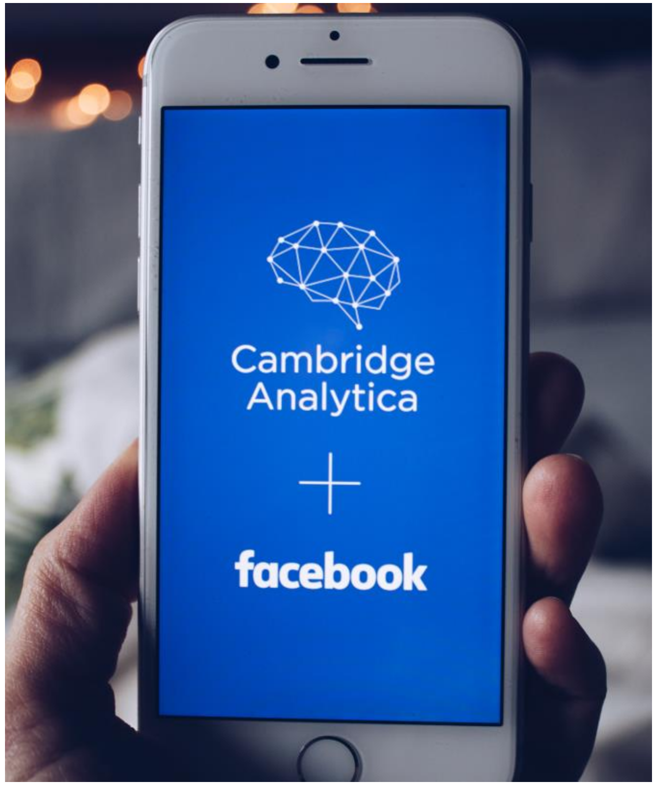
OLLSCOIL NA GAILLIMHE
UNIVERSITY OF GALWAY

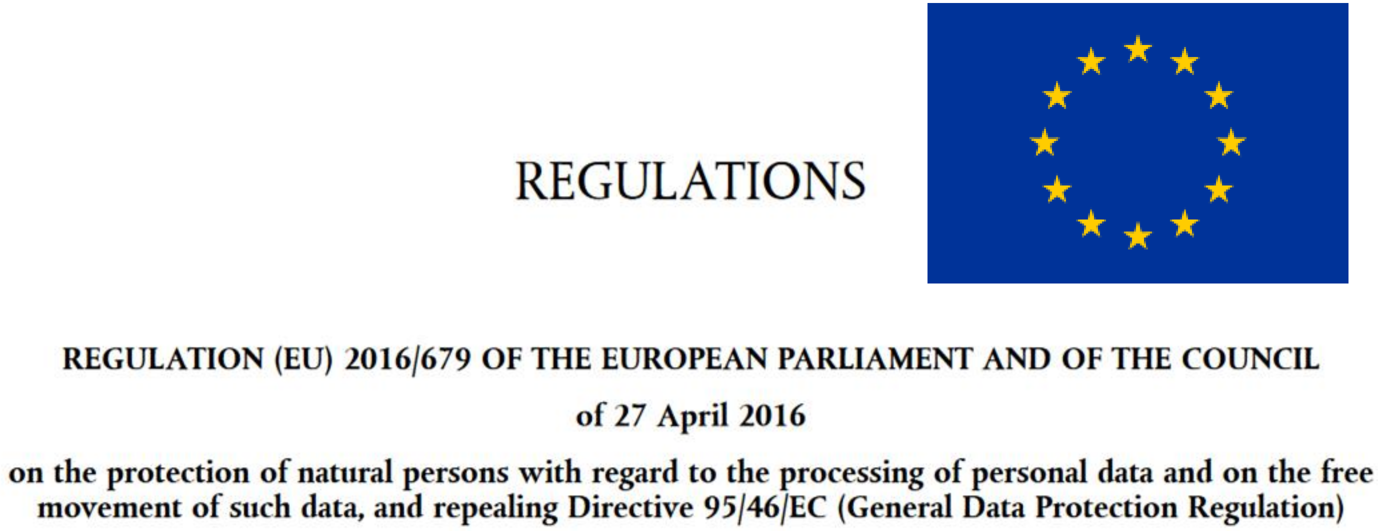
Ethics in NLP

Dr. Paul Buitelaar
Data Science Institute, University of Galway

Learning Objectives

- Understand ethical issues in NLP
- Understand some of the main approaches to address ethical issues in NLP through regulation and technology





REGULATIONS

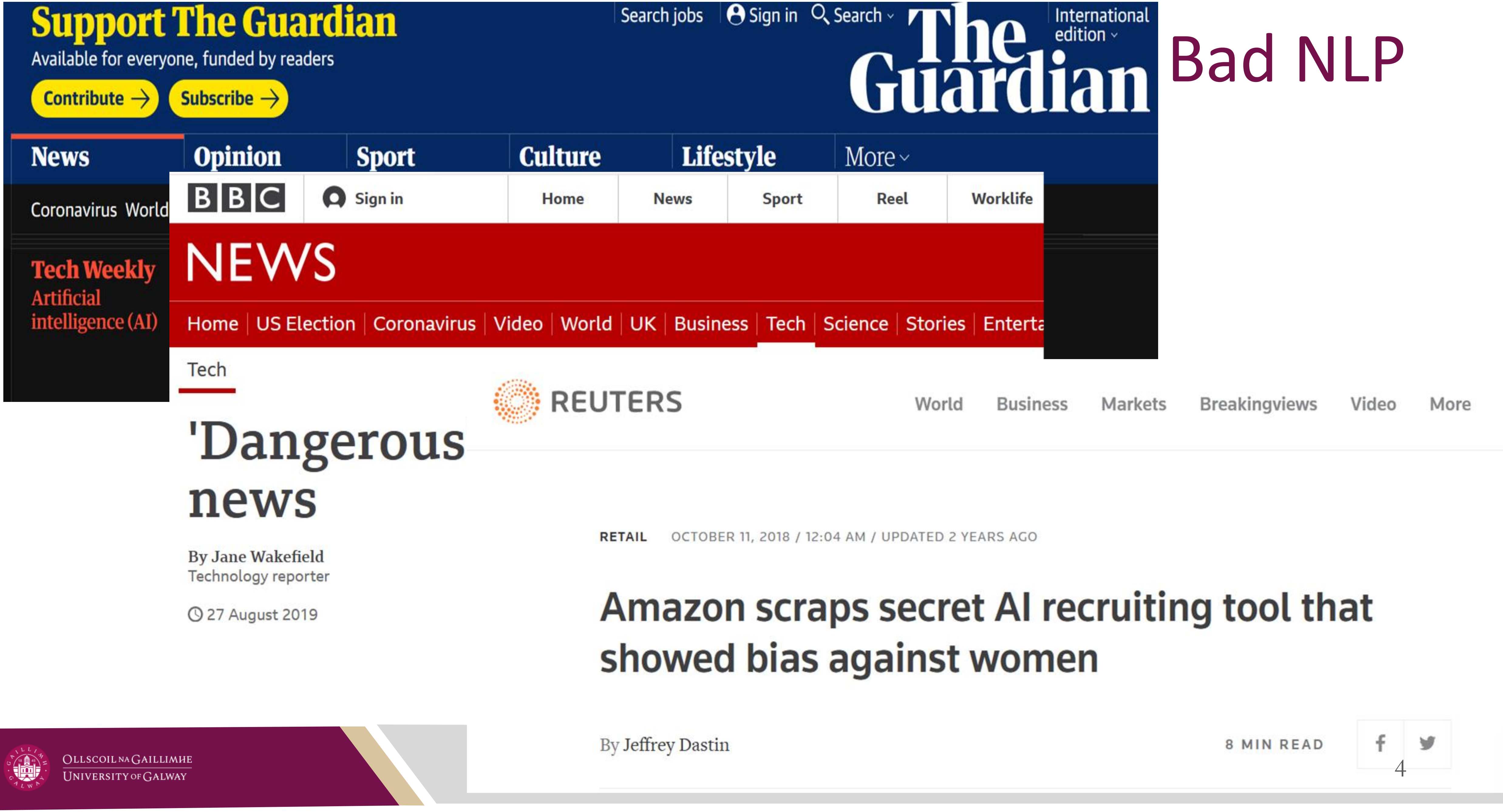
REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 27 April 2016

on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

Image sources: https://www.flickr.com/photos/bookcatalog/27162721118/ https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN

Bad NLP



Tech Weekly

Artificial intelligence (AI)

