Universität Stuttgart

IRIS

Language in Social Context: Bridging NLP and Sociolinguistics

Agnieszka Faleńska
Filip Miletić

Day 5:
Ethics, biases, applications

Image credit: DALLE (2024)

Recap: Language Variation in Interaction

- Language variation is influenced by & takes on meaning in a broader social context of interaction
- In sociolinguistics, stylistic variation has been explained as…
  - an effect of formality via attention to speech
  - an effect of the broader communicative situation
  - indexical positioning conveying social meaning
- Tricky to operationalize in social media, but possible through various proxy methods

UCL <https://www.ucl.ac.uk/research-ethics/do-i-need-ethical-approval>

Day 5: Outline

Ethical issues

Standard practices in sociolinguistics (Milroy & Gordon, 2003; Childs et al., 2011)

Social media data (Williams et al., 2017; Fiesler & Proferes, 2018)

Development of NLP systems (Bender et al., 2021)

Harms & biases

Biases in representations (Knupleš et al., 2024)

Downstream biases (Lalor et al., 2022)

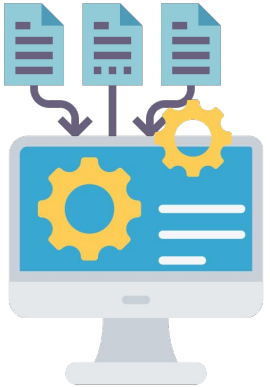
NLP applications

Personalization (Lynn et al., 2017; Rabinovich et al., 2017)

One size fits all? (Lucy et al., 2024)

Ethical issues

Background



- Sociolinguistics and NLP are empirical disciplines

DATA

- How is data collected?
- How is data used?
  - When is it justified to analyze a specific piece of information?
  - What does it entail for computational systems?

Check with your institution for standard practices and legal advice

Images: Flaticon.com

Ethics in sociolinguistic research

Basic principles

- “Being ‘ethical’ means acting in accordance with a set of core values and principles, in particular, integrity, compliance with the law, respect for human rights, and avoiding unnecessary risk to people’s safety and well-being.” (UCL)
- Work with human participants: **DO NO HARM**
- Informed consent:
  agreement to voluntary participation before any data collection
- Anonymization:
  - Substitution or removal of personally identifiable information
  - Storing the data in a secure manner
- Benefiting the community

Milroy & Gordon (2003); Childs et al. (2011); UCL <https://www.ucl.ac.uk/research-ethics/do-i-need-ethical-approval>

Seeking ethics approval

- Requirements for ethics approval vary across countries and institutions
- Ethics approval for non-interventional research might not be legally required, but is usually recommended and often de facto required
- Some work is not classified as research (e.g. literary criticism); some research is exempt (e.g. observational; publicly available data)

Why go through the process?

- Protect the participants: any research entails potential risks which should be considered ⇒ health, safety, privacy, environment …
- Protect the researcher ⇒ legal responsibility, funding, publication …
- Careful reflection about own work

UCL <https://www.ucl.ac.uk/research-ethics/do-i-need-ethical-approval>

Sample ethics review application

Context and scientific interest:

Healthcare application, industrial application, furthering scientific knowledge, for the benefit of society, etc.

Objectives:

General assumptions:

Conflicts of interest:

To the best of your knowledge, do any of the researchers involved in the project have a vested interest (either personal or institutional) or conflict of interest with regard to a partner, financial backer or any other institution?

University of Toulouse <https://www.univ-toulouse.fr/actualites/comite-d-ethique-de-recherche-cer>

Sample ethics review application

- **Participants**

Exact number of participants or approximate "range" and criteria used to determine this number:
The CER expects the project owner to justify the number of participants to be enlisted, based on the existing literature and use of a software application that estimates the number of participants required based on the estimated effect sizes

Recruitment:

Recruitment mode: *adverts, listings, "word of mouth" effect, etc.*
Place of recruitment: *specify the planned place of recruitment or the criteria that will determine the choice of place*
Selection criteria: *specify the selection criteria for participants, based on your research objectives. These criteria may include aspects such as age range, manual laterality, socio-cultural background, level of education, nationality, involvement in the process studied, etc.*
Non-inclusion criteria: *specify the non-inclusion criteria applied to potential participants, based on your research objectives. These criteria will come into play once you have selected your participants, i.e. they will lead to some of the pre-selected participants being excluded from your protocol. These criteria may include aspects such as visual or hearing impairments, neurological disorders, addictive behaviour, etc.*
Recommendations: *in order to minimise the risk of any invasion of privacy, we recommend that you list all the criteria (inclusion and/or exclusion) and ask participants whether they meet all the criteria, rather than asking them whether they meet each criterion one by one.*

Possible compensation for subjects:

Do you plan to offer compensation to the people involved in the research? If so, you need to specify this and describe the chosen form.

University of Toulouse <https://www.univ-toulouse.fr/actualites/comite-d-ethique-de-recherche-cer>

Sample ethics review application

Method

Description of the protocol: *tasks, questionnaires, etc.*

Materials used: *it is important that we are clearly informed about the materials you plan to use (e.g. questionnaires, interview framework, etc.) so that we are able to judge whether they pose a risk to your participants.*

Place where the study will be carried out:

Timetable of assessments or observations: *start and end dates (month and year), number of assessment sessions for each participant, duration of the assessment for each participant.*

Duration of the study:
Data analysis: *Description of the data analysis (quantitative and qualitative)*

- **Foreseeable or known benefits and risks to physical and mental health (self-esteem, etc.) and social life (reputation)**

- *Present the benefits of your study. These benefits may include scientific progress, improving quality of life for participants, boosting self-esteem, etc.*

University of Toulouse <https://www.univ-toulouse.fr/actualites/comite-d-ethique-de-recherche-cer>

Sample ethics review application

Yes/no	List of risks: **For each identified risk, please provide the reason why you need to take this risk and indicate all the precautions you will undertake to limit this risk**
	Does your protocol use a form of staging designed to conceal part of the objective or methodology from the subjects, or to make them believe that other objectives or methodologies are at stake?
	Questions or situations that may cause participants to feel uneasy?
	Materials likely to be perceived by the participants as threatening, shocking or repulsive?
	Possible invasion of participants' privacy, or that of their family, including the use of personal information?
	Use of physical stimuli (sound, visual, haptic, etc.) other than stimuli associated with everyday activities?
	Deprivation of physiological needs (drinking, eating, sleeping, etc.)
	Handling of psychological or social parameters such as sensory deprivation, social isolation or psychological stress?
	Physical exertion more demanding than a level that would be considered moderate for the participant?
	Exposure to drugs, chemical substances or potentially toxic agents?

University of Toulouse <https://www.univ-toulouse.fr/actualites/comite-d-ethique-de-recherche-cer>

Sample ethics review application

Confidentiality

Anonymisation process:

The notion of data anonymisation goes beyond simply hiding the person's name. It means that it should be impossible to match the subject's identity with the data, even using indirect means. Generally speaking, confidentiality will be guaranteed by the fact that each subject or group will be referred to by an identification code in the form of a random number in electronic and paper analyses and documents.

However, there are two different cases when it comes to protecting privacy and confidentiality.
Case 1 - The protocol is such that the data processed is anonymous or made anonymous via the use of random numbers. The person involved cannot be identified in any way, even indirectly; consequently, this data can no longer be considered to be personal data, and there is no table of correspondence between the identity of each person and a random number referring to a set of individual data.

Case 2 - The data is classified as personal data or there is a table of correspondence between each person and the random numbers identifying the set of data related to a given participant (this scenario may be justified by the research objectives).
In this second case, we talk about confidentiality rather than anonymity (people are identified or identifiable in the documents, even partially or temporarily). The principle of confidentiality rather than anonymity will hence be invoked. In such cases, please specify:

- *the reason why anonymisation is not possible*
- *a description of possible breaches of people's right to privacy resulting from the project or publication of the results,*
- *the precautions taken to handle this risk.*

University of Toulouse <https://www.univ-toulouse.fr/actualites/comite-d-ethique-de-recherche-cer>

Sample ethics review application

People authorised to access this data:

Please specify the people who will have access to the data: scientific coordinator, research associate(s), etc.

