The Alan Turing Institute

Queen Mary University of London

Learning at the Turing

“Bias” in Natural Language processing

Peiling Yi

About Me

- **Experience**

  Over 10 years in industry as software engineer and project manager
  Master in Software engineering at QMUL

  NOW-
  A part-time **PhD student** at cognitive science research group, QMUL
  A part-time **teaching fellow** at QMUL
  A **Turing Enrichment student**
  A Candidate of indoor climbing instructor

- **Research**

  Doing research in **Youth cyberbullying detection across different social medias.**
  Also interested in Novel deep **transfer learning** and **fair machine learning algorithms**.

Contents

Warning: This presentation contains examples of swear words!

- **What is the bias?**
- What is the bias in NLP?
- The sources of bias and Debias
- Debias
- Challenges
- My research (if time permits)

Bias

- **“In general, bias refers to a tendency or preference for or against something or someone.” - Cambridge Dictionary**

It can manifest in various forms, including personal beliefs, opinions, attitudes, or prejudices that influence how a person thinks, acts, or makes decisions. In the context of information or media, bias refers to a skewed presentation of facts, ideas, or perspectives that may favour a particular point of view or ideology.

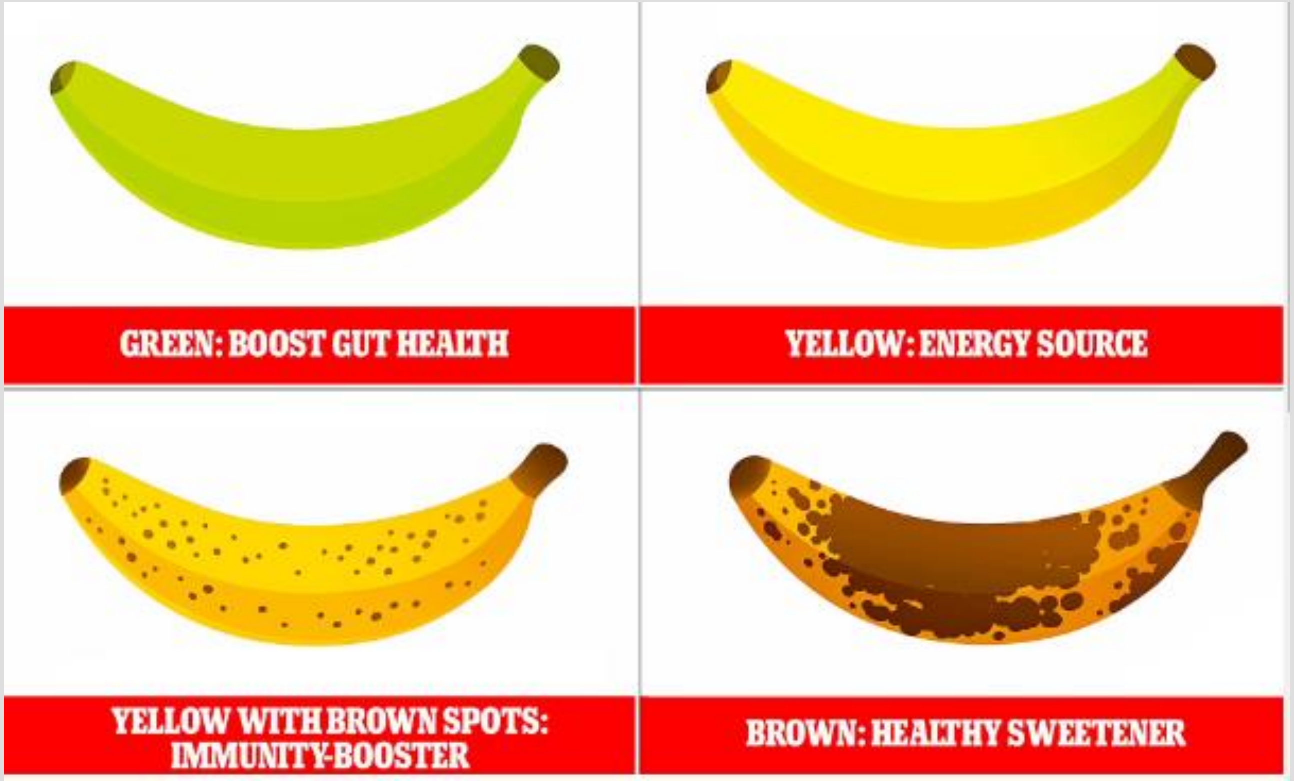
Especially in a way considered to be unfair.

