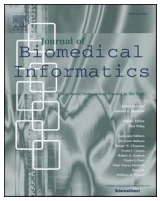
Journal of Biomedical Informatics 127 (2022) 103984

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing

Sifei Han [a], Robert F. Zhang [a,f], Lingyun Shi [a], Russell Richie [a], Haixia Liu [b], Andrew Tseng [c], Wei Quan [d], Neal Ryan [e], David Brent [e], Fuchiang R. Tsui [a,f,*]

[a] Tsui Laboratory, Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA, USA
[b] Central South University, Changsha, Hunan, CN
[c] Touro University Nevada, Henderson, NV, USA
[d] New York University Abu Dhabi, Abu Dhabi, AE
[e] Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA
[f] Perelman School of Medicine, University of Pennsylvania, PA, USA

ARTICLE INFO

Keywords
Social determinants of health
Natural language processing
Deep learning
Electronic health records

ABSTRACT

Objective: Social determinants of health (SDOH) are non-medical factors that can profoundly impact patient health outcomes. However, SDOH are rarely available in structured electronic health record (EHR) data such as diagnosis codes, and more commonly found in unstructured narrative clinical notes. Hence, identifying social context from unstructured EHR data has become increasingly important. Yet, previous work on using natural language processing to automate extraction of SDOH from text (a) usually focuses on an ad hoc selection of SDOH, and (b) does not use the latest advances in deep learning. Our objective was to advance automatic extraction of SDOH from clinical text by (a) systematically creating a set of SDOH based on standard biomedical and psychiatric ontologies, and (b) training state-of-the-art deep neural networks to extract mentions of these SDOH from clinical notes.

Design: A retrospective cohort study.

Setting and participants: Data were extracted from the Medical Information Mart for Intensive Care (MIMIC-III) database. The corpus comprised 3,504 social related sentences from 2,670 clinical notes.

Methods: We developed a framework for automated classification of multiple SDOH categories. Our dataset comprised narrative clinical notes under the "Social Work" category in the MIMIC-III Clinical Database. Using standard terminologies, SNOMED-CT and DSM-IV, we systematically curated a set of 13 SDOH categories and created annotation guidelines for these. After manually annotating the 3,504 sentences, we developed and tested three deep neural network (DNN) architectures – convolutional neural network (CNN), long short-term memory (LSTM) network, and the Bidirectional Encoder Representations from Transformers (BERT) – for automated detection of eight SDOH categories. We also compared these DNNs to three baselines models: (1) cTAKES, as well as (2) L2-regularized logistic regression and (3) random forests on bags-of-words. Model evaluation metrics included micro- and macro- F1, and area under the receiver operating characteristic curve (AUC).

Results: All three DNN models accurately classified all SDOH categories (minimum micro-F1 = 0.632, minimum macro-AUC = 0.854). Compared to the CNN and LSTM, BERT performed best in most key metrics (micro-F1 = 0.690, macro-AUC = 0.907). The BERT model most effectively identified the "occupational" category (F1 = 0.774, AUC = 0.965) and least effectively identified the "non-SDOH" category (F = 0.491, AUC = 0.788). BERT outperformed cTAKES in distinguishing social vs non-social sentences (BERT F1 = 0.87 vs. cTAKES F1 = 0.06), and outperformed logistic regression (micro-F1 = 0.649, macro-AUC = 0.696) and random forest (micro-F1 = 0.502, macro-AUC = 0.523) trained on bag-of-words.

Conclusions: Our study framework with DNN models demonstrated improved performance for efficiently identifying a systematic range of SDOH categories from clinical notes in the EHR. Improved identification of patient SDOH may further improve healthcare outcomes.

* Corresponding author at: 2716 South St., Philadelphia, PA 19146, USA.
E-mail address: tsuif@chop.edu (F.R. Tsui).

https://doi.org/10.1016/j.jbi.2021.103984
Received 16 September 2021; Received in revised form 28 December 2021; Accepted 29 December 2021
Available online 7 January 2022

1. Introduction

Social determinants of health (SDOH) are non-clinical factors (e.g., poverty, social environment, and unemployment) that have been shown to account for 30% to 55%[1] of mental and physical health morbidity. Diabetes, hypertension, depression, and suicidal behavior are all outcomes of SDOH-linked physical and mental morbidity [2]. SDOH also influences health care utilization. For example, including patient-specific SDOH improves risk prediction for outcomes such as 30-day hospital readmissions [3]. Moreover, SDOH account for 80% to 90% of modifiable health factors, whereas medical care accounts for only 10–20% of health factors [4]. Thus, addressing the SDOH within and outside of the healthcare system can potentially improve health and reduce gaps in the delivery of care [5]. However, extant research is limited in extracting SDOH from electronic health records (EHRs) such as narrative clinical notes to facilitate decision-making for patient care.

The most common approach to identifying SDOH leverages screening tools administered by frontline healthcare workers and caregivers, which aim to better understand patients' social status in clinical care [6]. However, most of these tools are ad-hoc and domain-specific (e.g., focusing on the social risks for intimate partner violence). Therefore, caregivers need to adapt questionnaires to the local clinical context. Moreover, developing and deploying these screening tools in clinical care require substantial resources and effort. Instead, as is the case with most clinical documentation, an alternate approach to systematically identifying SDOH is to leverage unstructured EHR data, i.e., narrative clinical notes. Navathe et al., for example, examined the prevalence of social risk factors in unstructured EHR data,[7] and found that social factors such as drug use, depression, housing instability, and poor social support could be found at much higher rates and oftentimes exclusively within clinical notes compared to structured EHR data (e.g., the prevalence of patients with identified social support increased by 40 times from 0.4% to 16% when clinical notes were reviewed). In addition, non-systematically reported SDOH such as ethnicity and marital status are often more missing-not-at-random in structured data compared to unstructured data, which introduces selection bias that must be addressed [8]. Hence, unstructured narrative clinical notes in the EHR can serve as a rich resource of SDOH information. However, manually extracting insights from these notes can be challenging and time-consuming due to the sheer quantity of reports and the unstructured and complex nature of language expressions.

Automating the extraction of SDOH information from text using natural language processing could therefore be useful, and indeed, many previous studies have shown the utility of NLP methods for this purpose [9,10]. While many of these studies have used conventional methods like regular expressions, dictionaries, or rule-based systems like cTAKES or Moonstone [11,12], many NLP applications from recent years, e.g., deep neural networks, which use layers of nonlinear processors to automatically learn successive layers of increasingly abstract features, often outperform such conventional approaches as well as simple machine learning methods based on, e.g., bag-of-words [13–15]. Despite this, Patra et al.'s [10] recent systematic review of studies using NLP for SDOH detection uncovered only seven studies using deep neural networks to detect SDOH information in narrative clinical notes (although DNNs are used elsewhere in clinical NLP [24,29–33]). Among these seven studies using DNNs, only one [16] used transformers (specifically, Bidirectional Encoder Representations from Transformers, BERT), a recent type of feedforward deep neural network with a self-attention mechanism that often demonstrated improved performance compared to convolutional and recurrent (e.g., LSTM) neural networks on many natural language processing tasks. However, to our knowledge, no studies compared all three types of networks on the same SDOH detection task. It is therefore unknown whether extraction of SDOH from clinical text could benefit from these latest advances in NLP. Further, in one study [17], its DNNs did *not* outperform simple machine learning methods, which raises questions about the circumstances in which DNNs are advantageous in SDOH detection.

However, even where DNNs (or other NLP tools) succeed at detecting SDOH information, the SDOH categories they are trained to detect tend to be chosen on a somewhat ad-hoc basis (such as in clinician-administered screening tools). That is, such systems either detect only one SDOH (the majority of studies on SDOH detection, according to Patra et al.'s review [10]), or, if a system extracts multiple SDOH, the SDOH categories appear not to be systematically chosen or based on any particular clinical conventions, and therefore may not represent the range of social determinants of health that most clinicians are primarily interested in. For example, although Feller et al. [17]had three clinicians generate a set of 30 SDOH, it is unclear whether this set reflects the importance perceived by these particular clinicians at their particular medical institutions, or the clinical field writ large. A possibly better strategy is to use clinical conventions as encoded in standard, widely used ontologies like SNOMED-CT or the DSM-IV [18,19] to identify SDOH categories recognized to be of clinical importance by a large set of clinician-researchers.