Archiving

Type of data archived (specify whether the data can lead to identification of the subjects, either directly or by cross-referencing):

Duration of archiving:
The CER advises an archiving period of 15 years after data collection. In any case, a duration of 5 years after publication is the absolute minimum. With regard to archiving of the consent forms (which inevitably enable identification of the subjects), the CER recommends that they be retained for 10 years from the publication date and 20 years if the results are not published, in a sealed envelope marked: "I certify that this envelope contains x (number) consent form(s) and x compliant information form(s), collected as part of the study xxx", followed by the name of the person responsible.

Place of archiving: With regard to digital archiving, the CER advises using only professional, encrypted, password-protected computers and secured remote servers. If you wish to use the digital environment of a university, please check with the IT department if it has an appropriate level of security. Always encrypt your research folders. Note that Huma-Num server has an appropriate level of security for long-term archiving.

University of Toulouse <https://www.univ-toulouse.fr/actualites/comite-d-ethique-de-recherche-cer>

Ethics and social media data

How do **you** feel about research on Twitter?

Let’s replicate Fiesler & Proferes (2018):
“Participant” Perceptions of Twitter Research Ethics

Ethics and social media data

Is using social media data legal?

- Check the Terms and Conditions
- Copyright issues
- Regionally specific legislation, e.g. EU Directive 2019/790
  ⇒ Text and data mining for the purposes of scientific research
- The right to publish collected data is often blurry

But is it ethical?

… integrity, compliance with the law, respect for human rights, and avoiding unnecessary risk to people's safety and well-being

EU 2019/790 <https://eur-lex.europa.eu/eli/dir/2019/790/oj>; UCL <https://www.ucl.ac.uk/research-ethics/do-i-need-ethical-approval>

Ethics and social media data

Does work on social media data require ethics review?

- In general, no, if it is understood as “information in the public domain [...] with no reasonable expectation of privacy”
  - “non-intrusive, does not involve direct interaction between the researcher and individuals through the Internet, and where there is no expectation of privacy”
  - “relies exclusively on cyber-material [...] to which the public is given uncontrolled access on the Internet and for which there is no expectation of privacy”
- Exceptions to the exemption
  - Data linkage
  - Online groups with an expectation of privacy, e.g. limited membership groups

Canada TCPS 2 <https://ethics.gc.ca/eng/policy-politique_tcps2-eptc2_2022.html>

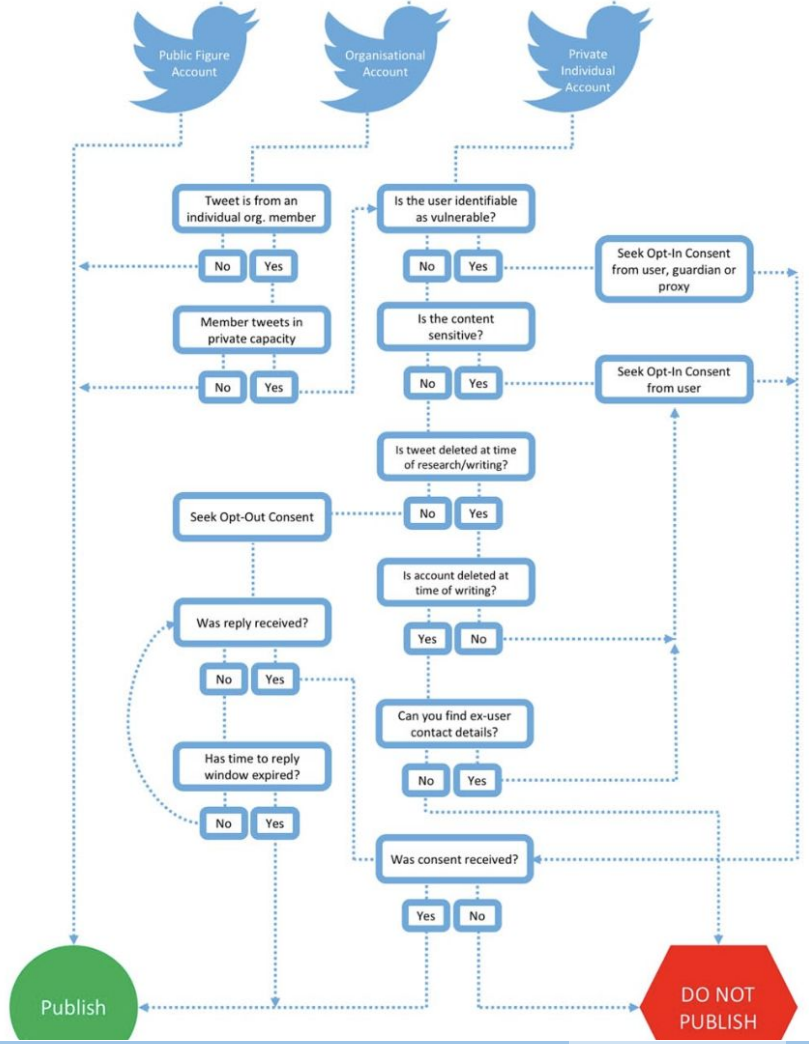
Fiesler & Proferes (2018): Implications for study design

- Consider asking for permission in a reasonable way,
  e.g. via opt-out or post-hoc notification
- Anonymize tweets and preferably do not quote them
- Attribute tweets only if benefits clearly outweigh harms
- Respect “privacy by obscurity”
- Pay attention in sensitive contexts,
  e.g. medical conditions, vulnerable populations

Williams et al. (2017): Implications for publication

- It should not be assumed that users have read and understood ToS
- Online survey of 564 Twitter users in UK
  - 76% aware of third party access provided by ToS
  - 16% quite or very concerned about **university** research
  - 49% quite or very concerned about **government** research
  - 51% quite or very concerned about **commercial** research
- Users may be communicating to an imagined limited audience but exposing content to the whole world ⇒ “context collapse” (Marwick & boyd, 2010)

Williams et al. (2017): Implications for publication



Ethics and the development of NLP systems

Bender et al. (2021): Can language models be too big?

- Widely used language technologies rely on pretrained language models, complex computational systems trained on string prediction tasks
- Increasing size of neural architectures and training datasets
- Ethical issues including:
  - Training data
  - Environmental and financial cost
  - Misdirected research efforts
  - Downstream risks and biases

Ethics & NLP systems: Training data

- Training data may include stereotypical and derogatory associations regarding gender, race, ethnicity, disability status
  - Unequal internet access ⇒ underrepresented populations
  - Spaces for underrepresented populations may still be undersampled
  - Coarse filtering practices
- Changes in societies may not be captured by static datasets
- Systems trained on data containing problematic associations may encode these associations and amplify them downstream

Bender et al. (2021)

Ethics & NLP systems: Cost

Environmental cost

- ≈ 650kg CO2e Training of a single BERT base model
- ≈ 284t CO2e Training of Transformer (big) with hyperparam search
- ≈ 900kg CO2e Trans-American flight
- ≈ 5t CO2e Average human in a year

Financial cost

- +0.1 BLEU for English to German translation ⇒ +$150,000 compute cost

General considerations

- Renewable energy: not everywhere and needed for other uses
- Most language technology serves those who already have the most privilege

Bender et al. (2021)