- “If you are human , you are biased”-Howard J. Ross

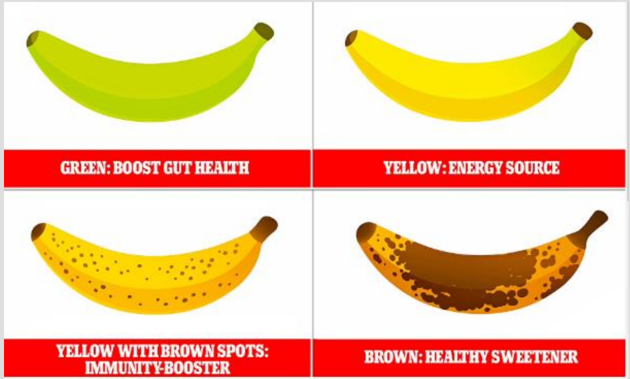
Psychological scientists define bias: Bias is just a **tendency** to respond one way **compared** to another when making some kind of a life choice.” Sometimes these biases can be completely neutral, like a bias for Coke over Pepsi, and can even be helpful in allowing you to **make decisions more rapidly.** - **Rose Miyatsu**

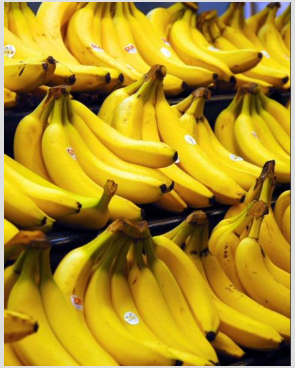
Good bananas-Bias

Good bananas-*Fairness*



Good bananas-***stereotyping-->quick decision***





What will you pick up?

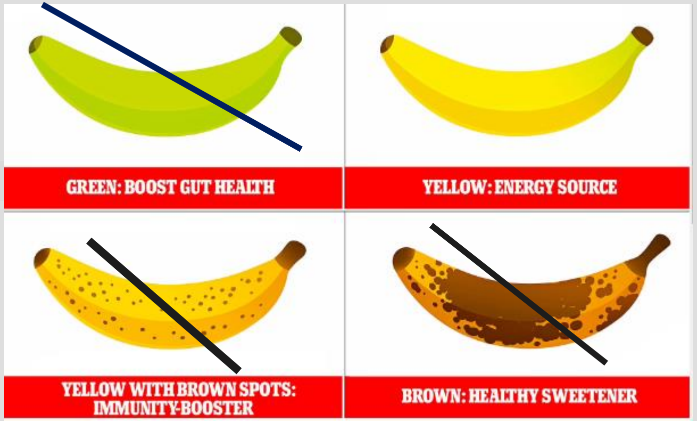
Good bananas-*Amplify bias*

Less ambiguous

A ripe banana with **a bright yellow color**. The peel should be mostly free from any green hues, indicating that it's not underripe. There might be **a few small brown spots** on the peel, which are natural and signify that the banana is at its peak ripeness. The banana should have a smooth texture and **be free from any significant bruises, cuts, or damage on the surface**.

Ambiguous

Ambiguous



Bias

- **“In general, bias refers to a tendency or preference for or against something or someone.” - Cambridge Dictionary**

It can manifest in various forms, including personal beliefs, opinions, attitudes, or prejudices that influence how a person thinks, acts, or makes decisions. In the context of information or media, bias refers to a skewed presentation of facts, ideas, or perspectives that may favor a particular point of view or ideology.

Especially in a way considered to be unfair.

- “If you are human , you are biased”-Howard J. Ross

Psychological scientists define bias: Bias is just a **tendency** to respond one way **compared** to another when making some kind of a life choice.” Sometimes these biases can be completely neutral, like a bias for Coke over Pepsi, and can even be helpful in allowing you to **make decisions more rapidly.** - **Rose Miyatsu**

Contents

Warning: This presentation contains examples of swear words!