Tech

'Dangerous news

By Jane Wakefield

Technology reporter

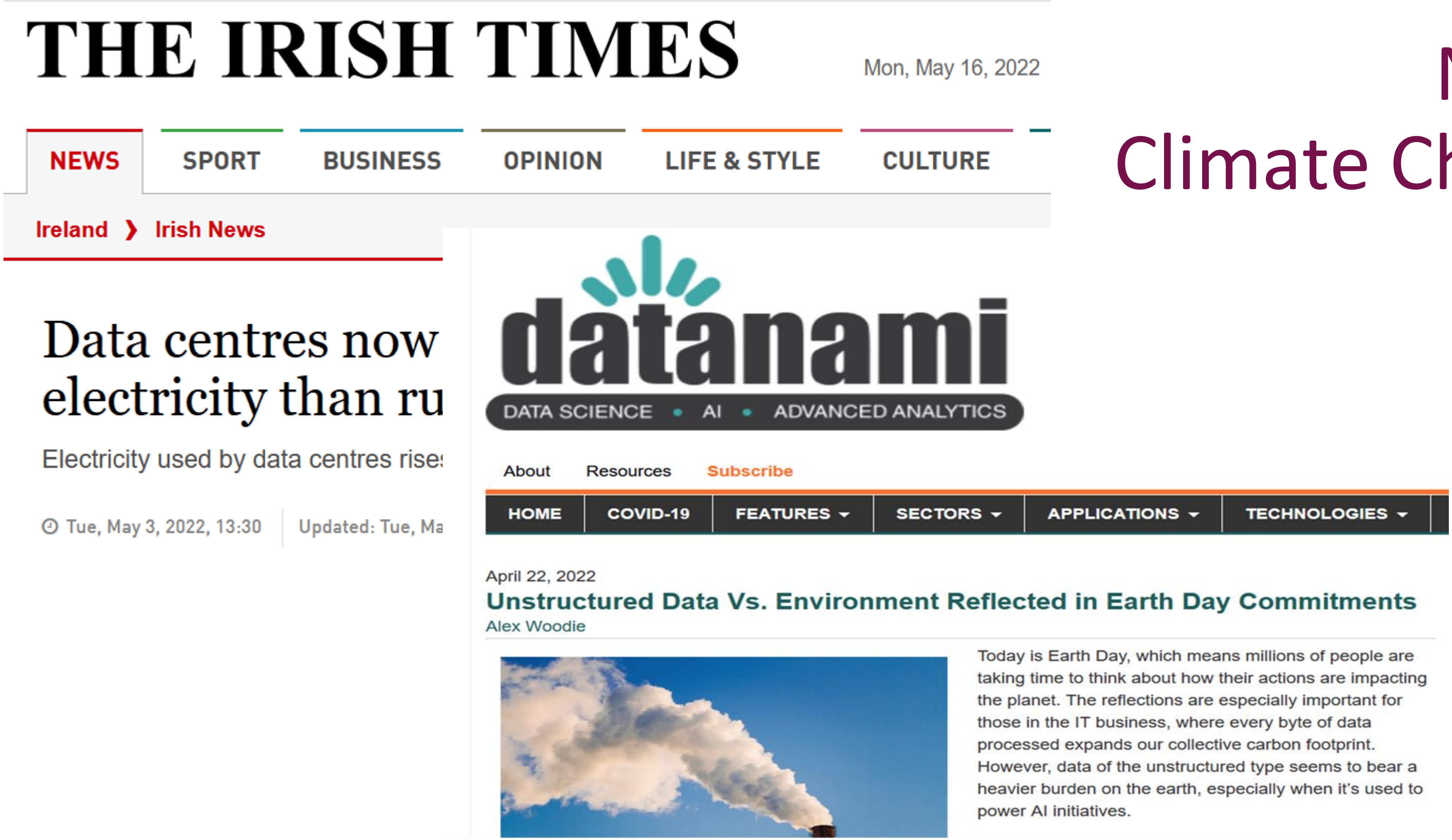
27 August 2019

RETAIL OCTOBER 11, 2018 / 12:04 AM / UPDATED 2 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

NLP & Climate Change



THE IRISH TIMES

Mon, May 16, 2022

Data centres now electricity than ru

Electricity used by data centres rise

Tue, May 3, 2022, 13:30 | Updated: Tue, Ma

datanami

DATA SCIENCE • AI • ADVANCED ANALYTICS

April 22, 2022

Unstructured Data Vs. Environment Reflected in Earth Day Commitments

Alex Woodie

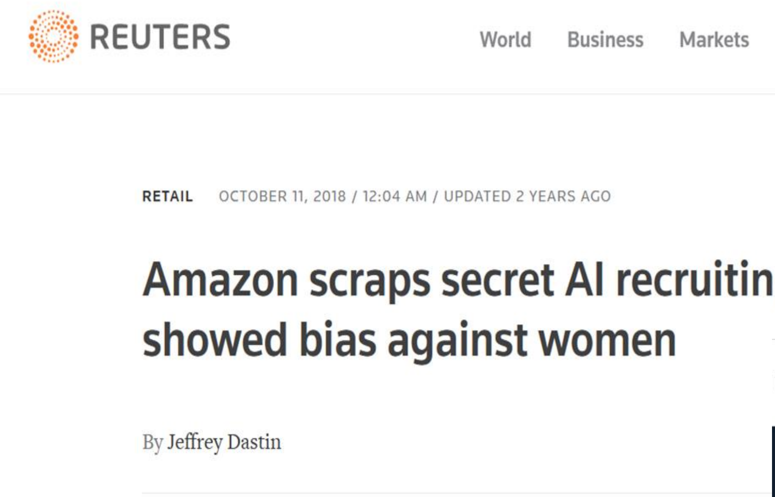
Today is Earth Day, which means millions of people are taking time to think about how their actions are impacting the planet. The reflections are especially important for those in the IT business, where every byte of data processed expands our collective carbon footprint. However, data of the unstructured type seems to bear a heavier burden on the earth, especially when it's used to power AI initiatives.

https://www.datanami.com/2022/04/22/unstructured-data-vs-environment-reflected-in-earth-day-commitments/

What can be done?

Regulation

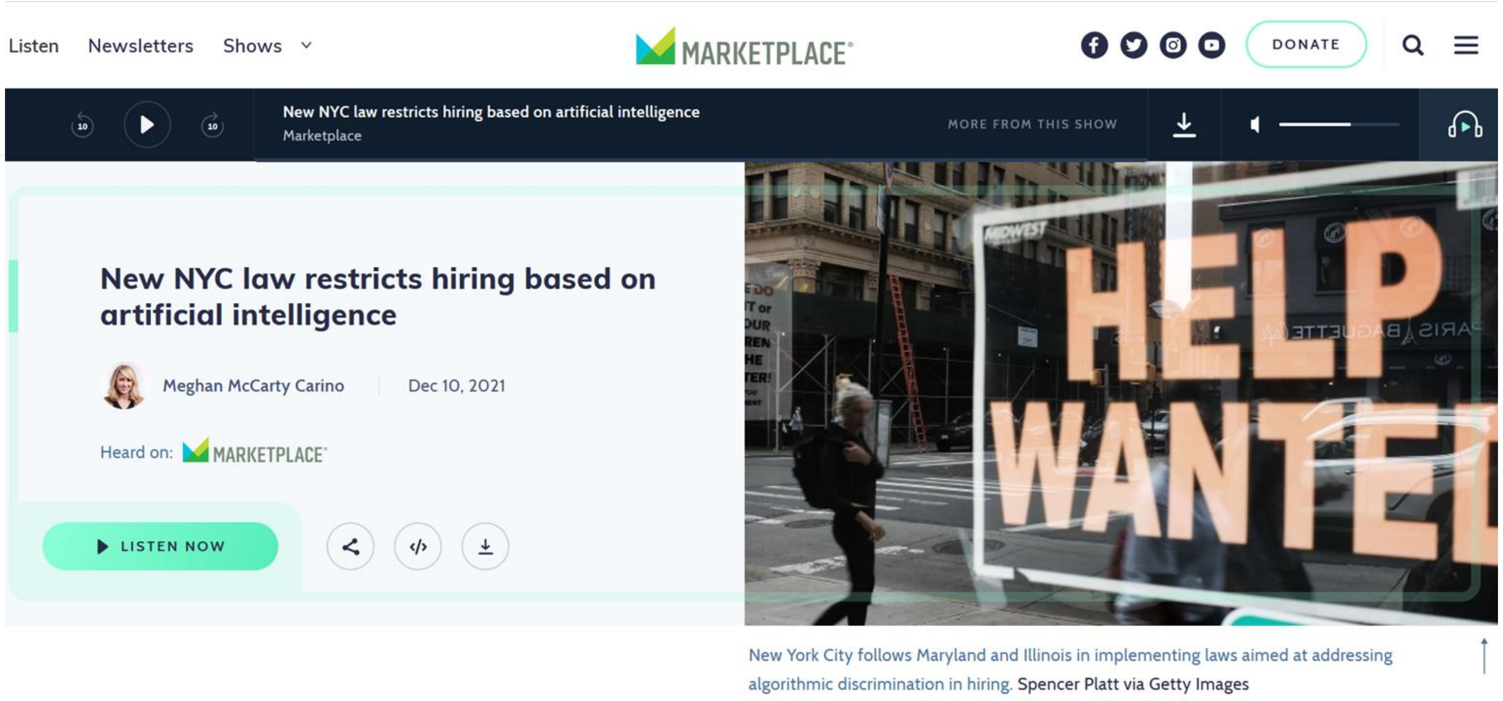
Regulate the use of AI



RETAIL OCTOBER 11, 2018 / 12:04 AM / UPDATED 2 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin



