Utrecht University

UMC Utrecht

Fairness in Clinical NLP: A Scoping Review of Challenges and Opportunities

Daniel Anadria

2024-05-23

Introduction

- Ongoing side-project
- Short version accepted at EWAF’24 (European Workshop on Algorithmic Fairness) in Mainz, Germany
- Goal: To map research gaps in clinical NLP – challenges and opportunities
- System safety discussion & contextualization of the topic

Agenda

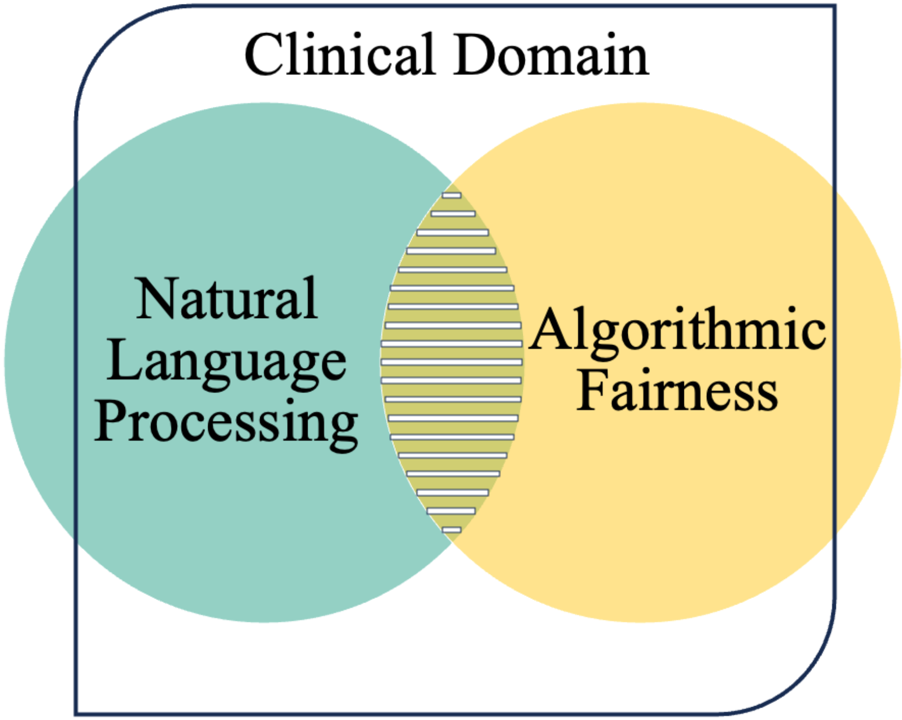
- Methodology
- Background
- Challenges & Opportunities
  - Protected Groups
  - Method Selection
  - Data Sharing & Privacy
  - Generalizability
  - Natural Language Generation
  - Multimodal Learning

Discussion

Methodology

Methodology

- Scoping review
- Extensive list of key query terms related to:
  - NLP
  - Algorithmic Fairness/Bias
  - Healthcare
- Preregistration on GitHub
- 7 scholarly databases and 3 search engines



