Food for thought about Ethics in AI

Marta R. Costa-jussà

Ethics in AI

- Motivation/Challenges
  - Robustness
  - Environmental costs
  - Biases

  in the Transformer models

- Towards Solving Biases
  - Evaluation
  - Algorithms
  - Datasets and Documentation

Timeline

2013	…	2016	2017	2018	2019	2020

Word2Vec
Mikolov et al.

Bias in Words Embeddings
Bolukbasi et al.

Compas Racial Bias

Transformer
Vaswani et al.

Ethics AI Regulation UE

Gender Shades
Buolamini et al.

Bias in Transformers
Basta et al.

Debias MT method
Escudé et al.

GeBioToolkit
Costa-jussà et al.

MT-Datasheets
Costa-jussà et al.

Environmental Costs Transformers
Gebru et al.

GPT-3
Brown et al.

Background: Transformer Models

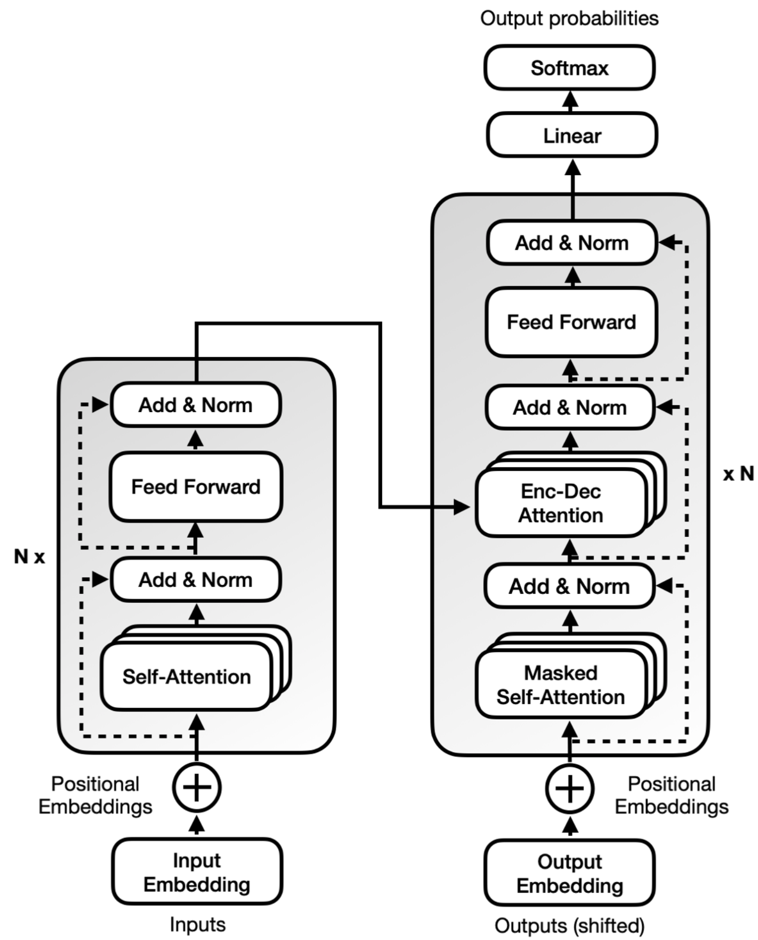


figure: gerard gallego

Robustness:
Sentiment Classification fails just with typos

Connoisseurs of Chinese film will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus.	Prediction: **Positive (77%)**
Aonnoisseurs of Chinese film will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus.	Prediction: **Negative (52%)**

Face recognition can fail with just glasses

Major flaws in facial recognition systems revealed: Bizarre 'face stealing' specs can fool them into thinking you are someone else (and can even turn a man into Milla Jovovich)

- Glasses allow wearer to dodge recognition or impersonate another person
- Method disrupts the system's ability to accurately read pixel colouration
- In experiments, it allowed a man to impersonate actress Milla Jovovich
- Researchers say it highlights the ways attackers might evade technology

ENVIRONMENTAL COSTS

T. Gebru pointed out the environmental cost of training large language models

Google widely criticized after parting ways with a leading voice in AI ethics

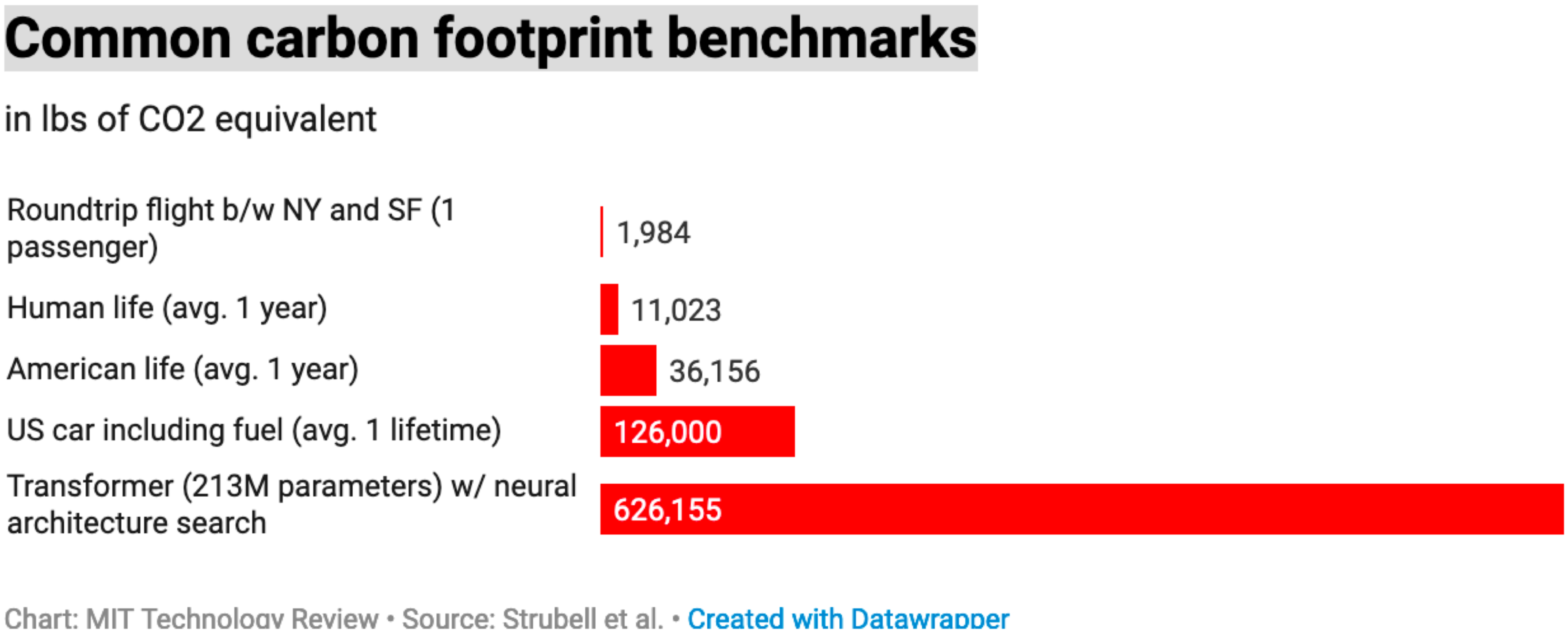
By Rachel Metz, CNN Business

Updated 0410 GMT (1210 HKT) December 5, 2020

Risks of deploying large language models

- The environmental cost
- The impossibility to audit the massive amount of training data as well as the model itself
- Research efforts concentrating towards these models at the expense of more environmentally-friendly ones or ones that attempt another approach at modelling language
- The very harmful mistakes these models make when they are trusted blindly

Common carbon footprint benchmarks



Common carbon footprint benchmarks

The estimated costs of training a model once

In practice, models are usually trained many times during research and development.

	Date of original paper	Energy consumption (kWh)	Carbon footprint (lbs of CO2e)	Cloud compute cost (USD)
Transformer (65M parameters)	Jun, 2017	27	26	$41-$140
Transformer (213M parameters)	Jun, 2017	201	192	$289-$981
ELMo	Feb, 2018	275	262	$433-$1,472
BERT (110M parameters)	Oct, 2018	1,507	1,438	$3,751-$12,571
Transformer (213M parameters) w/ neural architecture search	Jan, 2019	656,347	626,155	$942,973-$3,201,722
GPT-2	Feb, 2019	-	-	$12,902-$43,008

Note: Because of a lack of power draw data on GPT-2's training hardware, the researchers weren't able to calculate its carbon footprint.