Thus, there is a need for a more comprehensive comparison of the full range of NLP tools – from conventional, rule-based tools like cTAKES used in many previous studies of SDOH detection [10], to state-of-the-art DNNs – on their ability to extract SDOH information for a more comprehensive, systematically-derived set of categories. In the present study, we (a) systematically developed a comprehensive list of SDOH categories based on standard medical ontologies (SNOMED-CT and DSM-IV), (b) developed and used annotation guidelines to manually label clinical narrative notes for mentions of these SDOH categories, (c) trained and evaluated three classes of DNNs (CNN, LSTM, BERT) to detect mentions of these categories, and (d) compared these DNNs to more traditional NLP techniques, including supervised models (L2-regularized logistic regression and random forests) trained on bag-of-words and rule-based classification in the form of cTAKES.

2. Material and methods

This section describes our framework for accurate, automated identification of SDOH categories from unstructured EHR data. The framework includes a public dataset, annotated dataset sentences with SDOH labels, sentence preprocessing steps, and our three deep learning modeling architectures. This study follows the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guidelines [20].

2.1. Dataset

The Medical Information Mart for Intensive Care (MIMIC-III) Clinical Database is a relational database containing comprehensive clinical data relating to tens of thousands of patients who stayed within the Intensive Care Unit (ICU) at Beth Israel Deaconess Medical Center [21]. The dataset is publicly available and formatted in either comma-separated-value files or a single Postgres database backup file (Postgres 9.5). We used MIMIC-III v1.4, which comprises over 58,000 hospital admissions for 38,645 adults and 7,875 neonates. The data spans June 2001 to October 2012. This de-identified database contains detailed information regarding the clinical care of patients.

2.2. Annotation

In this study, we focused on the "Social Work" report type in MIMIC-III so as to retrieve the most sentences that mention SDOH. Table 1 lists 15 report types and the number of reports in each type.

To facilitate the understanding of a patient's social profile from free-text clinical narratives, we developed an ontology-based annotation scheme to label sentences in reports with 13 categories of social context. We developed this annotation scheme based on the sub-classes of "Social Context" and "Social and Personal History Finding" from Systematized

Table 1
Number of reports in MIMIC III for different note types.

Note Type	Number# of Reports
Nursing/other	822,497
Radiology	522,279
Nursing	223,556
ECG	209,051
Physician	141,624
Discharge summary	59,652
Echo	45,794
Respiratory	31,739
Nutrition	9,418
General	8,301
Rehab Services	5,431
Social Work	2,670
Case Management	967
Pharmacy	103
Consult	98
Total number of clinical reports	2,083,180

Nomenclature of Medicine-Clinical Terms (SNOMED-CT)[18], and the sub-class of "Axis IV: Psychosocial and Environmental Problems" from the Diagnostic and Statistical Manual of Mental Disorders version IV (although DSM-V is a newer edition of the DSM, it does not have the axes which provided a convenient taxonomy of SDOH) [19]. We also included a category for "substance abuse". Although substance abuse is not an SDOH, it is a health outcome heavily influenced by SDOH [22,23], and is a key contributor to decreased life expectancy in the US [24]. Table 2 shows the SDOH categories we developed and sentence inclusion criteria in the form of problems (negative SDOH indicators) or presence (other SDOH indicators, such as positive indicators) a sentence may have mentioned. We also provide an example sentence for each category. More details of our annotation process can be found in the supplemental materials, and researchers interested in our annotations can contact the corresponding author to discuss potential collaboration.

Two annotators (HL and AT) and one arbitrator (NR) annotated a total of 3,504 social sentences. Crucially, each sentence can mention multiple SDOH categories. The annotators went through a series of annotation stages, starting from guideline discussion to three rounds of inter-rater reliability tests. The initial guideline discussion used 20 sentences, followed by two iterations (50 sentences per iteration) to reach an inter-rater reliability (IRR) kappa value of 0.79 (Kappa $\kappa = \frac{p_0+p_e}{1-p_e}$ where p_0 is the observed agreement among raters, and p_e is the probability of chance agreement). Next, each annotator individually annotated another 300 sentences. Then, the third iteration with 50 more sentences gave a final IRR kappa of 0.7. Finally, each annotator individually annotated the remaining sentences. Any discrepancies between annotations were discussed between the two annotators until an agreement was reached. Fig. 1 gives a visualization of this annotation process.

2.3. Data preprocessing

Starting with a total of 14 initial annotation categories (12 SDOH categories + substance abuse category + 1 non-SDOH category), we condensed this set down to eight categories by preserving the seven most frequent categories and merging the rest of the labels into a single "other-social" category (we merged those categories due to their low prevalence). The final distribution of annotation category frequency is shown in Table 3. Even with the least frequent seven categories merged into one, these eight categories were still highly imbalanced.

Standard text pre-processing was applied to the 3,504 sentences: sentences were converted to all lowercase, foreign or uncommon symbols were removed, and contractions were separated (e.g., they've -> they 've).

Table 2
SDOH Category with Inclusion Criteria. Examples listed are from the "Social Work" reports in MIMIC III.

Economic	**Problems**: extreme poverty, inadequate finances, insufficient welfare support, lack of economic independence, bankruptcy, unstable income, insufficient income for needs **Presence**: wealthy, low/medium/high income, stable income, sufficient income for needs, manage their own finances **Example**: Other than finances, pt does not identify any other worry.
Education	**Problems**: illiteracy, academic problems, inadequate school environment, school attendance **Presence**: education level, currently receiving education, special education **Example**: She spends some time talking about his accomplishments, particularly his knowledge around nuclear energy and sciences, despite his lack of a college degree.
Healthcare	**Problems**: transportation to health care facilities unavailable, inadequate health insurance **Presence**: health insurance **Example**: She carries health insurance for family.
Housing	**Problems**: homelessness, inadequate housing, unsafe neighborhood, living alone, dangerous neighborhood, evicted, inadequate living space, live in a shelter, inadequate house sanitation **Presence**: lives with companion, lives in community, lives with family/friends, own/rent a house/apartment, adequate living space, lives with pets **Example**: She found their [**Last Name (un) 1470**] to be quite filthy and that was after the pts brother had done some preliminary cleaning.
Interaction with the legal system	**Problems**: arrest, incarceration, litigation, victim of crime, problems with police, legal problem, convicted of a crime, on probation, on remand, involved in illegal activities **Presence**: involved in legal procedures (e.g., a lawsuit) **Example**: The police have never been able to locate him.
Occupational	**Problems**: unemployment, threat of job loss, stressful work schedule, difficult work conditions, job dissatisfaction, job change, suspended from work. Unable to perform work activities due to medical condition **Presence**: has a job, does voluntary work, on leave, retired/veteran, receiving support to get a job **Example**: There was a discussion about returning to work.
Sexual Orientation	**Problems:** discrimination/ostracization from family/community due to sexual orientation **Risk Factors:** some indication of sexual orientation **Example**: It appears that he is constantly searching for justification/support that would validate him being a gay man but seems to have little success in so doing.
Social Environment	**Problems**: Death or loss of friend, inadequate social support, difficulty with acculturation, discrimination, adjustment to life-cycle transition, absence/disappearance of partner **Risk Factors**: Mother/father/siblings/other relatives/children/friends are alive and the patient has a relationship with them. Ethic group description, relationship status, divorced/elderly parents, single parent family. **Example**: It appears that the family has taken a great deal of control over the pt life, and he seems to have given it to him.
Spiritual Life	**Problems**: N/A **Risk Factors**: affiliated to a religious community, has religious beliefs, type of religion **Example**: Much of his presentation was of a philosophical/religious nature in that God never gives anyone "more than what a person can bear."
Substance Abuse	**Problems:** substance abuse history in patient, family, friends, or community

(*continued on next page*)

Table 2 (*continued*)

	Risk Factors: no substance abuse history in patient, family, friends, or community **Example**: Past addictions history: An extensive HX of EtOH addition.
Support Circumstances & Networks	**Problems**: needs assistance at home, lack of family/governmental support **Risk Factors**: Support available, cared by relatives/friends, patient providing full/part-time care to another person (e.g., disabled relative/friend), has governmental support **Example**: Patient's family is very supportive and sad and all are asking appropriate questions about how best to support him at this time.
Transportation	**Problems**: does not have a car, does not have means on their own to mobilize, needs help for mobilization, not able to drive **Risk Factors**: has a car, access to/use public transportation, has a bike, able to walk, needs special transportation **Example**: Provided contact info and 7 discount parking stickers as family reports financial hardships around travel and parking.
Other	**Problems**: exposure to disasters, war, other hostilities, access to weapons **Risk Factors**: has a cell phone, has internet, computer literacy **Example**: He spoke about ways in which his family feels alienated in this country because of American attitudes toward families as compared to Latinos.
Non-Social	**Problems:** N/A **Risk Factors**: Any sentence not classified in the above 13 categories was labelled 'non-social' **Example**: pt presented reclined in bed, with his dtr name at bedside visit.