Databases / Search Engines

PubMed
Embase®
Web of Science™

Scopus
Google scholar
ACM DL DIGITAL LIBRARY

IEEE Xplore® Digital Library
SEMANTIC SCHOLAR
ScholarAI
AI you can trust

Utrecht University

UMC Utrecht

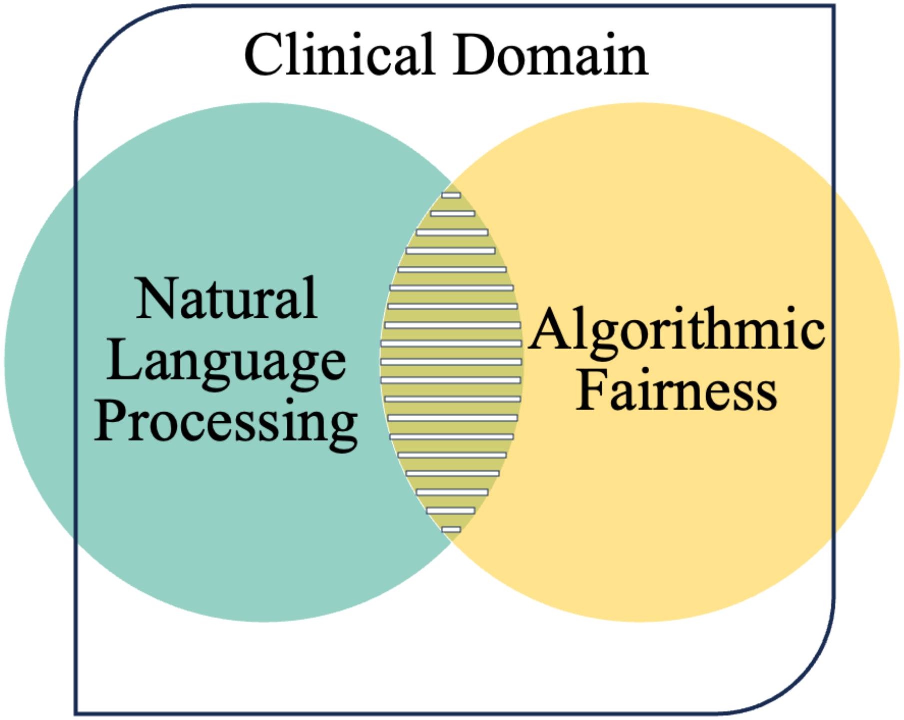
Methodology

Collection: 18-Oct to 25-Oct 2023

358 papers added for screening

24 key inclusions (i.e. applied studies in clinical NLP examining input or output fairness)

Various great discussion papers



Background

Utrecht University

UMC Utrecht

Background

- Clinical text as valuable input for automated decision-support systems
- Faithfulness to data means faithfulness to its biases
- Representational harms perpetuate social stigma and stereotyping of patient groups
- Allocative harms systematically deny patients access to opportunities and resources

Background

- Both technical and non-technical interventions are needed to mitigate harm in socio-technical systems (i.e. healthcare)
- Previous studies have proposed technical interventions – fairness auditing and bias mitigation methodologies
- Scarcity of evidence synthesis in fairness of clinical NLP pipelines
- Lack of clarity as to when a computationally feasible fairness intervention is clinically legitimate

Challenges & Opportunities

Challenges & Opportunities

Protected Groups

Method Selection

Data Sharing & Privacy

Generalizability

Natural Language Generation

Multimodal Learning

Protected Groups

- Establishing fairness across all demographic groups might not be feasible – whom to protect is a choice
- Studies examine a narrow scope of groups – sex, race/ethnicity, and to a lesser extent age
- Field is dominated by US-centric protected groups
- Consider groups that aren’t studied: individuals with mental health diagnoses, various forms of disability, individuals admitted to a hospital during the weekend vs. on a weekday, …

Utrecht University

UMC Utrecht

Protected Groups

- The difference in the geographical and cultural context on which local demographics should be considered protected remains under-examined

- The choice of whom to protect should be motivated by the local clinical and broader societal context

- Studies focus on more numerous protected groups (i.e. a utilitarian approach) and leave a gap w.r.t. protecting smaller-sized groups such as those at the intersection of multiple disenfranchised identities

Protected Groups

- The measurement of group membership is noisy - ranging from fully absent to the use of various proxies

- Majority of studies do not report how group labels were constructed

- When group information is fully absent, data imputation methods can estimate group membership. Robust indirect estimation methods such as Bayesian Improved Surname Geocoding are needed beyond the US context

Method Selection

- Fairness auditing and harm mitigation approaches carry many researcher degrees of freedom.
- Motivations behind decisions made are rarely motivated (e.g. why was this operationalization of fairness used?)
- Not every computationally feasible approach has clinical legitimacy.

