Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200145

Discovering New Social Determinants of Health Concepts from Unstructured Data: Framework and Evaluation

Joao H. BETTENCOURT-SILVA[a], Natalia MULLIGAN[a], Marco SBODIO[a], John SEGRAVE-DALY[b], Richard WILLIAMS[c], Vanessa LOPEZ[a], Carlos ALZATE[a]
[a] *AI for Health & Social Care Research Group, IBM Research, Ireland*
[b] *IBM Watson Health, Dublin, Ireland*
[c] *NIHR Greater Manchester Patient Safety Translational Research Centre, University of Manchester, Manchester, UK*

Abstract. Social determinants of health (SDoH) are the complex set of circumstances in which individuals are born, or with which they live, that impact their health. Integrating SDoH into practice requires that information systems are able to identify SDoH-related concepts from charts and case notes through vocabularies or terminologies. Despite significant standardisation efforts across healthcare domains, SDoH coverage remains sparse in existing terminologies due to the broad spectrum of this domain, ranging from family relations, risk factors, to social programs and benefits, which are not consistently captured across administrative and clinical settings. This paper presents a framework to mine, evaluate and recommend new multidisciplinary concepts that relate to or impact the health and well-being of individuals using a word embedding model trained from a large dynamic corpus of unstructured data. Five key SDoH domains were selected and evaluated by domain experts. The concepts resulting from the trained model were matched against well-established meta-thesaurus UMLS and terminology SNOMED-CT and, overall, a significant proportion of concepts from a set of 10,000 candidates were not found (31% and 28% respectively). The results confirm both the gaps in current terminologies and the feasibility and impact of the methods presented in this paper for the incremental discovery and validation of new SDoH concepts together with domain experts. This sustainable approach facilitates the development and refinement of new and existing terminologies and, in turn, it allows systems such as Natural Language Processing (NLP) annotators to leverage SDoH concepts across integrated care settings.

Keywords. Social Determinants of Health, Terminologies, Vocabularies

1. Introduction

Throughout human societies, health and illness are not homogeneously distributed due to disparities in social and economic aspects. The aspects that underpin such disparities are commonly referred to as the Social Determinants of Health (SDoH) and examples include intangible factors such as "political, socioeconomic, and cultural constructs, as well as place-based conditions including accessible healthcare and education systems, safe environmental conditions, well-designed neighbourhoods, and availability of

healthful food" [1]. The significant impacts on health and costs associated with SDoH have been reported [2] and in order for social aspects to be best utilised for service evaluation or outcome assessment and improvement, this information should be routinely collected by organisations across the integrated care continuum. However, despite technological advances, SDoH information is neither routinely nor systematically collected in Electronic Health Records (EHRs) and lacks standardisation [3,4]. The US Institute of Medicine was among the first to identify SDoH domains and recommend they be routinely collected and made available in EHRs [3]. When available, SDoH information is often stored in natural language as part of unstructured case notes which further hinders efforts to identify cases and select cohorts. Terminologies and meta-thesaurus such as SNOMED-CT and UMLS can be used to encode health-related observations or activities into a canonical form. A recent exploratory analysis [5] has confirmed the limitations of current healthcare vocabularies in documenting SDoH-related information and has called for more comprehensive, coherent, and user-friendly SDoH code sets [6] to support and facilitate the rapidly evolving health care use cases[5]. Techniques for bridging clinical and social domains using word embeddings have been proposed to explore cross-domain spaces [7] and calls for action to create standards for representing SDoH data have been suggested [3]. Recent initiatives include Health Level Seven's (HL7) Gravity Project [8] and LOINC's models for the representation of screening assessments and measures of SDoH [9]. Healthcare terminologies have begun adopting concepts in social domains through consultation with expert groups but sustainable processes and methodologies to support these efforts are lacking.