Table: MIT Technology Review • Source: Strubell et al. • Created with Datawrapper

Big Transformer Models

Positive

- it enables anyone building a machine learning model involving language processing to use this powerhouse as a readily-available component – saving the time, energy, knowledge, and resources that would have gone to training a language-processing model from scratch.

Negative

- energy-consuming
- ”dangerous”: it could easily help to generate “fake news”

Recommendations

- Authors should report training time and sensitivity to hyperparameters.
- Academic researchers need equitable access to computation resources.
- Researchers should prioritize computationally efficient hardware and algorithms.

EXAMPLES OF GENERAL BIASES

COMPAS is an assistive (biased) software and support tool used to predict *recidivism* risk

The prediction fails differently for the black defendants:

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

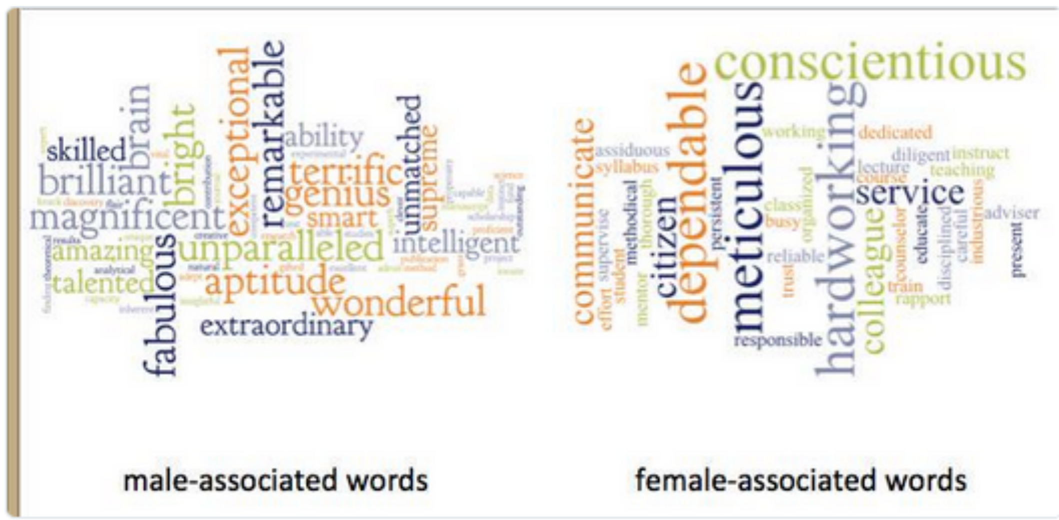
Algorithmic screening of Resumés can reproduce and even exacerbate human biases

REPORT

Challenges for mitigating bias in algorithmic hiring

Manish Raghavan and **Solon Barocas** · Friday, December 6, 2019

Male vs. Female Academic Reference Letters



Gender Shades showed face recognition is much less accurate on black people

AI TayTweets learnt from conversations held on social media and it turned to be racist

BBC

Taylor Swift 'tried to sue' Microsoft over racist chatbot Tay

10 September 2019

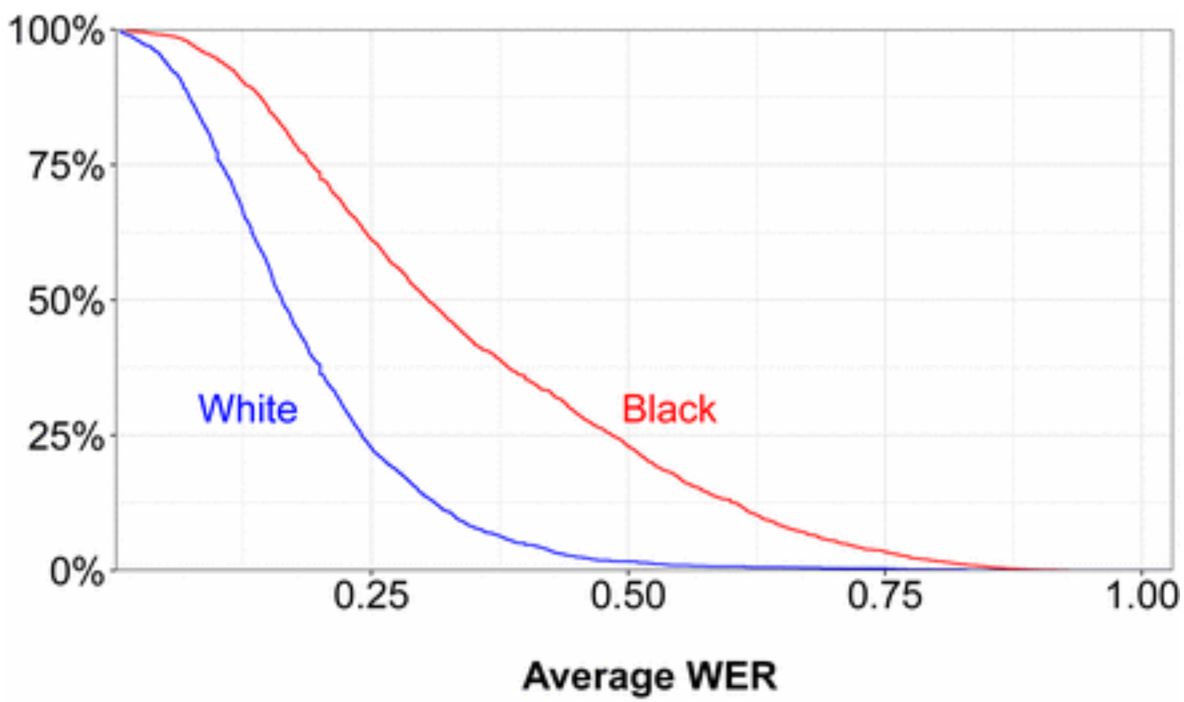
RESEARCH ARTICLE

Racial disparities in automated speech recognition

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel

PNAS April 7, 2020 117 (14) 7684-7689; first published March 23, 2020; https://doi.org/10.1073/pnas.1915768117

Edited by Judith T. Irvine, University of Michigan, Ann Arbor, MI, and approved February 12, 2020 (received for review October 5, 2019)



TOWARDS SOLVING BIASES

Evaluation of Gender Bias in Contextual Word Embeddings

Research together with Christine Raouf Basta and Noé Casas

Evaluation | Debiasing Algorithms | Balanced datasets

Words Embeddings

- Learned from raw data based on the Distributional Hypothesis:
  - ***"You shall know a word by the company it keeps" (Firth, 1957)***

- Each word in the vocabulary is represented by a low dimensional vector

Motivation for Contextual Word Embeddings

- Same word can have different meaning depending on the context. Example:
  - *Mary and Joanna* ***play*** *basketball in a wonderful way*
  - *John is the protagonist in this year's school* ***play***
- Classic word embeddings offer the same vector representation regardless of the context.
- Contextual Word Embeddings create **word representations** that **depend on the context.**

Evaluation | Debiasing Algorithms | Balanced datasets

Approaches for Contextual Word Embeddings

[credits Noe Casas]

Model Alias	Org.	Article Reference
ULMfit	fast.ai	*Universal Language Model Fine-tuning for Text Classification* Howard and Ruder
 ELMo	AllenNLP	*Deep contextualized word representations* Peters et al.
OpenAI GPT	OpenAI	*Improving Language Understanding by Generative Pre-Training* Radford et al.
 BERT	Google	*BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* Devlin et al.
xling BERT	Facebook	*Cross-lingual Language Model Pretraining* Lample and Conneau

Evaluation | Debiasing Algorithms | Balanced datasets

Why ELMO?

- Elmo was used for our experiments, as it provides word-level representations, as opposed to BERT’s subwords.

- This makes it possible to study the word-level semantic traits directly.