- What is the bias?
- **What is the bias in NLP?**
- The sources of bias and debias
- Challenges

Natural language processing

Natural Language Processing is a field of artificial intelligence (AI) that focuses on the interaction **between computers and human language**. started in the **1950s.**

Nowadays, it is not a toy with small dataset training. It started to impact our life.
--- Bias everywhere

- Handle negative comments before they escalate-**Swearing bais, Racial bias**
- Speech Recognition**-Gender bias**
- Aanalysing customer interactions like emails, chats, or social media posts**-Gender bias**
- It summarizes text, by extracting the most important information**-Social bias**
- Machine translation-**Gender bias**
- Chatbots**-Social bias, Gender bias**
- And……

Bias in Natural language processing

Bias refers to the presence of **unfair** and **discriminatory** behaviour or attitudes encoded in the language data, models, or algorithms used in NLP applications

- **Gender bias:** models might associate certain professions or roles more strongly with a **specific gender**, leading to biased outputs.
- **Racial bias:** models may show biased behaviour towards **certain racial or ethnic** groups, perpetuating stereotypes or discrimination.
- **Social bias:** models may reflect **societal biases and prejudices**, affecting the fairness and inclusivity of the generated content.
- **Swearing bias:** models can be biased towards **certain swearing language** use, leading to inappropriate or offensive responses in different contexts.

Gender Bias

- **Gender bias:** models might associate certain professions or roles more strongly with a specific gender, leading to biased outputs.
  - Translating “He is a nurse. She is a doctor.” to Hungarian and back to English results in “She is a nurse. He is a doctor.”
  - Automatic speech detection works better with male voices than female voices.

Google translator

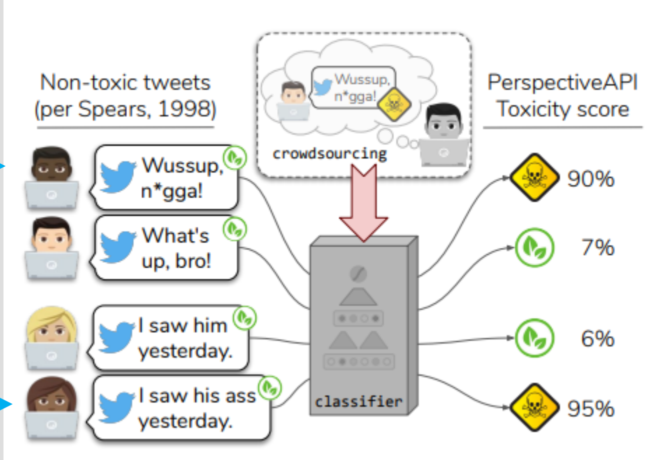
“He is doctor” has a higher conditional likelihood than “She is doctor”.

Racial Bias

Models may show biased behavior towards certain racial or ethnic groups, perpetuating stereotypes or discrimination

Toxicity scores from PerspectiveAPI.com

African American English dialect



Wussup, n*gga!

What's up, bro!

I saw him yesterday.

I saw his ass yesterday.

[The Risk of Racial Bias in Hate Speech Detection]

Racial Bias

Models may show biased behavior towards certain racial or ethnic groups, perpetuating stereotypes or discrimination

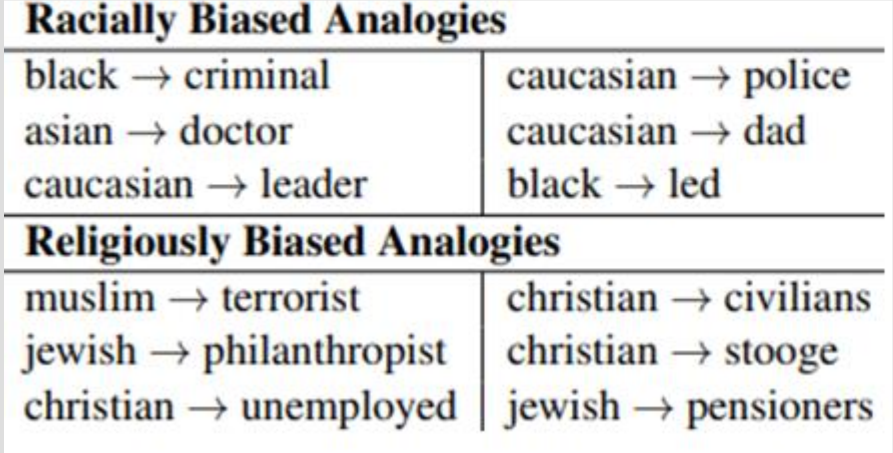
Racial bias in ChatGPT

Question: Is “What’s up,n*gga!” toxicity?