New NYC law restricts hiring based on artificial intelligence
Marketplace

New NYC law restricts hiring based on artificial intelligence

Meghan McCarty Carino | Dec 10, 2021

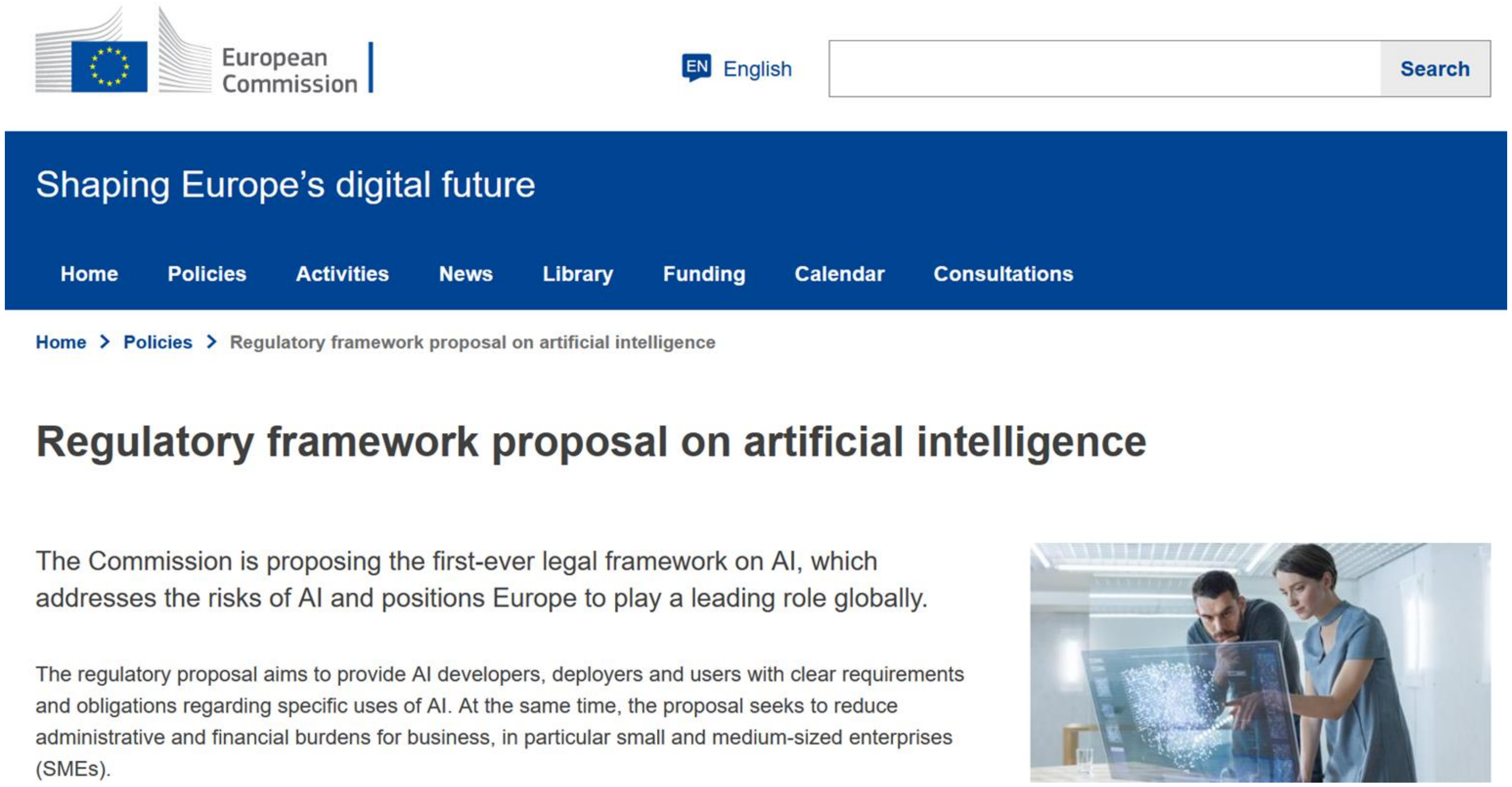
Heard on: MARKETPLACE

New York City follows Maryland and Illinois in implementing laws aimed at addressing algorithmic discrimination in hiring. Spencer Platt via Getty Images

“When a new law in New York City takes effect at the start of 2023, employers won’t be allowed to use artificial intelligence to screen job candidates **unless the tech has gone through an audit to check for bias**.” – Marketplace, Dec 2021

https://www.marketplace.org/2021/12/10/new-nyc-law-restricts-hiring-based-on-artificial-intelligence/

Regulation on AI Algorithms



European Commission
EN English
Search
Shaping Europe's digital future
Home
Policies
Activities
News
Library
Funding
Calendar
Consultations
Home > Policies > Regulatory framework proposal on artificial intelligence

Regulatory framework proposal on artificial intelligence

The Commission is proposing the first-ever legal framework on AI, which addresses the risks of AI and positions Europe to play a leading role globally.

The regulatory proposal aims to provide AI developers, deployers and users with clear requirements and obligations regarding specific uses of AI. At the same time, the proposal seeks to reduce administrative and financial burdens for business, in particular small and medium-sized enterprises (SMEs).

"The proposed rules will:

- address **risks** specifically created by AI applications;
- propose a list of **high-risk applications**;
- set clear **requirements** for AI systems for high risk applications;
- define specific **obligations** for AI users and providers of high risk applications;
- propose a **conformity assessment** before the AI system is put into service or placed on the market;
- propose **enforcement** after such an AI system is placed in the market;
- propose a **governance structure** at European and national level."

https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Regulation on Data Privacy

General Data Protection Regulation (GDPR)

“Regulation … 2016/679 of the European Parliament and of the Council ... regulates the processing by an **individual, a company or an organisation** of **personal data** relating to **individuals** in the EU.”

https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-does-general-data-protection-regulation-gdpr-govern_en

Personal Data

- “Personal data is **any information that relates to an identified or identifiable living individual**.
- Different pieces of **information, which collected together can lead to the identification of a particular person, also constitute personal data**.
- Personal data that has been **de-identified, encrypted or pseudonymised but can be used to re-identify a person remains personal data** and falls within the scope of the GDPR.
- **Personal data that has been rendered anonymous** in such a way that the individual is not or no longer identifiable **is no longer considered personal data**.
- For data to be truly anonymised, the **anonymisation must be irreversible**.”

https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-personal-data_en

Personal Data in NLP

Language data is also personal data as individuals can be identified by their language data use i.e., how they write or speak

Natural language processing can lead to ‘fingerprinting’ of individuals

Data Protection Impact Assessment

Identify **potential data privacy and data protection issues associated with the data collection for a specific (NLP) task** by answering questions such as:

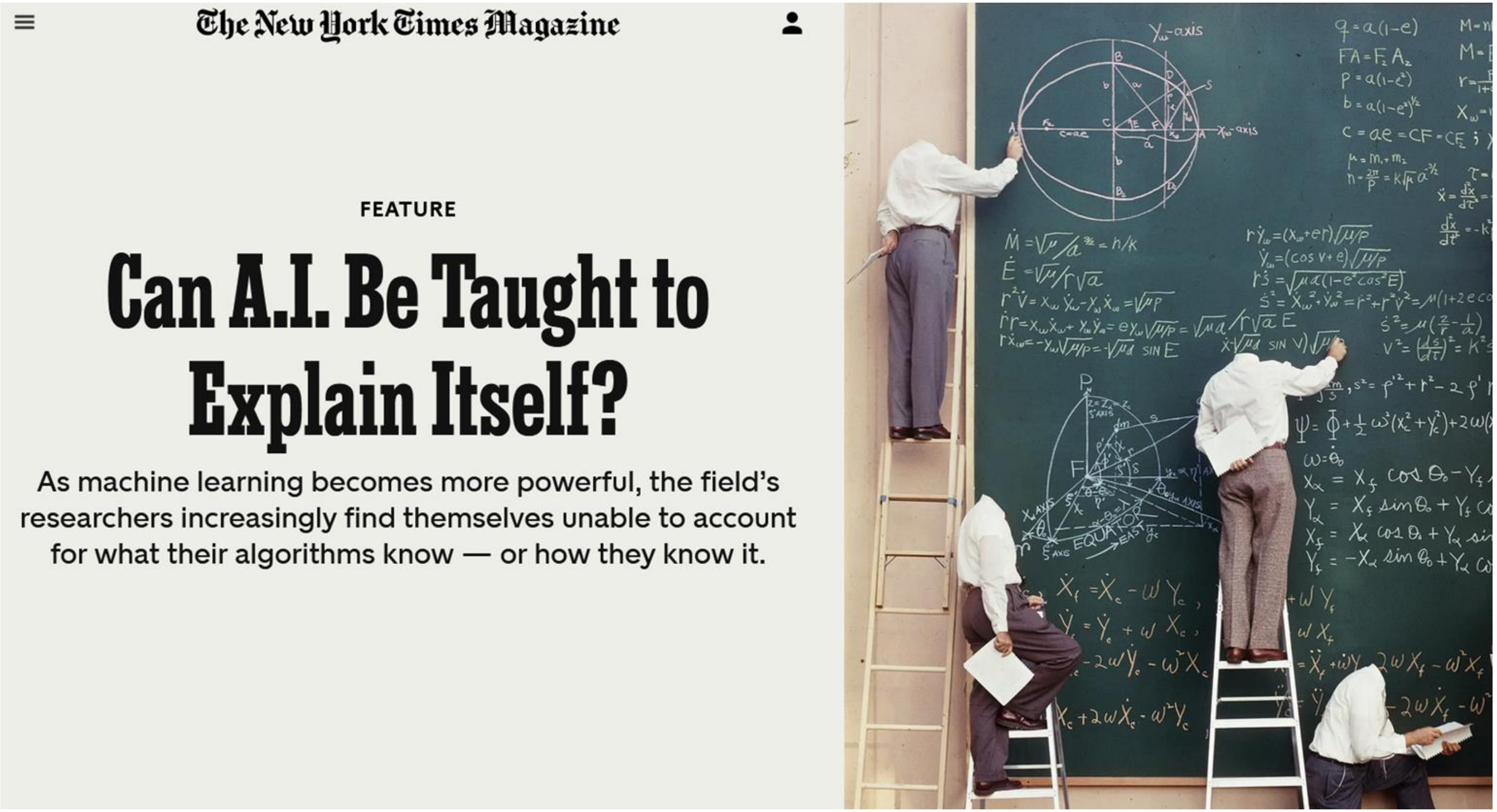
- *Have you identified the **minimally sufficient amount of data** for your purpose?*
- *Have you **informed affected individuals** what you are doing with their data?*
- *Have you **identified a secondary use for personal data** you collected?*
- *Have you established the **maximum time period required for keeping the data**?*
- *Have you developed **appropriate data anonymization strategies**?*

Data Protection Impact Assessment is to be discussed with and approved by the **Data Protection Officer** at the **Data Controller** (organization where the task is done).

https://www.dataprotection.ie/en/organisations/know-your-obligations/data-protection-impact-assessments

