# Deep Learning on a Healthy Data Diet

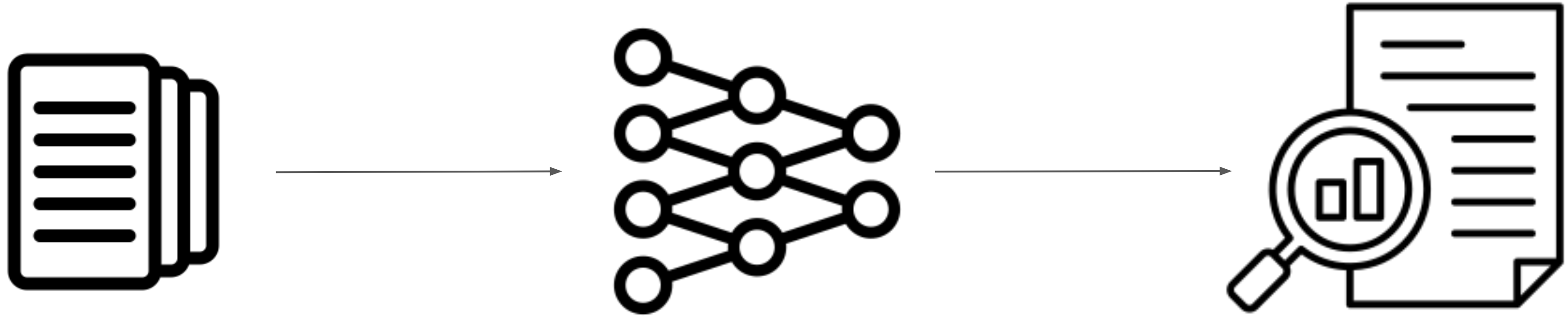
**Abdelrahman Zayed**, Prasanna Parthasarathi, Goncalo Mordido, Hamid Palangi,
Samira Shabanian, Sarath Chandar

AAAI, 2023

# What is bias?

| Abdel | Khaoula |
|---|---|
| Ph.D. from Mila<br>6 years of research experience | Ph.D. from Mila<br>6 years of research experience |
| Keep resume | Discard resume |

# How do we mitigate bias?



Pre-processing

In-processing

Post-processing

# Adding counterfactual examples

*“**He** is a doctor”* → *“**She** is a doctor”*

Including ***both*** in training data = Counterfactual Data augmentation (CDA)

Including ***one*** in training data = Counterfactual Data substitution (CDS)

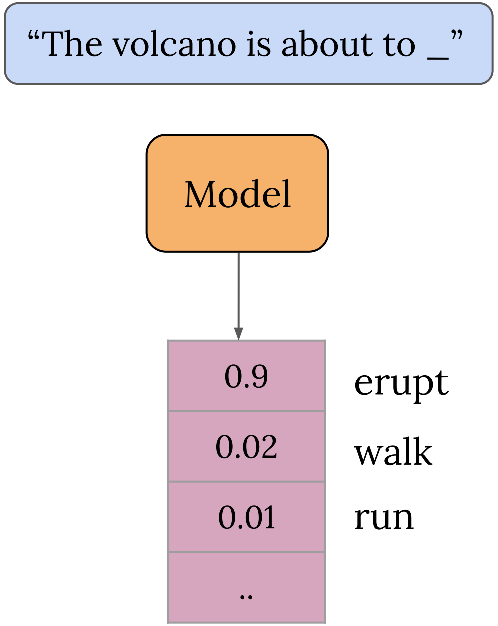
Zhao et al., EMNLP, 2018

Maudslay et al., ACL, 2020

# Can we outperform CDA and CDS?

# When do models become ***biased***?

## Pre-training

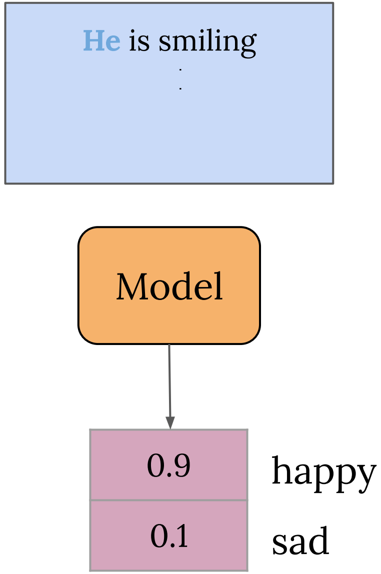


“The volcano is about to _”