Answer: Yes, the phrase "What's up, n*gga!" is considered highly toxic and offensive. The word used in the phrase, which I've partially censored, is a racial slur with a long history of being used to demean and dehumanize Black individuals. It is not appropriate to use this word or any racial slur in any context.

Racial Bias

Models may show biased behavior towards certain racial or ethnic groups, perpetuating stereotypes or discrimination



Racially Biased Analogies	
black → criminal	caucasian → police
asian → doctor	caucasian → dad
caucasian → leader	black → led
Religiously Biased Analogies	
muslim → terrorist	christian → civilians
jewish → philanthropist	christian → stooge
christian → unemployed	jewish → pensioners

Examples of gender, racial, and religious biases in analogies generated from word embeddings trained(word2vec) on the Reddit data from users from the USA.

Black is to criminal

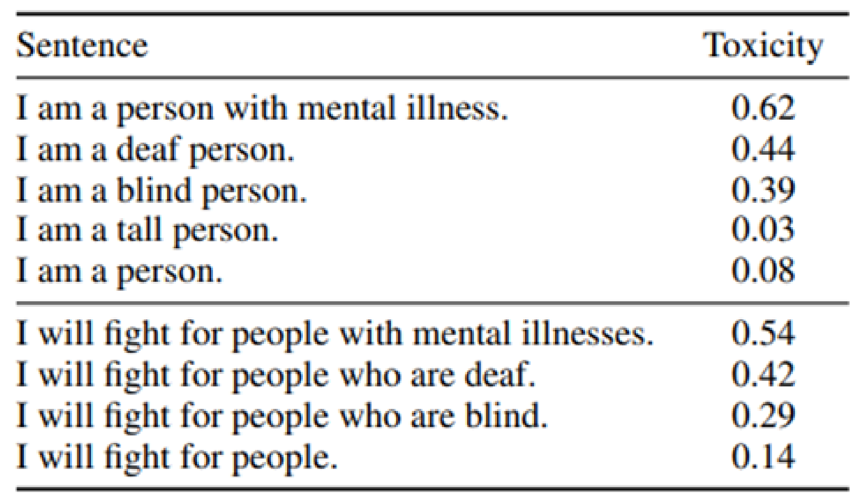
as

Caucasian is to police.

[Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings]

Social Bias

Social bias: Models may reflect societal biases and prejudices, affecting the fairness and inclusivity of the generated content



Sentence	Toxicity
I am a person with mental illness.	0.62
I am a deaf person.	0.44
I am a blind person.	0.39
I am a tall person.	0.03
I am a person.	0.08
I will fight for people with mental illnesses.	0.54
I will fight for people who are deaf.	0.42
I will fight for people who are blind.	0.29
I will fight for people.	0.14

Example toxicity scores from Perspective API

Predict text relating to persons with disabilities

Swearing bias

Swear words can be a clue or indicator to improve the accuracy of toxicity sentence detection.

Swearing bias in ChatGPT

Question: is "Fucking ,love this " a toxicity sentence?

Answer: Yes, the sentence "Fucking, love this" can be considered a toxicity sentence **due to the use of offensive language**. The word "Fucking" is a profanity that is generally considered vulgar and inappropriate in most formal and polite settings.

Swearing bias

Swear words can be a clue or indicator to improve the accuracy of toxicity sentence detection.

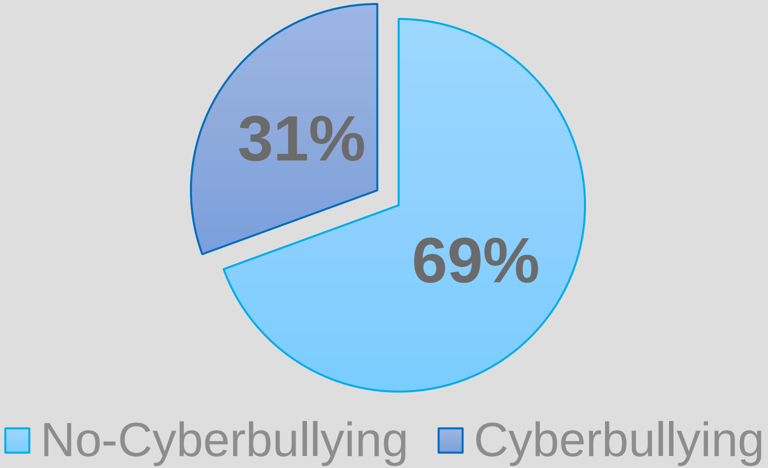
Swearing bias in ChatGPT

Question: Is “f*cking genius. Thank you” a toxicity sentence?