Technology

Explainable AI



The New York Times Magazine

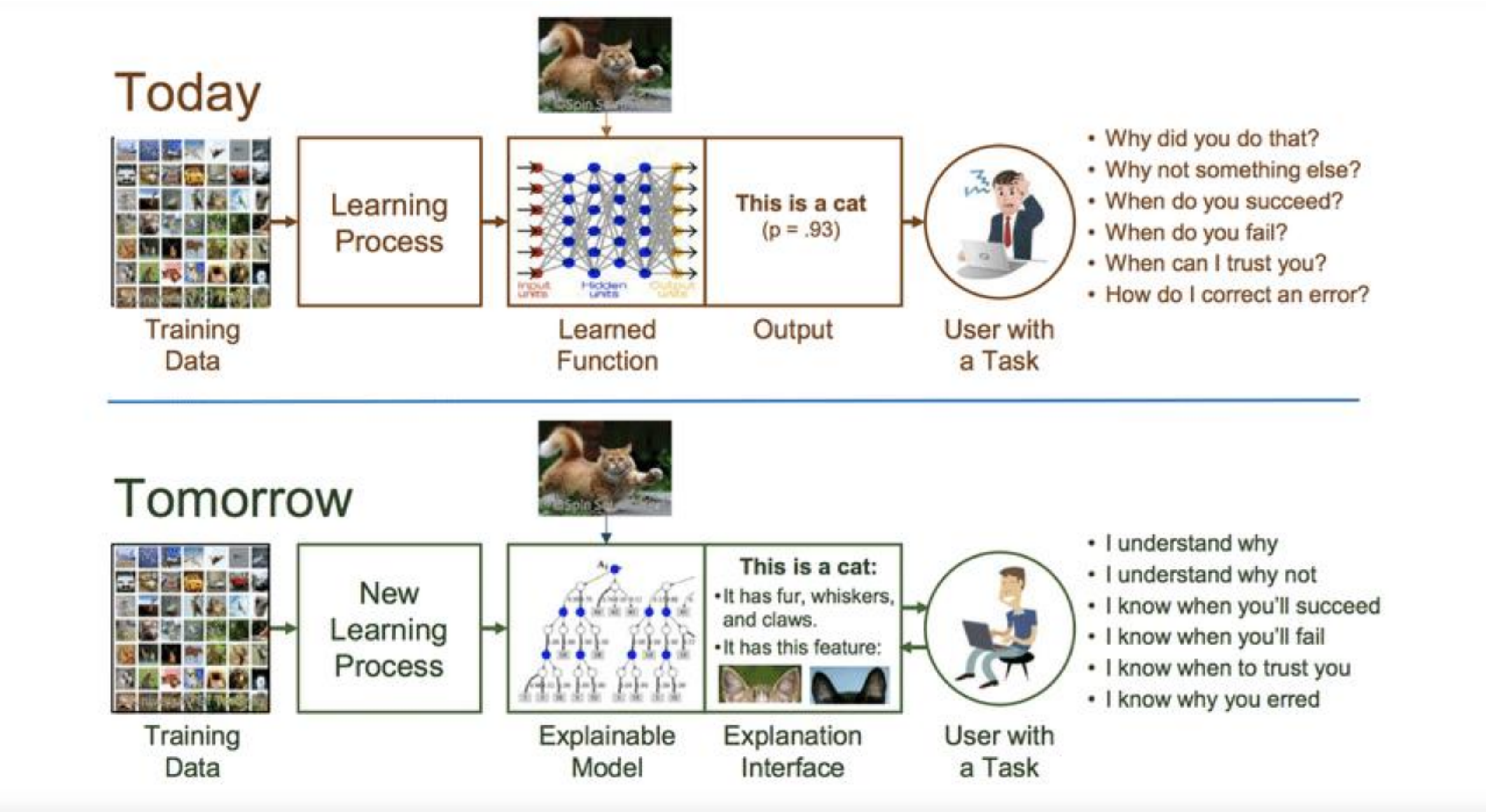
FEATURE

Can A.I. Be Taught to Explain Itself?

As machine learning becomes more powerful, the field's researchers increasingly find themselves unable to account for what their algorithms know — or how they know it.

https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html

XAI Concept by DARPA

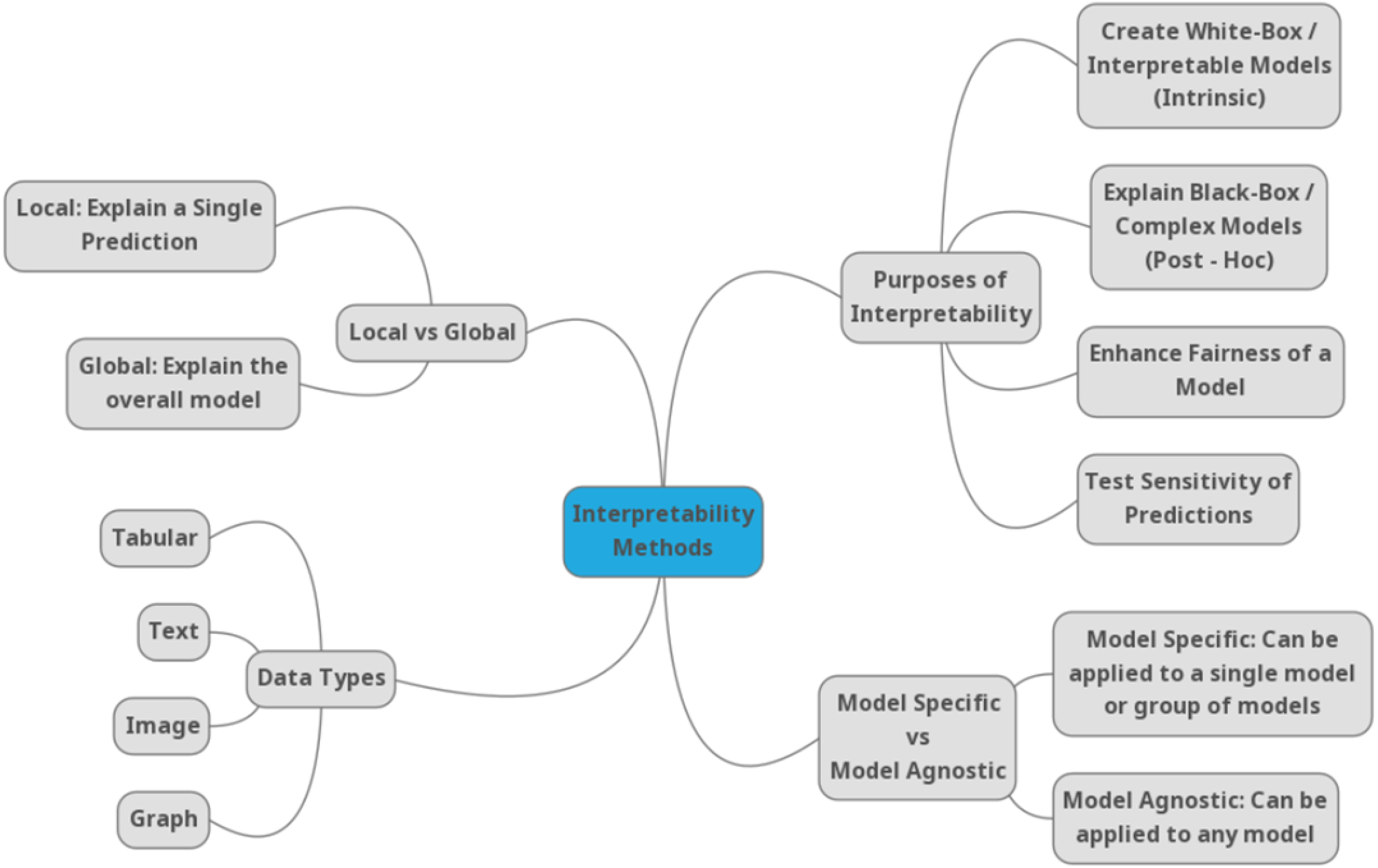


- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

- I understand why
- I understand why not
- I know when you'll succeed
- I know when you'll fail
- I know when to trust you
- I know why you erred

https://www.darpa.mil/program/explainable-artificial-intelligence

XAI Methods



Interpretability Methods
Local vs Global
Local: Explain a Single Prediction
Global: Explain the overall model
Data Types
Tabular
Text
Image
Graph
Purposes of Interpretability
Create White-Box / Interpretable Models (Intrinsic)
Explain Black-Box / Complex Models (Post - Hoc)
Enhance Fairness of a Model
Test Sensitivity of Predictions
Model Specific vs Model Agnostic
Model Specific: Can be applied to a single model or group of models
Model Agnostic: Can be applied to any model

Figure 2. Taxonomy mind-map of Machine Learning Interpretability Techniques.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, *23*(1), 18.

Limits to ‘transparency’ in Explainable AI

“Burrell (2016) distinguishes between three barriers to transparency:

*(1) **intentional concealment on the part of corporations or other institutions**, where decision-making procedures are kept from public scrutiny;*

*(2) **gaps in technical literacy**, which mean that, for most people, simply having access to underlying code is insufficient; and*

*(3) a “**mismatch between the mathematical optimization** in high-dimensionality characteristic of machine learning **and the demands of human-scale reasoning** and styles of interpretation.” …” - Goodman and Flaxman (2017)*

Goodman, B., & Flaxman, S. (2017). **European Union regulations on algorithmic decision-making and a “right to explanation”.** *AI magazine*, *38*(3), 50-57.

Burrell, J. 2016. **How the Machine “Thinks“: Understanding Opacity in Machine Learning Algorithms**. Big Data and Society 3(1)

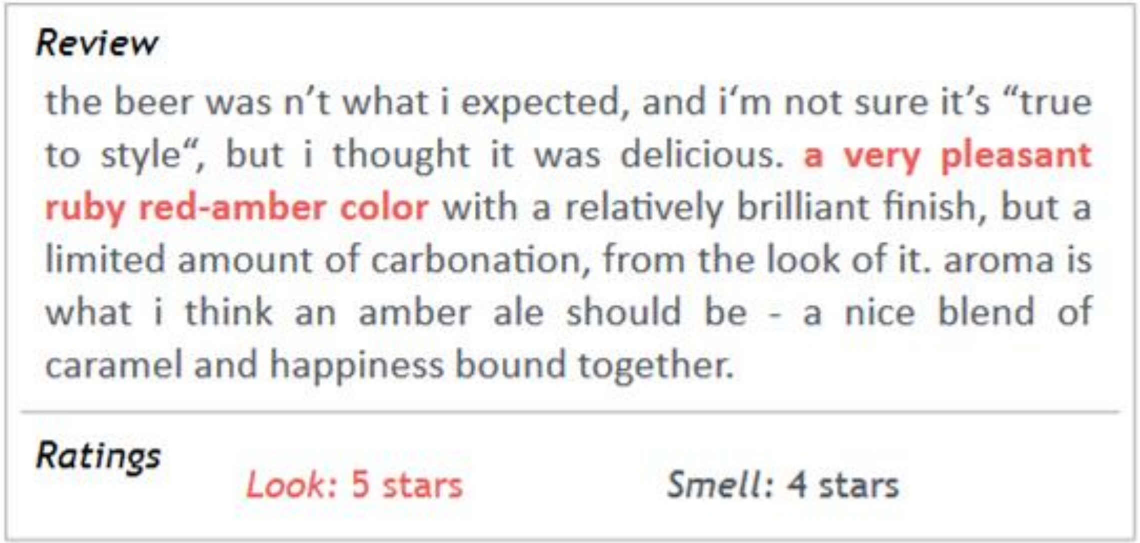
Rational Extraction

“ … we learn to extract pieces of input text as justifications – rationales – that are tailored to be short and coherent, yet sufficient for making the same prediction.