Ethics & NLP systems: Cost

TIME

SUBSCRIBE

BUSINESS • TECHNOLOGY

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

15 MINUTE READ

BY BILLY PERRIGO

JANUARY 18, 2023 7:00 AM EST

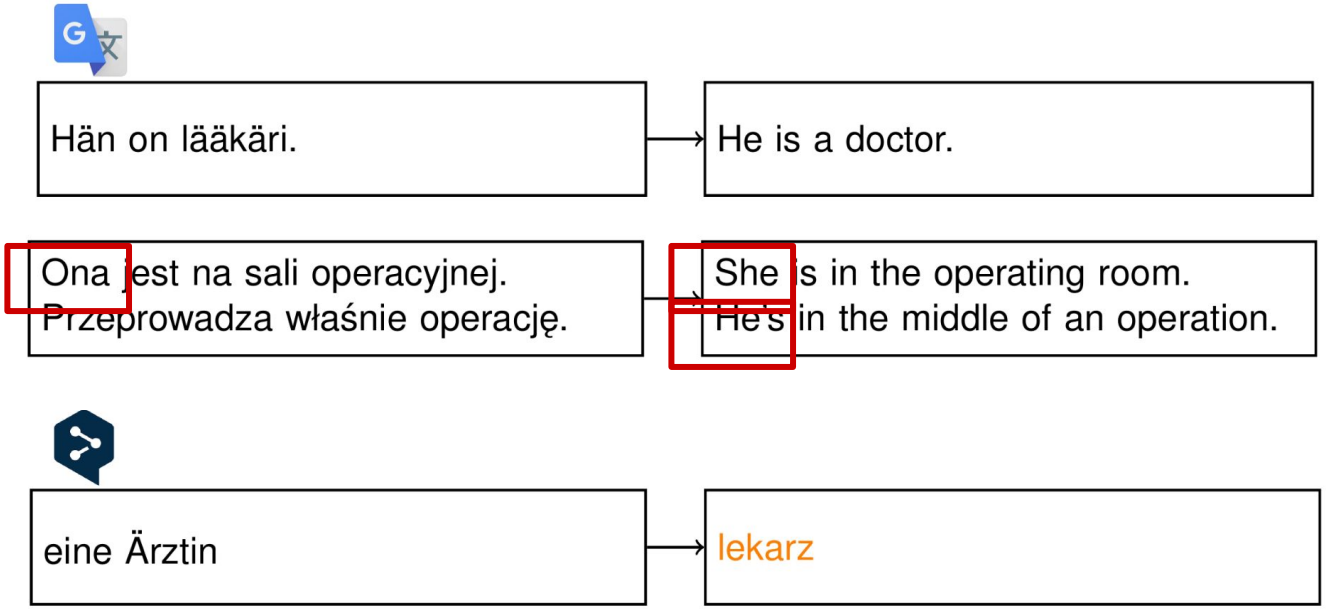
https://time.com/6247678/openai-chatgpt-kenya-workers/

Takeaways

- Empirical research on language use relies on data that attests it
- Whether obtained through direct interaction or public information, this data comes from human participants
- Duty to consider risks and benefits of our research
  - to protect the participants
  - to protect against downstream harms
  - …

Biases in NLP

Who can be a doctor?



Increasing concerns about NLP systems' harms

Amazon scraps secret AI recruiting tool showed bias against women

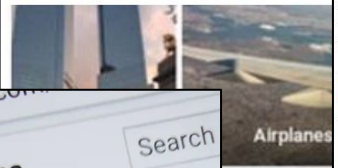
By Jeffrey Dastin

SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine specialists uncovered a big problem: their new recruiting engine did women.

Google Mistakenly Showing Limits of

By Alistair Barr Follow
Updated July 1, 2015 3:41 pm ET

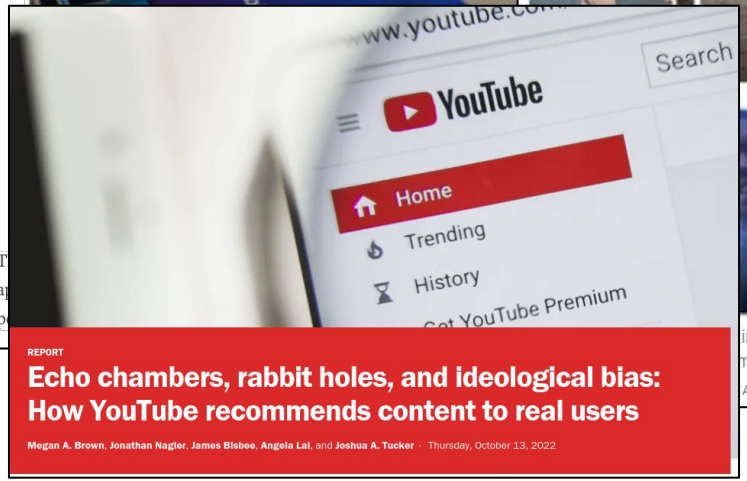
PRINT TEXT



AI language models show bias against people with disabilities: Study

Natural language processing (NLP) is a sort of artificial intelligence that allows machines to utilise written and spoken phrases in a variety of applications, such as smart assistants or email autocorrect and spam filters, to help automate and expedite activities for individual users and companies. However, the algorithms that power this technology frequently exhibit characteristics that might be insulting or discriminatory toward people with disabilities.

ANI | Pennsylvania | Updated: 14-10-2022 08:50 IST | Created: 14-10-2022 08:50 IST



REPORT

Echo chambers, rabbit holes, and ideological bias: How YouTube recommends content to real users

Megan A. Brown, Jonathan Nagler, James Bisbee, Angela Lai, and Joshua A. Tucker · Thursday, October 13, 2022

AND TWITTER

tions across two genders and five dialects of English. Speakers' dialect and gender was controlled for by using videos uploaded as part of the "accent tag challenge", where speakers explicitly iden-

been ex-
ert, 1989,
bust vari-
ers across
dialect regions. For instance, English varies dramatically between the United States (Cassidy and others, 1985), New Zealand (Hay et al., 2008) and Scotland (Milroy and Milroy, 2014).

Consequences of harmful behaviors

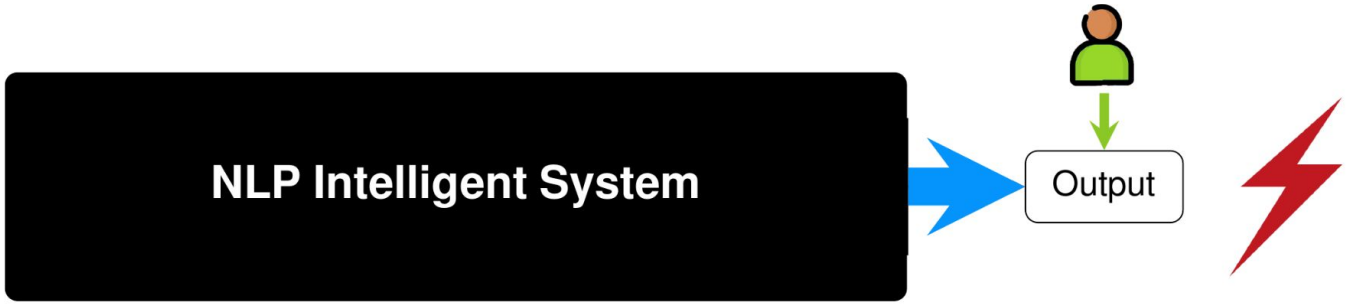
- Allocation and representational harms (Blodgett et al., 2020)

- Broad range of applications
  - healthcare, banking, judicial system, …

- We use them
  - by choice: translators, voice assistants, …
  - not by choice: CV filtering systems, political sentiment analyzers, ...