Related Work: Word Embeddings encode bias

[Caliskan et al. 2017] replicate a spectrum of biases from using word embeddings, showing text corpora contain several types of biases:

- morally neutral as toward insects or flowers
- problematic as toward race or gender ,
- reflecting the distribution of gender with respect to careers or first names

[credits to Hila Gonen]

Concepts 1	Concepts 2	Attributes 1	Attributes 2
Flowers: buttercup, daisy, lily	**Insects**: ant, caterpillar, flea	**Pleasant**: freedom, health, love	**Unpleasant**: abuse, crash, filth
European American names: Brad, Brendan	**African American names**: Darnell, Lakisha	**Pleasant**: joy, love, peace	**Unpleasant**: agony, terrible
Male attributes: male, man, boy	**Female attributes**: female, woman, girl	**Math words**: math, algebra, geometry	**Arts Words**: poetry, art, dance

Evaluation | Debiasing Algorithms | Balanced datasets

Evaluation of Contextual Word Embeddings

Contextual embeddings get a vector representation for the word according to its context, so we expect a different attitude towards the gender bias. [Zhao et al. 2019] show that contextualized word embeddings may inherit implicit gender bias. This motivates us to study **two main questions:**

- Do contextual word embeddings **exhibit gender?**
- Do different evaluation techniques identify similar biases?

Experiments For Evaluation Bias

Three experiments were carried out in our evaluation:

1. Detecting the gender space and the Direct bias
2. Male and female biased words clustering
3. Classification approach of biased words

Our comparison is based on pre-trained sets of all these options. For experiments, we use the English-German news corpus from WMT18

Lists for Definitional, Biased and Professional Terms

- **Definitional List** 10 pairs (e.g. he-she, man-woman, boy-girl)
- **Biased List**, which contains of 1000 words, 500 female biased and 500 male biased. (e.g. diet for female and hero for male)
- **Extended Biased List**, extended version of Biased List. (5000 words, 2500 female biased and 2500 male biased)
- **Professional List** 319 tokens (e.g. accountant, surgeon)

1. Gender Space and Direct Bias

1. Randomly sampling sentences that contain words from the Definitional List, swap the definitional word with its pair-wise equivalent from the opposite gender.

1. Gender Space and Direct Bias

1. Randomly sampling sentences that contain words from the Definitional List, swap the definitional word with its pair-wise equivalent from the opposite gender.
2. Get Elmo embeddings for the word and its swapped equivalence, compute their difference.

1. Gender Space and Direct Bias

1. Randomly sampling sentences that contain words from the Definitional List, swap the definitional word with its pair-wise equivalent from the opposite gender.
2. Get Elmo embeddings for the word and its swapped equivalence, compute their difference.
3. On the set of difference vectors, we compute their principal components to verify the presence of bias.

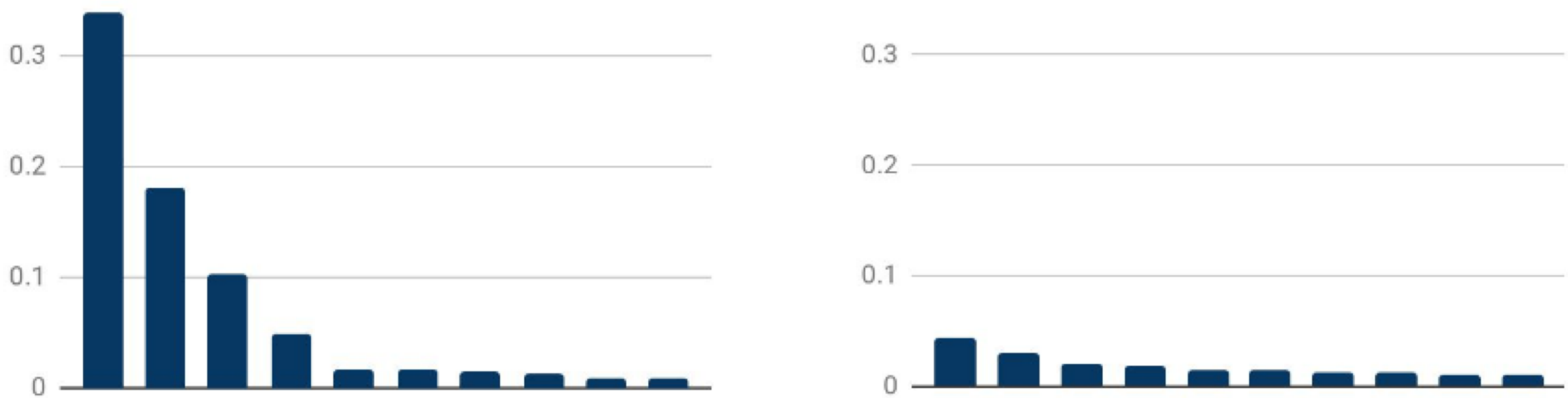
1. Gender Space and Direct Bias

1. Randomly sampling sentences that contain words from the Definitional List, swap the definitional word with its pair-wise equivalent from the opposite gender.
2. Get Elmo embeddings for the word and its swapped equivalence, compute their difference.
3. On the set of difference vectors, we compute their principal components to verify the presence of bias.
4. Repeat for an equivalent list of random words (skipping the swapping).

Evaluation | Debiasing Algorithms | Balanced datasets

1. Gender Space and Direct Bias

Percentage of variance in PCA: definitional vs random



(Left) Percentage of variance explained in the PCA of definitional vector differences.
(Right) The corresponding percentages for random vectors

Evaluation | Debiasing Algorithms | Balanced datasets

1. Gender Space and Direct Bias

- **Direct Bias** is a measure of how close a certain set of words are to the gender vector.
- Computed on list of (neutral) professions.

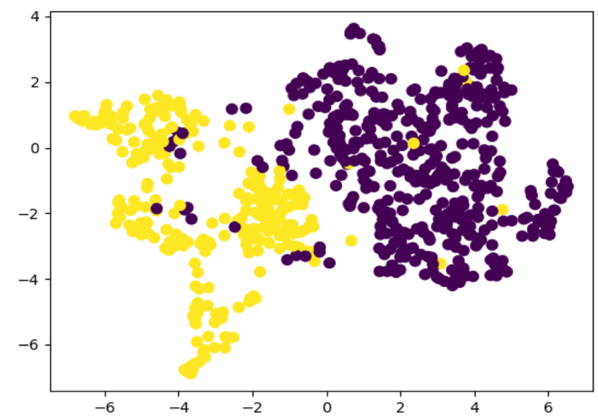
$$\frac{1}{|N|}\sum_{w \epsilon N} |cos(\vec{w}, g)|$$

	Direct Bias
WE	0.08
ELMO	0.03

2. Male and female-biased words clustering

- **k-means**

- Generate 2 clusters of the embeddings of tokens from the **Biased list** (e.g. diet for female and hero for male)



	Accuracy
WE	99,9%
ELMO	70,1%

3. Classification Approach

- **SVM**
- Classify **Extended Biased List** into words associated between male and female
- 1000 for training, 4000 for testing

	Accuracy
WE	98.25%
ELMO	85.56%

Visualization

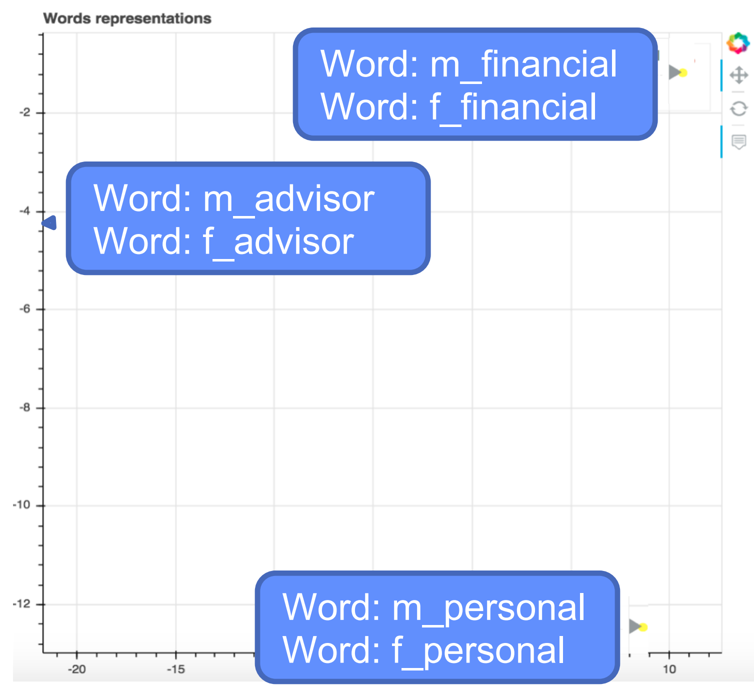
Research together with Carlos Escolano, Elora Lacroux, Pere-Pau Vàzquez

