Proceedings of the Seventeenth International AAAI Conference on Web and Social Media (ICWSM 2023)

HealthE: Recognizing Health Advice & Entities in Online Health Communities

Joseph Gatto*, Parker Seegmiller*, Garrett M Johnston, Madhusudan Basak, Sarah Masud Preum

Dartmouth College Department of Computer Science
{joseph.m.gatto.gr, matthew.p.seegmiller.gr,
garrett.m.johnston.22, madhusudan.basak.gr, sarah.masud.preum}@dartmouth.edu

Abstract

The task of extracting and classifying entities is at the core of important Health-NLP systems such as misinformation detection, medical dialogue modeling, and patient-centric information tools. Granular knowledge of textual entities allows these systems to utilize knowledge bases, retrieve relevant information, and build graphical representations of texts. Unfortunately, most existing works on health entity recognition are trained on *clinical notes*, which are both lexically and semantically different from public health information found in online health resources or social media. In other words, existing health entity recognizers vastly under-represent the entities relevant to *public health data*, such as those provided by sites like WebMD. It is crucial that future Health-NLP systems be able to model such information, as people rely on online health advice for personal health management and clinically relevant decision making.

In this work, we release a new annotated dataset, **HealthE**, which facilitates the large-scale analysis of online textual health advice. HealthE consists of 3,400 health advice statements with token-level entity annotations. Additionally, we release 2,256 health statements which are *not* health advice to facilitate health advice mining. HealthE is the first dataset with an **entity-recognition label space designed for the modeling of online health advice**. We motivate the need for HealthE by demonstrating the limitations of five widely-used health entity recognizers on HealthE, such as those offered by Google and Amazon. We additionally benchmark three pretrained language models on our dataset as reference for future research. All data is made publicly available.

Introduction

Online health information or *health advice* found on health websites (e.g., WebMD, CDC website) plays an important role in improving health literacy, medical information search, and patient empowerment (Calixte et al. 2020; Masoni et al. 2013; Kubb, Foran et al. 2020). A 2019 survey showed that 89% of patients in the U.S. would google their health symptoms before going to their doctor, a proportion that has been increasing for over a decade, indicating that the computational modeling of online health advice has become a critical public health task (Eligibility 2020; Gualtieri 2009).

This study defines health advice as any information with an explicit or implicit suggestion regarding a personal healthcare decision. To model health advice at scale, we propose that a system must be able to 1) Detect if a given text contains actionable health advice and 2) Recognize which entities in the text pertain to relevant health concepts. The first task, Health Advice Classification (HAC), aims to identify texts from which users can make personal healthcare decisions. Automatic recognition of health advice will enable more advanced patient-centric health applications and facilitate the collection of large-scale health advice datasets.

Once a text has been classified as health advice, the next step is Health Entity Recognition (HER), i.e., extracting and classifying health entities from natural language. We note that solutions to HER are formulated the same as the popular task of Medical Named Entity Recognition (MedNER). However, we define the task as HER, which includes entities important to modeling informal health texts where health concepts are mentioned in layman's terms and thus fall outside academic/clinical vocabularies. For example, in the first row of Table 1, all of the popular HER models fail to extract "moving your body" (i.e., physical activity), which is crucial to accurately modeling the relationship between the relevant entity and other health concepts in that text.

Extraction of health entities from online textual health advice can inform various downstream tasks relevant to online health communities (OHCs) and large-scale health advice mining, including misinformation detection (Cui et al. 2020), medical dialogue systems (Chintagunta et al. 2021), and patient-centric information tools (Dai, Sun, and Wang 2020; Preum et al. 2017a; Gatto, Basak, and Preum 2023; Beaunoyer et al. 2017; Preum et al. 2017b).

For example, Wührl and Klinger (2022) use health entities to reduce the complexity of health misinformation statements on social media, allowing a fact-checking algorithm to focus solely on the relevant health entity relationships needed to determine claim veracity. Chintagunta et al. (2021) show how a health entity extractor can improve the quality of AI-generated medical dialogue summaries. Roy and Pan (2021) illustrate how health entity extraction is key to infusing medical knowledge bases with large language models. Thus, the automatic extraction of health entities is

*These authors contributed equally to this work.

Health Advice	Google	Amazon	Spacy	HealthE Ground Truth
Did you know moving your body can help ease the fatigue of multiple sclerosis? See how walking swimming yoga and other exercises do the trick.	fatigue: PROBLEM multiple sclerosis: PROBLEM	fatigue: MEDICAL CONDITION	fatigue: DISEASE multiple sclerosis: DISEASE	moving your body: Exercise multiple sclerosis: Disease other exercises: Exercise swimming: Exercise walking: Exercise yoga: Exercise
Do not underestimate calories of snacks.	None	None	None	snacks: Food
Order fruit for dessert instead of ice cream or cake.	None	None	None	cake: Food dessert: Food fruit: Food ice cream: Food

Table 1: Comparing outputs of three existing HER systems to HealthE labels. Specifically we highlight the outputs from Google's Healthcare Natural Language API (Google), Amazon Comprehend Medical (Amazon), and SciSpacy (Spacy). We see that while the existing HER systems are capable of extracting some correct health entities, the HealthE labels offer increased coverage of health-related entities.

valuable to solutions that aim to manage large-scale public health information and promote health literacy.

