From Notes to Needs: Extracting SDoH from CHW Narratives with NLP

Ankita Mishra[1], Charmi Patel[1], Navika Maglani[1], Kelly McCabe[2], Roselyne Tchoua[1], Jacob Furst[1], Daniela Raicu[1]
[1]College of Computing and Digital Media, DePaul University, Chicago, IL
[2]Sinai Urban Health Institute, Sinai Chicago, Chicago, IL

Introduction

- **Social Determinants of Health** (SDoH) such as food, housing, and transportation strongly influence health outcomes[1] but are under-documented in structured Electronic Health Records.
- Community Health Workers (CHWs) record rich SDoH information in free-text contact notes, but their unstructured format limits large-scale analysis.
- Automating SDoH extraction from CHW documentation can help health systems better identify and address social needs. This approach minimizes reliance on extensive manual annotation[2].

Goals

- Develop an **automated Natural Language Processing framework** to detect food insecurity — a key SDoH — from CHW contact notes.
- Improve detection accuracy using **semantic similarity** and **domain-specific vocabulary refinement**.
- Validate the model's predictions using **human-annotated labels**.

Dataset Overview

- Data provided by **Sinai Urban Health Institute (SUHI)**, Chicago, IL.
- Includes **2,613 patient records** (July 2020- Feb 2024).
- **1,298 contact notes** and **92 resource notes** focused on food insecurity.
- Each record contains:
  - **129 variables** (demographics, SDOH indicators, referrals, and CHW logs).
  - Resource notes represent structured documentation of resources provided to each patient; contact notes provide narrative context.

Fig. 1: Overlap among patient records with a food need, food resource notes, and contact notes.

Methodology

Using the 92 food resource notes and 1,298 contact notes, a three-stage framework was developed:

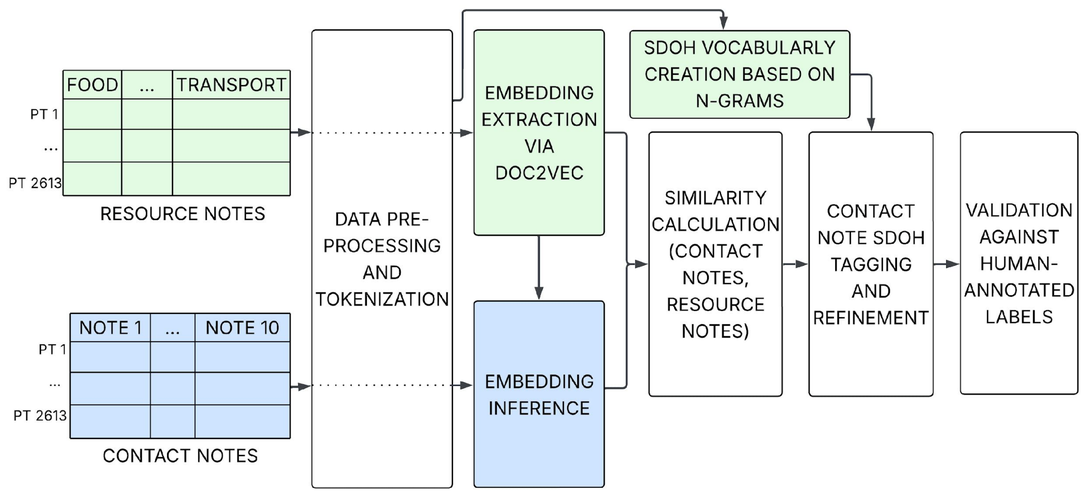

Fig. 2: Label Assignment Pipeline for classifying contact notes culminating in the final Label Assignment step.

A) Semantic Similarity: Compute Doc2Vec embeddings for contact and resource notes; measure cosine similarity.

Optimal similarity threshold (0.85) was established based on manual review of the top 10 notes explicitly referencing food insecurity.

B) Vocabulary Refinement: Create a domain-specific food vocabulary using N-grams from resource notes.

Created three levels of vocabularies:

1. Frequency-based: N-grams with frequency ≥ 2
2. Refined-based: remove N-grams like "provided assistance"
3. Rule-based: exclude names, geographic references, adjectives

C) Validation: Compare predicted labels with human-annotated ground truth of 1,298 contact notes using Fleiss' Kappa statistic.