2.4. Classification modeling

In this study, we chose three deep neural network architectures: the commonly used CNN and LSTM architectures, as well as Google's more recent pre-trained BERT-base model [25]. All models were trained on a server with two Intel® Xeon® Silver 4214 2.20 GHz processors, an Nvidia Tesla V100-PCIE 32 GB graphic processing unit (GPU), and 754 GB of physical memory. It took 8 min to fine tune a BERT model and 2.5 min to train CNN and LSTM models. After the BERT model (our best model) was trained, it processed 350 sentences per second for SDOH classification.

The Convolutional Neural Network (CNN) is a deep neural network architecture that uses convolutional layers to extract key features from data with spatial or temporal information (e.g., images, text, audio). Our CNN model followed Kim's text classification CNN architecture [26]. Kim's model achieved improved classification performance across a wide range of text classification tasks and has since become a standard baseline for new text classification architectures. We preprocessed input sentences to be the same length by applying right padding using the </PAD > token, e.g., *He met with the family today </PAD>…</PAD>*. This length was equal to the number of words in the longest sentence (82 words). We then used the pre-trained GloVe word embeddings to embed each word in a sentence into a 200x1 vector, $\mathbb{R}^{200x1}$ [27]. Any out of vocabulary words, such as "micu6" and "superego", were assigned a word vector by randomly sampling from the near-zero uniform distribution between −0.01 and 0.01. Hence, the input to the CNN is an $\mathbb{R}^{200x82}$ matrix. Next, we used a single convolutional layer that used the ReLU activation function. This layer used three common filter sizes, with 512 filters for each size. The filter sizes – (3x200), (4x200), (5x200) – were used such that each filter was applied to a window size of n words where $n \in [3, 4, 5]$: a "convolution over time." Next, we performed max pooling from the result of the convolutional layer, selecting the maximum value from each filter. Then, we concatenated and flattened all max-pooling results into a single feature vector. Dropout regularization (i.e., using only a random subset of the neural network's nodes at each training step to reduce overfitting) with a dropout rate = 0.5 was applied to this vector during training, and passed to a fully connected 8x1 output layer with sigmoid activation. This output layer provided the eight classification categories, as shown in Table 3. We chose the sigmoid activation because it allows each output node to independently yield a value between 0 and 1, which addressed our multi-label classification setting, where multiple SDOH categories can be present in a single sentence. Fig. 2 summarizes the classification process in our CNN model.

The long short-term memory (LSTM) network is a recurrent deep neural network architecture commonly used in classifying, processing, and making predictions based on sequential data [28]. Our LSTM model followed a many-to-one (i.e., the model uses the many words of a sentence to return a single predicted category) bidirectional architecture, which allows for sentence classification. Like our CNN model, we applied the same padding method and GloVe word embedding to the input data. The LSTM cell had 512 hidden nodes with the hyperbolic tangent activation function. Dropout regularization with a dropout rate = 0.5 was applied to the final LSTM output during training and passed to a fully connected 8x1 output layer with sigmoid activation. Fig. 3 shows this LSTM model architecture.

Our BERT model used the BERT-base preprocessing and pre-trained transformer layers from the TensorFlow Hub [29]. These layers handled all BERT preprocessing (i.e., BERT-specific tokenization) and computation. We added a 256-node hidden layer to the BERT pooled output (the output of the BERT [CLS] classifier token) [30]. Dropout regularization with dropout rate = 0.5 and the GELU activation function was used on the hidden layer, which mirrors the original BERT paper. The hidden layer activation was passed to an output layer with eight nodes using a

Table 3
Distribution of annotated social factor categories in the dataset. Note the percentages do not add up to 1, because each sentence can be assigned to one or more categories (428 sentences have more than one SDOH category).

Annotation Category	Count (% Frequency)	Category ID
Social environment	1783 (50.9%)	C1
Non-Social	572 (16.3%)	C2
Support circumstances and networks	515 (14.7%)	C3
Substance abuse	385 (11.0%)	C4
Housing	236 (6.7%)	C5
Occupational	191 (5.5%)	C6
Other-Social	208 (5.9%)	C7
Transportation	68 (1.9%)	C8

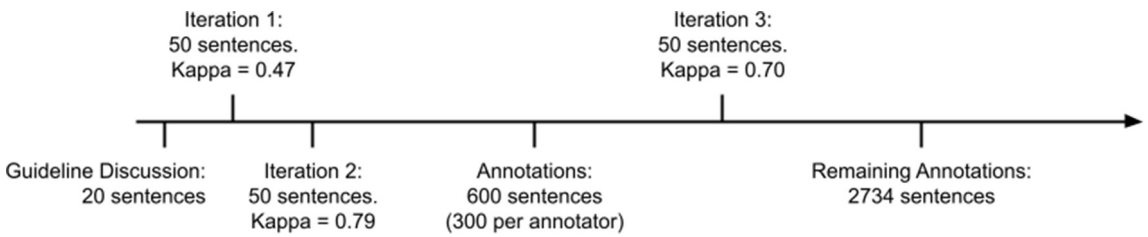


Fig. 1. Annotation Process.

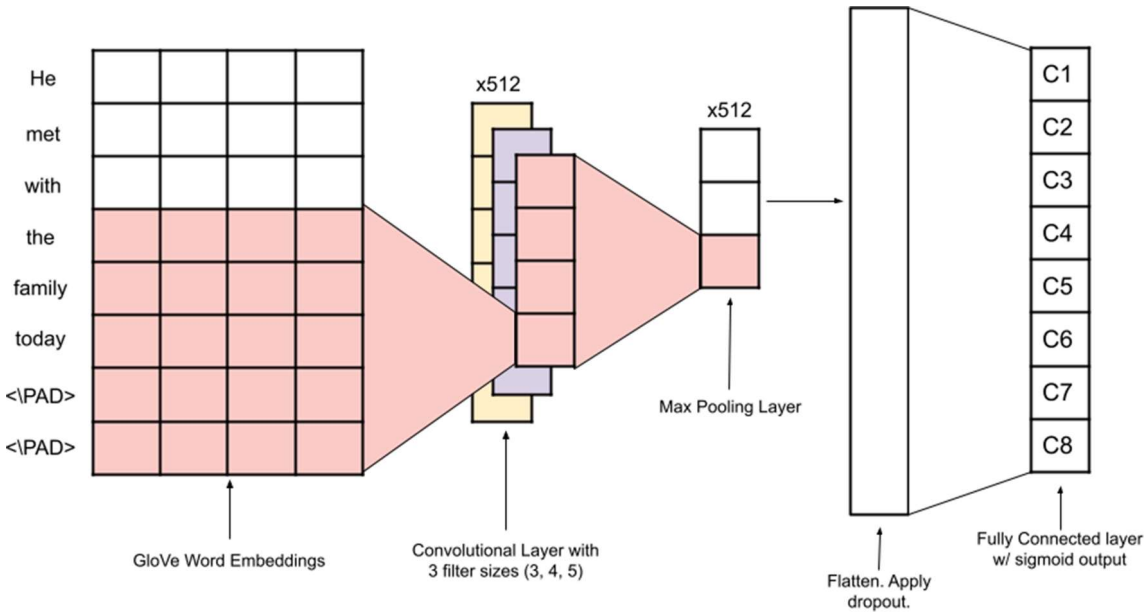


Fig. 2. Overview of Convolutional Neural Network (CNN) model, Ci represents the *i-th* category.

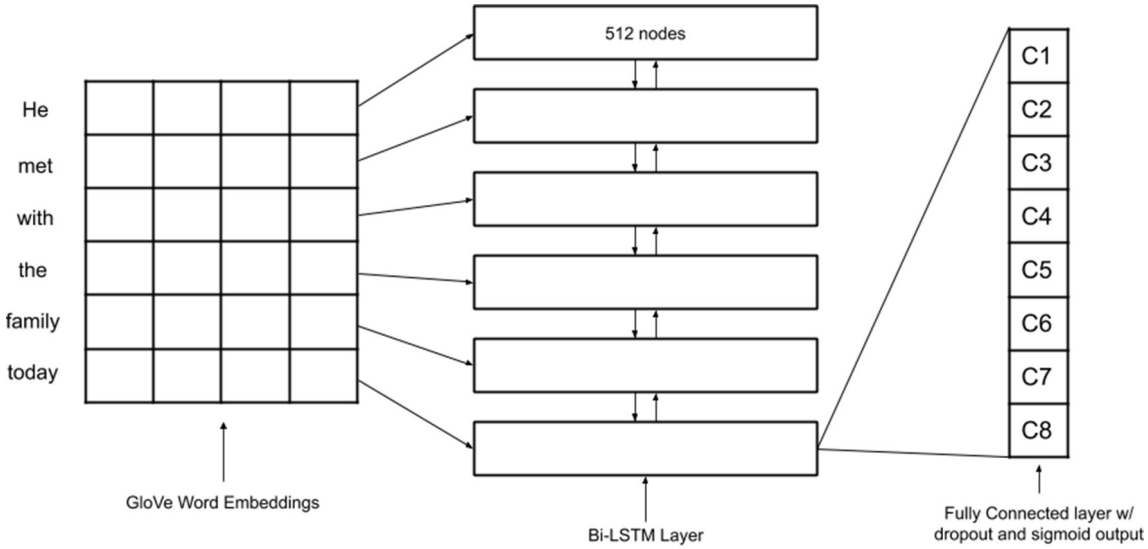


Fig. 3. Overview of the Long Short-Term Memory (LSTM) model, Ci represents the *i-th* category.

sigmoid activation function. The training was applied to all parameters of the entire network, with no "frozen" layers (i.e., no pre-trained BERT parameters were held out from standard fine-tuning). Fig. 4 shows this BERT model architecture.