Same representation for *personal financial advisor* (in a male/female context)

https://github.com/elorala/interlingua-visualization

I've known **him** for a long time, my friend works as a **personal financial advisor**

I've known **her** for a long time, my friend works as a **personal financial advisor**

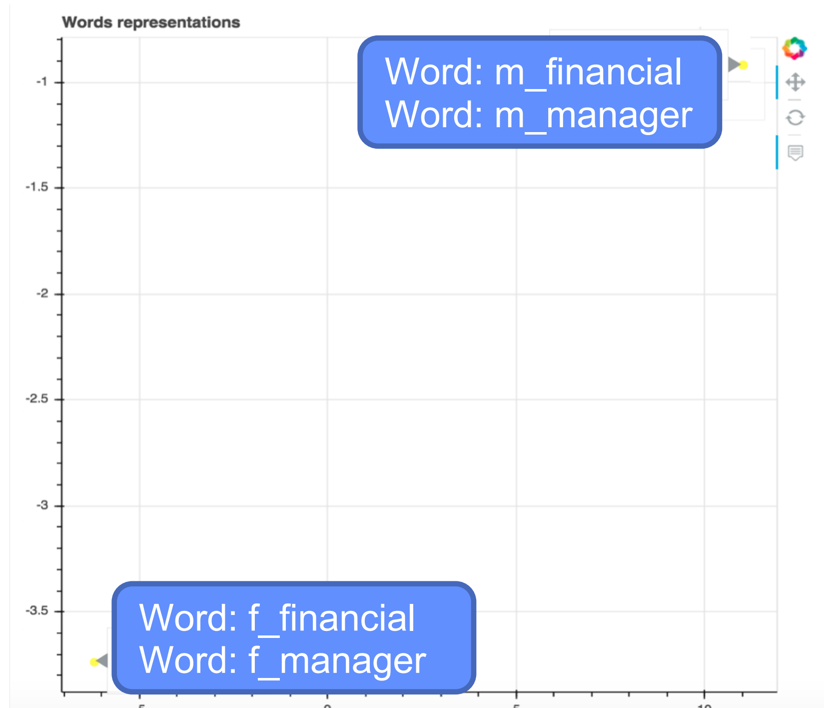


Different representation for *financial manager* (in a male/female context)

https://github.com/elorala/interlingua-visualization

I've known **him** for a long time, my friend works as a **financial manager**

I've known **her** for a long time, my friend works as a **financial manager**



Evaluation Debiasing Algorithms Balanced datasets

Conclusions on evaluating gender bias in contextual word embeddings

😃 Contextual word embeddings seems to **mitigate bias** in when measuring in the following aspects:

- ↓ gender **space and direct bias**
- ↓ male/female **clustering**,
- ↓ **classification** experiment

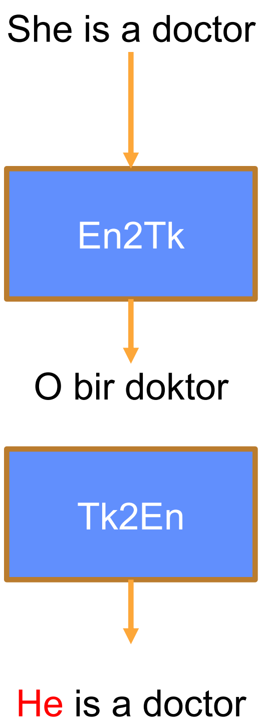
😞 Contextual word embeddings **preserve** gender bias

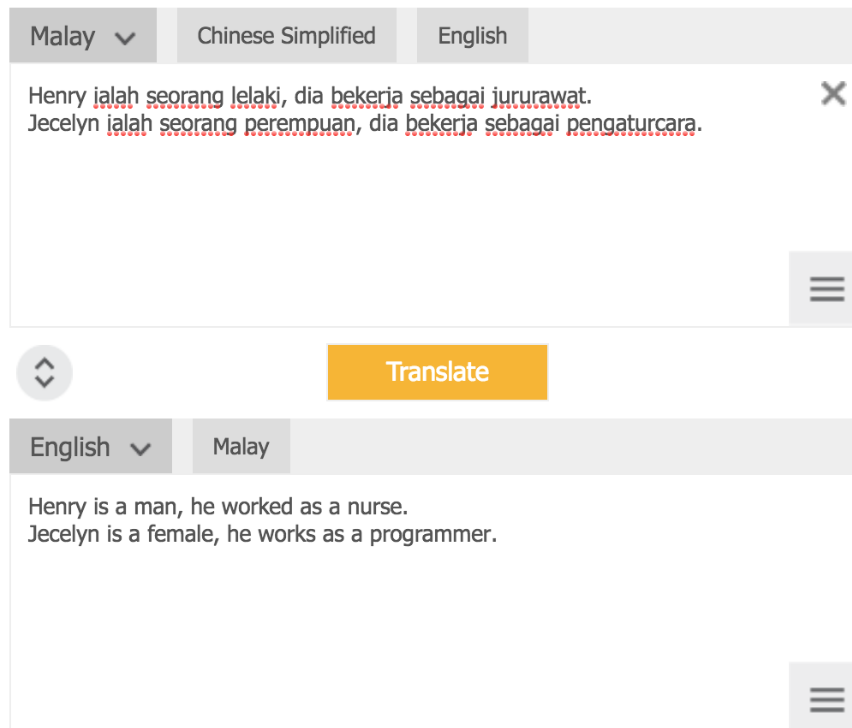
Evaluation | Debiasing Algorithms | Balanced datasets