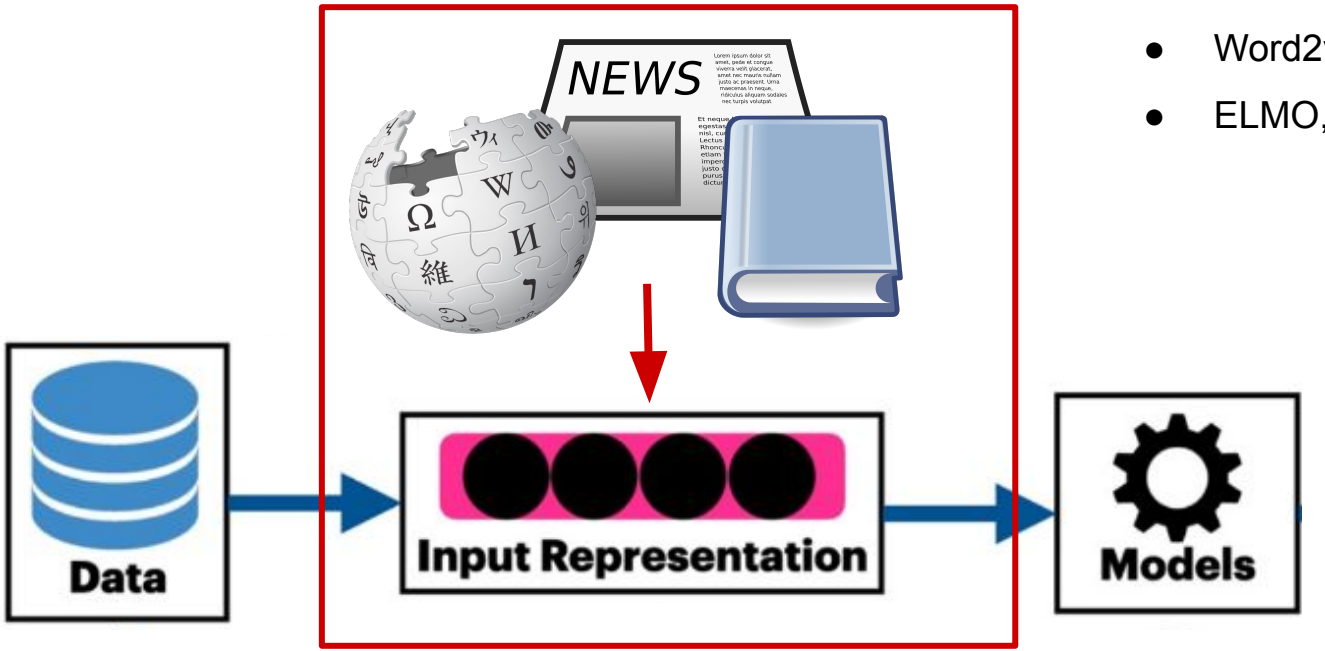
Harmful behaviors are symptoms of bias

Bias – systematic preference or discrimination against certain groups of users (Friedman and Nissenbaum, 1996)



- Models reflect demographic imbalances (Hovy and Yang, 2021)
- Speech recognizers with lower accuracy for female voices (Tatman, 2017)

Sources of bias

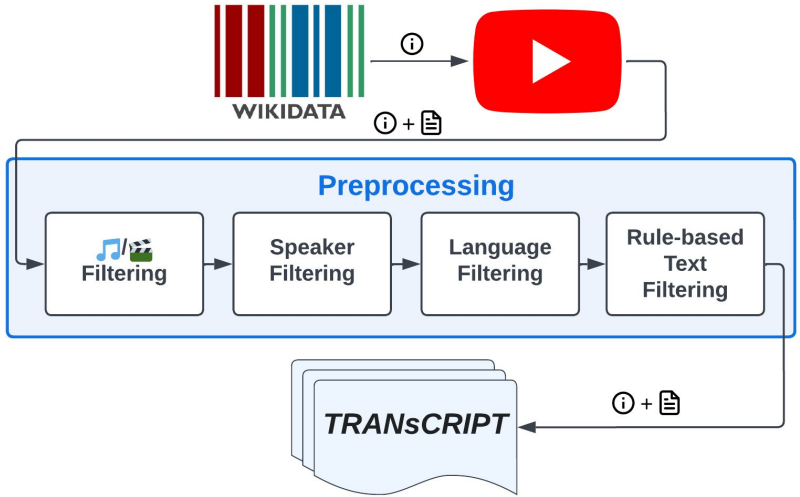


- Word2vec
- ELMO, BERT, …

Source: Hovy and Prabhumoye, 2021

Do word representations capture demographic signals?

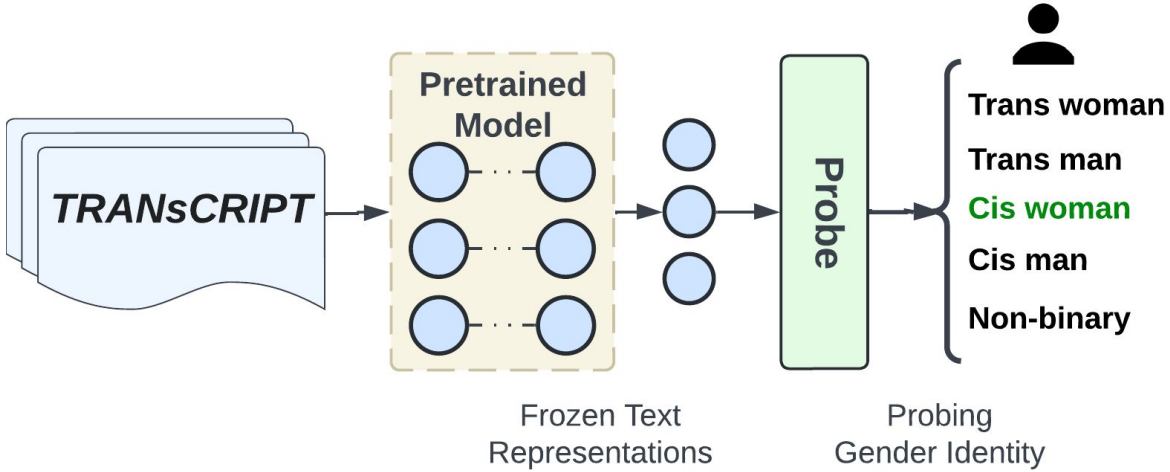
Knupleš et al., (2024): “Gender Identity in Pretrained Language Models: An Inclusive Approach to Data Creation and Probing”



Gender Ident.	Users	Transc.	Segm.	Tokens
Trans woman	13	546	15,121	1,646,4121
Trans man	6	192	6,478	548,579
Cis woman	55	2,446	92,436	4,986,314
Cis man	79	2,474	80,902	4,309,761
Non-binary	15	514	9,397	960,224
Total	168	6,172	204,334	12,451,290

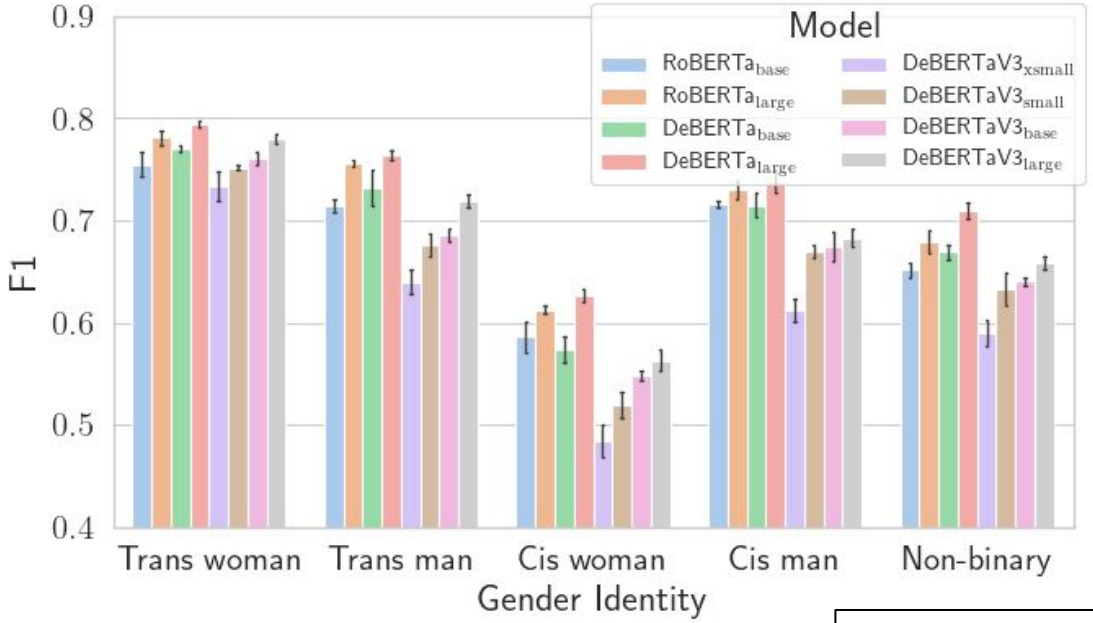
Source: Knupleš et al., (2024)

Probing word representations for gender



Source: Knupleš et al., (2024)

Do word representations capture gender identity?