… In order for the subset to qualify as a rationale it should satisfy two criteria:

1) the selected words should be interpretable and

2) they ought to suffice to reach nearly the same prediction (target vector) as the original input. ...”



Review

the beer was n’t what i expected, and i‘m not sure it’s “true to style“, but i thought it was delicious. **a very pleasant ruby red-amber color** with a relatively brilliant finish, but a limited amount of carbonation, from the look of it. aroma is what i think an amber ale should be - a nice blend of caramel and happiness bound together.

Ratings

Look: 5 stars

Smell: 4 stars

Figure 1: An example of a review with ranking in two categories. The rationale for Look prediction is shown in bold.

Lei, T.; Barzilay, R.; Jaakkola, T.S. (2016) **Rationalizing Neural Predictions**. *In Proceedings EMNLP, (pp. 107–117).*

Data Analysis
- Trustworthy AI -

Definition of Trustworthy AI

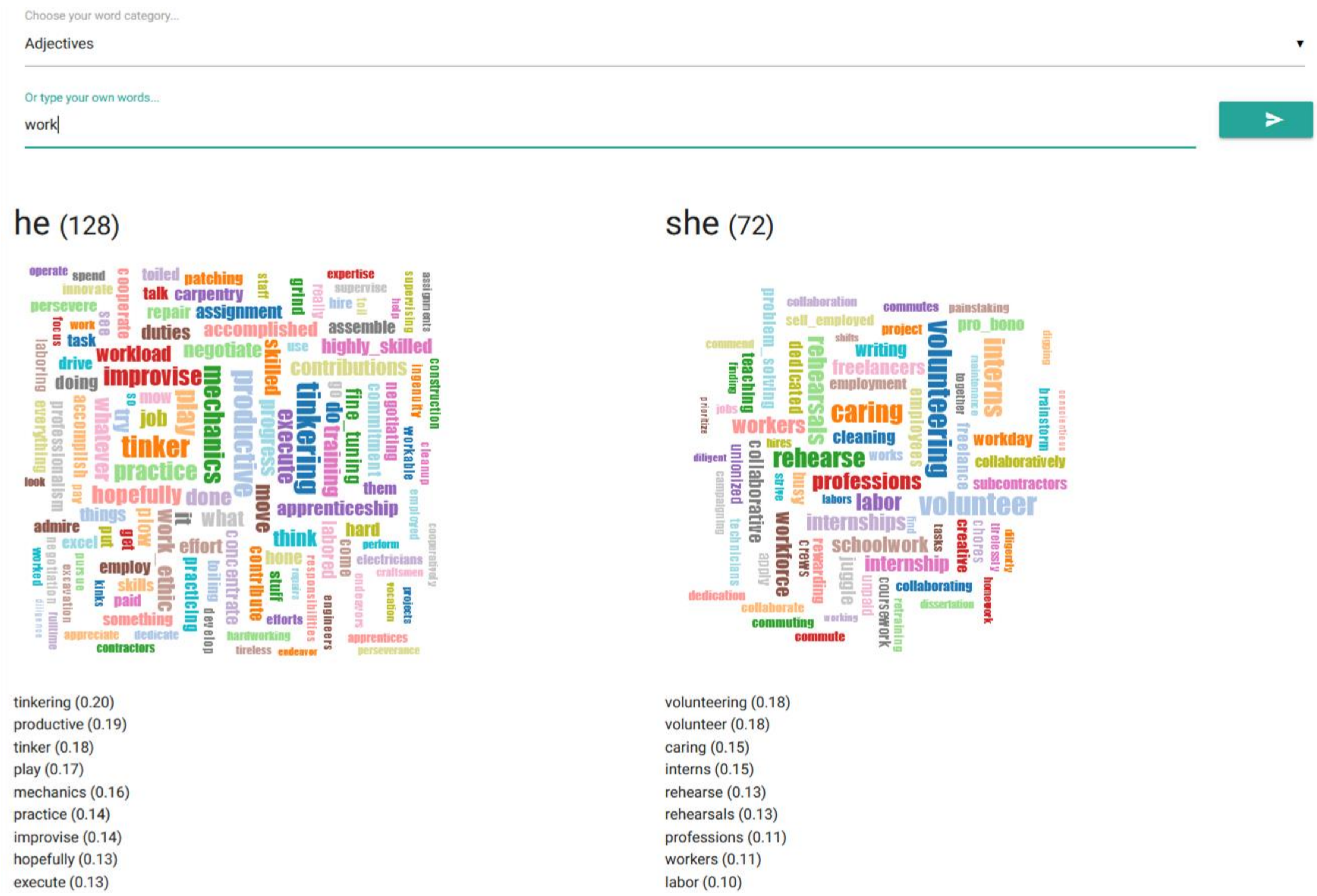
ETHICS GUIDELINES FOR TRUSTWORTHY AI: High-Level Expert Group on Artificial Intelligence

“Trustworthy AI has three components …:
1. it should be **lawful**, complying with all applicable laws and regulations ;
2. it should be **ethical**, ensuring adherence to ethical principles and values; and
3. it should be **robust**, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm.“

“In an AI context, equality entails that the system’s operations cannot generate unfairly biased outputs (e.g. the **data used to train AI systems should be as inclusive as possible, representing different population groups**).”

https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf

Bias in Data



Choose your word category...

Adjectives ▾

Or type your own words...

work

➤

he (128)

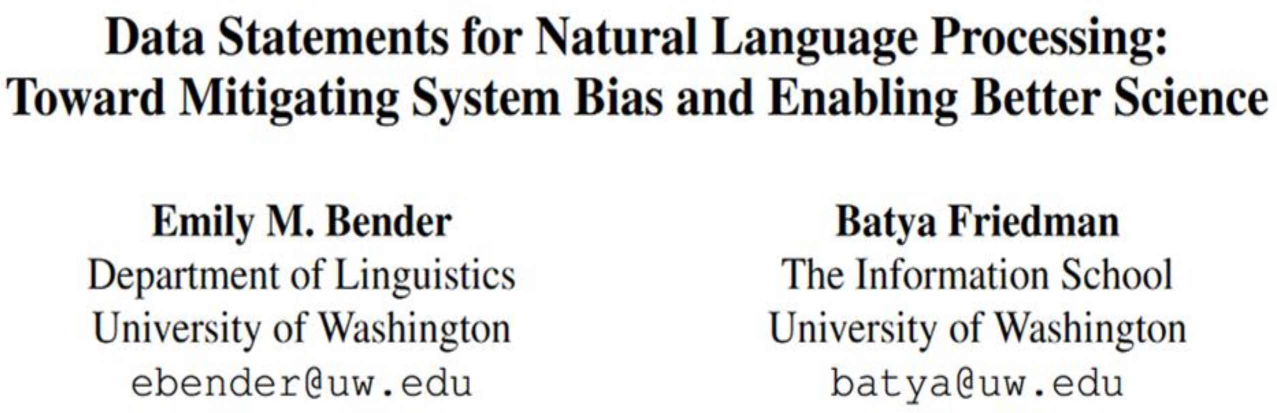
tinkering (0.20)
productive (0.19)
tinker (0.18)
play (0.17)
mechanics (0.16)
practice (0.14)
improvise (0.14)
hopefully (0.13)
execute (0.13)

she (72)

volunteering (0.18)
volunteer (0.18)
caring (0.15)
interns (0.15)
rehearse (0.13)
rehearsals (0.13)
professions (0.11)
workers (0.11)
labor (0.10)

http://wordbias.umiacs.umd.edu/

Data Statements



Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science

Emily M. Bender
Department of Linguistics
University of Washington
ebender@uw.edu

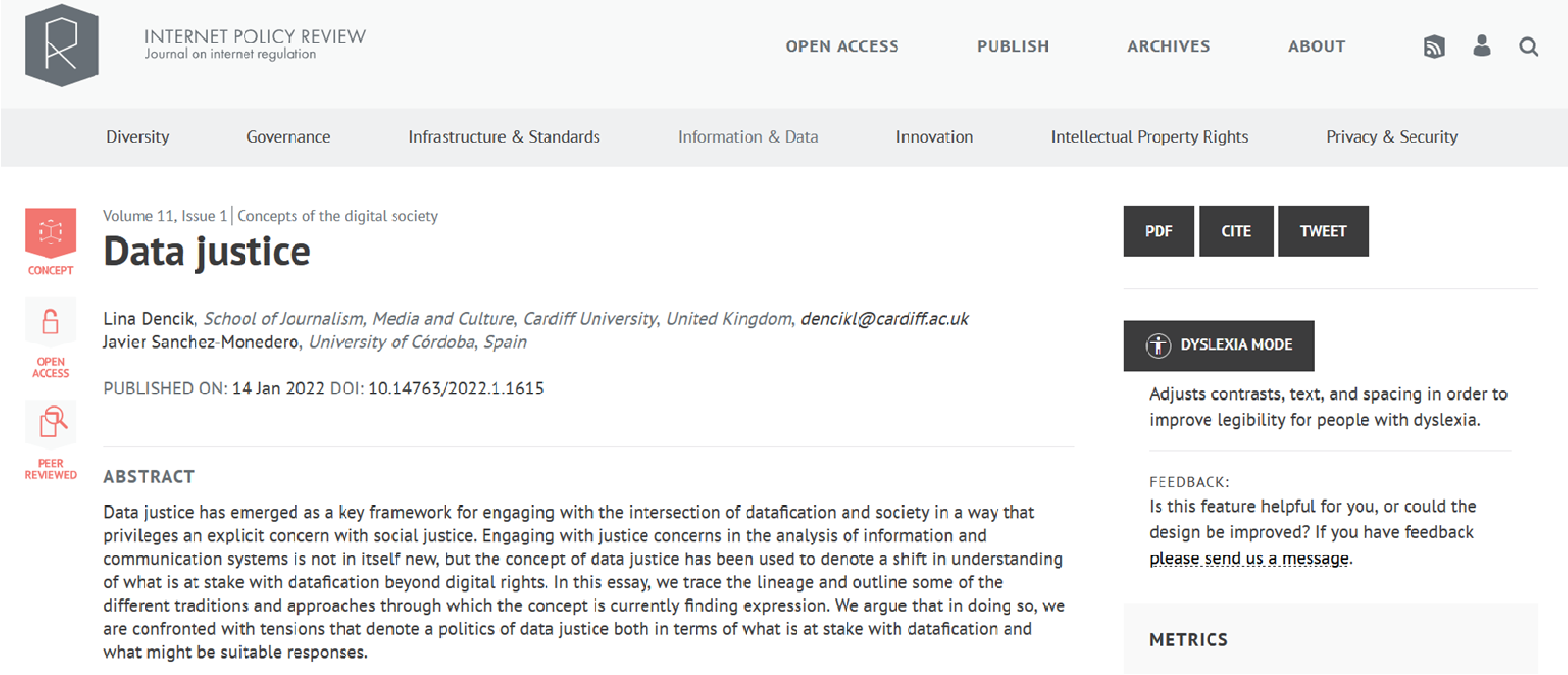
Batya Friedman
The Information School
University of Washington
batya@uw.edu

“A **data statement** is a characterization of a dataset that provides context to allow developers and users to better understand ... **what biases might be reflected in systems built on the software**.”