Debiased algorithm for Machine Translation

Research together with Joel Escudé

Gender Bias in MT: Example

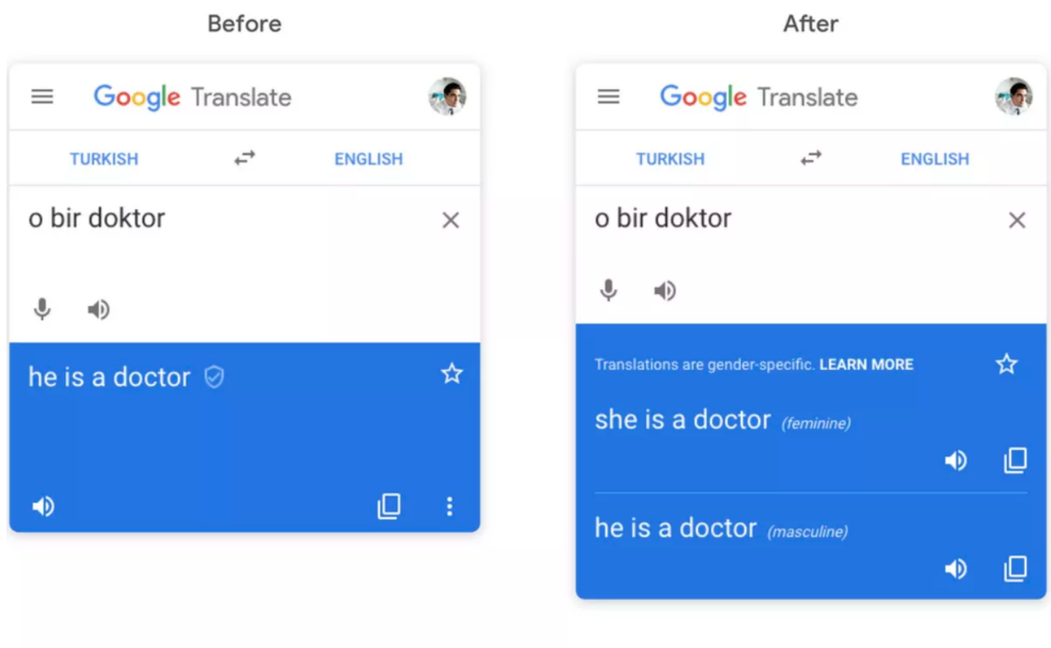




Evaluation | Debiasing Algorithms | Balanced datasets

Related work: Providing Gender-Specific Translations

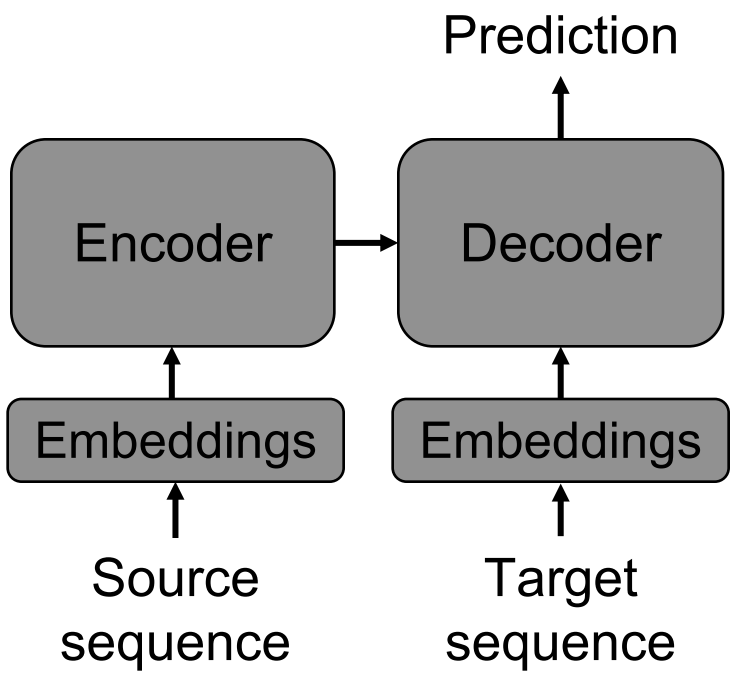
[Johnson et al., 2018]



Evaluation | Debiasing Algorithms | Balanced datasets

How to reduce gender bias in a neural MT?

- Neural MT system
  - Transformer
- Word embeddings
  - GloVe
  - GloVe Debias-WE
  - GN-GloVe
- Data
  - EN->ES WMT



Evaluation | Debiasing Algorithms | Balanced datasets

Techniques to Debias Word Embeddings

(1) Debias **After** Training [Bolukbasi et al. 2016] ---> Debias WE

Define a gender direction

Define inherently neutral words (nurse as opposed to mother)

Zero the projection of all neutral words on the gender direction

Remove that direction from words

(2) Debias **During** Training [Zhao et al. 2018] ---> GN-Glove

Train word embeddings using GloVe (Pennington et al., 2014)

Alter the loss to encourage the gender information to concentrate in the last coordinate (use two groups of male/female seed words, and encourage words from different groups to differ in their last coordinate)

To ignore gender information –simply remove the last coordinate

Small Impact on Translation Quality

Pre-trained emb.	BLEU
Baseline	29.78
GloVe	30.62
GloVe Debias-WE	29.95
GN-GloVe	30.74

Dataset for Explicitly Testing Gender Bias

4 test sets of 1000 sentences, on the patterns

Test1/Test2

(En) I've known her/him for a long time, my friend works as a/an .... [OCCUPATION]

(Es) La/Lo conozco desde hace mucho tiempo, mi amiga/amigo trabaja como .... [OCCUPATION]

Test3/Test4

(En) I've known Mary/John for a long time, my friend works as a/an .... [OCCUPATION]

(Es) Conozco a María/Juan desde hace mucho tiempo, mi amiga/amigo trabaja como .... [OCCUPATION]

List of 1000 occupations [U.S. Bureau of Labor Statistics].

(En) accounting clerk : (Es) contable

Impact on Equalizing Gender Bias: Accuracy

Pre-trained emb.	her : amiga	him : amigo	Mary : amiga	John : amigo
Baseline	99.8	99.9	69.5	99.9
GloVe	**100.0**	100.0	90.0	100.0
GloVe Debias-WE	**99.9**	100.0	**100.0**	100.0
GN-GloVe	99.6	100.0	56.4	100.0

Evaluation | Debiasing Algorithms | Balanced datasets

Conclusions on Equalizing Gender Bias in MT

Using equalized word embeddings on a MT system show:

- Similar translation quality
- Less biased gender predictions

Limitations

- Based on “debiased” word embeddings (Gonen and Goldberg 2019)
- Re-learning biases during MT training

Generating "Fair" Datasets

Research together with Pau Li Lin, Cristina España

Related Work: Getting Gender Right in NMT

[Vanmassenhove, et al., 2018]

(Source) … I am happy that …

(Translation 1) ... je suis heureuse que...

(Translation 2) ... je suis heureux que …

→ Creation of a multilingual dataset with utterances labelled for speaker gender and other demographic information.

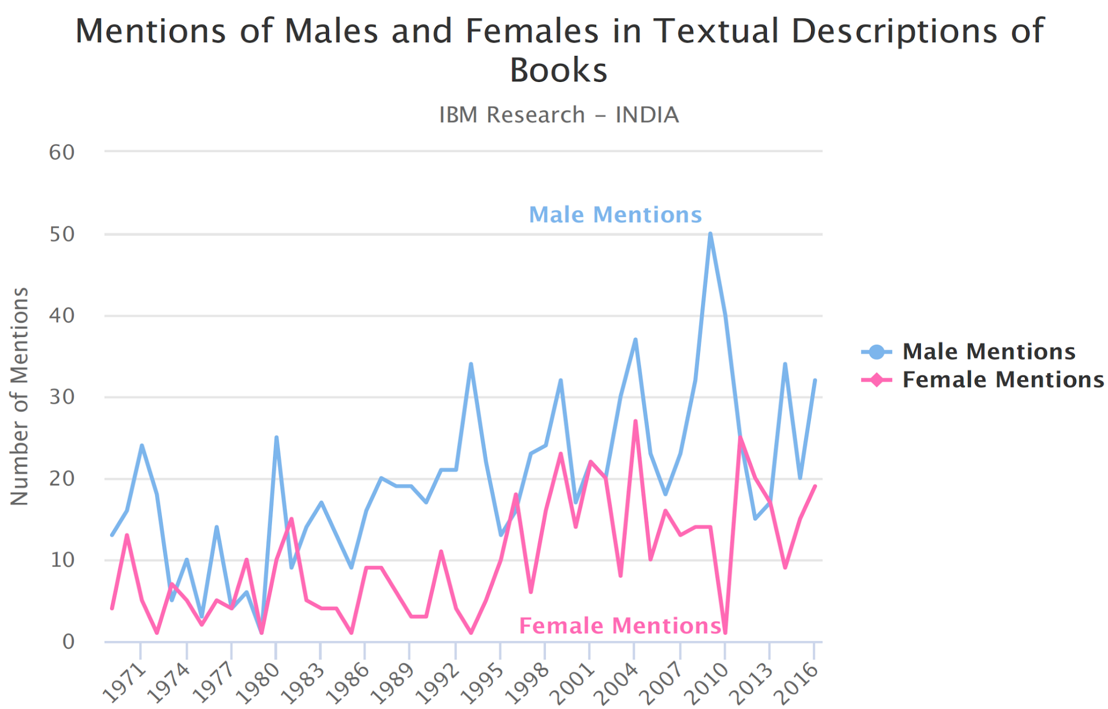
→ Experiments with NMT systems tagged for speaker gender.