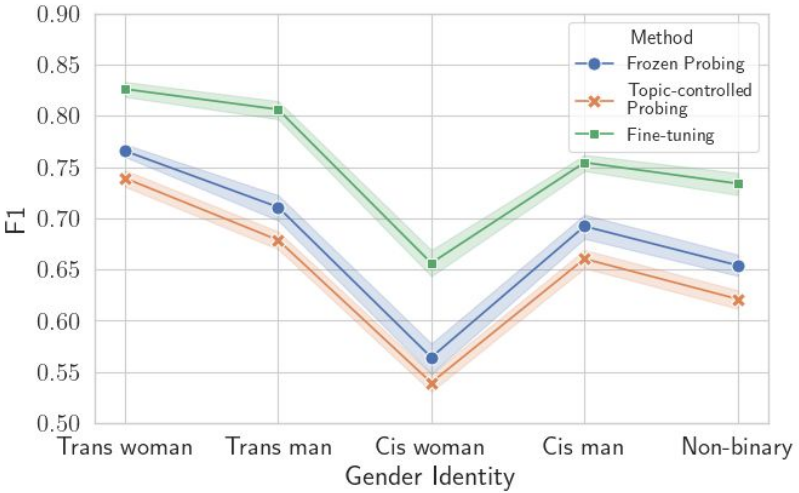
Content or style?

Source: Knupleš et al., (2024)

Do word representations capture gender identity?

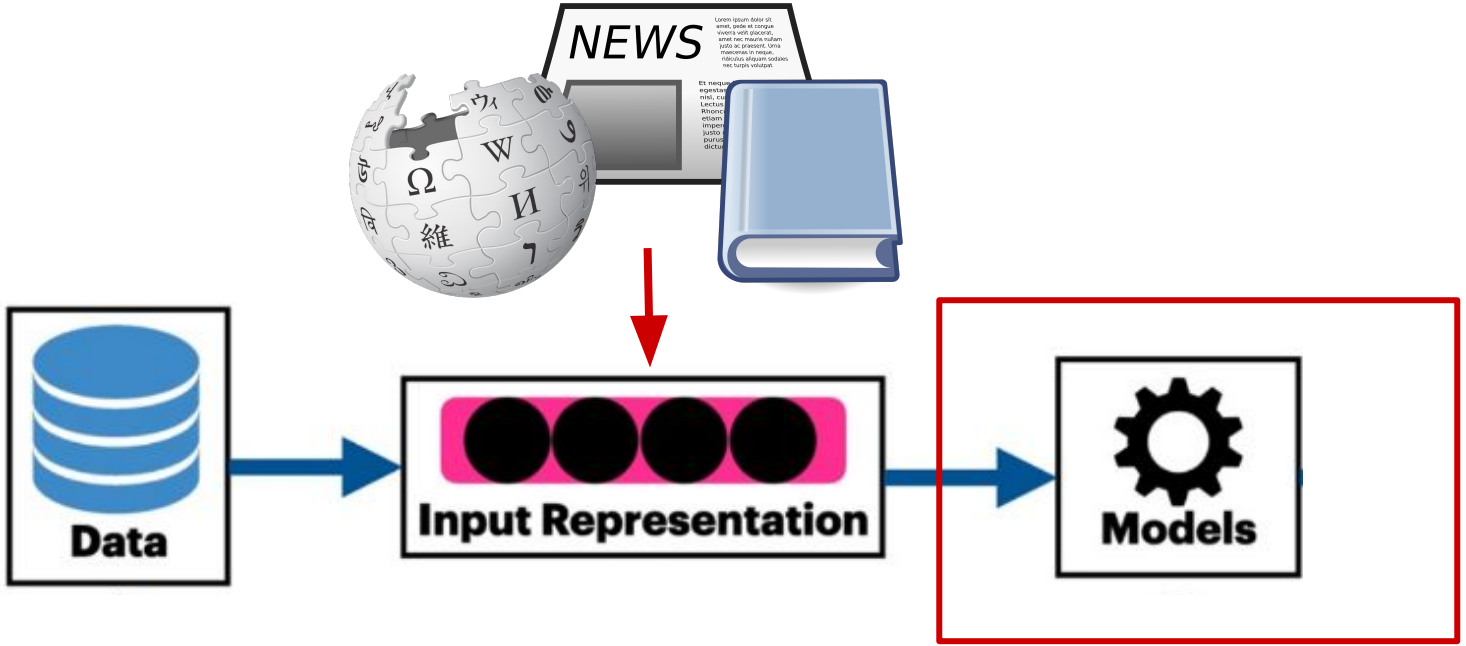
- Use SAGE to find the most indicative terms for each author
  - underachiever, teehee, …
- Remove them from texts

- Substantial amount of gender identity information encoded within representations
- Error discrepancies in model behaviors



Source: Knupleš et al., (2024)

Sources of bias



Source: Hovy and Prabhumoye, 2021

How to measure bias?

Bias – systematic preference or discrimination against certain groups of users (Friedman and Nissenbaum, 1996)

Fairness metrics (Czarnowska et al., 2021)

$$DI = \frac{p(\hat{y} = 1|A = 0)}{p(\hat{y} = 1|A = 1)}$$

protected group

privileged group

Disparate Impact
(Friedler et al., 2019)

- a ratio of **1** indicates *demographic parity*
- metric comes from the legal field, where certain regulations require DI above **0.8** and below **1.2**

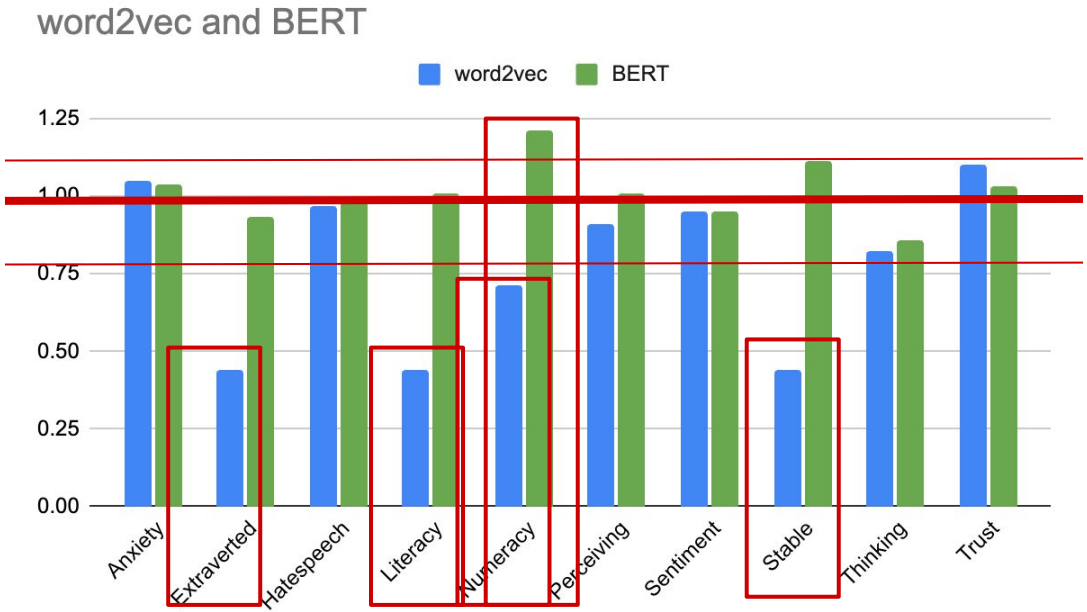
In which tasks can we see downstream biases?