“We propose here that **foregrounding the characteristics of our datasets** can help, by allowing reasoning about what the **likely effects** may be and by making it clearer **which populations are and are not represented**, for both training and test data.”

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, *6*, 587-604.

Data Justice



Volume 11, Issue 1 | Concepts of the digital society

Data justice

Lina Dencik, *School of Journalism, Media and Culture, Cardiff University, United Kingdom*, *dencikl@cardiff.ac.uk*
Javier Sanchez-Monedero, *University of Córdoba, Spain*

PUBLISHED ON: 14 Jan 2022 DOI: 10.14763/2022.1.1615

ABSTRACT

Data justice has emerged as a key framework for engaging with the intersection of datafication and society in a way that privileges an explicit concern with social justice. Engaging with justice concerns in the analysis of information and communication systems is not in itself new, but the concept of data justice has been used to denote a shift in understanding of what is at stake with datafication beyond digital rights. In this essay, we trace the lineage and outline some of the different traditions and approaches through which the concept is currently finding expression. We argue that in doing so, we are confronted with tensions that denote a politics of data justice both in terms of what is at stake with datafication and what might be suitable responses.

Adjusts contrasts, text, and spacing in order to improve legibility for people with dyslexia.

FEEDBACK:
Is this feature helpful for you, or could the design be improved? If you have feedback please send us a message.

METRICS

Dencik, L., & Sanchez-Monedero, J. (2021). Data justice. *Internet Policy Review*, *11*(1), 1-16.

Linguistic Justice

"… equitable access to social, economic, and political life regardless of linguistic repertoire"

"… linguistically just ***NLP tools must****:*

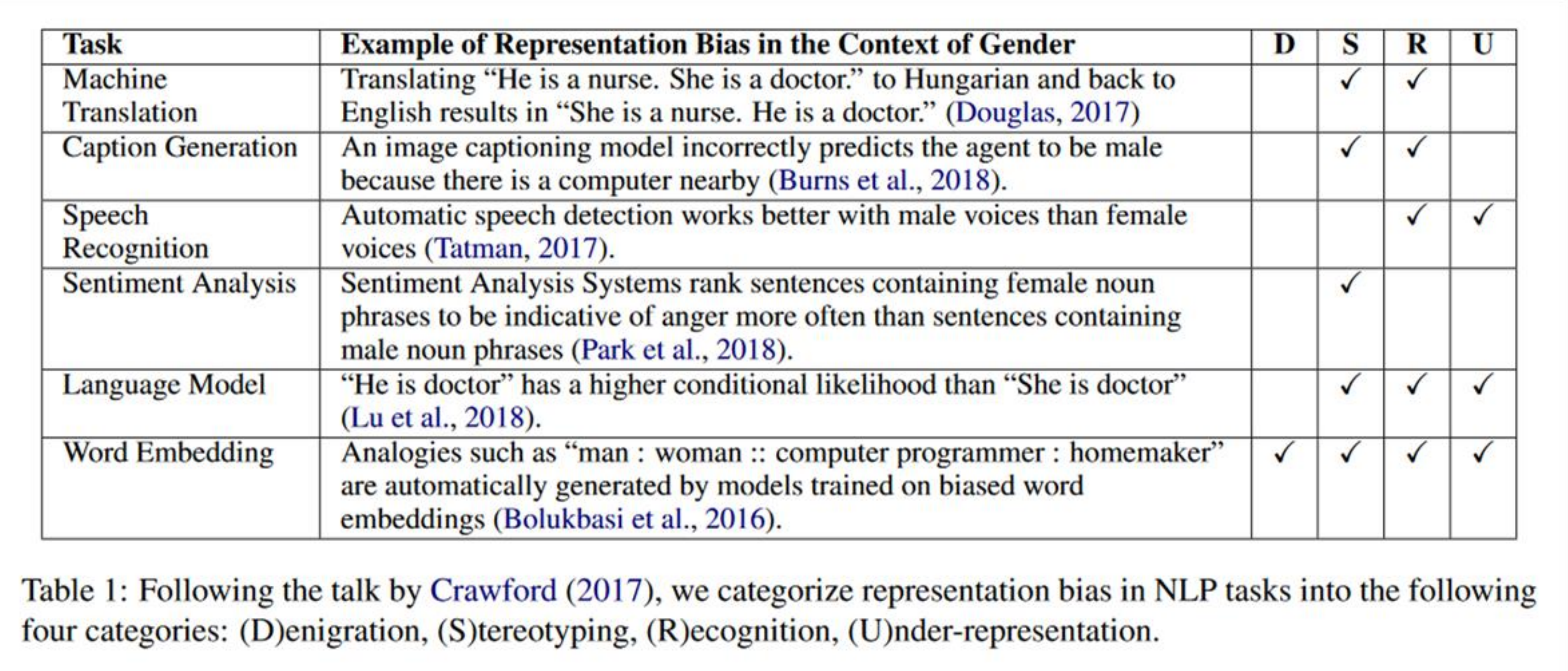
(1) ***work well for users regardless of language variety*** *they use and*

(2) work to ***counteract inequities based on language use*** *in decision-making and resource allocation.*

To build ***linguistically just NLP tools****, we must recognize and address power inequities such as* ***over/underrepresentation of linguistic patterns*** *and discourses within datasets.“*

Nee, J., Smith, G. M., Sheares, A., & Rustagi, I. (2022). **Linguistic justice as a framework for designing, developing, and managing natural language processing tools.** *Big Data & Society*, *9*(1). ; GAZZOLA, MICHELE; WICKSTRÖM, BENGT-ARNE; FETTES, Mark. **Towards an index of linguistic justice**. *Working Paper no. 20-1, Research Group "Economics and Language"(REAL)*, 2020.

Bias in NLP Data Sets

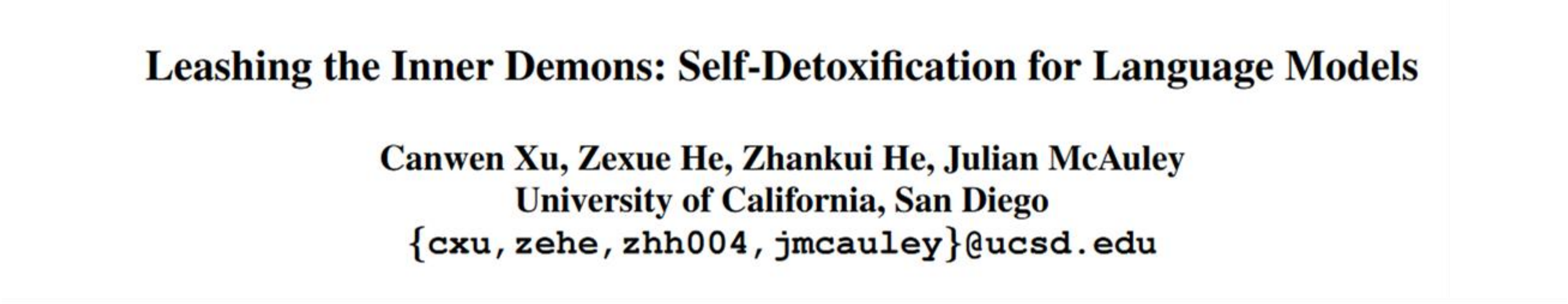


Task	Example of Representation Bias in the Context of Gender	D	S	R	U
Machine Translation	Translating "He is a nurse. She is a doctor." to Hungarian and back to English results in "She is a nurse. He is a doctor." (Douglas, 2017)		✓	✓	
Caption Generation	An image captioning model incorrectly predicts the agent to be male because there is a computer nearby (Burns et al., 2018).		✓	✓	
Speech Recognition	Automatic speech detection works better with male voices than female voices (Tatman, 2017).			✓	✓
Sentiment Analysis	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).		✓		
Language Model	"He is doctor" has a higher conditional likelihood than "She is doctor" (Lu et al., 2018).		✓	✓	✓
Word Embedding	Analogies such as "man : woman :: computer programmer : homemaker" are automatically generated by models trained on biased word embeddings (Bolukbasi et al., 2016).	✓	✓	✓	✓

Table 1: Following the talk by Crawford (2017), we categorize representation bias in NLP tasks into the following four categories: (D)enigration, (S)tereotyping, (R)ecognition, (U)nder-representation.

Sun, T., Gaut, A., Tang, S., *Computational Linguistics (ACL 2019)*. Huang, Y., ElSherief, M., Zhao, J., Mirza D., Belding E., Chang K-W, & Wang, W. Y. (2019). **Mitigating Gender Bias in Natural Language Processing: Literature Review**. In *Proceedings of the 57th ACL Conference* (pp. 1630-1640).