The parameter configuration of CNN and LSTM models, such as word embedding size, filter size, etc., followed previous studies [26,31]. For all three models, we used a weighted binary cross-entropy loss function, where each label is inversely weighted based on their frequency in the dataset to account for class imbalance. The learning rate was set to 1e-4. We trained using the Adam optimizer with a momentum of 0.9. We trained the CNN and the LSTM models for 100 epochs, and BERT for ten epochs (only ten epochs were necessary since BERT is pre-trained). We used the Keras software package with the TensorFlow backend to implement our models [32].

2.5. Model training and testing

All three DNN models were trained and evaluated over 10-fold stratified cross-validation in our dataset, stratifying instances using the iterative stratification method for multi-label problems [33]. Stratification ensures that each category's cases and controls are evenly distributed across the ten dataset partitions. As is common, we used 0.5 as the threshold to decide the predicted labels for all three models.

2.6. Baseline models

We implemented two baseline models for comparison against our deep neural network models: L2-regularized logistic regression (LR) and random forest (RF) classifiers trained on the bag of words. For text preprocessing, we first removed stopwords from the NLTK [34] stopwords list, removed non-ASCII characters, and converted contractions with ' into full terms (e.g., what's -> what is; can't -> can not, 'll -> will, etc.). Next, we applied TF-IDF feature weighting on the bag-of-words. The multi-label classifier was trained using the OneVsRestClassifier with RandomForestClassifier and LogisticRegression from Python's Sci-kit Learn machine learning package [35]. To simultaneously select and

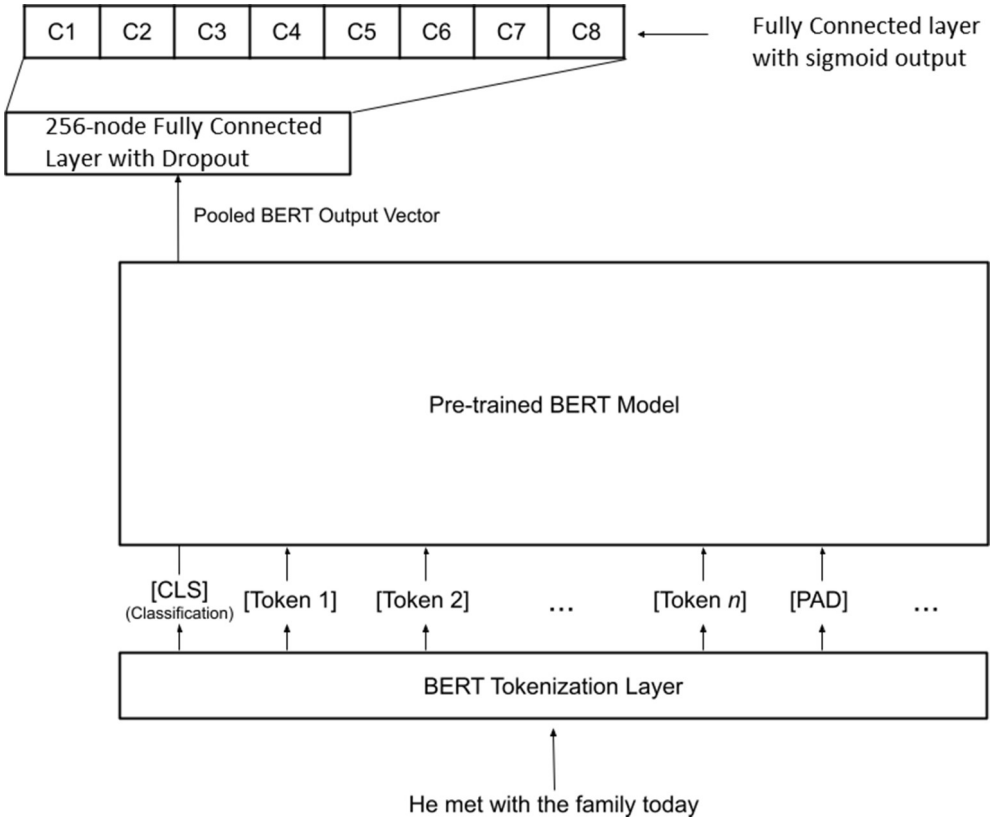


Fig. 4. Overview of BERT model, Ci represents the *i-th* category.

evaluate models without overfitting [36], we performed nested cross-validation (CV), which consists of an outer CV loop to evaluate models and an inner CV loop to select models. More specifically, in one iteration of our 10-fold outer CV loop, one fold serves as a test set on which a selected model is evaluated, while the other 9 folds are used as a training set to select a model. This training set was then subjected to a 5-fold, inner CV loop with random search for hyperparameters (as random search can be more efficient than grid search [37]), where a model with a particular hyperparameter combination was trained on four folds and evaluated on the remaining fold. The model (trained under particular hyperparameters) that performed best across all five folds was then retrained on the entire training set (i.e., the 9 folds of the current iteration of the outer CV loop), and then evaluated on the test set of the outer CV loop. This process is repeated for all 10 outer folds.

2.7. Secondary analysis: cTAKES performance for classifying SDOH sentences

We also examined performance of cTAKES, a well-known natural language processing system for extracting information from clinical free-text notes, in distinguishing social and non-social sentences. We ran cTAKES on each annotated sentence in the cohort. As a result, cTAKES identified a set of Unified Medical Language System® (UMLS) Concept Unique Identifiers (CUI) code(s) (2020AB release of UMLS knowledge sources) in each sentence. As defining cTAKES-based rules for all eight SDOH categories would have been prohibitively difficult, we simply grouped the eight categories (Table 3) into just two classes: social and non-social. The social class comprised six SDOH categories (C1, C3, C5-C8), and the general non-social class comprised two categories: non-Social (C2) and substance abuse (C4). We created two rules for classifying social sentences from cTAKES results, with any sentence not satisfying either rule classified as non-social:

1. We used the UMLS API to retrieve extracted CUI codes' ancestors in SNOMED CT that refer to social concepts. We identified two social-related SNOMED CT categories (i.e., *284490008*: finding relating to complex and social behaviors or *22032002*: family-related social factor).
2. We used the PyMedTermino2 Python package to (a) map CUI codes to ICD-10-CM codes and (b) find the ICD-10 codes in DSM V that refer to social concepts [38].

Note that while we used DSM-IV (and SNOMED-CT) to develop our annotation scheme, in the second rule above we were forced to use DSM-V as DSM-IV is no longer included in the UMLS Metathesaurus.

2.8. Evaluation metrics

Since the task of assigning multiple SDOH categories to a sentence is a multi-label classification problem, there are multiple methods for evaluating the model performance. In this study, we evaluated the performance of our models in aggregate, as well as by category. For evaluation metrics, we used Hamming loss, which is the average measure of difference between the actual and predicted values for labels, i.e., $Hammingloss = \frac{1}{NL}\sum_{i=1}^{N}\sum_{j=1}^{L} xor(y_{i,j}, z_{i,j})$ where N is the total number of instances, and L is the number of labels, $y_{i,j}$ and $z_{i,j}$ are the ground truth and predicted results, respectively. Precision (also known as positive predictive value), recall (also known as sensitivity), F1 score ($=\frac{2*precision*recall}{precision+recall}$), area under the receiver operating characteristic curve (AUC), and average precision ($\sum_n (R_n - R_{n-1})P_n$) where R_n and P_n are recall and precision at the nth threshold. To evaluate the models in aggregate, we evaluated both the micro (total average) and macro (category average) metrics. The macro-averaged F1-score is an extension of the F1-score for binary classification and is simply the average of the F1-score from each category/label, and is defined as $MacroF1 =$

$\frac{1}{L}\sum_{i=1}^{L} F1_i$ where L is the number of labels. In contrast to Macro-F1, Micro-F1 computes statistics globally for all labels and instances, i.e., $MicroF1 = \frac{\sum_{n=1}^{N}\sum_{i=1}^{L}\widehat{y_i}y_i}{\sum_{n=1}^{N}\sum_{i=1}^{L}\widehat{y_i}+\sum_{n=1}^{N}\sum_{i=1}^{L}y_i}$, where $\widehat{y_i^n}$ and (y_i^n) are predicted and actual values, respectively, and N is the total number of instances, and L is the number of labels. For measuring cTAKES performance, we used sensitivity, specificity, positive predictive value (PPV), accuracy, and F1-score.