Lalor et al. 2022, Benchmarking Intersectional Biases in NLP

Dataset	Task	Demographics	Data source	*N*
Psychometrics	Anxiety, Literacy, Numeracy, Trust	Gender, Race, Age, Income, Education	Survey	8,395
Multilingual Twitter Corpus (MTC)	Hate Speech Identification	Gender, Race, Age	Twitter	83,078
Five Item Personality Inventory (FIPI)	Extraverted, Stable	Gender, Race, Age, Income, Education	Survey	6,805
AskAPatient	Sentiment	Gender, Age	Forums	20,000
Myers–Briggs Type Indicator (MBTI)	Perceiving, Thinking	Gender, Age	Reddit	7,406 (1,584)

Source: Lalor et al. 2022

Biases in downstream models



$$DI = \frac{p(\hat{y} = 1|A = 0)}{p(\hat{y} = 1|A = 1)}$$

Source: Lalor et al. 2022

Takeaways

- NLP models can cause harms
- Harms come from biases learned from the data
  - biases can appear already in word representations and surface in downstream tasks
  - they depend on the models, data, and tasks
- Importance of using fairness metrics

Source: Lalor et al. 2022

Applications in NLP

So far…

- Methods to discover and analyze linguistic variation
- Methods that learn demographically-related spurious correlations and end in biases
- How to **benefit** from the linguistic, demographically-related variables?
  - Which NLP applications **use** linguistic variation?
  - Which NLP tasks **improve from using** linguistic variation
  - How to define the **improvement**?

Which NLP applications use linguistic variation?

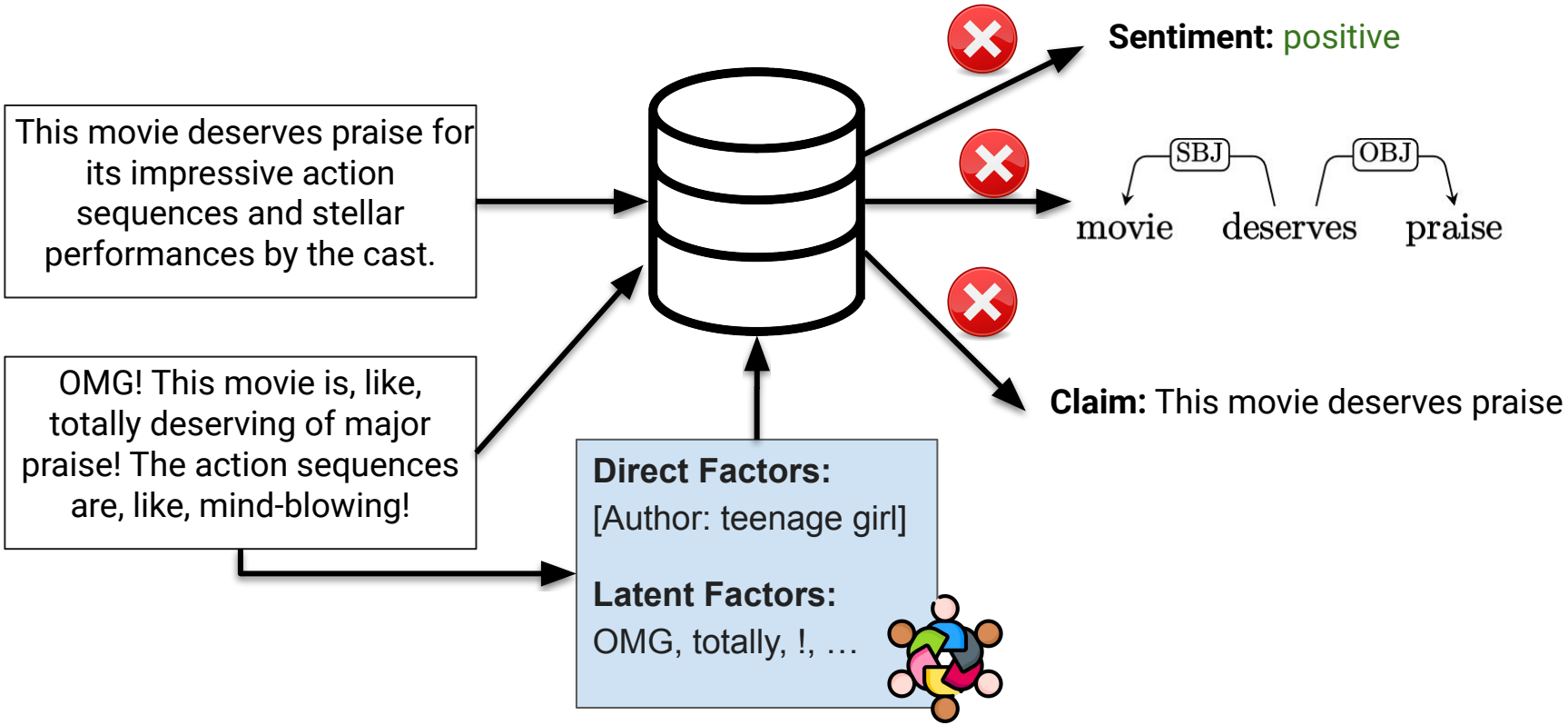
Which NLP applications **use** linguistic variation?

Authorship profiling (Mishra et al. 2018)

- forensics and abuse detection
- targeted advertisement
- enriching the data → computational social science
- (...)

Which NLP tasks improve from using linguistic variation

Which NLP tasks **improve from using** linguistic variation



Which NLP tasks **improve from using** linguistic variation

- Lynn et al., 2017: “Human Centered NLP with User-Factor Adaptation”
  - POS Tagging → POS tags
  - Sentiment → positive, neutral, negative
  - Sarcasm → sarcastic, not sarcastic
  - Stance → for, against, neutral

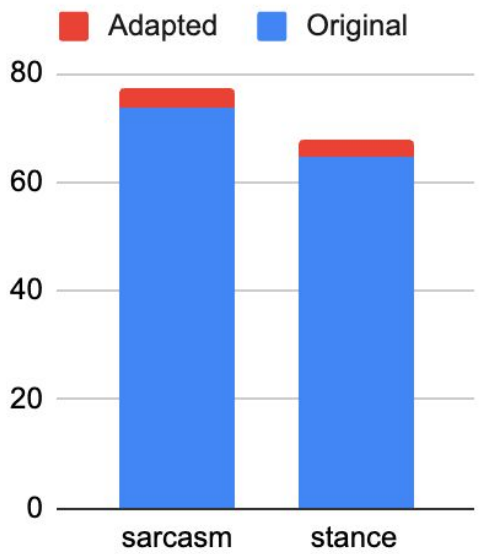
- Hovy, (2015): “Demographic Factors Improve Classification Performance”
  - Topic classification
  - Sentiment → positive, neutral, negative

- Hung et al. (2023) Can Demographic Factors Improve Text Classification? Revisiting Demographic Adaptation in the Age of Transformers

How to define the improvement?

How to define the improvement?

- **Accuracy** regardless of the group
- **Invariance**: systems are expected to behave identically for social groups
  - debiasing, mitigating biases…
- **Adaptation**: system behaviors are expected to vary across social groups
  - personalization



Source: Lynn et al., (2017)

Demographic adaptation

Rabinovich et al. 2017: “Personalized Machine Translation: Preserving Original Author Traits”

- What happens to personality and demographic textual markers during the translation process?
- Data:
  - parallel corpus of the European Parliament Proceedings (Koehn, 2005)
  - transcripts of TED talks
- Annotate gender of speakers using WikiData

Source: Rabinovich et al. 2017

Which stylistic features are present in the data?