Data Detoxing



Leashing the Inner Demons: Self-Detoxification for Language Models

Canwen Xu, Zexue He, Zhankui He, Julian McAuley
University of California, San Diego
{cxu,zehe,zhh004,jmcauley}@ucsd.edu

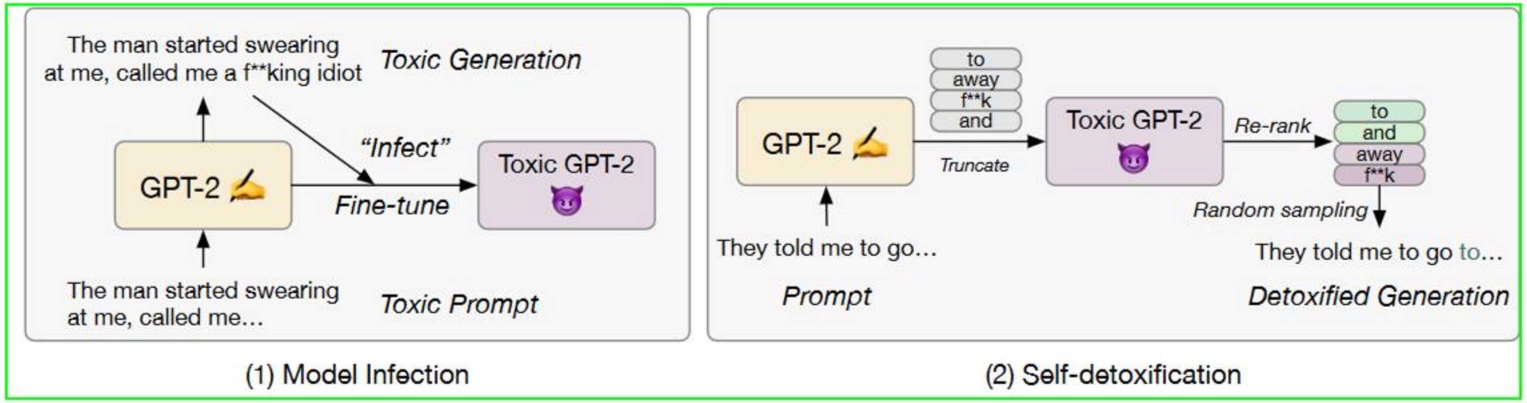
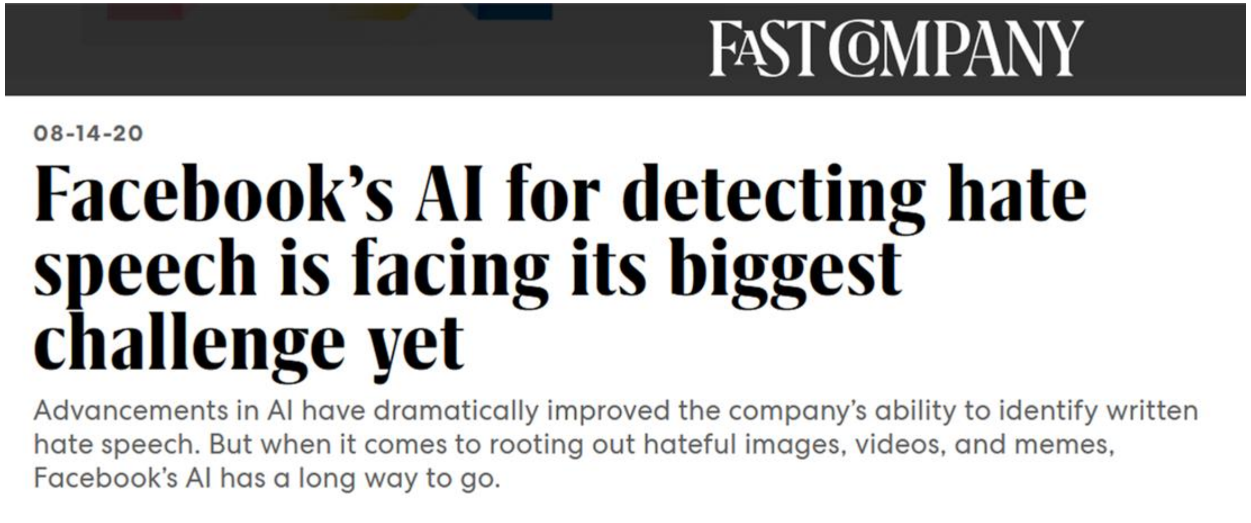


Figure 1: The workflow of self-detoxification. (1) We feed toxic prompts to the pretrained GPT-2 model to encourage toxic content to be generated. Then, we fine-tune a GPT-2 model on the generated toxic content and obtain an "infected" toxic GPT-2. (2) When doing self-toxification, the original GPT-2 model generates a probability distribution for the next token. After applying top-k truncation, we use the toxic GPT-2 to score the token candidates and re-rank. Therefore, the words that are less favored by the toxic GPT-2 would have a better chance to be generated.

Xu C, He Z, He Z, McAuley J. (2022) **Leashing the Inner Demons: Self-Detoxification for Language Models.** *AAAI Conference on Artificial Intelligence.*

AI for Good
- NLP in Ethical Use Cases -

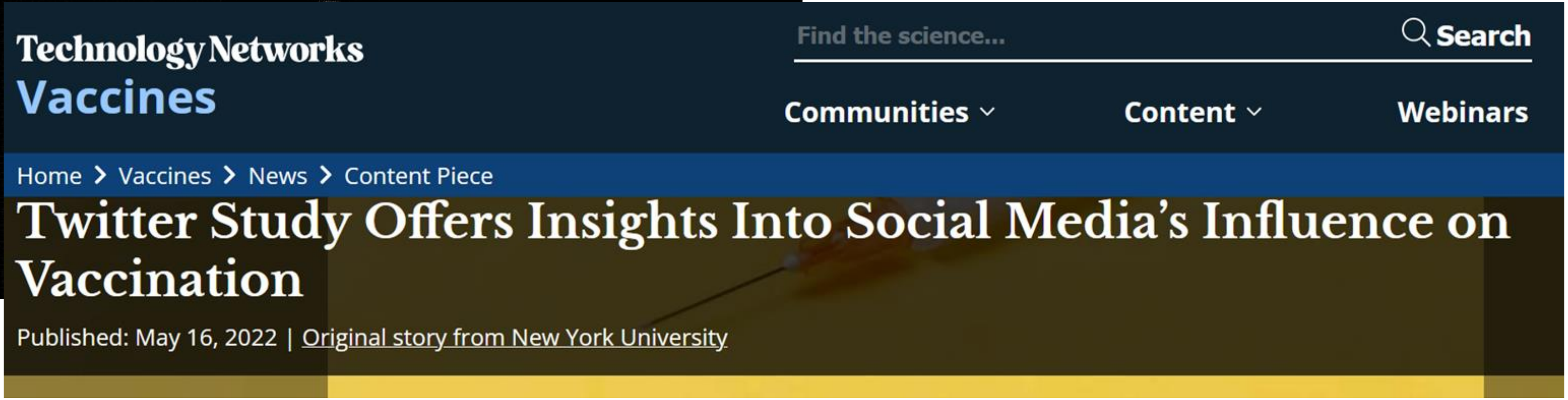
Good NLP


FAST COMPANY

08-14-20

Facebook’s AI for detecting hate speech is facing its biggest challenge yet

Advancements in AI have dramatically improved the company’s ability to identify written hate speech. But when it comes to rooting out hateful images, videos, and memes, Facebook’s AI has a long way to go.

