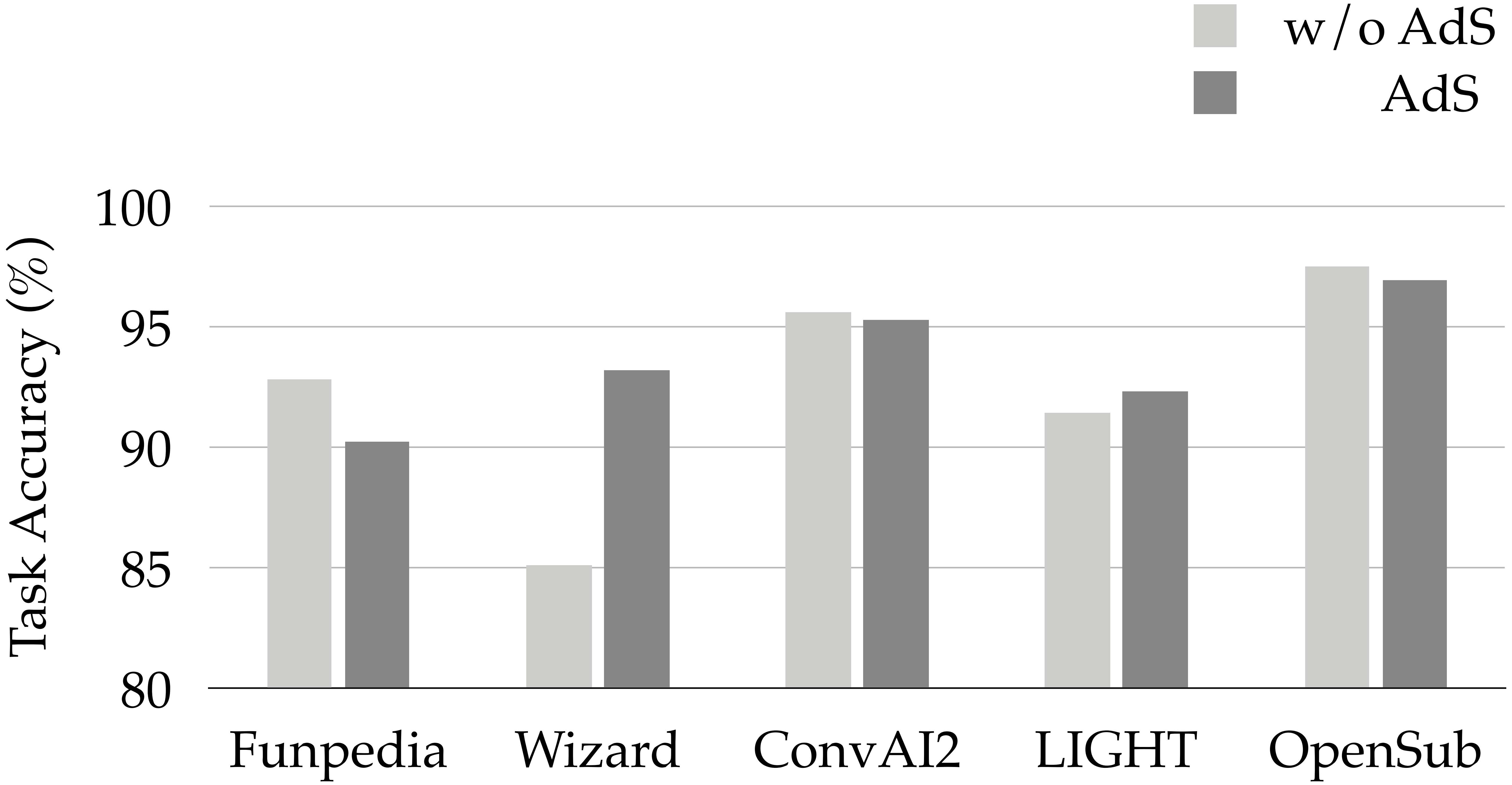


# Results - Dialogue datasets

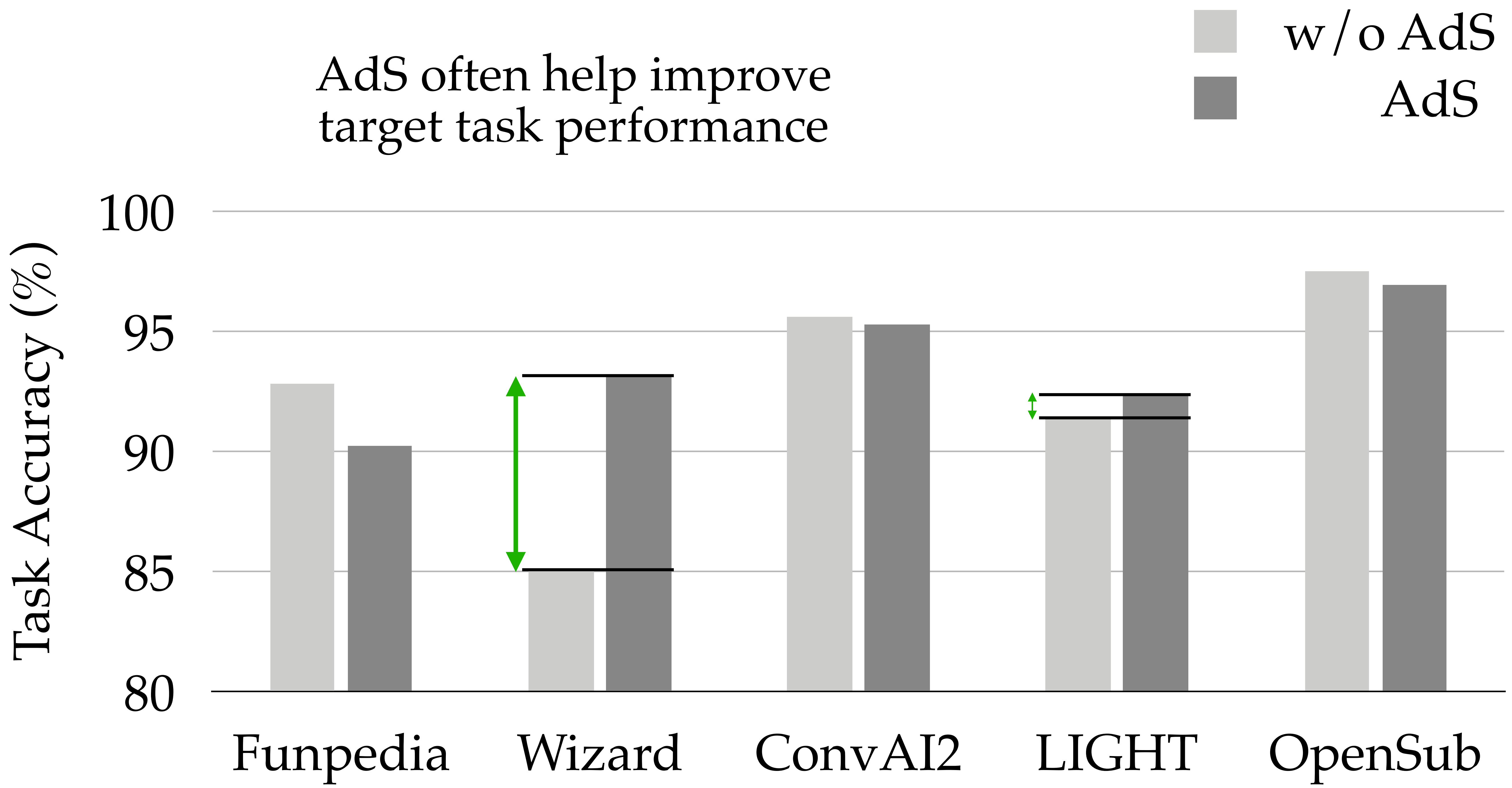


Higher $\overrightarrow{\mathrm{MDL}}_z$ signifies increased difficulty in extracting the protected variable.