3. Results

A total of 3,504 sentences from 2,670 reports and 1,337 patients were annotated through a multi-stage annotation process with an average Kappa value of 0.745 (excluding the initial round). Table 4 summarizes average performance metrics across the three models from the 10-fold stratified cross-validation. The BERT model significantly outperformed the CNN and LSTM models in most of the metrics we reported, particularly in recall, as shown in Table 4. However, the CNN model yielded slightly higher precision, and significantly higher AUC of the micro-ROC. The LSTM model did not perform best in any of our metrics. (See Supplemental Materials for average macro and micro-ROC and precision-recall curves of the 3 DNN models.)

This pattern of results generally holds for individual SDOH categories as well. Table 5 reports performance for each of the eight categories, averaged over all 10 CV folds, and shows that BERT significantly outperformed the CNN and LSTM in most SDOH categories and most metrics, particularly in recall. The CNN occasionally performed better in Hamming loss and precision, while the LSTM had the lowest performance compared to the other models. (See Supplemental Materials for ROC curves for the individual SDOH categories for both CNN and BERT.)

Table 6 summarizes average performance metrics between our best DNN model (BERT) and our baseline machine learning models (LR and RF) from the 10-fold stratified cross-validation. As shown in Table 6, while the LR model had better performance on precision (with statistical significance) and hamming loss (without statistical significance), BERT outperformed the conventional machine learning models on all other metrics, including the critical F1 and AUC scores.

Table 4
10-fold average of Hamming loss, micro/macro F1/precision/recall/AUC/average precision. Bold-faced numbers represent the best performance among the three models, while numbers in parentheses represent 95% confidence intervals. P-values were calculated based on the Mann-Whitney U tests comparing the best performing model and the other two models. An asterisk indicates that the bolded model outperforms the others on that metric, to a statistically significant degree ($\alpha = 0.05$).

	CNN (95% CI)	LSTM (95% CI)	BERT (95% CI)
Hamming Loss	0.097 (0.094, 0.100)	0.106 (0.102, 0.110)	**0.095 (0.088, 0.102)**
Macro F1	0.550 (0.527, 0.572)	0.555 (0.536, 0.574)	**0.642 (0.623, 0.662)***
Micro F1	0.649 (0.639, 0.659)	0.632 (0.619, 0.644)	**0.690 (0.670, 0.710)***
Macro Precision	**0.587 (0.567, 0.606)**	0.546 (0.533, 0.559)	0.583 (0.555, 0.612)
Micro Precision	**0.667 (0.657, 0.676)**	0.620 (0.606, 0.634)	0.640 (0.613, 0.667)
Macro Recall	0.529 (0.502, 0.556)	0.574 (0.544, 0.603)	**0.756 (0.742, 0.770)***
Micro Recall	0.633 (0.619, 0.647)	0.644 (0.628, 0.659)	**0.750 (0.735, 0.765)***
Macro AUC	0.854 (0.845, 0.864)	0.847 (0.834, 0.859)	**0.907 (0.899, 0.915)***
Micro AUC	**0.907 (0.903, 0.912)***	0.865 (0.858, 0.872)	0.890 (0.879, 0.902)
Macro Avg Precision	0.579 (0.557, 0.602)	0.556 (0.530, 0.581)	**0.698 (0.680, 0.715)***
Micro Avg Precision	0.695 (0.683, 0.708)	0.643 (0.624, 0.662)	**0.718 (0.694, 0.741)***

*: $p < .05$ according to Mann-Whitney U Tests.

3.1. Secondary analysis results

Table 7 shows the sensitivity, specificity, PPV, and accuracy for BERT and cTAKES. In contrast to our primary analyses with our three DNNs, the secondary analysis used only BERT and cTAKES to predict whether a sentence is social or non-social. Therefore, although the BERT model's output is multi-label, if the highest probability prediction was one of the social categories, we classified it as a social sentence for BERT results. We found that cTAKES performed very poorly in identifying social sentences (F1 = 0.06) compared to BERT (F1 = 0.87). This is because, while cTAKES had very high specificity (0.997), it had very low sensitivity (0.03).

3.2. Error analysis

To better understand why cTAKES performed poorly, we conducted an error analysis. To this end, we randomly selected 100 sentences that cTAKES failed to assign to one of our SDOH categories (i.e., randomly sampled false negatives). Table 8 lists six of these misclassified sentences. Overall, we found that cTAKES focused on the individual words instead of considering the context of the whole sentence and was limited by the quality of its lexicons, which were built to find medical rather than social words and phrases. For example, in the sentence in the first row of the table, the term 'blood' does not refer to the bodily fluid (a medical concept), but is instead part of the compound 'blood relative' (a social concept). Likewise, in the third row, cTAKES falsely categorized the sentence as non-social as it did not detect the term 'divorcing', which is not in SNOMED-CT.

4. Discussion

4.1. Main findings

Social determinants of health account for enormous variability in health outcomes, and for the majority of modifiable health factors. They are therefore key targets for interventions, but most SDOH information is locked away in unstructured clinical notes, making detection and intervention difficult. While previous research has shown NLP systems can detect SDOH in clinical text with some accuracy, (a) such systems usually were not designed or trained to detect a comprehensive, systematically chosen range of SDOH, and (b) these systems tend not to exploit the state of the art in deep neural networks for natural language processing. The contribution of the present work is thus two-fold:

First, using standard biomedical and psychiatric ontologies (SNOMED-CT and DSM-IV), we curated a list of 13 categories of social determinants of health, such as sexual orientation, support circumstances and networks, housing, and transportation. We then developed annotation guidelines and manually annotated 3,504 sentences from the social work reports in MIMIC-III on each of these categories, through a multi-stage, multi-rater annotation process, throughout which we measured inter-annotator agreement to ensure annotation consistency and accuracy ($\kappa = 0.7$). Our annotation scheme and guidelines thus cover a broad range of SDOH and enable consistent annotation.

Second, with these annotations of SDOH, we trained and evaluated three contemporary deep neural networks for text classification – a convolutional neural network (CNN), a long short-term memory (LSTM) network, and the recently developed Bidirectional Encoder Representations from Transformers (BERT). We found that fine-tuning a pre-trained BERT outperformed our CNN and LSTM on most key metrics, achieving Hamming loss = 0.095, Macro F1 = 0.64, Micro F1 = 0.69, and Macro AUC-ROC = 0.91 (but only Micro AUC-ROC = 0.89 compared to 0.91 for CNN). Likewise, BERT outperformed conventional machine learning models (L2-regularized logistic regression and random forest) trained on bags of words for all these same metrics, except

Table 5
10-fold average of Hamming loss, F1, precision, recall, AUC, and average precision. Numbers in parentheses represent 95% confidence intervals.