Model

| | |
|---|---|
| 0.9 | erupt |
| 0.02 | walk |
| 0.01 | run |
| .. | |

## Fine-tuning



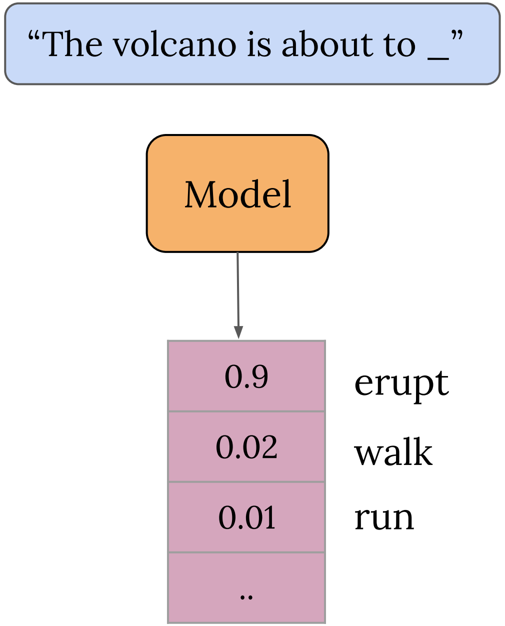
**He** is smiling

Model

| | |
|---|---|
| 0.9 | happy |
| 0.1 | sad |

# How does data augmentation help?

## Pre-training

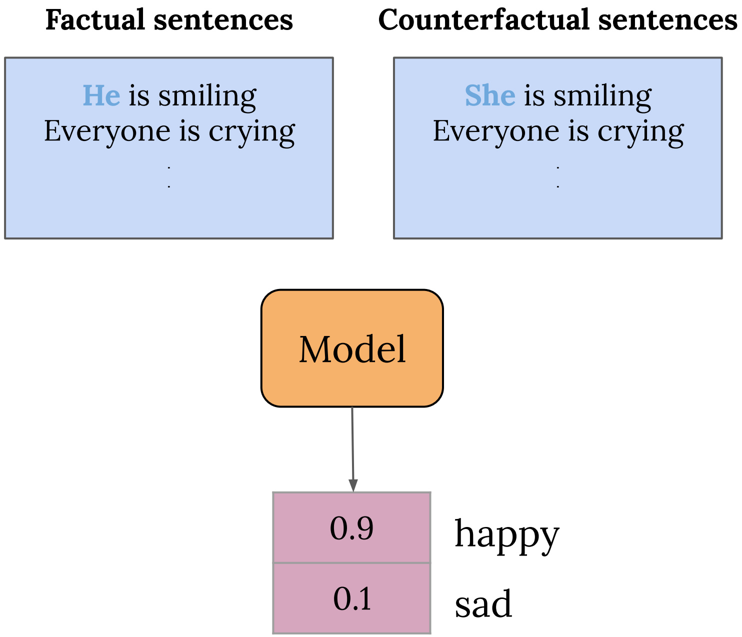


“The volcano is about to _”

Model

| | |
|---|---|
| 0.9 | erupt |
| 0.02 | walk |
| 0.01 | run |
| .. | |

## Fine-tuning



**Factual sentences**

He is smiling
Everyone is crying
.
.

**Counterfactual sentences**

She is smiling
Everyone is crying
.
.

Model

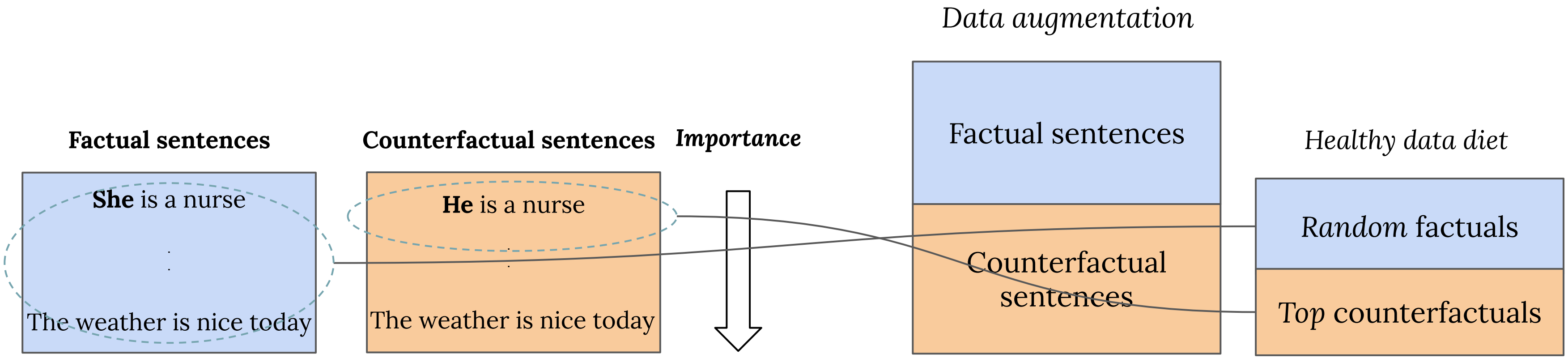
| | |
|---|---|
| 0.9 | happy |
| 0.1 | sad |