Method Selection

- For example, Minot et al. (2022) propose a method to ‘debias’ clinical text by removing the most-gendered tokens. While the approach removed terms such as “he”, “his”, “she”, “her”, it had also erased medically valuable terms such as “urinal”, “prostate”, “hysterectomy”, “vaginal”, and “osteoporosis”

- While there is a plethora of fairness metrics and bias mitigation methodologies, there is a lack of clarity as to when an approach is appropriate. Authors avoid deliberating on their choices.

Data Sharing and Privacy

- Acquisition of diverse clinical datasets as challenge

- Group information is often omitted due to anonymization or institutional blindness. This renders many public datasets unusable for for fairness auditing and harm mitigation.

- Accurate prediction tools require comprehensive datasets which include sensitive information such as social-determinants of health.

Data Sharing and Privacy

- Lack of gold standard datasets - construction of accurate outcome labels for supervised learning tasks is costly, especially for large datasets.

- Opportunities to address this:
  - synthetic data
  - transfer learning
  - weak supervision approaches

Generalizability

- Lack of diversity in the datasets used in clinical NLP studies – MIMIC and MIMIC-derived datasets represent the majority of publicly available clinical text data
- We identified only three publicly available English language datasets not based on MIMIC notes
- Some studies had access to non-public hospital data, however, all these hospitals were based in the US

Generalizability

- Further motivated by this finding, we searched PhysioBank for public medical databases
- The only languages with representation other than English were Spanish and Brazilian Portuguese, each with a single database
- Gap in research on fair NLP in languages other than English and countries other than the US
- We know little about generalizability of proposed methodologies

Natural Language Generation

- In supervised learning tasks, outcome fairness metrics are derived from breaking down the global model performance (i.e. confusion matrix) by group

- Due to the lack of ground truth labels for generated output, validation of the output is challenging. The same holds for fairness audits.

- Probabilistic models prone to hallucination. Healthcare domain-enhanced LLMs appear to be more factual.

Natural Language Generation

- Particular use-case of the generated content will influence its fairness

  - Generation of Patient Discharge Letters – data completion task, relatively low harm if that data would otherwise be missing, but also beware of automation bias, and model collapse if generated summary used for future system input

  - Medical chatbots – inherently riskier as giving direct advice to the patient

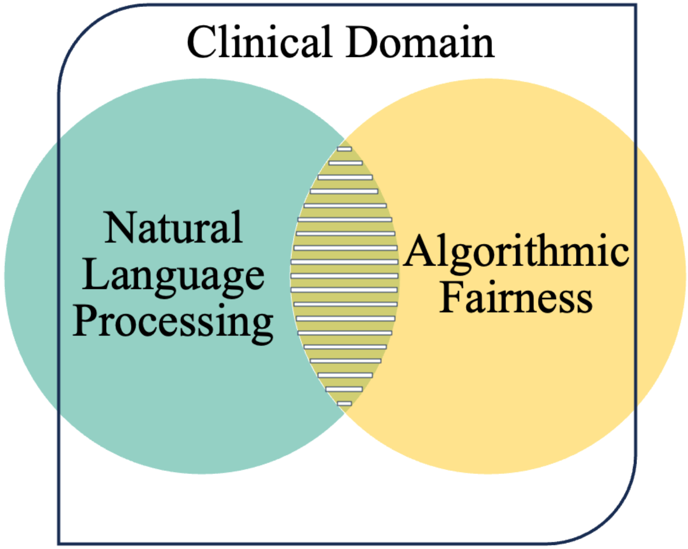
Multimodal Learning

- Zhou et al. found that multimodal models appear to show improvements in model performance, robustness, and fairness compared to single modality methods
- This finding remains to be demonstrated with text data
- Multimodal public datasets are limited
- GPT-4o are large multimodal models are already out!

Next Steps

- Adding system safety and design science methodology perspective in the introduction and discussion
- Turn into a long version
- Journal or conference

Discussion



Challenges & Opportunities

- Protected Groups
- Method Selection
- Data Sharing & Privacy
- Generalizability
- Natural Language Generation
- Multimodal Learning