TRACING RACISM IN TEXT DATA – A CSS PERSPECTIVE

HOW WERE, ARE AND CAN COMPUTATIONAL APPROACHES BE APPLIED TO MEASURE RACISM IN TEXTUAL DATA?

KEY FINDINGS FROM A DOCTORAL THESIS

HOW DO RESEARCHERS DETECT RACISM IN TEXTUAL DATA?

RESEARCH AIM

Taking stock of the ***state-of-the-art*** research using automated detection of racism and related concepts such as racist hate speech, bias, and so on.

RESEARCH DESIGN

Systematic Literature Review
N = 115 articles from 2004-2023 (racism + computational methods + text as data)
Variables: Measured concept, data type, source and language, utilized method, validation, open science

KEY FINDINGS

- Strong increase in studies over time
- Strong focus on hate speech
- Interchangeable use of concepts, not much on theory
- Dominance of studies using social media data
- Heavy use of secondary datasets, but focus on a few
- Supervised classification is most used method
- Validation of data collection rarely done

MEASUREMENT PIPELINES

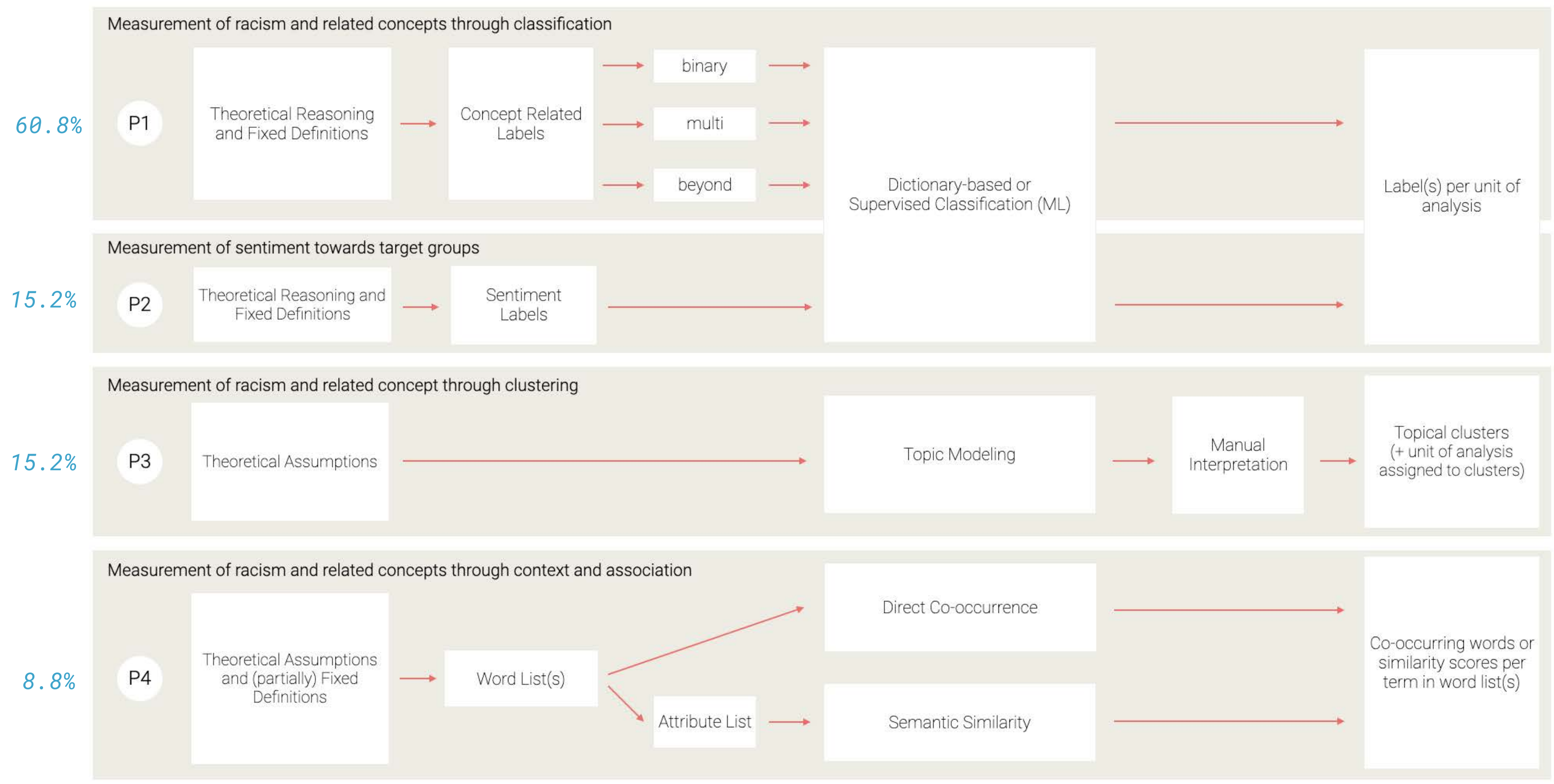
Measurement of racism and related concepts through classification
60.8%
P1
Theoretical Reasoning and Fixed Definitions
Concept Related Labels
binary
multi
beyond
Dictionary-based or Supervised Classification (ML)
Label(s) per unit of analysis
Measurement of sentiment towards target groups
15.2%
P2
Theoretical Reasoning and Fixed Definitions
Sentiment Labels
Measurement of racism and related concept through clustering
15.2%
P3
Theoretical Assumptions
Topic Modeling
Manual Interpretation
Topical clusters (+ unit of analysis assigned to clusters)
Measurement of racism and related concepts through context and association
8.8%
P4
Theoretical Assumptions and (partially) Fixed Definitions
Word List(s)
Attribute List
Direct Co-occurrence
Semantic Similarity
Co-occurring words or similarity scores per term in word list(s)