This paper proposes a computational approach to facilitate the process of seeking new terminology concepts as well as refreshing existing terminologies. The framework proposed in this paper leverages existing word embedding techniques [10, 11, 12] in order to mine new candidate terms from large corpora of unstructured data. Other approaches have focused on developing ontologies and taxonomies [11] yet the work presented in this paper aims to incorporate expert humans-in-the-loop in order to recommend new concepts not included in existing terminologies. This approach can also be used to refine existing terminologies with categories for specific domains.

2. Methods and Experiments

The proposed framework takes a list of seeded terms as input and returns a list of relevant concepts with domain categories assigned. This process relies on training a model to discover relationships between terms in relevant corpora and is supported by a human-in-the-loop evaluation step to ascertain the relevance of the discovered terms. Framework steps are summarised in Figure 1 and described in detail in the next sections.

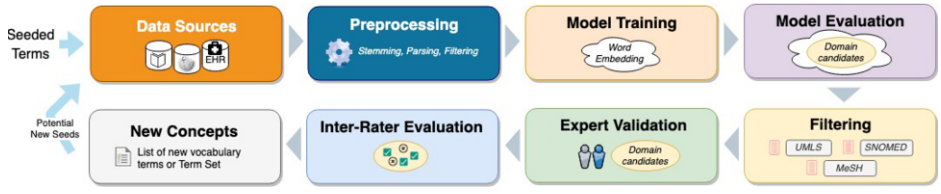


Figure 1. Proposed framework for recommending new terminology concepts.

Training a comprehensive SDoH model requires a list of relevant seeded term(s) for the first iteration of the process. Together with domain experts (DEs), 57 distinct seeded terms were selected for our experiments. The terms were extracted from relevant literature definitions of SDoH (WHO, Marmot & Wilkinson [13], CDC, Wikipedia).

2.1. Data Sources and Preprocessing

The first step consists of obtaining a corpus of unstructured data suitable for model training. A collection of case notes may be obtained from EHRs, case notes, books, or other sources of knowledge. Data sources should ensure coverage of the relevant domain areas in the seeded terms and this can be achieved in consultation with the DEs.

In order to train the comprehensive SDoH model presented in this paper, we prepared a corpus of text by crawling the set of Wikipedia pages hierarchically linked through DBpedia's taxonomy to the seeded terms. We curated the collected corpus by removing irrelevant pages, and obtained 2,134 distinct pages, with 3,282,508 non-unique terms. Finally, we pre-processed the collected corpus by removing metadata, images, and tables, and by applying lemmatisation and filtering of non-textual characters.

2.2. Model Training and Model Evaluation

This step focuses on the application of word embedding techniques on the prepared corpus of text. The purpose is to train a model able to capture contextual similarities between terms and term classes. The SDoH model was built using state-of-the-art embedding techniques to discover domain-specific n-grams from the corpus of text [10]. We trained a skip-gram model with negative sampling and subsampling of the frequent words in the training corpus. Experiments to tune the model hyperparameters were undertaken and found 100-dimension embeddings, computed with a context window of 5, and n-grams of size 3, provided a rich domain representation for our task. Other techniques such as pre-trained word embedding models may be used in this step yet these typically only yield unigrams. The resulting trained model contained 77,232 terms.

For evaluation, the DEs selected five terms each defining specific SDoH domains of interest: *Housing, Criminal Justice, Financial Services, Health Behaviours* and *Homeless*. For each of the five terms, a set of similar (candidate) terms was computed by selecting the closest 2000 neighbours in the model's high-dimensional space using Euclidean distance. The ranked set of 2000 candidate terms was considered large enough to be evaluated by DEs in a timely manner. The set was later randomised for the DEs.

2.3. Filtering and Expert Validation

In the filtering step, candidate terms from each of the five domains of interest were programmatically checked against UMLS and SNOMED-CT terminology. Each term was labelled as *matching* when either a partial or exact match was found in the terminology (via its API), or *no match*. It is possible to use more complex matching algorithms and to filter out terms that already exist in terminologies (e.g. synonyms), however, for the purpose of developing and evaluating the framework proposed in this paper, no filters were applied, and partial and exact matching were combined.