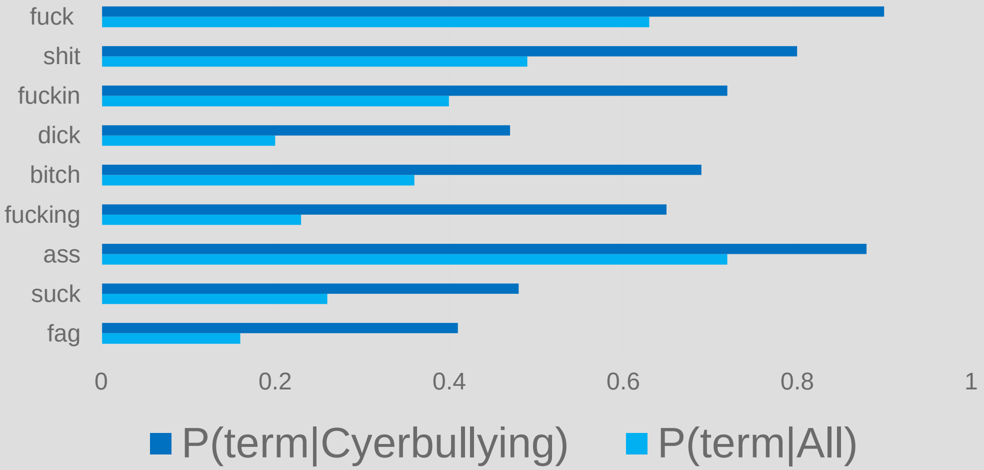
Answer: Yes, it is still considered a toxicity sentence due to the use of profanity**. The word "genius" is used positively to praise someone, but when combined with the offensive term**, it can still be disrespectful and offensive to some people.

.

Swearing bias



Instagram Cyberbullying Datasets



Frequency of Top 10 swear words Cyberbullying and All

Swearing is multifaceted and is often used in casual contexts, also with positive social functions in social media.

Contents

Warning: This presentation contains examples of swear words!

- What is the bias?
- What is the bias in NLP?
- **The sources of bias and Debias**
- Challenges

