NLP System for Mining Social Determinant of Health From Clinical Notes and its Fairness Evaluations

Zhecheng Sheng
Insititute of Health Informatics
University of Minnesota, Twin Cities
Minneapolis, USA
sheng136@umn.edu

2022 IEEE 10th International Conference on Healthcare Informatics (ICHI) | 978-1-6654-6845-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICHI54592.2022.00076

***Abstract*—This research proposal talks about recognizing SDoH components from clinical text using machine learning approaches. Upon developing a best model in practice for the task, existing tools are employed to assess and recalibrate machine biases. This study plans to use housing related entities as start point.**

Keywords—SDoH, NLP, AI bias

I. Introduction

Social determinant of health (SDoH) includes a wide range of components in the environments that closely related to patient health conditions and outcomes. SDoH is also known to potentially contribute to social challenges such as health disparities, discrimination, and inequality. Leveraging the study of SDoH through modern informatics tools and machine learning techniques enables physicians to better understand patient characteristics and provide better patient care. It is often costly to collect social information at a large scale, but these information are typically enriched in the unstructured clinical notes. It is reported SDoH factors can be successfully identified using NLP techniques through either rule-based methods, unsupervised or supervised learning approaches. [1] However, the variety of the information could be narrowed in some SDoH categories due to the limitation of lexicon curations. In addition, it is known that machine learning models trained by Electronic Health Records without calibration can lead to bias and fairness issues [2], and the model itself may learned that hidden feature to escalate that prejudice. Such issues in SDoH identification system are largely unexplored. Thus, in this research project, our objectives are to first develop a learning-based NLP framework to extract certain SDoH information from unstructured text. Next, we will examine the bias and fairness of our models.

II. Methods

Data collection: we plan to use the medical records of the patient cohort that visited for COVID-19 starting from March, 2020 in the Fairview database. The cohort currently contains 170,252 encounter records and we will first focus on SDoH categories related to individual financial stability such as housing status and income. Those categories are also most exposed and studied so we can have benchmarks to compare with.

Manual annotation: In order to develop gold standard corpus with mentioned SDoH, we have first gathered housing related keywords to curate lexicons for the category through two approaches. The first one, we collect a list of key terms and synonyms by consulting with domain experts and searching associated terms in UMLS Metathesaurus. The second one is to use locally pre-trained word embedding models to select semantically similar terms and misspellings. We will combine the terms list to sample notes for annotations. Alternatively, we may use relevant ICD codes to identify patients for selecting notes.

Methods: We will compare different models using the annotated corpus. We plan to use a subset (80%) of corpus to train traditional machine learning models or fine-tuned BERT models. We will also use existing NLP systems (e.g., cTAKES, MetaMap, CLAMP) as baselines to compare the results. To evaluate the fairness of these models, we will compare their performances on characteristics (e.g., race, gender, ethnicity) using some developed tools (Aequitas [3], LIME [4], etc.) and recalibrate our models accordingly. We further consider conducting cross-site evaluation to test the generalizability of our approach.

Expected results: We will compare various methods using 20% of manual annotated corpus using standard evaluation metrics (e.g., prediction, recall, F-score). We expect learning-based models to outperform existing baselines on the SDoH recognition tasks. Also, we expect to see the metrics evaluations vary among social subgroups initially and the imbalance is mitigated after the model recalibrations.

References

[1] Braja G Patra, Mohit M Sharma, Veer Vekaria, Prakash Adekkanattu, et al., Extracting social determinants of health from electronic health records using natural language processing: a systematic review. https://doi.org/10.1093/jamia/ocab170

[2] Hale M Thompson, Brihat Sharma, Sameer Bhalla, et al., Bias and fairness assessment of a natural language processing opioid misuse classifier: detection and mitigation of electronic health record data disadvantages across racial subgroups. https://doi.org/10.1093/jamia/ocab148

[3] Pedro Saleiro and Benedict Kuester, et al., Aequitas: A Bias and Fairness Audit Toolkit, 2019, https://arxiv.org/pdf/1811.05577.pdf

[4] Ribeiro, Marco Tulio and Singh, Sameer and Guestrin, Carlos, "Why Should I Trust You?": Explaining the Predictions of Any Classifier. https://doi.org/10.1145.2939672.2939778

2575-2634/22/$31.00 ©2022 IEEE
DOI 10.1109/ICHI54592.2022.00076