	Ham. Loss	F1	Precision	Recall	AUC_ROC	Avg Prec.
Social Environment						
CNN	0.241 (0.225,0.257)	0.765 (0.751,0.779)	0.761 (0.741,0.781)	0.770 (0.750,0.790)	0.836 (0.820,0.852)	0.825 (0.811,0.839)
LSTM	0.246 (0.228,0.264)	0.759 (0.741,0.777)	0.755 (0.739,0.771)	0.763 (0.740,0.786)	0.824 (0.805,0.843)	0.812 (0.790,0.834)
BERT	**0.193 (0.172,0.214)***	**0.807 (0.781,0.833)***	**0.821 (0.806,0.836)***	**0.796 (0.753,0.839)**	**0.874 (0.856,0.892)***	**0.855 (0.834,0.876)***
Non-Social						
CNN	**0.186 (0.177,0.195)**	0.382 (0.355,0.409)	0.421 (0.396,0.446)	0.356 (0.316,0.396)	0.752 (0.739,0.765)	0.395 (0.363,0.427)
LSTM	0.217 (0.204,0.230)	0.357 (0.327,0.387)	0.348 (0.316,0.380)	0.369 (0.332,0.406)	0.696 (0.677,0.715)	0.327 (0.298,0.356)
BERT	0.187 (0.168,0.206)	**0.491 (0.464,0.518)***	**0.449 (0.419,0.479)**	**0.552 (0.501,0.603)***	**0.788 (0.768,0.808)***	**0.454 (0.422,0.486)***
Support Circumstances and Networks						
CNN	**0.125 (0.117,0.133)***	**0.576 (0.547,0.605)**	**0.575 (0.548,0.602)***	0.581 (0.541,0.621)	**0.845 (0.825,0.865)**	**0.578 (0.537,0.619)**
LSTM	0.149 (0.142,0.156)	0.513 (0.486,0.540)	0.493 (0.472,0.514)	0.538 (0.491,0.585)	0.803 (0.788,0.818)	0.484 (0.459,0.509)
BERT	0.155 (0.136,0.174)	0.550 (0.523,0.577)	0.493 (0.450,0.536)	**0.643 (0.575,0.711)**	0.844 (0.828,0.860)	0.573 (0.517,0.629)
Substance Abuse						
CNN	**0.025 (0.023,0.027)**	**0.885 (0.872,0.898)**	**0.893 (0.872,0.914)***	0.878 (0.858,0.898)	0.982 (0.976,0.988)	0.929 (0.912,0.946)
LSTM	0.029 (0.024,0.034)	0.869 (0.847,0.891)	0.852 (0.823,0.881)	0.888 (0.861,0.915)	0.981 (0.970,0.992)	0.920 (0.898,0.942)
BERT	0.029 (0.022,0.036)	0.877 (0.847,0.907)	0.821 (0.775,0.867)	**0.945 (0.929,0.961)***	**0.984 (0.974,0.994)**	**0.946 (0.926,0.966)**
Housing						
CNN	**0.073 (0.066,0.080)**	0.428 (0.374,0.482)	0.466 (0.400,0.532)	0.406 (0.343,0.469)	0.857 (0.838,0.876)	0.437 (0.387,0.487)
LSTM	0.074 (0.066,0.082)	0.457 (0.399,0.515)	0.456 (0.404,0.508)	0.464 (0.389,0.539)	0.830 (0.800,0.860)	0.398 (0.33,0.466)
BERT	0.077 (0.065,0.089)	**0.552 (0.504,0.600)***	**0.474 (0.416,0.532)**	**0.707 (0.613,0.801)***	**0.903 (0.883,0.923)***	**0.548 (0.503,0.593)***
Occupational						
CNN	0.036 (0.031,0.041)	0.634 (0.570,0.698)	0.704 (0.647,0.761)	0.586 (0.503,0.669)	0.924 (0.903,0.945)	0.692 (0.636,0.748)
LSTM	0.036 (0.029,0.043)	0.668 (0.610,0.726)	0.685 (0.609,0.761)	0.659 (0.600,0.718)	0.900 (0.869,0.931)	0.677 (0.609,0.745)
BERT	**0.028 (0.021,0.035)**	**0.774 (0.728,0.820)***	**0.707 (0.645,0.769)**	**0.869 (0.818,0.920)***	**0.965 (0.948,0.982)***	**0.835 (0.785,0.885)***
Other-Social						
CNN	0.067 (0.060,0.074)	0.330 (0.271,0.389)	0.417 (0.343,0.491)	0.279 (0.226,0.332)	0.775 (0.738,0.812)	0.351 (0.281,0.421)
LSTM	0.070 (0.061,0.079)	0.413 (0.357,0.469)	0.424 (0.354,0.494)	0.413 (0.348,0.478)	0.818 (0.787,0.849)	0.421 (0.378,0.464)
BERT	**0.057 (0.043,0.071)**	**0.607 (0.530,0.684)***	**0.549 (0.441,0.657)***	**0.708 (0.644,0.772)***	**0.921 (0.896,0.946)***	**0.688 (0.620,0.756)***
Transportation						
CNN	**0.020 (0.018,0.022)***	0.399 (0.277,0.521)	**0.455 (0.330,0.580)**	0.376 (0.244,0.508)	0.865 (0.815,0.915)	0.428 (0.307,0.549)
LSTM	0.027 (0.023,0.031)	0.404 (0.285,0.523)	0.353 (0.247,0.459)	0.493 (0.336,0.650)	0.921 (0.876,0.966)	0.406 (0.299,0.513)
BERT	0.037 (0.026,0.048)	**0.482 (0.411,0.553)**	0.352 (0.288,0.416)	**0.829 (0.714,0.944)***	**0.976 (0.958,0.994)***	**0.680 (0.568,0.792)***

*: $p < .05$ according to Mann-Whitney U Tests.

Table 6
10-fold average of Hamming loss, micro/macro F1/precision/recall/AUC_ROC for Logistical Regression (LR), Random Forest (RF), and BERT.

	LR (95% CI)	RF (95% CI)	BERT (95% CI)
Hamming Loss	**0.089 (0.086, 0.092)**	0.121 (0.115, 0.126)	0.095 (0.088, 0.102)
Macro F1	0.519 (0.499,0.540)	0.098 (0.089, 0.107)	**0.642 (0.623, 0.662)***
Micro F1	0.643 (0.632, 0.653)	0.501 (0.490, 0.512)	**0.690 (0.670, 0.710)***
Macro Precision	**0.727 (0.694, 0.760)***	0.138 (0.068, 0.208)	0.583 (0.555, 0.612)
Micro Precision	**0.738 (0.722, 0.754)***	0.607 (0.568, 0.645)	0.640 (0.613, 0.667)
Macro Recall	0.433 (0.415, 0.451)	0.122 (0.121, 0.123)	**0.756 (0.742, 0.770)***
Micro Recall	0.570 (0.558, 0.581)	0.430 (0.420, 0.440)	**0.750 (0.735, 0.765)***
Macro AUC	0.693 (0.684, 0.702)	0.520 (0.514, 0.526)	**0.907 (0.899, 0.915)***
Micro AUC	0.768 (0.762, 0.774)	0.691 (0.687, 0.696)	**0.890 (0.879, 0.902)***

*: $p < .05$ according to Mann-Whitney U Tests.

Table 7
Sensitivity, Specificity, Positive Predictive Value (PPV), Accuracy, and F1 scores for BERT and cTAKES, in classifying sentences as social or non-social.

	BERT	cTAKES
Sensitivity	**0.8901**	0.0321
Specificity	0.6273	**0.9969**
PPV	0.85	**0.9608**
Accuracy	**0.8122**	0.3181
F1	**0.8696**	0.0621

Table 8
Samples of sentences falsely categorized as non-social by cTAKES.

Sentence	cTAKES Annotation	True Label	Identified Issue
she does not have close blood relatives with the exception of two nephews, according to mr name s	blood	Social environment	Failed to catch the full term "blood relatives" but only caught "blood."
Pt states not feeling depressed but being sad because she is not with her children	depressed; sad	Social environment	Failed to catch the terms "with her children."
Psychiatrically hospitalized at name following a suicide attempt in response to her husband divorcing her	suicide attempt; suicide	Social environment	Failed to catch "divorcing," not available in the SNOMED CT
pt states feeling fine, with the exception of feeling sad today from having seen her son cry because of her admission	feeling sad; feeling; sad	Social environment	Failed to catch "son cry"
she notes her drinking caused the divorce	Drinking; divorce	Substance abuse (None-social)	Failed to associate the "drinking" with alcohol abuse.

Hamming Loss (logistic regression Hamming Loss = 0.087). Further, BERT vastly outperformed cTAKES, a common NLP tool applied in unstructured EHR data, even in just simply classifying sentences as social vs non-social (BERT F1 = 0.87 vs. cTAKES F1 = 0.06). Error analysis of cTAKES revealed that it failed because (a) it is not suited to considering the meaning of an entire sentence, and (b) its lexicon is mostly designed to extract medical rather than social concepts.

In the rest of the discussion, we consider each of these contributions more closely, as well as limitations and future directions.

4.2. A systematically curated set of SDOH categories

As we have noted elsewhere, most NLP systems that detect SDOH in clinical notes focus on a single SDOH (see Patra et al review [10]), and when such systems *are* designed to detect multiple SDOH, the categories that the system can detect are selected (by developers) on a somewhat ad hoc basis (e.g., Feller et al [17]). Our first (methodological) contribution was therefore to derive a more comprehensive, systematic set of SDOH categories, through SNOMED-CT and DSM-IV, two conventional medical and psychiatric ontologies. With these ontologies, we derived 13 categories: economic, education, healthcare, housing, interaction with the legal system, occupational, sexual orientation, social environment, spiritual life, substance abuse, support circumstances & networks, transportation, other, and non-social. Although we condensed these 13 categories down to 8 categories for the present modeling owing to small sample sizes for some of the categories, future SDOH detection systems could be trained to detect all 13 SDOH categories where data quantity and quality allow. Our annotation guidelines can also be useful to other researchers who wish to annotate corpora for our 13 categories, and we have therefore uploaded the guidelines as supplemental materials.

