n2c2 2022 Track 2 Data Overview

This document presents a summary of the data used in the n2c2 2022 Track 2 extraction challenge (https://n2c2.dbmi.hms.harvard.edu/2022-track-2). Track 2 explores the extraction of social determinant of health (SDOH) information from clinical text and utilizes an updated and deidentified version of the Social History Annotated Corpus (SHAC) [1]. SHAC includes annotated clinical text from MIMIC-III and the University of Washington (UW) and is the divided into training, development (dev), and test splits. Table 1 summarizes the number of documents by source and split in SHAC.

Track 2 consists of three subtasks, A, B, and C. Subtask A focuses on in-domain extraction, Subtask B focuses on generalizability to a new domain, and Subtask C focuses on learning transfer. Table 2 summarizes the train and test data associated with each subtask. Subtasks A and B share a common training set ($\mathcal{D}_{train}^{mimic}, \mathcal{D}_{dev}^{mimic}$). The Subtask C train set consists of the union of the Subtask A train set and Subtask B test set ($\mathcal{D}_{train}^{mimic}, \mathcal{D}_{dev}^{mimic}, \mathcal{D}_{train}^{uw}, \mathcal{D}_{dev}^{uw}$).

Table 1. SHAC data partition counts (# of documents)

Source	Train	Dev	Test
MIMIC-III	1,316 ($\mathcal{D}_{train}^{mimic}$)	188 ($\mathcal{D}_{dev}^{mimic}$)	373 ($\mathcal{D}_{test}^{mimic}$)
UW	1,751 ($\mathcal{D}_{train}^{uw}$)	259 ($\mathcal{D}_{dev}^{uw}$)	518 ($\mathcal{D}_{test}^{uw}$)

Table 2. Extraction challenge data partitions counts (# of documents)

Subtask	Train	Test
A – Extraction	1,504 ($\mathcal{D}_{train}^{mimic}, \mathcal{D}_{dev}^{mimic}$)	373 ($\mathcal{D}_{test}^{mimic}$)
B – Generalizability	1,504 ($\mathcal{D}_{train}^{mimic}, \mathcal{D}_{dev}^{mimic}$)	2,010 ($\mathcal{D}_{train}^{uw}, \mathcal{D}_{dev}^{uw}$)
C – Learning Transfer	3,514 ($\mathcal{D}_{train}^{mimic}, \mathcal{D}_{dev}^{mimic}, \mathcal{D}_{train}^{uw}, \mathcal{D}_{dev}^{uw}$)	518 ($\mathcal{D}_{test}^{uw}$)

1. Kevin Lybarger, Mari Ostendorf, and Meliha Yetisgen, "Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction," Journal Biomedical Informatics, vol. 113, p. 103631, 2021. DOI: 10.1016/j.jbi.2020.103631