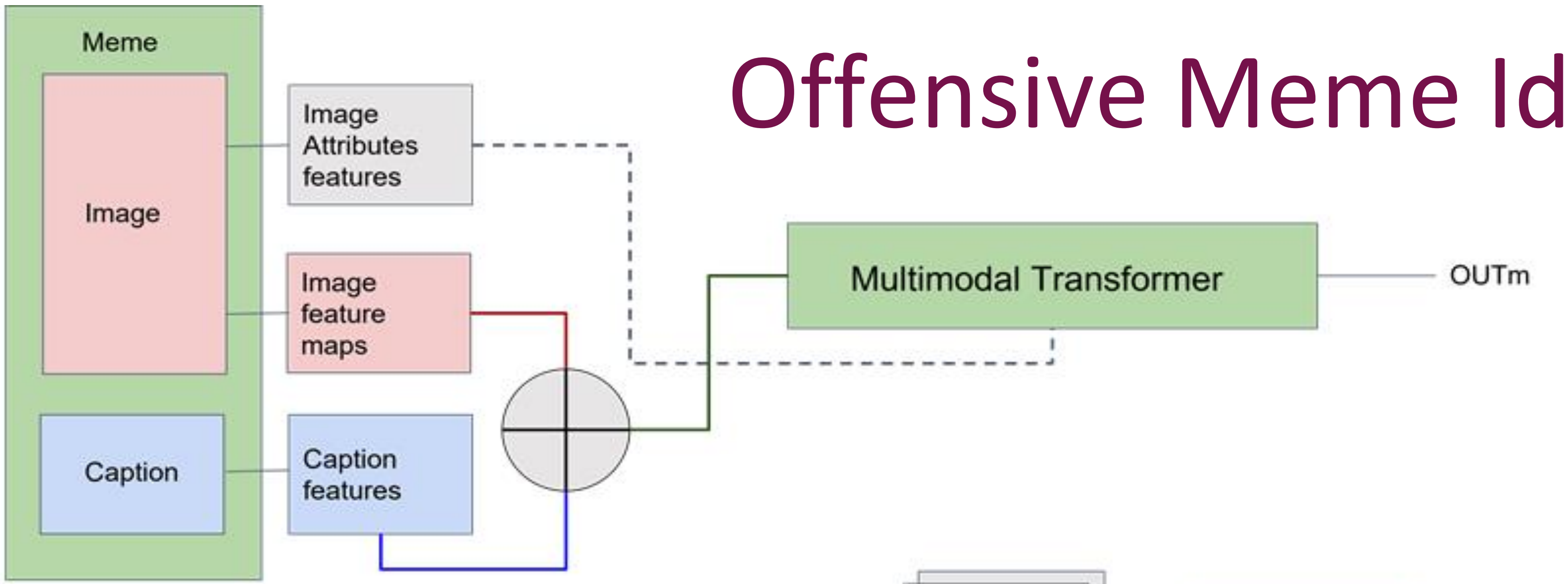


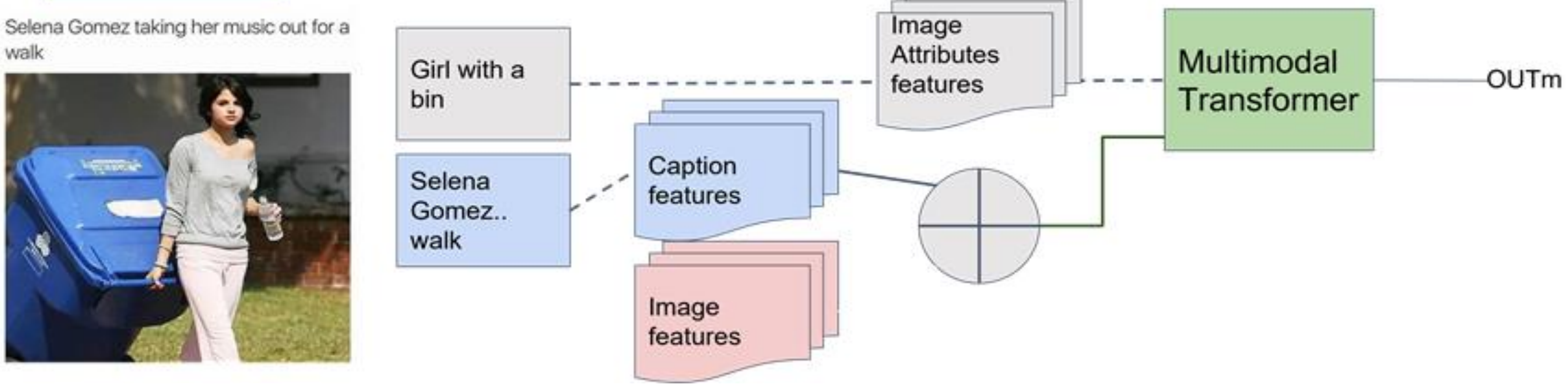
Twitter Study Offers Insights Into Social Media’s Influence on Vaccination

Published: May 16, 2022 | Original story from New York University

https://www.fastcompany.com/90539275/facebooks-ai-for-detecting-hate-speech-is-facing-its-biggest-challenge-yet
https://www.technologynetworks.com/vaccines/news/twitter-study-offers-insights-into-social-medias-influence-on-vaccination-361603

Offensive Meme Identification





Suryawanshi, S., Chakravarthi, B. R., Arcan, M., & Buitelaar, P. (2020). **Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text**. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying* (pp. 32-41).

Data Ethics Challenges

Bias in training data, resulting in **biased hate speech predictions**

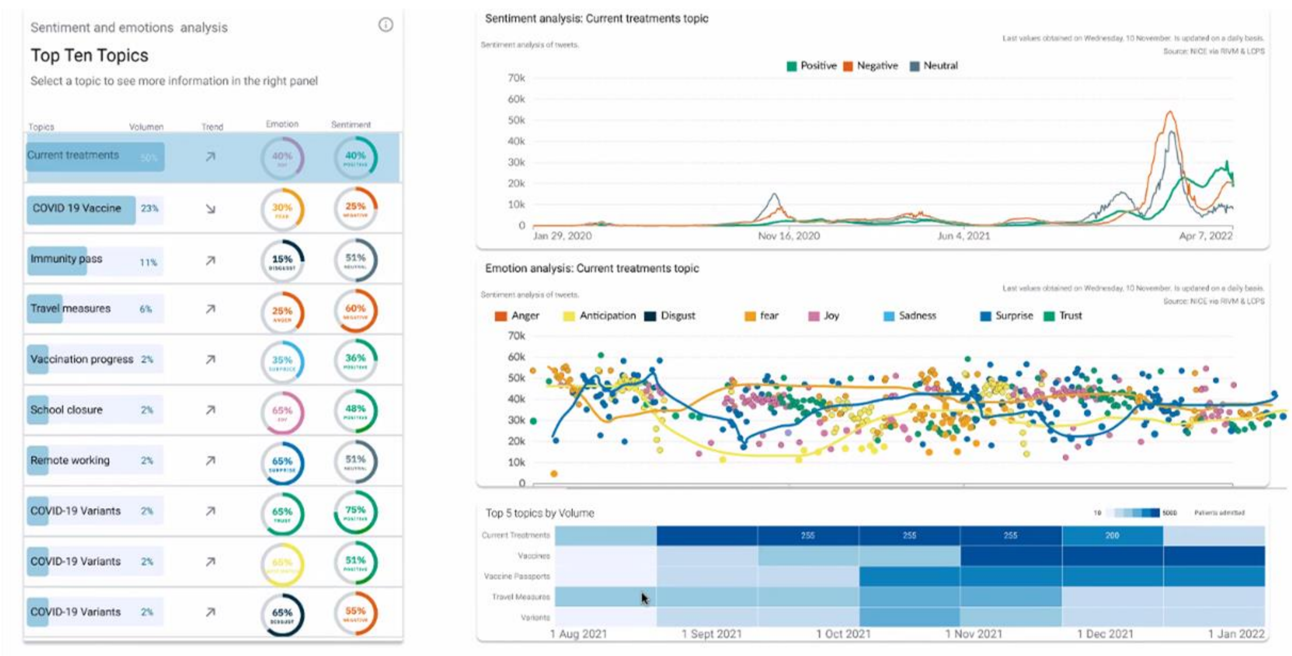
- *"... **dialect can lead to racial bias** in automatic hate speech detection models … unexpected correlations between surface markers of African American English …" – Sap et al 2019*
- *"... **impact of political bias** on hate speech classification by constructing three politically-biased data sets (left-wing, right-wing, politically neutral) … political bias negatively impairs the performance of hate speech classifiers …" – Wich et al 2020*

Sap, Maarten, Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, A. N. **The risk of racial bias in hate speech detection**. *ACL*. 2019. ; Wich, M., Bauer, J., & Groh, G. (2020). **Impact of politically biased data on hate speech classification**. In *Proceedings of the 4th workshop on online abuse and harms* (pp. 54-64).

PANDEM-2
PANDEMIC PREPAREDNESS AND RESPONSE

NLP for Public Health

Social media analysis (emotion/sentiment, suggestions) to identify **‘two-way’ communication aspects around pandemics**, such as reactions to government measures



Sentiment and emotions analysis

Top Ten Topics

Select a topic to see more information in the right panel

Topics	Volumen	Trend	Emotion	Sentiment
Current treatments	50%	↗	40% JOY	40% POSITIVE
COVID 19 Vaccine	23%	↘	30% FEAR	25% NEGATIVE
Immunity pass	11%	↗	15% DISGUST	51% NEUTRAL
Travel measures	6%	↗	25% ANGER	60% NEGATIVE
Vaccination progress	2%	↗	35% SURPRISE	36% POSITIVE
School closure	2%	↗	65% JOY	48% POSITIVE
Remote working	2%	↗	65% SURPRISE	51% NEUTRAL
COVID-19 Variants	2%	↗	65% TRUST	75% POSITIVE
COVID-19 Variants	2%	↗	65% ANTICIPATION	51% POSITIVE
COVID-19 Variants	2%	↗	65% DISGUST	55% NEGATIVE

Sentiment analysis: Current treatments topic

Sentiment analysis of tweets.

Last values obtained on Wednesday, 10 November. Is updated on a daily basis.
Source: NICE via RIVM & LCPS

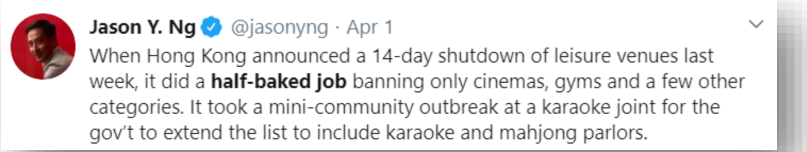
Emotion analysis: Current treatments topic

Sentiment analysis of tweets.

Last values obtained on Wednesday, 10 November. Is updated on a daily basis.
Source: NICE via RIVM & LCPS

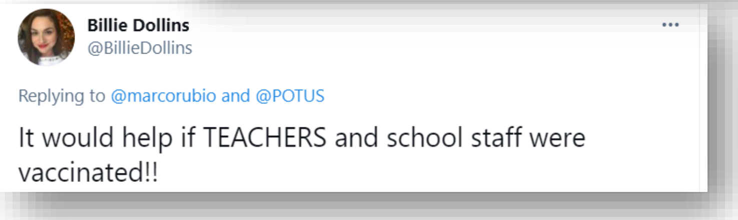
■ Anger ■ Anticipation ■ Disgust ■ fear ■ Joy ■ Sadness ■ Surprise ■ Trust
70k
60k
50k
40k
30k
20k
10k
0

Top 5 topics by Volume



Jason Y. Ng @jasonyng · Apr 1

When Hong Kong announced a 14-day shutdown of leisure venues last week, it did a **half-baked job** banning only cinemas, gyms and a few other categories. It took a mini-community outbreak at a karaoke joint for the gov’t to extend the list to include karaoke and mahjong parlors.



Billie Dollins
@BillieDollins

Replying to @marcorubio and @POTUS

It would help if TEACHERS and school staff were vaccinated!!

Dashboard image by University of Galway Pandem-2 team

Data Ethics Challenges

- *“… **availability, quality and nature of the training data** …*
- *… ability to **de-bias data** … **specific age and socioeconomic groups** … data from Facebook is likely to be **biased towards health data and linguistic quirks specific to a population** older than one trained on data from Snapchat …*
- *… **public perception of privacy and data access** … recent **survey of social media users** found that the majority considered analysis of their social media data to identify mental health issues “intrusive and exposing” and they **would not consent to this** …*
- *… **assessment and evaluation of NLP models** to ensure that they are working as intended … important **not to equate high scores with true language understanding** …”*

Baclic, O., Tunis, M., Young, K., Doan, C., Swerdfeger, H., & Schonfeld, J. (2020). **Artificial intelligence in public health: Challenges and opportunities for public health made possible by advances in natural language processing**. *Canada Communicable Disease Report*, *46*(6), 161.

Learning Outcomes

After completing this topic, you should be able to:

- understand the challenge of ethical issues in NLP
- be aware of major regulatory frameworks that address Ethics in AI
- understand some of the main concepts in research on ethical issues in NLP

Lab of this week

Exercises on bias in NLP

Industry Talk - Genesys

Dr. Maciej Dąbrowski

Chief Data Scientist, Digital & AI

Genesys, Galway

https://www.advertiser.ie/galway/article/119094/bonham-quay-bound-genesys-to-create-100-new-jobs-in-city-and-remotely