Evaluation | Debiasing Algorithms | Balanced datasets

Unbalanced gender representation in data

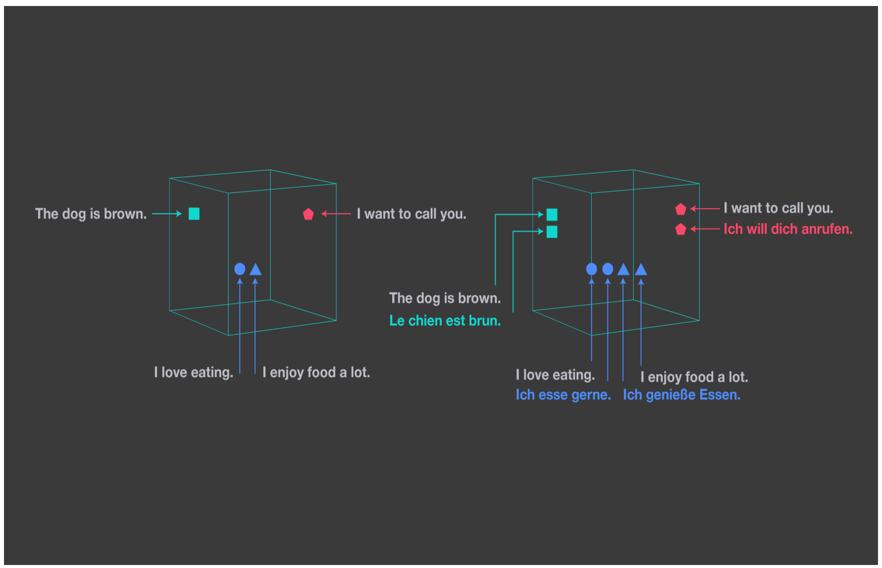
Under-representation of females in text books

[Maadan et al., 2018]



Evaluation | Debiasing Algorithms | Balanced datasets

GeBioToolkit: Built on-top of LASER used to extract wikimatrix



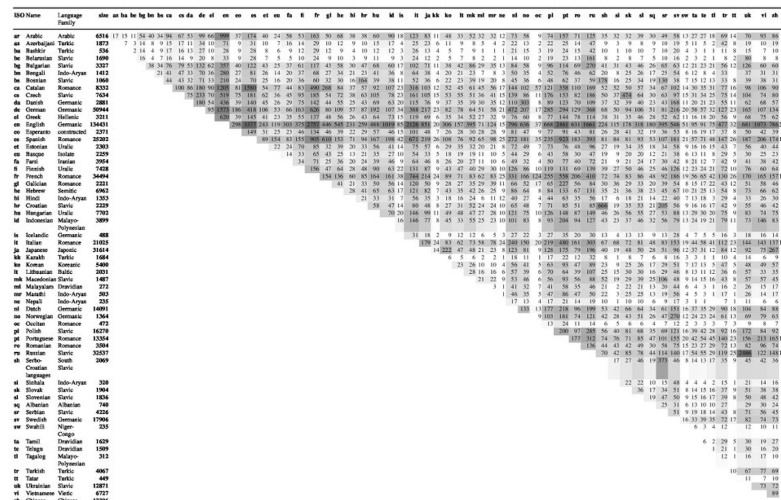
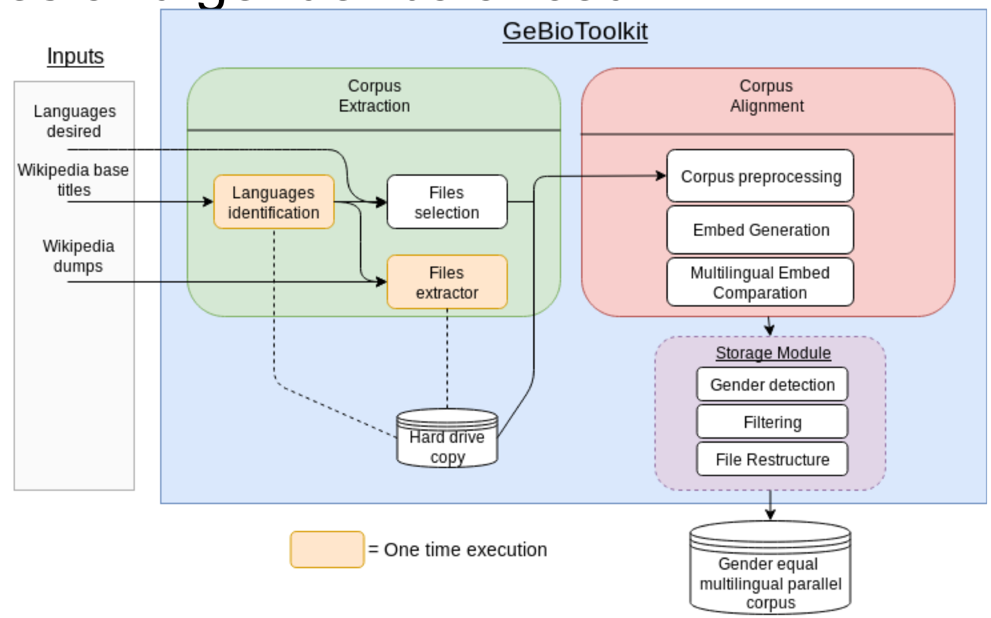


Table 1: WikiMatrix: size of mined sentences (in thousands) for each langauge pair.

GeBioToolkit: Extracting Balanced data (female/male) data from Wikipedia Biographies

- Based on LASER,
- Customizable for languages and gender balanced
- Document information
- Gender information



Evaluation | Debiasing Algorithms | Balanced datasets

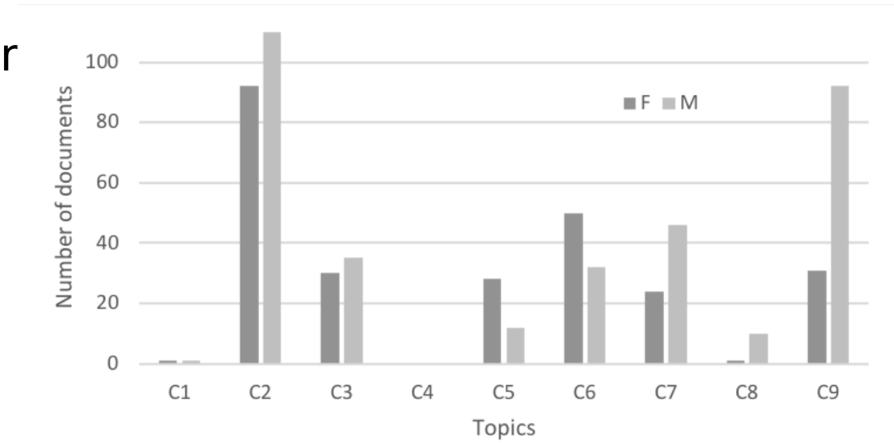
GeBioToolkit: accuracy of 96%

- We randomly select **50 sentences** in **3 languages** (English,Spanish and Catalan).
- **7 different native/fluent speakers** (annotators) were asked to score a tuple (3 sentences) with 1 if it conveys the same meaning and 0 otherwise.
- When computing the majority vote among the evaluators, we reached 96% accuracy.
- We computed **Fleiss' kappa** which resulted in **0.67**, which is considered a substantial agreement

GeBioCorpus: Gender-Balanced Test Dataset

- 2000 sentences in English, Spanish and Catalan (1000 male, 1000 female)
- Allow the evaluation of machine translation outputs in: distant morphologies for a high-resourced language pair (English–Spanish); low-resourced pair (English–Catalan); and closely related languages (Spanish– Catalan)
- Topic information

(C1) Healthcare and medicine
(C2) Arts
(C3) Business
(C4) Industrial and manufacturing,
(C5) Law enforcement, social movements ar
(C6) Science, technology and education
(C7) Politics
(C8) Religion
(C9) Sports



Evaluation | Debiasing Algorithms | Balanced datasets

GeBioCorpus: Example

<doc **docid**=”Aurelia Arkotxa ” **wpid**=”51690640” **language**=”en” **topic**=”C6” **gender**=”Female” >
<title>Aurelia Arkotxa </title>
<seg id=”1”>She teaches classics at the University of Bayonne; she was co-founder of the literary magazine and a new newspaper.<\seg>
</doc>
<doc **docid**=”Catriona Gray ” **wpid**=”51838666” **language**=”en” **topic**=”C2” gender=”Female”>
<title>Catriona Gray </title>
<seg id=”1”>In addition, she obtained a certificate in outdoor recreation and a black belt in Choi Kwang-Do martial arts.<\seg >
<seg id=”2>Catriona Elisa Magnayon Gray (born 6 January 1994) is a Filipino-Australian model, singer, and beauty pageant titleholder who was crowned Miss Universe 2018.<\seg> <seg id=”3”>Gray was born in Cairns, Queensland, to a Scottish-born father, Ian Gray, from Fraserburgh, and a Filipina mother, Normita Ragas Magnayon, from Albay.<\seg > </doc>