The Expert Validation step consists of gathering information from DEs about the candidate terms' relevance in their specific domains. This was achieved by carrying out

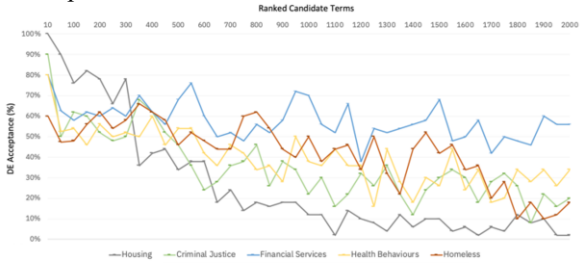
90 minute workshop sessions with social workers where at least 3 different DEs of the same specialty evaluated the same set by responding *accept*, *reject* or *pass* for each term.

2.4. Inter-Rater Evaluation and Recommendation of New Concepts

In this step the results are analysed in order to determine consensus among DEs (inter-rater agreement based on simple majority was used due to small number of DEs, however, other methods are possible). Decisions are made for all candidate terms and a curated list of recommended new concepts is the output of the framework (Figure 1). This list may serve to refine the input list of seeded terms in order to further build or refine models.

3. Results and Discussion

The framework proposed in this paper enabled the training of a model based on SDoH seeded terms. The model was evaluated on five key SDoH domains through workshops with DEs (each domain seen by 3 distinct DEs with the exception of *Housing*). Figure 2 summarises the DEs evaluation results for each domain and depicts a trend where lower ranked terms were accepted less often than those nearest to each of the five domain terms.



Domain	DEs	Accepted (Consensus) Full Set	Accepted (Consensus) Top 100	Accepted (Any) Full Set	Accepted (Any) Top 100	Linear Trend (R^2)
Housing	2	24.65% (493)	83%	68.85% (1377)	96%	0.7236
Criminal Justice	3	34.25% (685)	59%	68.10% (1362)	87%	0.6261
Financial Services	3	57.25% (1145)	61%	90.00% (1800)	84%	0.2384
Health Behaviours	3	38.55% (771)	56%	66.20% (1324)	76%	0.5965
Homeless	3	42.40% (848)	48%	63.90% (1278)	67%	0.6572
All Domains (Average)		39.42% (3942)	61.4%	71.41% (7141)	82%	

Figure 2. Proportion of terms accepted by DEs. For each domain, terms were ranked using Euclidean distance in the SDoH embedding model high-dimensional space.

Table 1 shows the results of matching each of the terms against SNOMED-CT and UMLS. Overall our findings show a large number of new terms that did not match either (on average 76% in SNOMED-CT and 83% in UMLS), with SNOMED-CT showing a better coverage throughout. Across all domains, on average, 40% of terms are accepted by DEs consensus of which 30% of them were not found in either (no match). *Financial Services* was the domain with most new relevant terms found across both SNOMED-CT (48%) and UMLS (50%) suggesting this area may be most lacking. *Housing* and *Homeless* are the two domains with the least number of new terms found and accepted by DEs (respectively 16% and 23% for SNOMED-CT, and 19% and 30% UMLS). Out of all 10,000 terms evaluated in this paper, the framework yielded 30% as concepts to be suggested for inclusion in terminologies and highlights terms from other terminologies of relevance to this domain. A limitation of the modelling approach is yielding less relevant n-gram combinations (e.g. '*housing provides' versus* '*housing programs')*. However, the human-in-the-loop step helps filter out such cases. A subset of the concepts found could also be synonyms of existing terminology concepts and terms could appear

in both UMLS and SNOMED-CT. Since no partial or exact matches were found, however, it is still relevant to suggest bringing these terms in to terminologies.

Table 1. Results of matching terms against SNOMED-CT and UMLS.