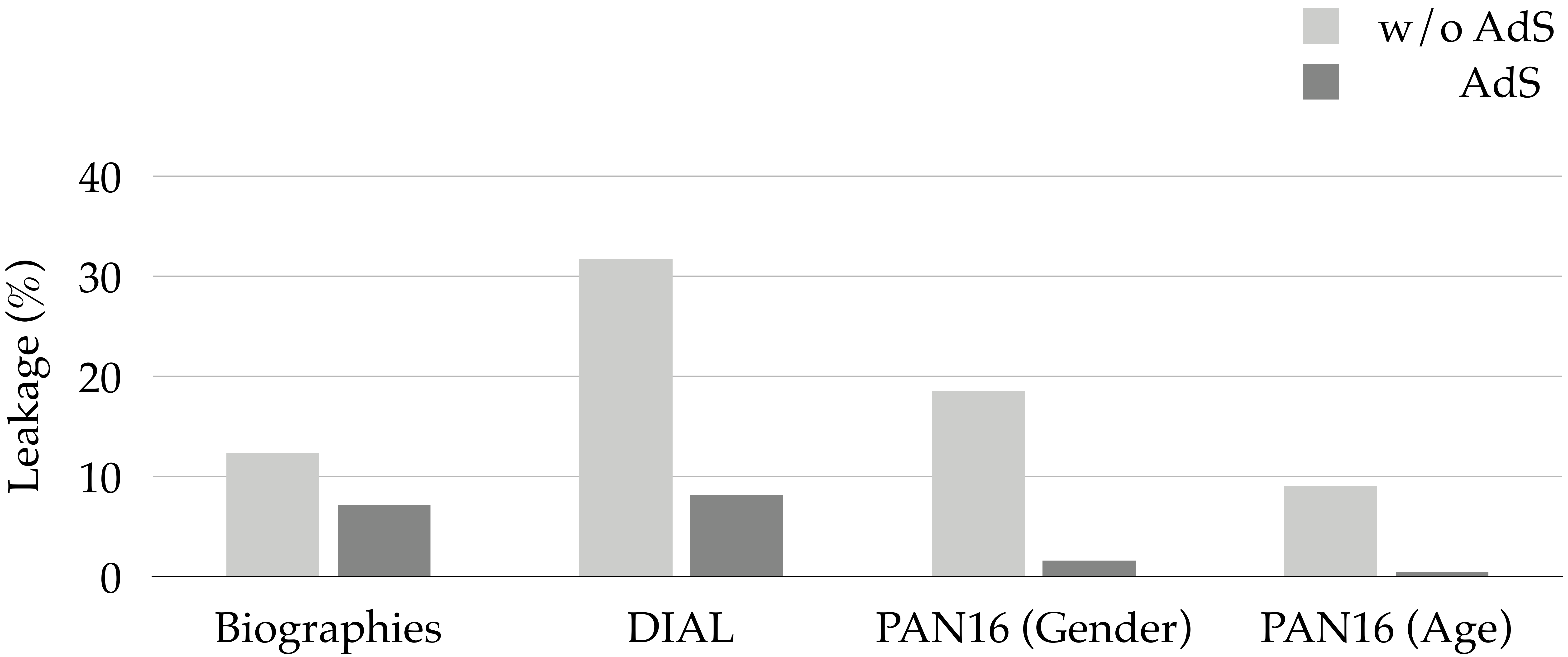
# Results - Dialogue datasets



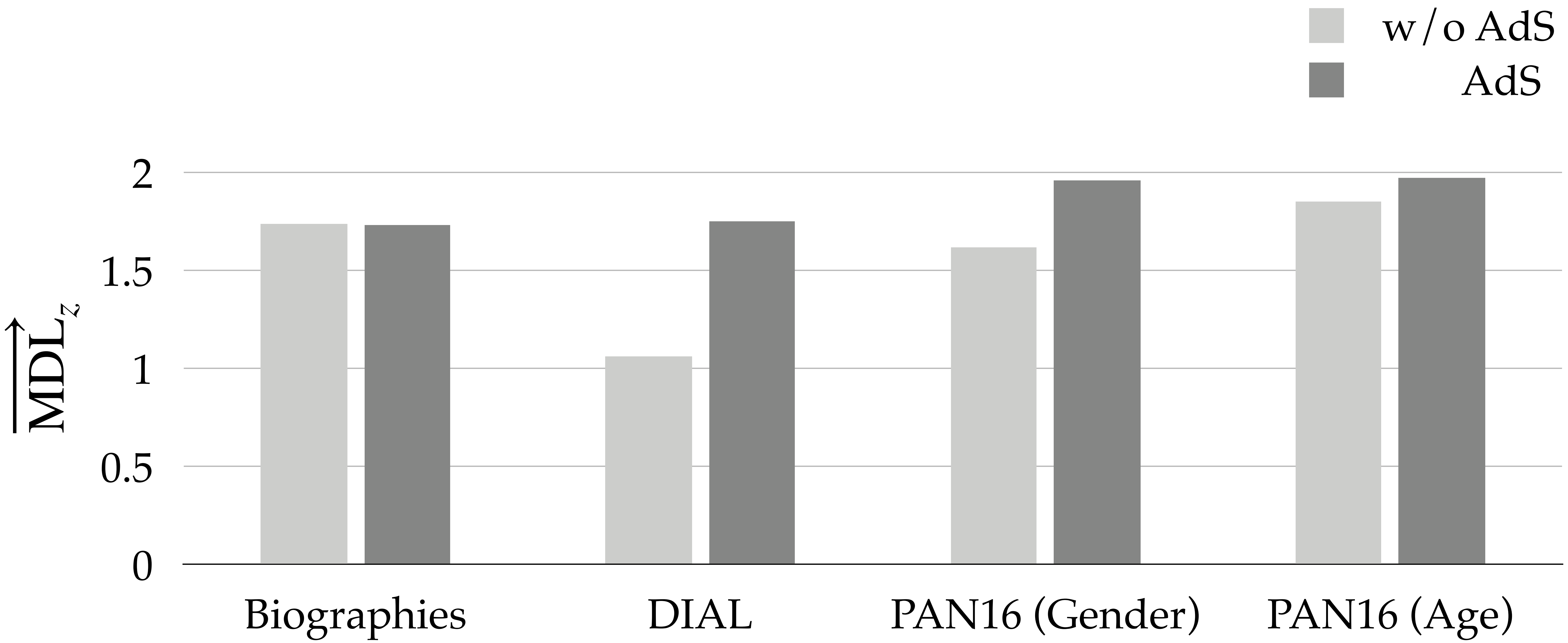
# Results - Dialogue datasets


AdS often help improve target task performance

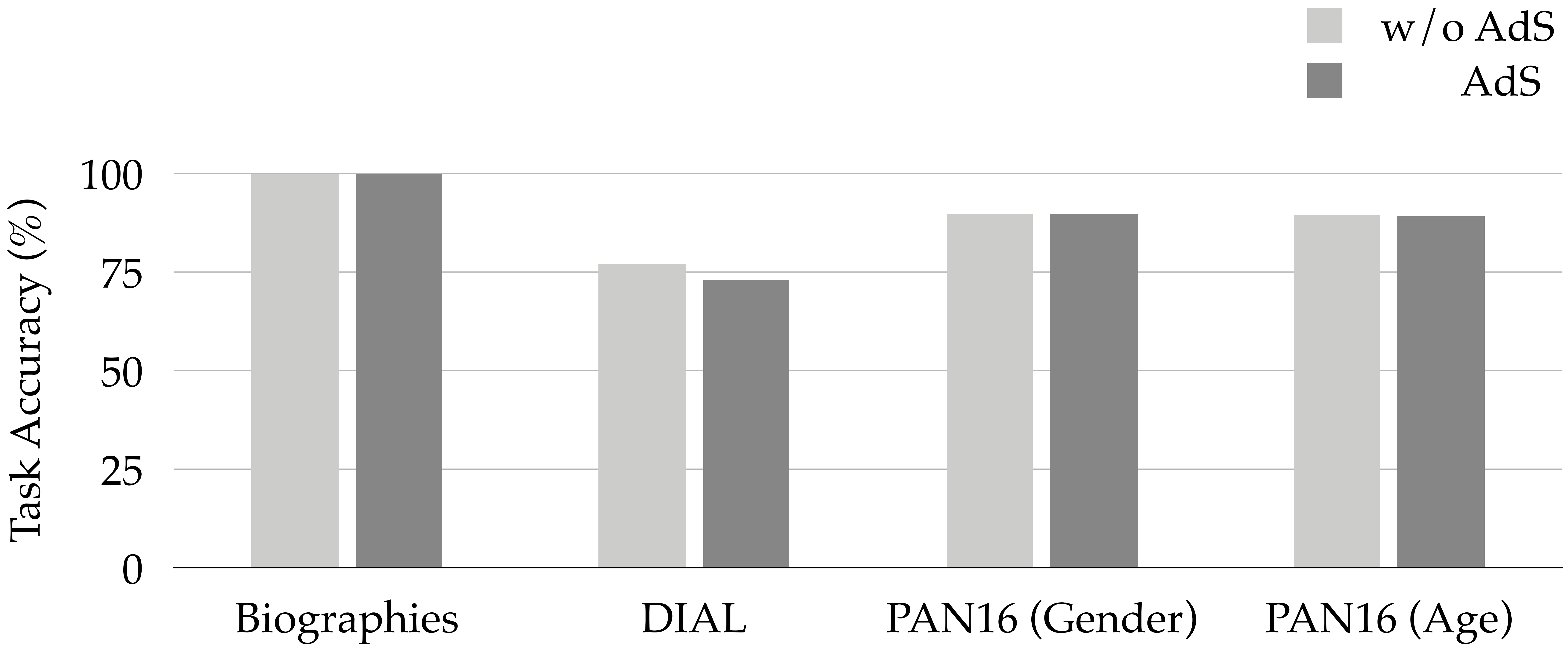
# Results - Tweet/biography classification



# Results - Tweet/biography classification



# Results - Tweet/biography classification



# Takeaways

★ Text classification systems can encode demographic information even without having direct access to them

# Takeaways

- ★ Text classification systems can encode demographic information even without having direct access to them
- ★ To ensure fairness, intermediate representations should have *zero leakage* about demographic attributes

# Takeaways

★ Text classification systems can encode demographic information even without having direct access to them

★ To ensure fairness, intermediate representations should have *zero leakage* about demographic attributes

★ Empirical evaluation on 8 datasets show AdS is able to prevent leakage while maintaining target task performance

# Takeaways

★ Text classification systems can encode demographic information even without having direct access to them

★ To ensure fairness, intermediate representations should have *zero leakage* about demographic attributes

★ Empirical evaluation on 8 datasets show AdS is able to prevent leakage while maintaining target task performance

★ We are far away from achieving complete fairness in AI, therefore such systems should be used in the real world *with caution*

# Takeaways

★ Text classification systems can encode demographic information even without having direct access to them

★ To ensure fairness, intermediate representations should have *zero leakage* about demographic attributes

★ Empirical evaluation on 8 datasets show AdS is able to prevent leakage while maintaining target task performance

★ We are far away from achieving complete fairness in AI, therefore such systems should be used in the real world *with caution*

@SomnathBrc

brcsomnath/adversarial-scrubber

somnath@cs.unc.edu