Figure 1 Simplified illustration of common measurement pipelines in studies detecting racism in text and their relative usage in percent (*N* = 115)

GAPS & RECOMMENDATION FOR FUTURE RESEARCH

- More fine-grained concepts and intersections
- Multilingual and comparative research
- Other data types than social media, other languages than English
- Validate data collection
- More open science, but also diversified use of published datasets
- More consideration on bias and ethics needed

HOW DO HUMAN AND LLM CODER BIAS INFLUENCE RACISM DETECTION?

RESEARCH AIM

Understanding how ***coder-level characteristics*** and ***textual properties*** contribute to annotation decisions of ***human*** and ***persona-assigned LLMs*** when detecting racism

RESEARCH DESIGN

Annotation task: Binary classification of racism in 360 German traditional and far-right alternative news media articles

STUDY 1:
HUMAN CODERS

- Crowdcoding with survey
- Socio-demographics, political attitudes, task-specific variables (such as being affected, awareness and attitudes towards racism)
- Definition and examples for task
- 164 Participants * 15 tasks each = 2.460 annotation decisions

STUDY 2:
PERSONA-ASSIGNED LLM CODERS

- Default and persona-assigned prompts
- One-shot (definition and examples as with human coders)
- GPT-3.5 vs. GPT-4o
- 16 personas: Being affected, contact with affected, education, age
- 2 different temperature settings
- 5 iterations = 61.200 annotation decisions

COMPARISON OF HUMAN & LLM

Inductive study of texts with high deviation between human and LLM annotation decisions

KEY FINDINGS

Human coders

- Contact with affected people and education have positive effects

Persona-assigned LLMs

- Differences between GPT-3.5 and GPT-4o
- All persona variables have effects
- Default in most cases ‚significantly' different to persona including being affected