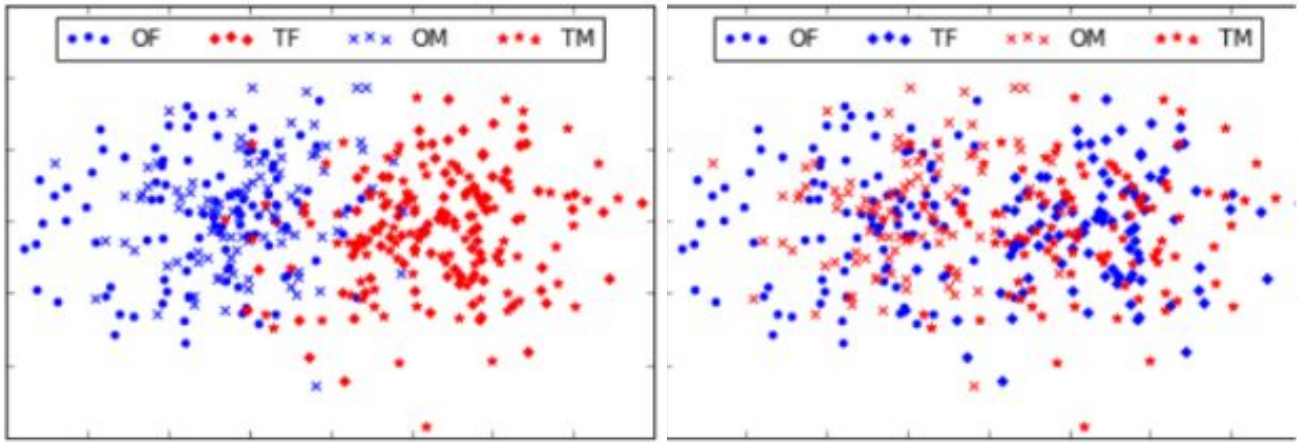
- Filter function words and cluster the data

OF – original female

TF – translated female

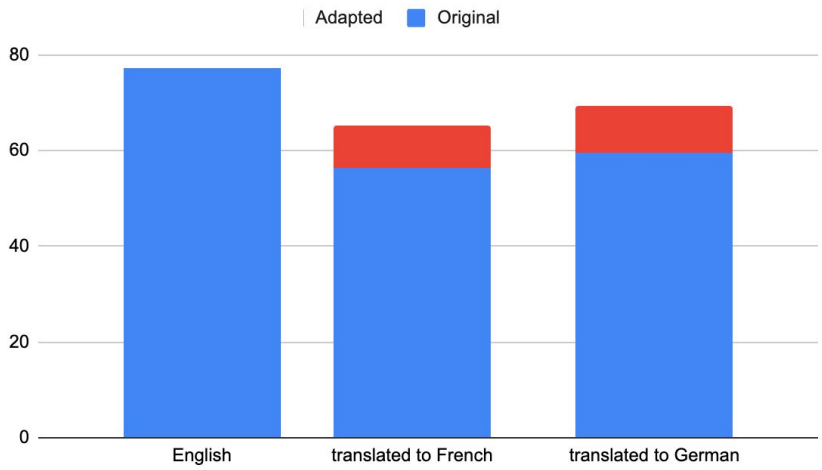
OM – original male

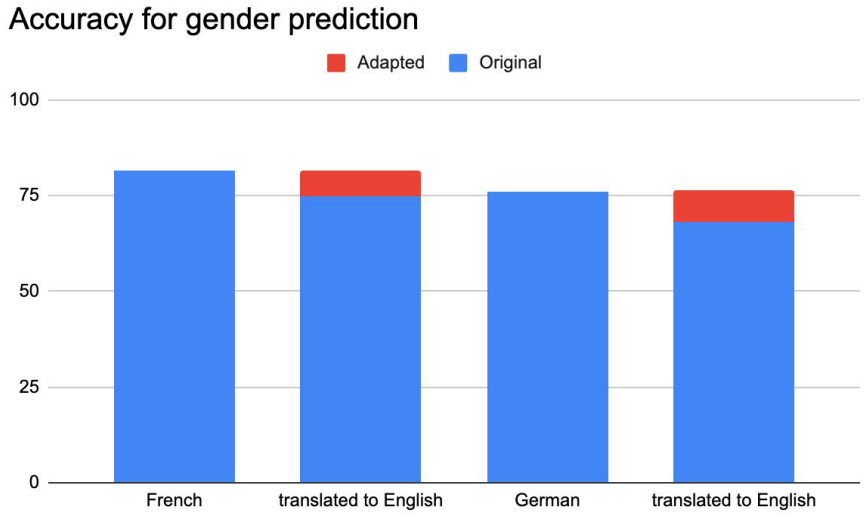
TM – translated male



Source: Rabinovich et al. 2017

Gender markers in translations





- Gender traits in translation are diminished compared to their originals.
- Some can be restored through demographic adaptation

Source: Rabinovich et al. 2017

Do you want machine translation to maintain your demographic characteristics?

The future…

Demographic adaptation in NLG models

Invariance: systems that behave identically for social groups

Hey, so, like, I'm seriously frustrated about this delayed flight situation. Ugh! Anyway, do you, have any info on possible connections? Gotta, like, figure out my next move ASAP!

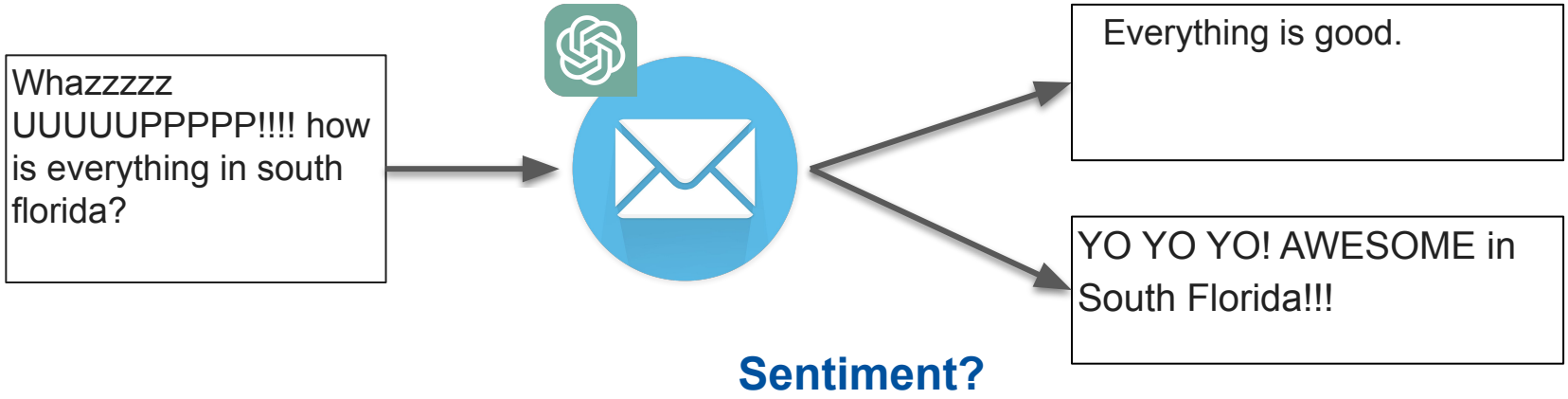
Oh no! That's, like, a downer! 😱💔 We are working on it ASAP!

[Author: teenage girl]

We apologize for the inconvenience caused. We will assist you in finding possible connections.

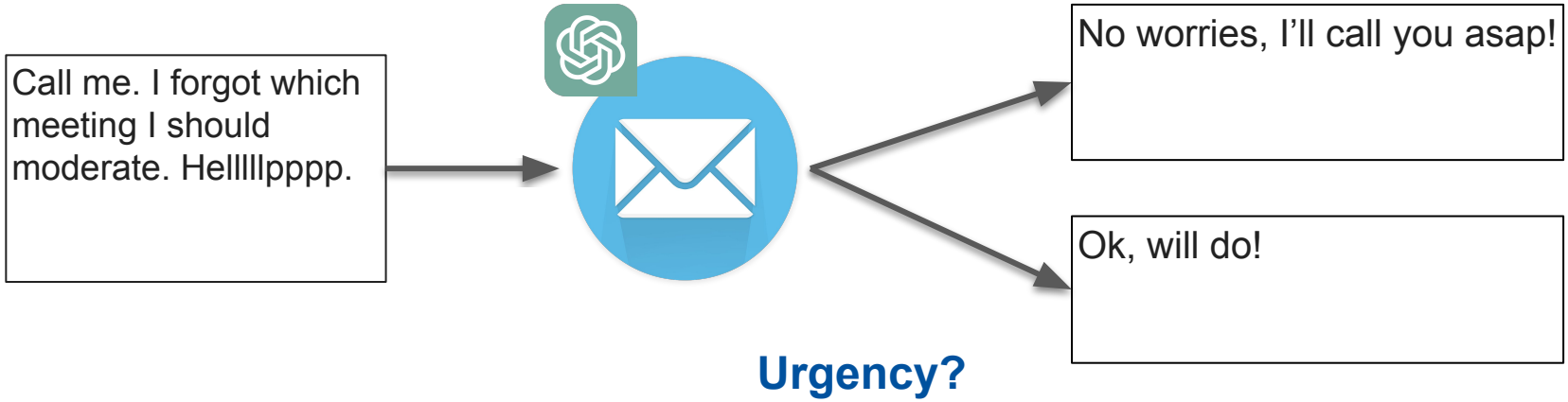
People's expectations of NLG behaviors

Lucy et al. 2024: "One-Size-Fits-All"? Examining Expectations around What Constitute "Fair" or "Good" NLG System Behaviors