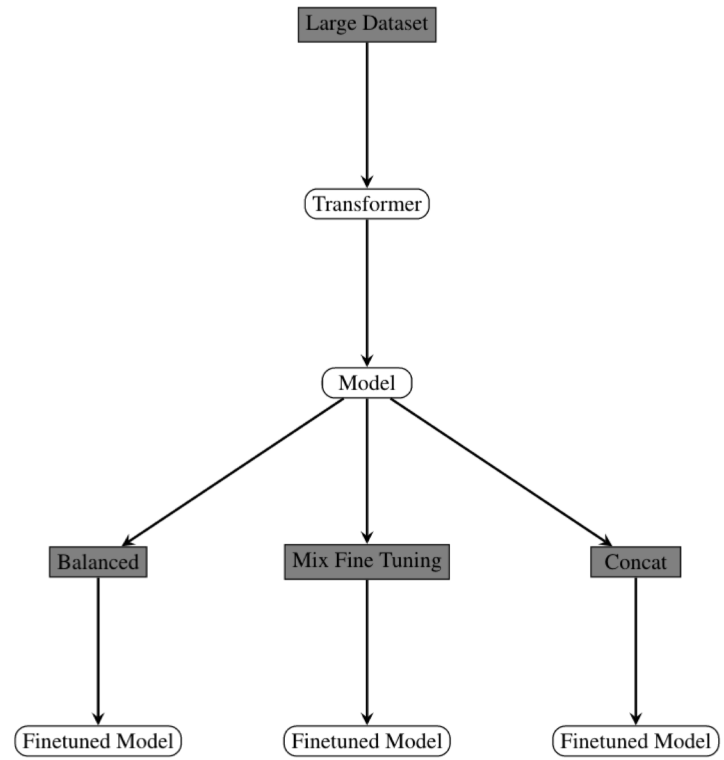
Using Gender Balanced Corpus to Mitigate Bias in MT

GeBioCorpus balanced set is used to mitigate gender biases in MT.

We perform fine-tuning techniques from a bigger model trained on unbalanced datasets with the balanced set.

Fine-tuning Machine Translation on Gender-Balanced Datasets

Marta R. Costa-jussà* and Adrià de Jorge*
TALP Research Center
Universitat Politècnica de Catalunya, Barcelona
marta.ruiz@upc.edu,adria.de.jorge@estudiantat.upc.edu



Evaluation | Debiasing Algorithms | Balanced datasets

MT-DataSheets for Datasets: Template

MT-Adapted Datasheet for Datasets Template

Open as Template | View Source | Download PDF

Author Marta R. Costa-jussà and Roger Creus and Oriol Domingo and Albert Domínguez and Miquel Escobar and Cayetana López and Marina Garcia and Margarita Geleta

Abstract This template is inspired by the already proposed datasheet template by Gebru et al. (2018) and slightly adapted to serve two main purposes: dataset usage in Machine Translation (MT) and dataset consumer-oriented. By doing so, we are making a call to the community to work on these datasheets, independently of being the dataset author.

MT-Adapted Datasheet for Datasets Template

I. Disclaimer

This Datasheet has been inspired by [1] and modified as proposed by [2] and it is not filled out by the dataset creator. Therefore it is strongly recommended to only make use of this if the creator has not filled in a proper datasheet or to use it in combination. It is required that writers indicate their personal and contact data as well as the date this datasheet was last reviewed hereunder. Please, also remember to change the datasheet title to the name of the datatset in question.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

II. Motivation

A. Who created the dataset(e.g., which team, research group) and on behalf of which entity (e.g. company, institution, organization)?

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

B. Did they fund it themselves? If there is an associated grant, please provide the name of the grant and/or the grant name and number).

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

C. For what purpose was the data set created? Was there a specific task in mind? If so, please specify the result type (e.g. unit) to be expected.

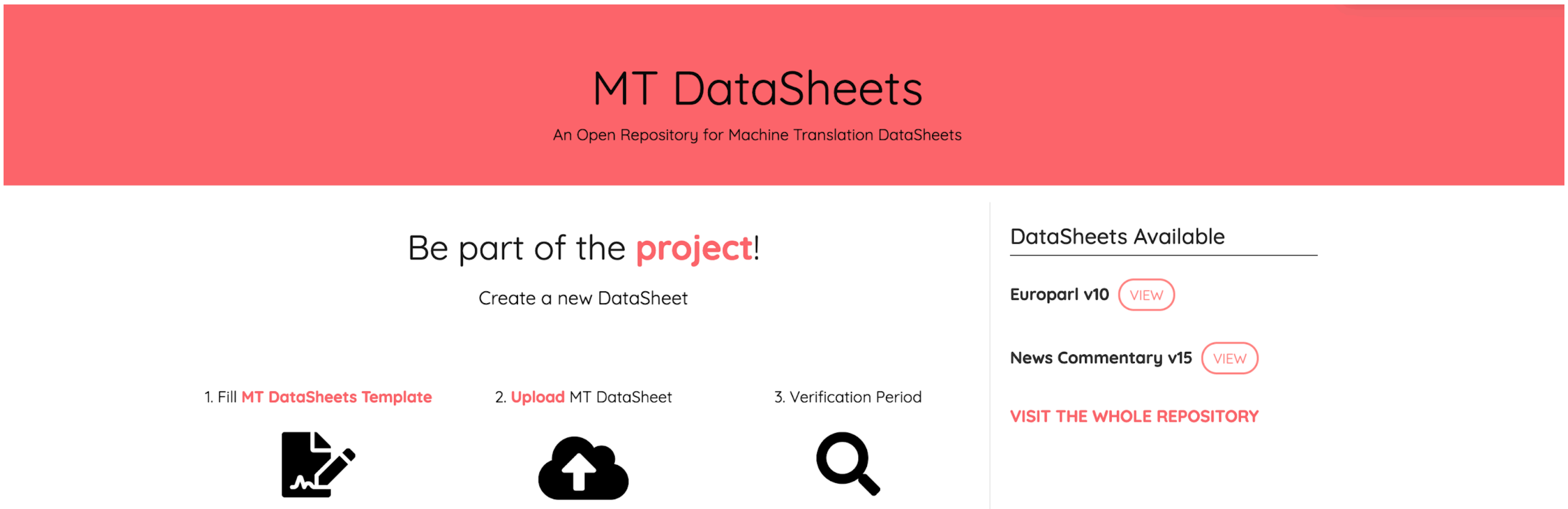
Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

D. Could any of these uses, or their results, interfere with human will or communicate a false reality?

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus

Evaluation | Debiasing Algorithms | Balanced datasets

MT-DataSheets for Datasets: Repository



Evaluation | Debiasing Algorithms | Balanced datasets

Conclusions in Datasets and Documentation

- Gender balanced datasets allow to produce fairer systems
- Documentation allows to analyse our training material and knowing more about our systems

This is more than biases, robustness or environmental costs...

GENERAL CONCLUSIONS

Is debiasing even (always) desirable?

- ML is about learning biases. Removing attributes removes information.

BUT...

- Gender information in NLP systems becomes harmful when the use of the system has a negative impact on people’s lives.

Bias comes from data... but algorithms can amplify this bias in a different amount