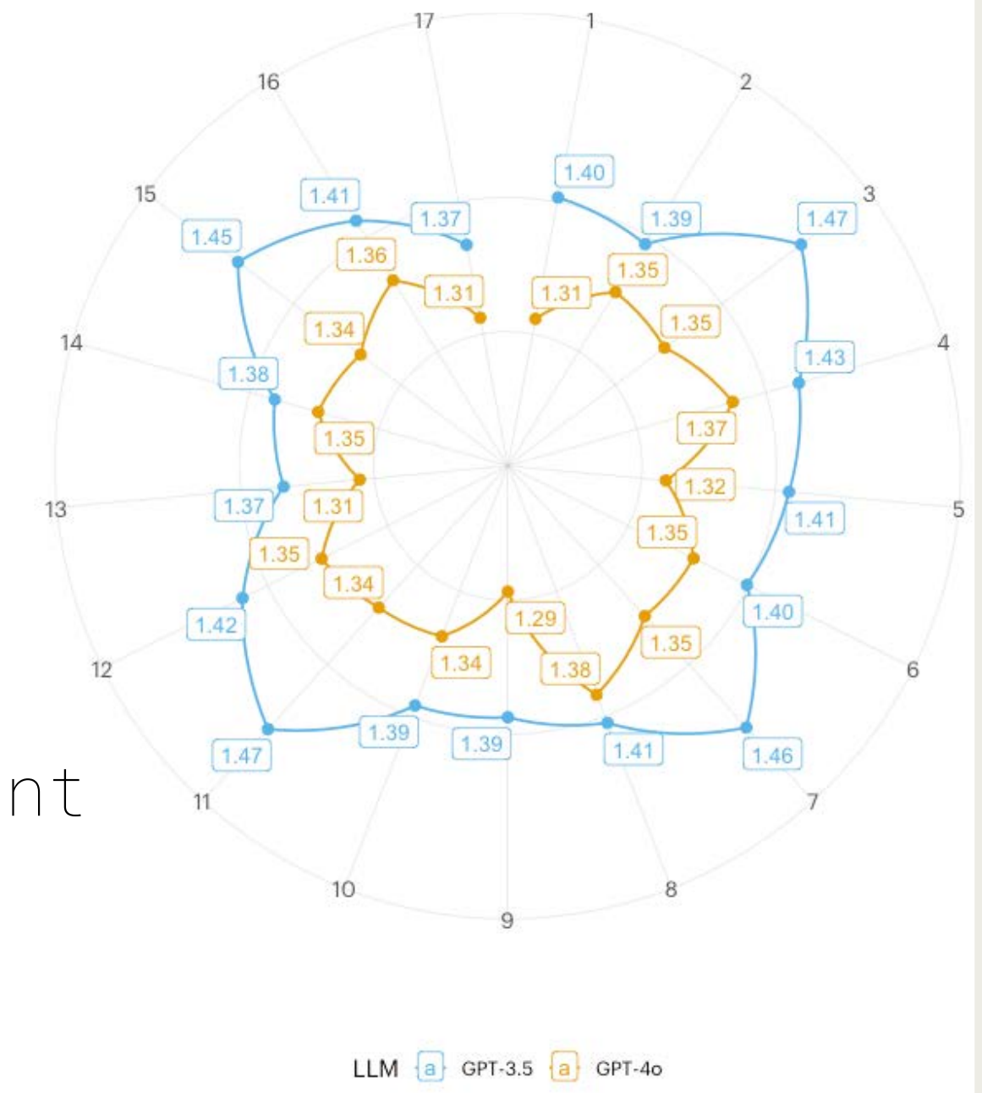

Figure 2 Mean annotation decision per persona

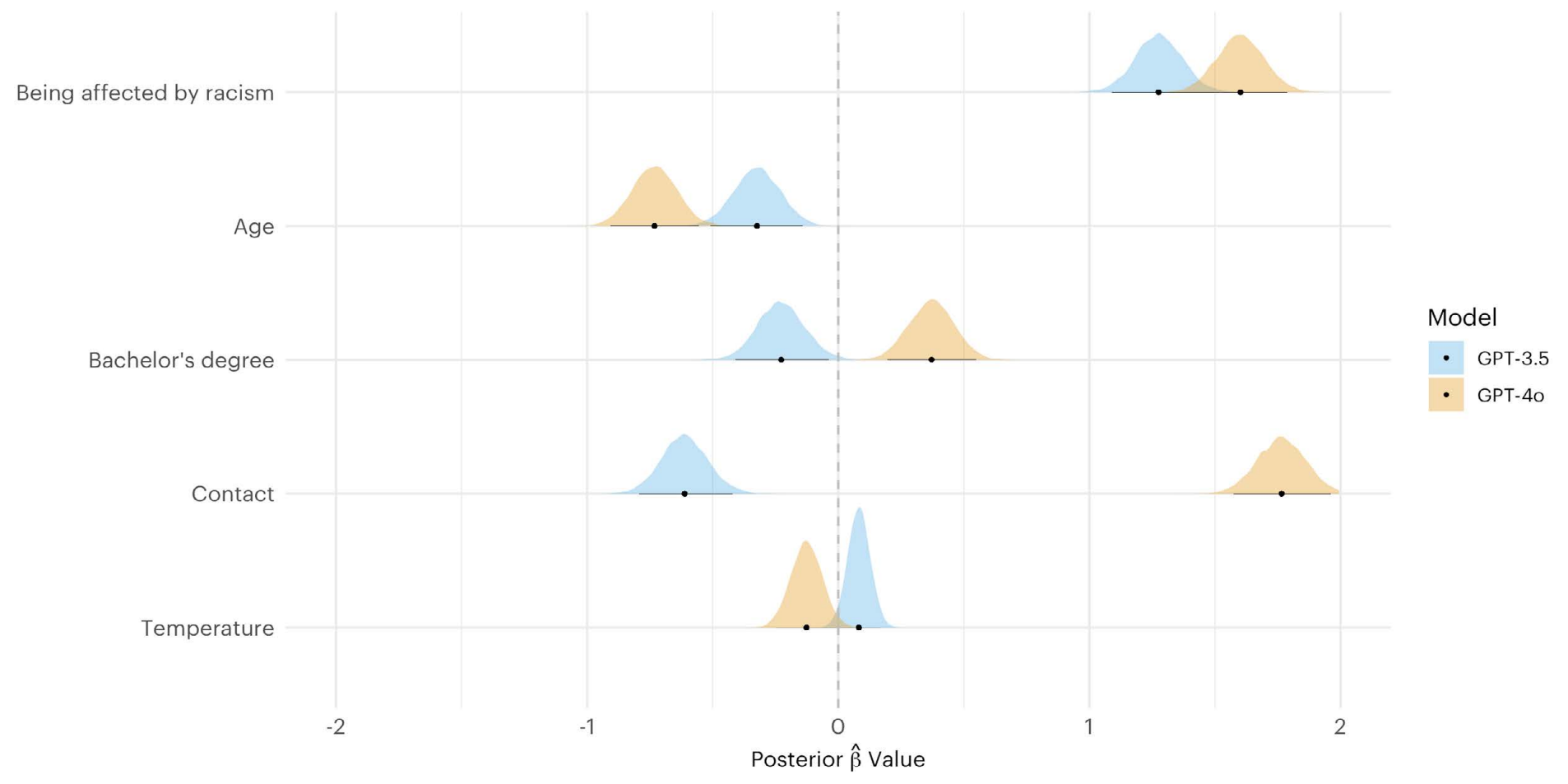

Figure 3 Posterior distribution for persona-assignment
Bayesian multilevel logistic regression, dv: annotation decision
levels: prompt and task

Human vs. LLMs

- LLMs coded more text as racist
- LLMs more sensitive to
  - coverage on crime + outgroup
  - numbers + migration
  - racist slurs

CONCLUSION & RECOMMENDATIONS

- If intersubjective truth matters, the subjects matter
- Be aware of coder selection and prompting
- Learn from and accept variance or disagreement
- Consider polling for potentially relevant attitudes
- Consider inclusive annotation for constructs of marginalization

FIND OUR PREPRINT :)

HOW CAN WE INCLUDE MARGINALIZED VOICES IN CSS METHODS?

WHY DICTIONARIES?

Chances	Transparent, scalable, efficient, accessible, etc. Remains important part of data collection Can be included as feature to ML
Challenges	Polysemy, lack of context, domain dependency and top-down selection bias: ***whose words count?***

TURNING DICTIONARIES INTO PARTICIPATORY METHOD?

Using surveys for dictionary creation to mitigate selection bias by including perspectives underrepresented in academia

Examples	Constructs of marginalization, self-descriptions, microaggressions, algospeak, slang, etc.
Chances	Equitable, sensitive, culture and context-aware
Challenges	Ethical considerations,resource intensive, introducing other bias, power dynamics