Although we have suggested that, by using standard ontologies, we have developed a systematic, non-ad hoc set of SDOH categories, we acknowledge that we have not captured *all* social determinants of health, e.g., sexual activity or gender identity. To some extent, we feel that collecting *all* social determinants of health is an inherently impossible task, as the organization of social systems, and their impact on health, is constantly varying across time and space, making the definition of social determinants of health a moving target. Thus, we do not claim that our list of SDOH is complete. Rather, our claim is that we are capturing the SDOH that a critical mass of clinicians and psychiatrists view as salient enough to be included in SNOMED-CT and the DSM. To the extent that this method of collecting SDOH is incomplete, it suggests ways in which current standard ontologies could be augmented.

4.3. Natural language processing for SDOH detection

Our second contribution lies in the natural language processing models we used to detect SDOH. First, whereas some other studies (e.g., Feller et al [17], but see Stemerman et al [39] for multi-label classification with 6 SDOH) built separate models to predict separate SDOH even when a note could contain multiple SDOH, we treated SDOH detection as a multi-label prediction problem, allowing us to build a single model (for a given ML or DNN architecture) that can predict multiple SDOH categories at the same time. Having a single model that can predict multiple categories may be more efficient in clinical application than having separate models for each category. We also speculate that, to the extent that all sentences regardless of meaning must undergo similar (implicit) linguistic processing (e.g., syntactic parsing or semantic role labeling) in a deep neural network to achieve accurate classification, then it is helpful to consolidate the training signals from the different SDOH labels into learning parameters for a single model. Put differently, a multilabel setting allows for a kind of transfer learning from one SDOH category to another. Thus, we suspect that higher performance can be achieved by a single multi-label model than by separate models for each category, although this is a suspicion that should be empirically tested.

Second, as noted in the introduction, compared to previous work, we evaluated a larger range of NLP techniques for SDOH classification, from cTAKES, to conventional machine learning on bags of words, to three DNNs: CNN, LSTM, and finally the very recent BERT. According to Patra et al.'s very recent systematic review [10], our paper is only the second to test BERT (or any other transformer) on its ability to detect SDOH, and the first paper to compare this architecture to older deep neural networks (CNN and LSTM) for SDOH classification. As in other tasks of NLP, the order of our models' performance largely mirrors that in other tasks [23,28–32]: cTAKES < machine learning on bags of words < LSTM < CNN < BERT. We believe this pattern of performance can be interpreted reasonably well. First, our error analysis of cTAKES suggest it failed because (a) it is not suited to considering the meaning of an entire sentence, and (b) its lexicon is mostly designed to extract medical rather than social concepts. While the few rules it has to extract SDOH are accurate, they are few in number, leading to high precision but low recall of SDOH. Bags of words performed better because they allow models to learn a greater range of individual words that predict SDOH (or in the case of nonlinear models like RF, combinations of words). Yet DNNs outperformed bags of words because (a) they utilized pre-trained GloVe word embeddings (taking advantage of transfer learning, which is key in a small data setting), and (b) they incorporated more information from word order and syntax. Moreover, our CNN outperforming our LSTM is consistent with prior work [40], which shows that CNNs outperformed recurrent neural networks (e.g., LSTM) when the relevant classification features were more local than global, i.e., the classification task was essentially keyword/phrase recognition. Evidence that our task can be at least partially solved with keyword/phrase recognition comes from preliminary work (also mentioned in the next section), where we found that many SDOH categories are reliably cued by individual words and phrases. For example, 'substance abuse' was often reliably cued by words/phrases like 'cocaine' or the phrase 'substance abuse' itself. Finally, as in other domains, we suspect BERT out-performed other DNNs because of its self-attention mechanism and its more extensive pre-training (i.e., the word embedding layer *and all subsequent layers of the network* were pre-trained). Again, pre-training and transfer learning are keys in our relatively small-data setting.

We note that our finding that DNNs outperformed other ML approaches contrasts with Feller et al. [17], who found the opposite finding with CNN's and feedforward networks (presumably just multilayer perceptrons without recurrence, convolutions, attention, or other sophisticated techniques). They attribute their finding to a small dataset, but our dataset (2,670 notes) is actually considerably smaller than theirs (4,663 notes). Although they do not provide much detail about how they used DNNs, we suspect the difference comes down to the different preprocessing of the inputs: we trained our DNNs just on raw text, whereas they seemed to have trained theirs on bags of words (sometimes with structured EHR data). As noted above, the bag-of-words representation obliterates the spatiotemporal information in raw text that DNNs are designed to exploit. It is unclear, however, why their feedforward network performed poorly, although we suspect this is 'simply' a matter of finding the right hyperparameters (number of layers, dropout rate, etc.). In any case, our work demonstrates quite clearly that DNNs, when used appropriately, can outperform traditional machine learning techniques in detection of SDOH.

4.4. Limitations and future work

We recognize that this study has limitations, which present opportunities for future work. First, we performed a single-center study, and did not split training and test data by time period. Future work therefore could externally validate our models on a second center, and/or on notes collected from a different time period from those our model was trained on. Relatedly, we annotated, and trained and evaluated models on, only social work notes. These are only added for patients that potentially require social support (we thank a reviewer for this point), which may introduce selection bias into our models, i.e., our models may have learned a higher base rate of SDOH mentions than is realistic in other note types. Therefore, it will be useful to know if our model can generalize to other note types. Second, as is well known, deep neural networks are typically difficult to interpret, yet interpretability is paramount in a sensitive domain like clinical decision-making. To address this, we have begun testing a standard post-hoc explainable AI

technique, SHapley Additive exPlanations (SHAP, Lundberg [41]), to find the words and phrases that drive DNN decisions in particular instances. Such work could also address the fact (observed in Patra et al.'s systematic review [10]) that most papers on automated SDOH detection do not identify the spans of text containing SDOH [16]. Third, our models only detected whether an SDOH was mentioned; they did not detect whether the SDOH was about the patient, or someone else. Although in most cases a note describes the SDOH of the patient, this is not always the case. For example, the sentence "a doctor hospitalized his [the patient's] former girlfriend for two months with endocarditis secondary to injecting heroin" was categorized as "substance abuse," since it mentioned heroin usage, even though this drug usage does not refer to the patient. In this case, it would be helpful to have a model that can further identify *who* has the substance abuse disorder (i.e., the girlfriend rather than the patient). Likewise, our models did not distinguish between protective (e.g., being currently employed) and risk (e.g., being currently *un*employed) factors. It is of course possible to address these limitations by expanding the annotation scheme and retraining our models, but given the cost of this, methods entailing less manual effort may be preferred. Devising such methods could therefore be an opportunity for future work (e.g., automatic semantic role labeling, which identifies *who* did *what* to *whom,* could be useful for determining whether an SDOH is about the patient or someone else).

Other future work that we plan includes testing different model architectures and problem structures to further improve performance. Specifically, we look to test versions of BERT that have been trained on clinical datasets (e.g., BioBERT, ClinicalBERT), as well as more advanced transformer architectures (e.g., XLNet) [42–44]. By strengthening our models in this way, and addressing the limitations above, we believe this work will lead to an effective, general-purpose model to accurately identify patients with SDOH, allowing researchers to better integrate these factors into their analyses and further improve healthcare outcomes.

5. Conclusion

In this study, we developed an innovative framework for accurate automated classification of SDOH categories from unstructured clinical notes in EHR data. We (1) systematically identified a broad range of SDOH categories from SNOMED-CT and (2) evaluated the abilities of three deep learning models (CNN, LSTM, and BERT), as well as cTAKES and bags of words, to label sentences in MIMIC III clinical reports with six SDOH categories, one substance abuse category, and one non-SDOH category. Although all three deep learning models had accurate SDOH classification, BERT outperformed the CNN and LSTM models by a wide margin in aggregate, as well as in individual categories, and also outperformed conventional machine learning trained on bags of words. Compared to the three models, the commonly used NLP tool, cTAKES, was very limited in identifying SDOH sentences. Our framework can be of value in identifying SDOH in practice, which may further improve healthcare outcomes.

