Drew Walker
Large HW 3

Gold Standard Creation for Detecting Bias

Following the data preprocessing from previous labs and working from the coding ontology outlined here , which is derived heavily from Beach et. al's studies and provider recommendations surrounding scare quotes in medical charts.[1,2] I randomly sampled 600 tokenized sentences from

- 3,842 quoted sentences
- 2,909 unique charts
- 545 caregivers
- 928 patients
- 1479 Admissions

I realized (late in the game) that the time required to gain access to the MIMIC-III dataset would be too large of a burden in combination with the coding for undergrad assistance, so in this case I was the only one able to code the quoted sentences. On the bright side, my kappa score was 1. (lol). Still, I made sure to write code to calculate inter-rater agreement when I'm able to get my wife (who is a PA student) set up in the MIMIC-III dataset. Undergrad coders (Sam Lucius, Ariana Gassel, Eugene Lee) worked on a categorization task for a Delta 8 THC Twitter study.

In this corpus, I was able to identify:

- 252 instances of "Probably Useful", (Code = 0) or medically appropriate quotes, defined as quotes that:
    - Provides important contextual clues for clinical info
    - Conveys effect of illness on patient's life
    - Conveys patient values or preferences
- 110 instances of "Probably Harmful" quotes (code=1). These are quotes that:
    - Cast doubt on integrity of patient to provide reliable testimony or
    - Convey ridicule, contempt, or frustration by highlighting unsophisticated language or limited knowledge
- 36 instances of "Possibly Harmful" quotes (code =2) which are defined as:
    - Neutral phrases where quotes serve no clear purpose but could be read as scare quotes conveying doubt or judgment.
- 190 instances where the quotes were not being used to directly quote the patient (code =3). This was defined as:
    - Was someone else's quotes (family, other providers)
    - Used to describe the title of procedures or medical terminology
    - At first, this was included as a separate column for "not_applicable", but I wanted to include it in the classifier to help determine the not applicable notes in one round of classification instead of trying to classify out non-applicable quotations and then classify the corpus a second time for types of quotes in two steps.

References

Drew Walker
Large HW 3

1. Beach MC, Saha S, Park J, et al. Testimonial Injustice: Linguistic Bias in the Medical Records of Black Patients and Women. *J GEN INTERN MED*. Published online March 22, 2021. doi:10.1007/s11606-021-06682-z
2. Beach MC, Saha S. Quoting Patients in Clinical Notes: First, Do No Harm. *Ann Intern Med*. 2021;174(10):1454-1455. doi:10.7326/M21-2449