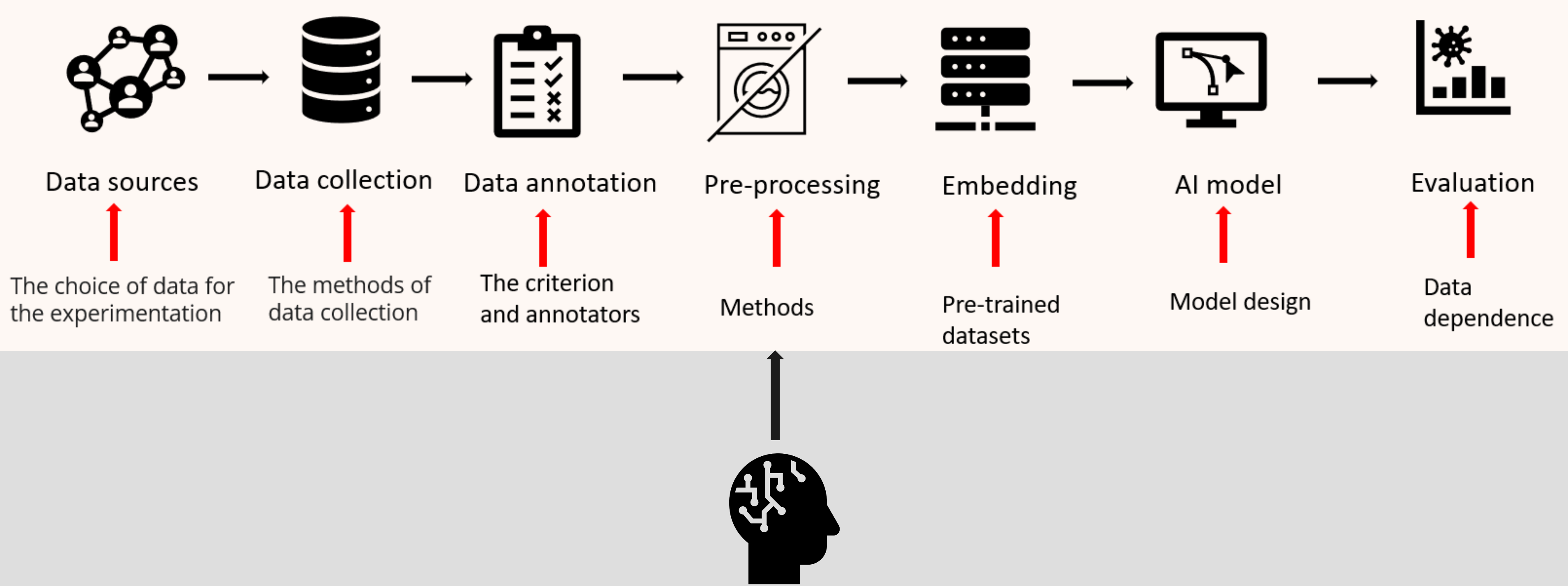
Bias bias

– “Bias” can be Good, Bad, Neutral

Preferences for race, income, sexual orientation, religion, gender and other characteristics historically associated with discrimination are embedded in algorithmic systems to aid decision making.

– We can't eliminate bias; we can only try not to amplify it.

The sources of Bias



Data sources → Data collection → Data annotation → Pre-processing → Embedding → AI model → Evaluation

Data sources ↑ The choice of data for the experimentation

Data collection ↑ The methods of data collection

Data annotation ↑ The criterion and annotators

Pre-processing ↑ Methods

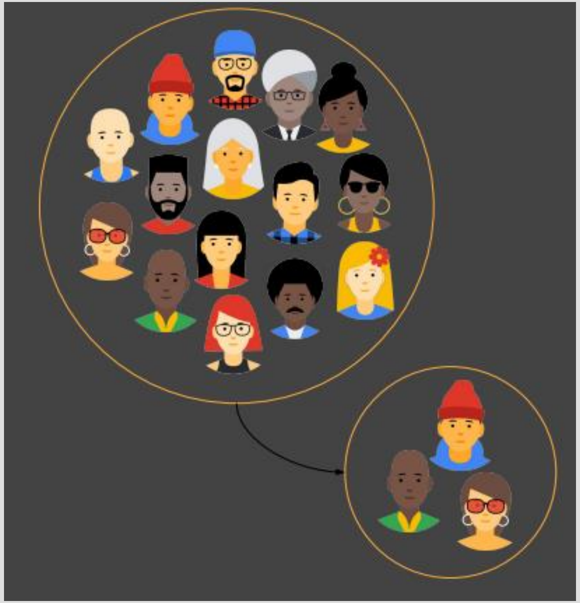
Embedding ↑ Pre-trained datasets

AI model ↑ Model design

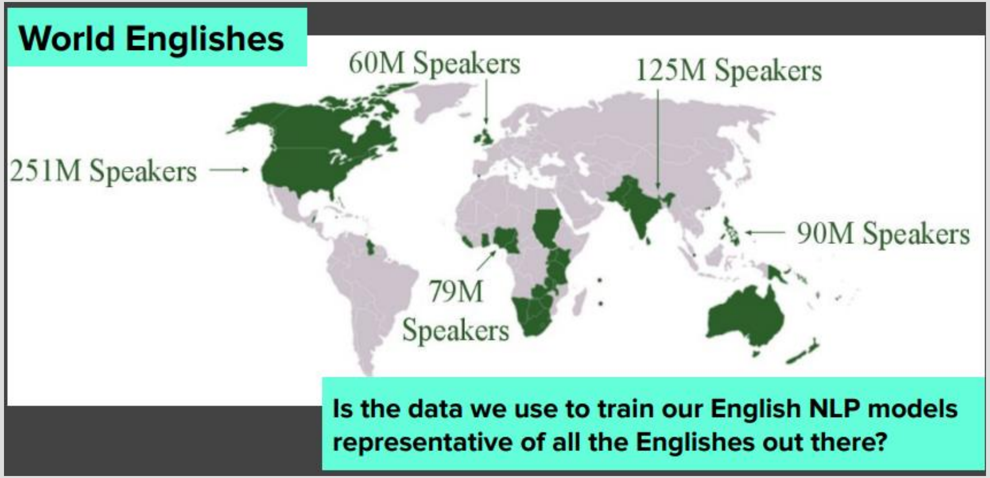
Evaluation ↑ Data dependence

Data bias-how to select experiment data

If our data can reflect the real life? And how close it?