# Two observations!

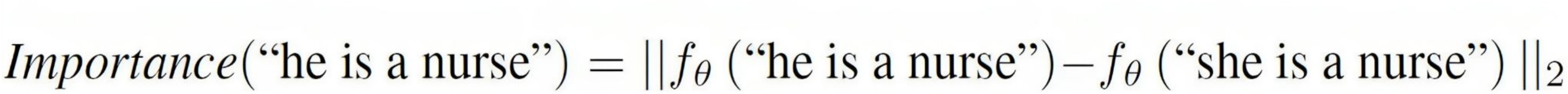
1- We only need the ***important*** counterfactuals

2- Removing some factuals also ***helps***

# Healthy data diet

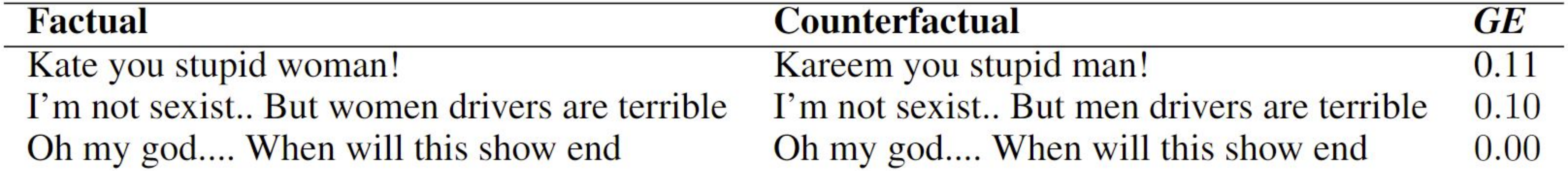


# How to find ***important*** counterfactuals?



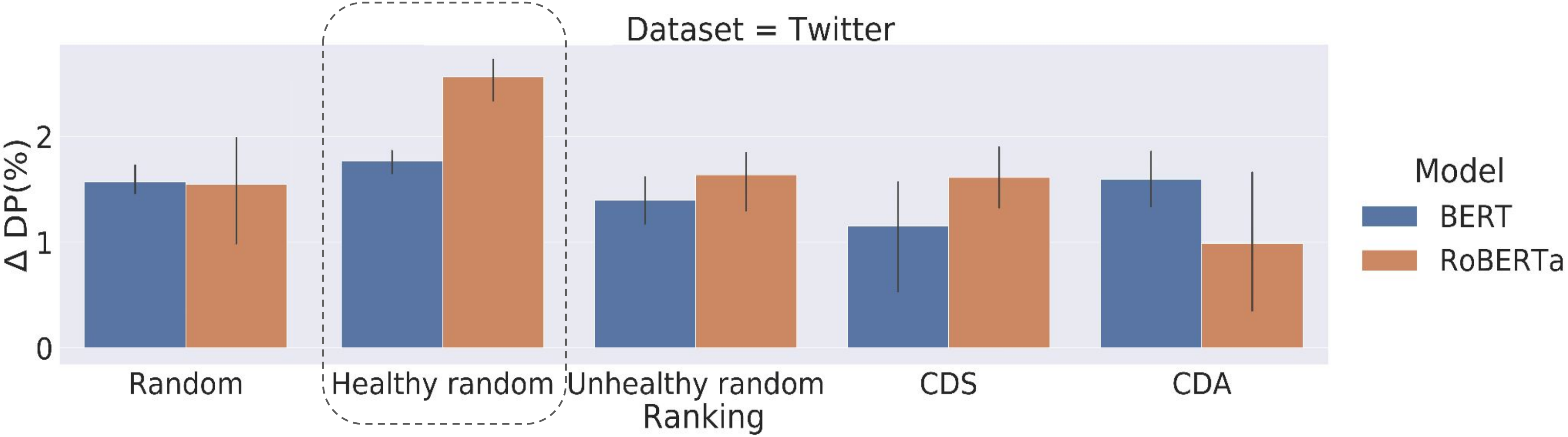
$$Importance(\text{“he is a nurse”}) = ||f_\theta\,(\text{“he is a nurse”}) - f_\theta\,(\text{“she is a nurse”})\,||_2$$

# How does the importance score look like?



| Factual | Counterfactual | *GE* |
|---|---|---|
| Kate you stupid woman! | Kareem you stupid man! | 0.11 |
| I’m not sexist.. But women drivers are terrible | I’m not sexist.. But men drivers are terrible | 0.10 |
| Oh my god.... When will this show end | Oh my god.... When will this show end | 0.00 |

# Healthy data diet is ***better*** than CDA and CDS!



# To summarize

Our recipe adds only **important** counterfactuals and removes **harmful** factuals