Funding

This research was supported by the National Institute of Mental Health (R21MH123916), the Richard King Mellon Foundation (MWRIF 3659), and the Children's Hospital of Philadelphia. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

Author contributions

Study concept and design: FRT, LS, NR, and DB. Analysis and interpretation of data: SH, RZ, LS, and FRT. Collection or annotation of data: RZ, WQ, HL, AT, NR, and FRT. Drafting of the manuscript: RZ, WQ, and FRT. Critical revision of the manuscript for important intellectual content: SH, RZ, LS, RR, HL, AT, WQ, NR, DB, and FRT. Funding: FRT. Study Supervision and Coordination: FRT.

8. Research Ethics Approval

This study did not receive nor require ethics approval, as it does not involve human & animal participants.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors would also like to thank Edward Gruver in the Tsui Laboratory and anonymous reviewers for their valuable feedback.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2021.103984.

References

[1] Datto A. Social determinants of health. https://www.who.int/health-topics/social-determinants-of-health#tab=tab_(accessed 26 Jul 2021).
[2] Halfon N, Larson K, Russ S. Why social determinants? *Healthc Q* 2010;**14**:8–20.
[3] Chen M, Tan X, Padman R. Social determinants of health in electronic health records and their impact on analysis and risk prediction: A systematic review. *J Am Med Informatics Assoc* 2020;**27**:1764–73.
[4] Magnan S. Social determinants of health 101 for health care: five plus five. *NAM Perspect* 2017.
[5] D. Williams, M.V. Costa, A.O. Odunlami, S.A. Mohammed, Moving upstream: how interventions that address the social determinants of health can improve health and reduce disparities, J. Public Heal Manag. Pract. JPHMP 14 (6) (2008) S8–S17.
[6] A. Andermann, Screening for social determinants of health in clinical care: moving from the margins to the mainstream, Public Health Rev. 39 (2018) 1–17.
[7] A.S. Navathe, F. Zhong, V.J. Lei, F.Y. Chang, M. Sordo, M. Topaz, S.B. Navathe, R. A. Rocha, L. Zhou, Hospital readmission and social risk factors identified from physician notes, Health Serv. Res. 53 (2) (2018) 1110–1136.
[8] S.M. Goodday, A. Kormilitzin, N. Vaci, Q. Liu, A. Cipriani, T. Smith, A. Nevado-Holgado, Maximizing the use of social and behavioural information from secondary care mental health electronic health records, J. Biomed. Inform. 107 (2020) 103429, https://doi.org/10.1016/j.jbi.2020.103429.
[9] Bompelli A, Wang Y, Wan R, *et al.* Social determinants of health in the era of artificial intelligence with electronic health records: A systematic review. *arXiv Prepr arXiv210204216* 2021.
[10] Patra BG, Sharma MM, Vekaria V, *et al.* Extracting social determinants of health from electronic health records using natural language processing: a systematic review. *J Am Med Informatics Assoc* 2021.
[11] E.S. Chen, E.W. Carter, I.N. Sarkar, et al., Examining the use, contents, and quality of free-text tobacco use documentation in the electronic health record, AMIA Annual Symposium Proceedings. 366 (2014).
[12] J.L. Greenwald, P.R. Cronin, V. Carballo, G. Danaei, G. Choy, A novel model for predicting rehospitalization risk incorporating physical function, cognitive status, and psychosocial support using natural language processing, Med. Care 55 (3) (2017) 261–266.
[13] Chauhan S, Vig L, De Filippo De Grazia M, *et al.* A comparison of shallow and deep learning methods for predicting cognitive performance of stroke patients from MRI lesion images. *Front Neuroinform* 2019;**13**:53.
[14] E. Kanjo, E.M.G. Younis, C.S. Ang, Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection, Inf. Fusion 49 (2019) 46–56.
[15] Feuerriegel S, Fehrer R. Improving decision analytics with deep learning: the case of financial disclosures. arXiv Prepr arXiv150801993 2015.
[16] K. Lybarger, M. Ostendorf, M. Yetisgen, Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction, J. Biomed. Inform. 113 (2021) 103631, https://doi.org/10.1016/j.jbi.2020.103631.
[17] D.J. Feller, O.J. Bear Don't Walk IV, J. Zucker, M.T. Yin, P. Gordon, N. Elhadad, Detecting social and behavioral determinants of health with structured and free-text clinical data, Appl. Clin. Inform. 11 (01) (2020) 172–181.
[18] SNOMED. No Title. https://www.snomed.org/snomed-ct.
[19] C.C. Bell, DSM-IV: diagnostic and statistical manual of mental disorders, JAMA 272 (10) (1994) 828, https://doi.org/10.1001/jama.1994.03520100096046.

[20] K.G.M. Moons, D.G. Altman, J.B. Reitsma, J.P.A. Ioannidis, P. Macaskill, E. W. Steyerberg, A.J. Vickers, D.F. Ransohoff, G.S. Collins, Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration, Ann. Intern. Med. 162 (1) (2015) W1–W73.
[21] A.E.W. Johnson, T.J. Pollard, L. Shen, L.W. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Sci. data 3 (1) (2016), https://doi.org/10.1038/sdata.2016.35.
[22] H.M.E. Belcher, H.E. Shinitzky, Substance abuse in children: Prediction, protection, and prevention, Arch. Pediatr. Adolesc. Med. 152 (1998) 952–960.
[23] V. Knerich, A.A. Jones, S. Seyedin, C. Siu, L. Dinh, S. Mostafavi, A.M. Barr, W. J. Panenka, A.E. Thornton, W.G. Honer, A.R. Rutherford, L. Palinkas, Social and structural factors associated with substance use within the support network of adults living in precarious housing in a socially marginalized neighborhood of Vancouver, Canada, PLoS One 14 (9) (2019) e0222611.
[24] S.H. Woolf, H. Schoomaker, Life expectancy and mortality rates in the United States, 1959–2017, JAMA 322 (20) (2019) 1996, https://doi.org/10.1001/jama.2019.16932.
[25] Devlin J, Chang M-W, Lee K, *et al.* Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv Prepr arXiv181004805* 2018.
[26] Kim Y. Convolutional neural networks for sentence classification. arXiv Prepr arXiv14085882 2014.
[27] Pennington J, Socher R, Manning C.D. Glove: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014. 1532–43.
[28] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural. Comput. 9 (8) (1997) 1735–1780.
[29] Mart\'\in~Abadi, Ashish~Agarwal, Paul~Barham, *et al.* {TensorFlow}: Large-Scale Machine Learning on Heterogeneous Systems. 2015.https://www.tensorflow.org/.
[30] M.C. Chen, R.L. Ball, L. Yang, N. Moradzadeh, B.E. Chapman, D.B. Larson, C. P. Langlotz, T.J. Amrhein, M.P. Lungren, Deep learning to classify radiology free-text reports, Radiology 286 (3) (2018) 845–852.
[31] W. Quan, Z. Chen, J. Gao, et al., Comparative study of CNN and LSTM based attention neural networks for aspect-level opinion mining, in: 2018 IEEE International Conference on Big Data (Big Data), IEEE, 2018, pp. 2141–2150.
[32] Chollet F, others. Keras [Internet]. GitHub; 2015. Available from: https://github.com/fchollet/keras.
[33] K. Sechidis, G. Tsoumakas, I. Vlahavas, On the stratification of multi-label data, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2011, pp. 145–158.
[34] S. Bird, E. Klein, E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit, ' O'Reilly Media, Inc`. (2009).
[35] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., Scikit-learn: Machine learning in Python, J Mach Learn Res 12 (2011) 2825–2830.
[36] G.C. Cawley, N.L.C. Talbot, On over-fitting in model selection and subsequent selection bias in performance evaluation, J Mach Learn Res 11 (2010) 2079–2107.
[37] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, J Mach Learn Res 13 (2012).
[38] J.B. Lamy, A. Venot, C. Duclos, PyMedTermino: an open-source generic API for advanced terminology services, in: Digital Healthcare Empowering Europeans, IOS Press, 2015, pp. 924–928.
[39] Stemerman R, Arguello J, Brice J, *et al.* Identification of social determinants of health using multi-label classification of electronic health record clinical notes. *JAMIA Open* 2021.
[40] Yin W, Kann K, Yu M, *et al.* Comparative study of CNN and RNN for natural language processing. arXiv Prepr arXiv170201923 2017.
[41] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st international conference on neural information processing systems, 2017, pp. 4768–4777.
[42] J. Lee, W. Yoon, S. Kim, et al., BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240.
[43] Alsentzer E, Murphy JR, Boag W, et al. Publicly available clinical BERT embeddings. arXiv Prepr arXiv190403323 2019.
[44] Z. Yang, Z. Dai, Y. Yang, et al., Xlnet: Generalized autoregressive pretraining for language understanding, Adv Neural Inf Process Syst 32 (2019).