	SNOMED-CT				UMLS			
	Match % (N)		No Match % (N)		Match % (N)		No Match % (N)	
Domain	**Accepted**	**Rejected**	**Accepted**	**Rejected**	**Accepted**	**Rejected**	**Accepted**	**Rejected**
Housing	8.50% (170)	25.75% (515)	16.15% (323)	49.60% (992)	6.05% (121)	17.80% (356)	18.60% (372)	57.55% (1151)
Criminal Justice	9.95% (199)	5.75% (115)	24.30% (486)	60.00% (1200)	7.55% (151)	3.75% (75)	26.70% (534)	62.00% (1240)
Financial Services	9.40% (188)	3.40% (68)	47.80% (956)	39.40% (788)	7.65% (153)	4.20% (84)	49.55% (991)	38.60% (772)
Health Behaviours	10.85% (217)	11.15% (223)	27.70% (554)	50.30% (1006)	6.80% (136)	7.50% (150)	31.75% (635)	53.95% (1079)
Homeless	19.30% (386)	17.25% (345)	23.05% (461)	40.40% (808)	12.75% (255)	10.90% (218)	29.60% (592)	46.75% (935)
Total (Average)	11.60%	12.66%	27.80%	47.94%	8.16%	8.83%	31.24%	51.77%
	24.26%		75.74%		16.99%		83.01%	

4. Conclusions

The lack of standardisation, terminologies and code sets in complex multidisciplinary domains remains a barrier to the successful documentation and use of SDoH in practice. This work assessed the coverage of existing terminologies in selected SDoH domains using expert humans-in-the-loop and proposed a framework for the discovery of SDoH concepts from unstructured data. Further work is needed to refine model training, to include additional categories for DEs to assign to terms and allowing new term/code sets to be defined in the process, and to inspect SDoH in other terminologies (e.g. ICD-11).

References

[1] New England Journal of Medicine – Catalyst; 2017. Social Determinants of Health (SDOH). Retrieved from https://catalyst.nejm.org/social-determinants-of-health

[2] McGovern L, Miller G, Hughes-Cromwick P. The relative contribution of multiple determinants to health. Health Affairs Health Policy Briefs. 2014 Aug 21;21.

[3] Hripcsak G, Forrest CB, Brennan PF, Stead WW. Informatics to support the IOM social and behavioral domains and measures. J Am Med Inform Assoc. 2015 Jul;22(4):921-4. doi: 10.1093/jamia/ocv035.

[4] Adler NE, Stead WW. Patients in context--EHR capture of social and behavioral determinants of health. N Engl J Med. 2015 Feb 19;372(8):698-701. doi: 10.1056/NEJMp1413945.

[5] Arons A, DeSilvey S, Fichtenberg C, Gottlieb L, Documenting social determinants of health-related clinical activities using standardized medical vocabularies, JAMIA Open. 2019. 2(1) 81–88

[6] Williams R, Kontopantelis E, Buchan I, Peek N. Clinical code set engineering for reusing EHR data for research: A review. J Biomed Inform. 2017 Jun;70:1-13. doi: 10.1016/j.jbi.2017.04.010.

[7] Bettencourt-Silva J, Mulligan N, Cullen C, Kotoulas S. Bridging Clinical and Social Determinants of Health Using Unstructured Data. Stud Health Technol Inform. 2018; 255:70-74.

[8] Health Level Seven, Gravity Project; 2019. Retrieved from http://www.hl7.org/gravity

[9] Vreeman D. 2018; Advancing the interoperability of social and behavioral determinants of health. LOINC. Retrieved from https://loinc.org/sdh/webinar

[10] Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and their Compositionality. arXiv preprint arXiv:1310.4546. 2013 Oct 16.

[11] Grefenstette G, Muchemi L. Determining the Characteristic Vocabulary for a Specialized Dictionary using Word2vec and a Directed Crawler. arXiv preprint arXiv:1605.09564. 2016 May.

[12] Gu Y, Leroy G, Pettygrove S, Galindo MK, Kurzius-Spencer M. Optimizing Corpus Creation for Training Word Embedding in Low Resource Domains: A Case Study in Autism Spectrum Disorder (ASD). AMIA Annu Symp Proc. 2018 Dec 5

[13] Marmot M, Wilkinson R. *Social Determinants of Health*, Oxford University Press, Oxford. 2005.