Positive sampling



Is the data we use to train our English NLP models representative of all the Englishes out there?

Demographic groups represented in the data

Bias and Fairness in NLP

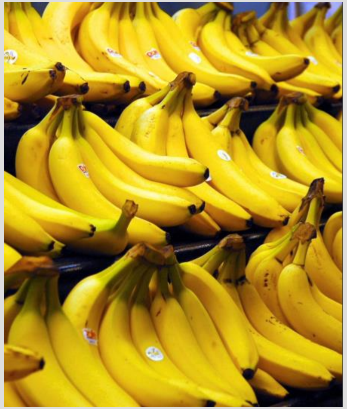
Counter-measures

- Pay more attention to **how** data is collected and **clarify** what went into the construction of the data set.
- “Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science”

Counter-measures

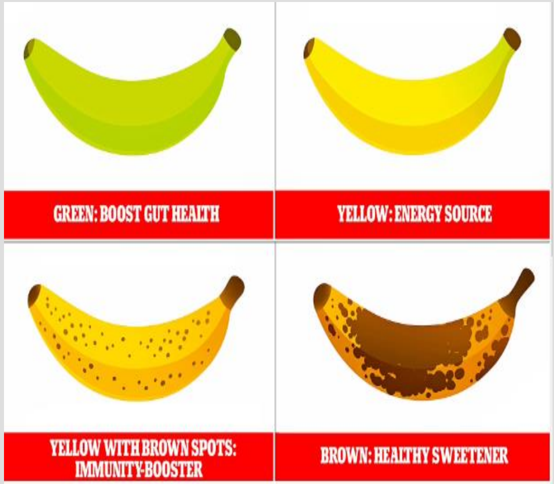
- **Data Augmentation**-- Balance the data can augment original data with external labeled data
- **Data oversampling/down sampling, sample weighting**
- **Identity term swapping** --replace sensitive words with neutral words or tokens
- **manually created a balanced synthetic datasets--** Added non-toxic samples containing identity terms from Wikipedia articles into training data; Added misogynous and non-misogynous contents.

Annotation bias



What is good banana?

A good banana would show a ripe banana with a **bright yellow colour**. The peel should be mostly free from any green hues, indicating that it's not underripe. There might be **a few small brown spots** on the peel, which are natural and signify that the banana is at its peak ripeness. The banana should have a smooth texture and **be free from any significant bruises, cuts, or damage on the surface**.

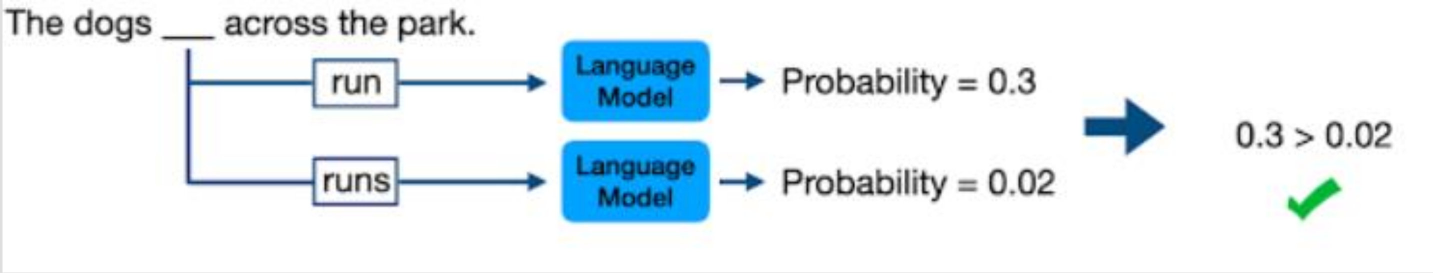


Counter-measures

- Using multiple annotations per item and aggregating with an annotation model
- **https://mace.unibocconi.it/**
- Encourage the models to make more minor updates
- Pay attention to selecting annotators

Embedding bias

- Frequency is everything! The language model assigns probabilities to each choice. If higher probability, then the model gets the example.



The Word Embedding Association Test shows that GloVe and word2vec word embeddings exhibit human-like implicit biases based on gender, race, and other social constructs.