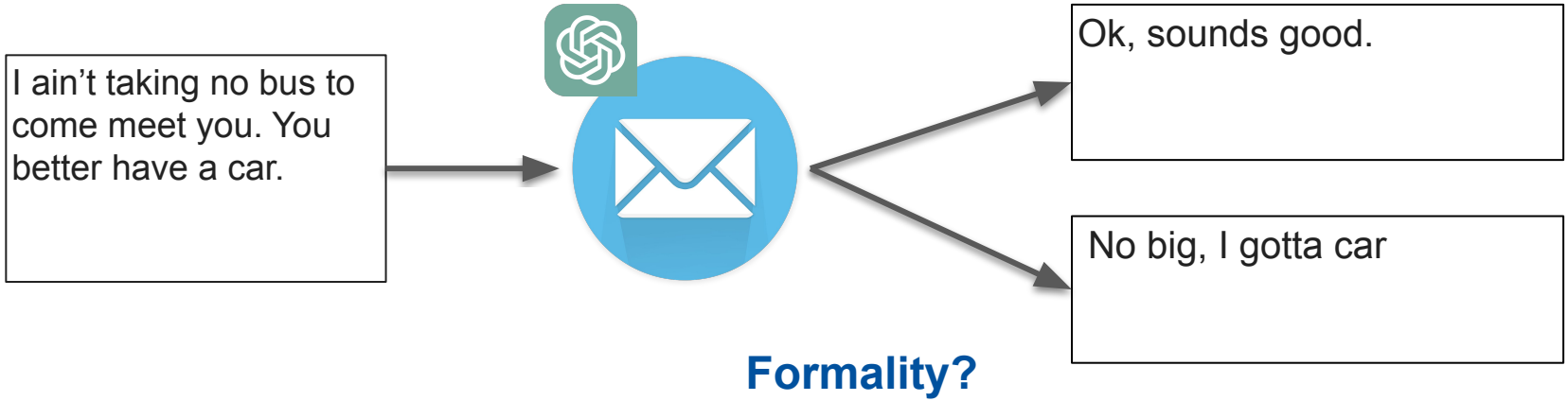
People’s expectations of NLG behaviors

Lucy et al. 2024: “One-Size-Fits-All”? Examining Expectations around What Constitute “Fair” or “Good” NLG System Behaviors



People's expectations of NLG behaviors

Lucy et al. 2024: "One-Size-Fits-All"? Examining Expectations around What Constitute "Fair" or "Good" NLG System Behaviors



People’s expectations of NLG behaviors

Lucy et al. 2024: “One-Size-Fits-All”? Examining Expectations around What Constitute “Fair” or “Good” NLG System Behaviors

- Adaptation can make replies more realistic, natural, authentic or genuine
- People’s expectations vary for different types of features
- Acceptability of making identity-related assumptions from a language feature varies among judges

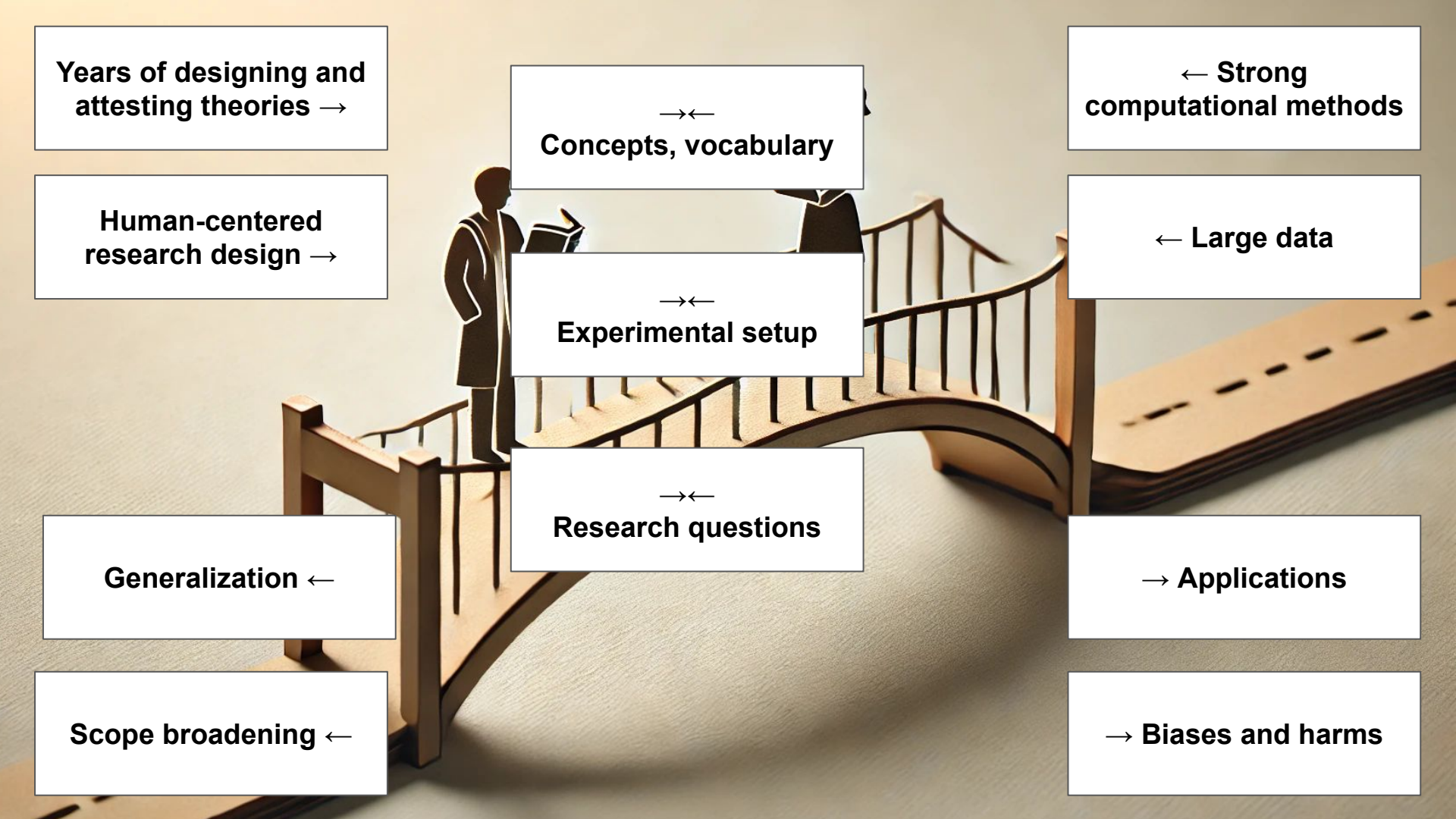
Takeaways

- Broad range of applications for demographically-related signals
  - using linguistic variation to approximate people's demographics
  - using linguistic variables to improve models' accuracy

- Which applications require demographic adaptation?
  - we should first ask people what they expect
  - opinions vary among judges

Source: Lalor et al. 2022

Takehomes


Years of designing and attesting theories →
Human-centered research design →
Generalization ←
Scope broadening ←
→←
Concepts, vocabulary
→←
Experimental setup
→←
Research questions
← Strong computational methods
← Large data
→ Applications
→ Biases and harms

Universität Stuttgart

IRIS

Thank you for your attention!

Image credit: DALLE (2024)