Results

- Optimal similarity threshold: > 0.85, identifying 462 semantically relevant food-related contact notes.
- High inter-rater agreement among human annotators (Fleiss' κ = 0.955).
- The **rule-based vocabulary** approach **minimized false positives** and provided high-precision tagging.

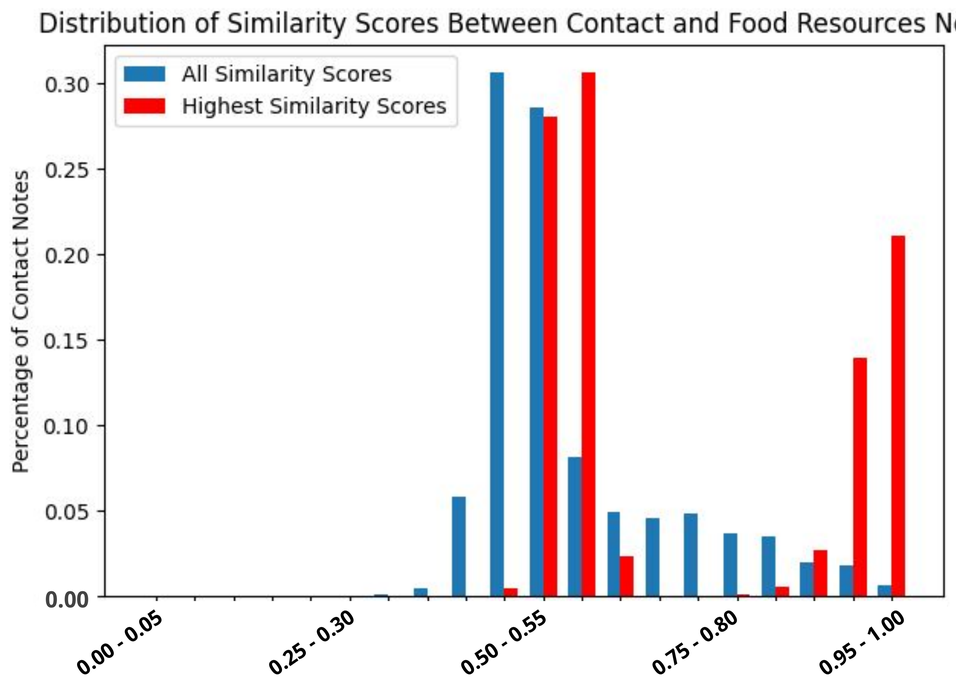

Fig. 3: Cosine Similarity Scores of Contact Note Vectors showing the distribution of contact notes based on their similarity to food resource notes.

Table 1: Evaluation of vocabulary list approaches for tagging food-related social needs in contact notes, using human-annotated labels as ground truth.

Vocabulary List	Accuracy	Sensitivity	Specificity
Frequency-based	0.694	0.958	0.679
Refined-based	0.790	0.958	0.781
Rule-based	**0.990**	**0.915**	**0.994**

Conclusions

- **Rule-based vocabulary** framework achieved **99% accuracy** in identifying food-related SDoH from CHW notes.
- Semantic similarity, enhanced with domain vocabulary, enables **interpretable and scalable NLP for SDoH**.
- Reduces dependency on manual annotation, saving time and resources.

Future Work

Model Enhancements:

- Compare baseline Doc2Vec with transformer-based models (BERT, BlueBERT).
- Test alternative similarity metrics (e.g., Euclidean, Jaccard).

Scope Expansion:

- Generalize to other SDoH domains: housing, employment, transportation.
- Integrate negation detection and patient-level temporal modeling across multiple contact notes.

Acknowledgements

This work was supported by the **National Science Foundation (DBI 2326676)** and the **Medical Informatics experiences (MedIX) lab at DePaul University.**

We thank the **Sinai Urban Health Institute** for providing data access and the Community Health Workers whose dedication makes this research possible.

References

[1] Braveman P, Gottlieb L. *The social determinants of health: it's time to consider the causes of the causes*. Public Health Rep. 2014 Apr;129 Suppl 2:19-31.

[2] Lybarger K, Dobbins NJ, Long R, Singh A, Wedgeworth P, Uzuner O, et al. *Leveraging natural language processing to augment structured social determinants of health data in the electronic health record*. J Am Med Inform Assoc. 2023 Aug 16;30(8):1389-97. doi: 10.1093/jamia/ocad073.