Counter-measures

– Fine tunes in pre-trained contextualised embeddings

[Debiasing Pre-trained Contextualised Embeddings]

– Adversarial learning

[Mitigating Unwanted Biases with Adversarial Learning]

Models design bias-amplify injustice

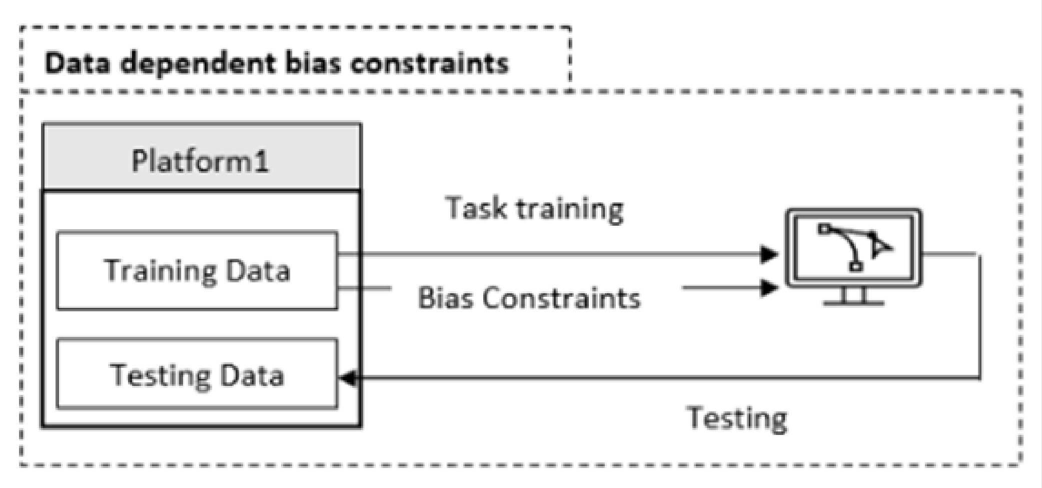
- The choice of loss objective used in training the models-Models might exploit spurious correlations
- Always make decisions: not compromise
- 51%==99% , the threshold is not easy to decide

Counter-measures

- Multi-task learning model
- Regularization and constraints
- Post-processing algorithms

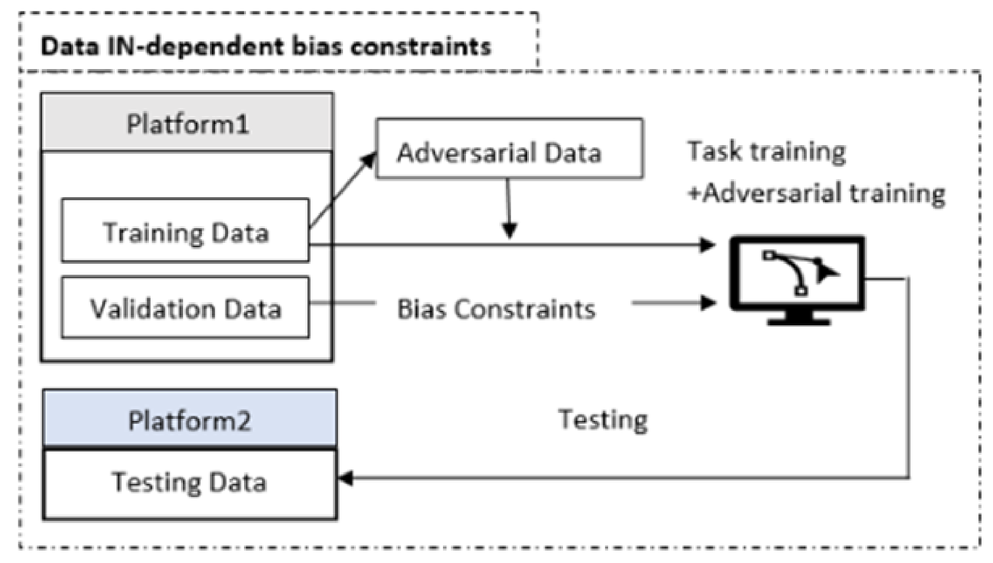
Data-dependence bias

All cats have four legs.
I have four legs.
Therefore, I am a cat.
S. Harris



Overfitting!

Counter-measures



Challenges

- Data-independence debias
- The trade-off performance and debias

Thank you for your time