A MODULAR PROCESS PROTOTYPE

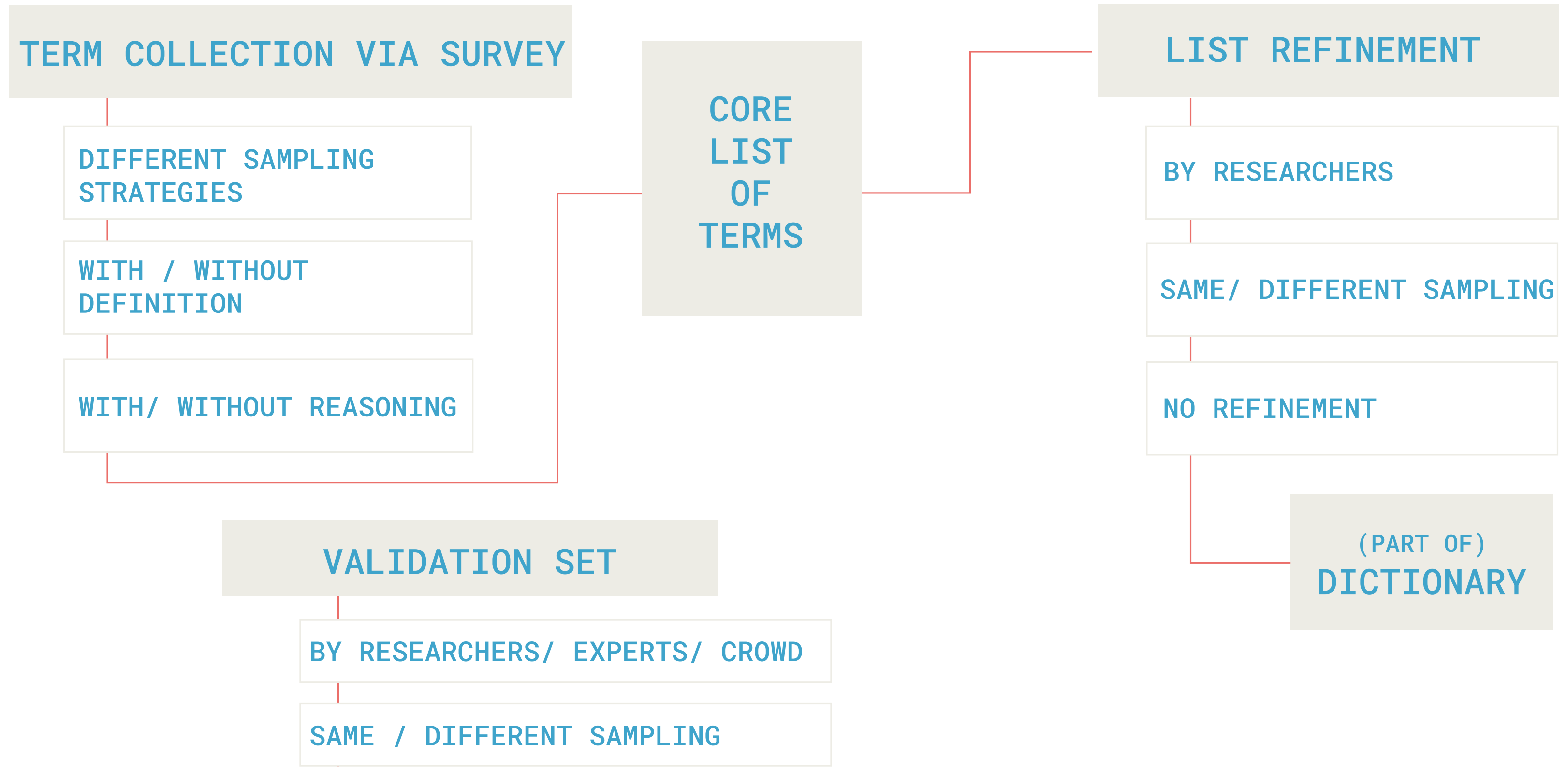

Figure 4 Draft process prototype for bottom-up design of dictionary creation

CASE STUDY: RACIST LANGUAGE

Research Aim Studying racist language in German mainstream and far-right alternative news media

- + two surveys: snowball sampling, quota-based sampling
- + with definition of racist language
- + refined and validated by researchers
- + combined with other dictionary creation methods (manual coding, glossaries)
- + pipeline with direct co-occurrence and topic modeling

RECOMMENDATION & OPEN QUESTIONS

- (Re)consider usefulness of dictionaries in text-as-data studies
- Consider how your dictionary creation might have selection bias
- How can we ensure inclusion of participants in most ethical way?
- What other bias do we introducing with this approach?

MEASURING RACISM AND RELATED CONCEPTS USING COMPUTATIONAL METHODS – A SYSTEMATIC LITERATURE REVIEW
KATHIRGAMALINGAM, LIND & BOOMGAARDEN

AGREE TO DISAGREE? HUMAN AND LLM CODER BIAS FOR CONSTRUCTS OF MARGINALIZATION
KATHIRGAMALINGAM, LIND, BERNHARD-HARRER & BOOMGAARDEN

TURNING DICTIONARIES INTO BOTTOM-UP METHODS? KATHIRGAMALINGAM & BOOMGAARDEN
PUSHING THE LIMITS? RACIST LANGUAGE USE IN GERMAN NEWS MEDIA KATHIRGAMALINGAM

AHRABHI KATHIRGAMALINGAM
ahrabhikat.bsky.social | ahrkat.github.io

SUPERVISED BY HAJO BOOMGAARDEN & FABIENNE LIND