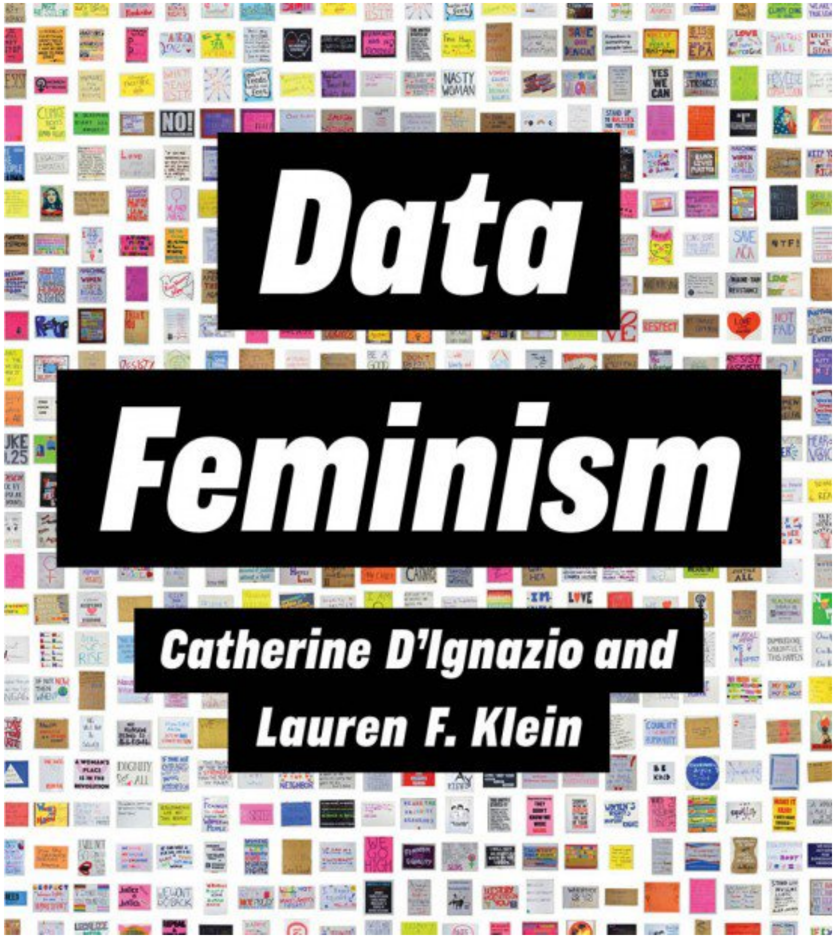
- Algorithms trained with the same data can have different amount of data...
  - e.g. the more generalization an algorithm gets from biased data, the less amount of bias that

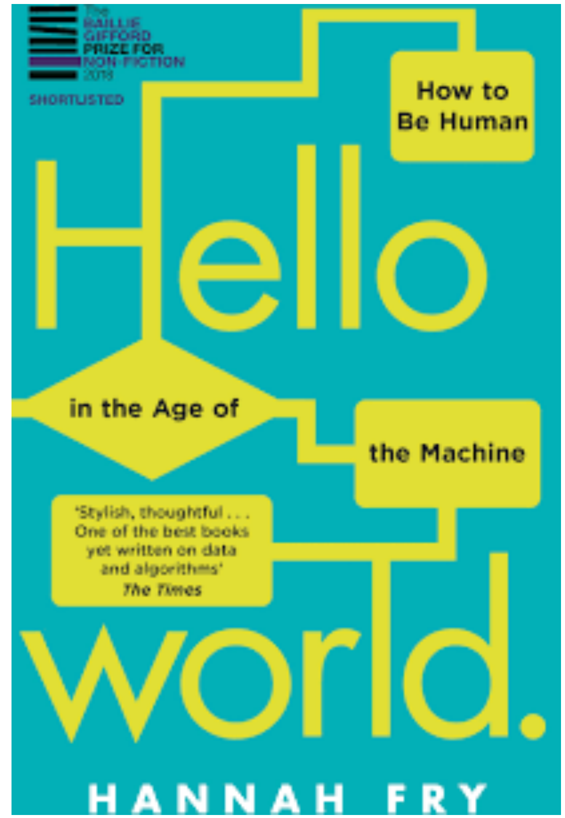
This is more than biases, robustness or environmental costs...

- This is about how do we want our society to be: debiasing computer systems may help in debiasing society
- This is about critical thinking, inclusiveness and co-operation: gender bias is a social phenomenon that can't be solved with mathematical methods alone. Discussions among politics, philosofers, sociologists, computer scientists... are required!
- This is about **continuing being human** in the algorithmic era

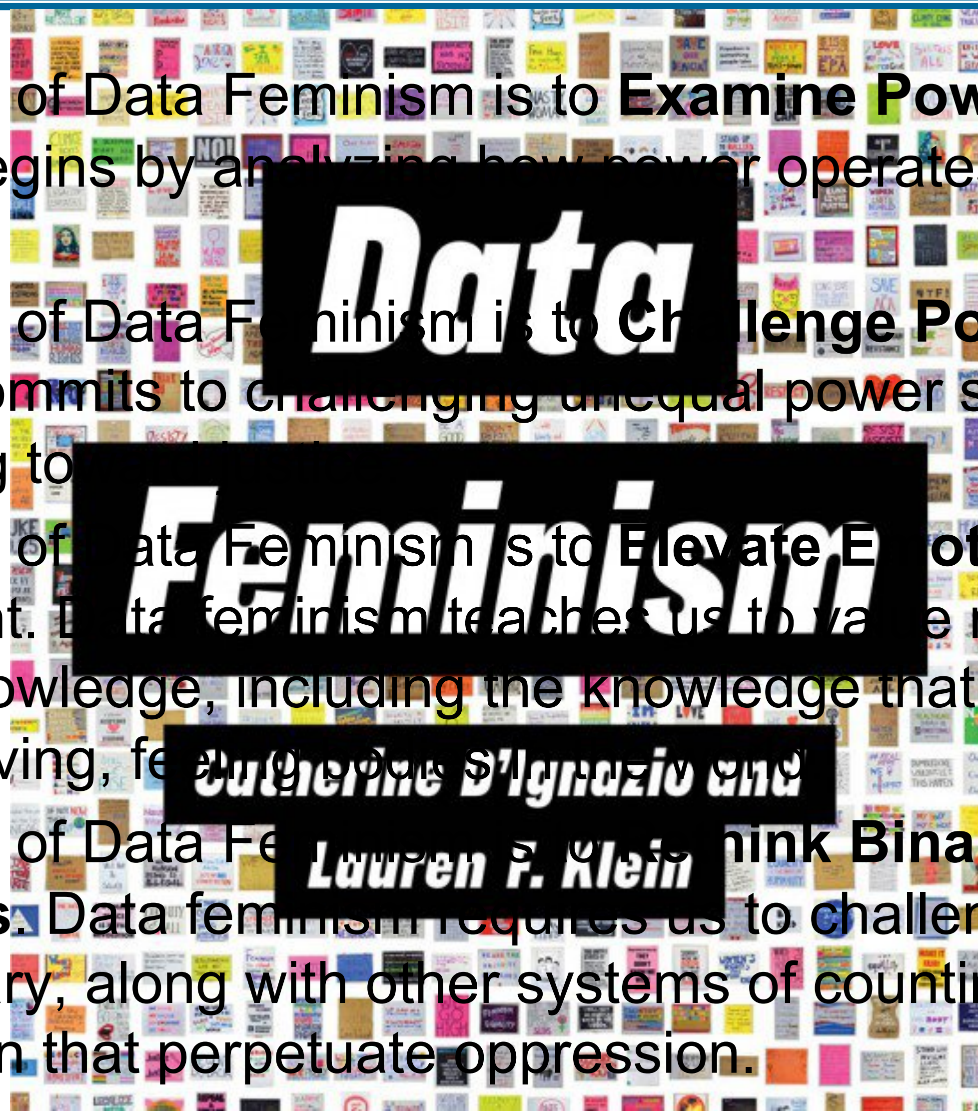
BONUS SLIDES: INSPIRING READINGS

Data Feminism and Hello World





8 Principles of Data Feminism



- Principle #1 of Data Feminism is to **Examine Power**. Data feminism begins by analyzing how power operates in the world.
- Principle #2 of Data Feminism is to **Challenge Power**. Data feminism commits to challenging unequal power structures and working to [illegible]
- Principle #3 of Data Feminism is to **Elevate Emotion** and Embodiment. Data feminism teaches us to value multiple forms of knowledge, including the knowledge that comes from people as living, feeling bodies in the world.
- Principle #4 of Data Feminism is to **Rethink Binaries and Hierarchies**. Data feminism requires us to challenge the gender binary, along with other systems of counting and classification that perpetuate oppression.

8 Principles of Data Feminism

- Principle #5 of Data Feminism is to **Embrace Pluralism**. Data feminism insists that the most complete knowledge comes from synthesizing multiple perspectives, with priority given to local, Indigenous, and experiential ways of knowing.
- Principle #6 of Data Feminism is to **Consider Context**. Data feminism asse[illegible]jective. They are the products of unequal social relations, and this context is essential for conducting accurate, ethical analysis.
- Principle #7 of Data Feminism is to **Make Labor Visible**. The work of data science, like all work in the world, is the work of many hands. Data fem[illegible] labor visible so that it can be recognized and [illegible]
- Principle #8 of Data Feminism is to **Multiply**