Unfortunately, most existing health entity extractors cover only a small subset of the entity classes commonly found in online textual health advice, e.g., medication and other formal treatment options, disease, symptoms, and side effects. **This is because these models are almost exclusively trained on clinical and biomedical texts** (i.e., EHR notes, biomedical articles) (Neumann et al. 2019; Lee et al. 2020). Thus, the training sets of current state-of-the-art HER models are largely void of health data with entity annotations critical to computationally represent health advice in layman's terms, such as food, exercise, and informal mentions of physiological status, and treatment options. The inclusion of entity classes such as food and exercise, for example, promotes the development of patient-centric tools for personalized health management. Consider people with diabetes who are prescribed to follow restrictive dietary guidelines and exercise programs. Existing HER models cannot model the full scope of health entity relationships in such texts, as they are historically clinician-centered algorithms.

It is also important to note that existing HER models may struggle with the lexical and semantic differences between their training data and the informal, lay-person-targeted structure of online health advice. This can make the predictive performance of classes found in an HER model's label space *low* when detection needs to occur outside of technical content, such as on public health websites or on social media. Challenges in Health-NLP brought about by the stylistic mismatch between user-generated health content and technical or synthetic health content has been discussed previously in the literature (Wührl and Klinger 2022). Evidence of the limitations of existing HER models can be found in Table 1.

Our Contribution

1. We introduce a new annotated dataset for HER, **HealthE**[1], containing 3,400 pieces of health advice with hand-labeled health entities. To the best of our knowledge, this is the largest HER dataset designed for online health communities (OHC) and patient-centric information systems. In addition to the 3,400 advice statements contained in HealthE, we release 2,256 pieces of non-advice statements to facilitate the related task of HAC. Both HealthE and the non-advice statements are collected from a set of popular and reliable health websites, including WebMD[2], MedlinePlus[3], CDC[4], and MayoClinic[5] where the textual content is carefully reviewed by health professionals.
2. We explore the problem of health entity recognition by benchmarking 3 language models fine-tuned to HealthE for future research. We additionally evaluate 5 off-the-shelf pre-trained HER models on HealthE, namely, Amazon Comprehend Medical, Google Healthcare NLP Pipeline, Apache cTAKES (Savova et al. 2010), SciSpacy (Neumann et al. 2019), and BioBERT (Lee et al. 2020). The results demonstrate the limitations of these widely-used existing solutions in the HealthE label space and underlines the need to develop a customized solution to capture the informal nature of the language found in online health advice. Additionally, we benchmark the HAC task to facilitate future work on public health advice mining.

Related Works

Medical Named Entity Recognition

HealthE is the largest health entity recognition dataset designed for OHCs. However, many public benchmark datasets have been released for the related task of MedNER. Several MedNER datasets contain abstracts or articles from PubMed with named entities tagged as either *DISEASE* or *CHEMICAL*. (Krallinger et al. 2015; Li et al. 2016). For

[1]https://zenodo.org/record/7539392#.Y8SY2HbMKM8

[2]https://www.webmd.com

[3]https://www.medlineplus.gov

[4]https://www.cdc.gov

[5]https://www.mayoclinic.org

example, the popular NCBI corpus contains 793 PubMed abstracts annotated only for disease mentions (Doğan, Leaman, and Lu 2014). Similarly, the more recent NLM-Chem dataset contains 150 full PubMed texts with annotation for chemical entities (Islamaj et al. 2021). Other works in MedNER differ in how they classify the entities. MedMentions, for example, contains over 4,000 PubMed abstracts but with entity tags which link named entities to a subset of the 3 million concepts in the UMLS concept ontology (Mohan and Li 2019; Bodenreider 2004). We note that downstream applications on non-technical health texts often leverage UMLS when entity modeling is required (Chintagunta et al. 2021; Kalyan and Sangeetha 2020). Unlike the aforementioned MedNER datasets, HealthE offers a wider variety of entity types with samples sourced from layperson-targeted content. Both the differing label space and linguistic properties demand explicit consideration in model design as existing works struggle to adapt to this shift in data distribution.

HER has also been explored in the context of social media on sites such as Twitter. A common set of entity classes on social media HER datasets includes *DISEASE*, *DRUG*, and *SYMPTOM* (Batbaatar and Ryu 2019; Jimeno-Yepes et al. 2015). METS-CoV, for example, contains 10,000 tweets annotated for 7 types of entities, including the health-related *DISEASE*, *DRUG*, *SYMPTOM*, and *VACCINE* entity tags (Zhou et al. 2022). Both HealthE and social media HER datasets contain non-technical language. However, social media posts are user-generated. The linguistic nature of online health advice from health websites differ in style and intention from advice on social media, e.g., user-generated content on social media might contain shorthands, slang terms, colloquial terms, and special characters like emojis.

Health Advice Classification

To the best of our knowledge, the task of Health Advice Classification (HAC) has not been explored outside the realm of biomedical research texts. The prevalence of health advice in research literature has been studied in a clinical context, where it was shown a large percentage of medical publications *do* make suggestions regarding medical practice (Prasad et al. 2013). In the NLP domain, Li, Wang, and Yu (2021) explore a BERT-based solution for classifying health advice found in PubMed articles. Conversely, our study collects advice texts from various popular health websites that support patients with health-related information seeking and decision making. The language on these platforms is stylistically and semantically different from PubMed articles, distinguishing HealthE from works like Li, Wang, and Yu (2021).

Identifying if a text contains advice or not is generally related to the problem of suggestion mining, where the goal is to identify texts containing tips or advice. The popular SemEval-2019 dataset provides suggestion annotation for online review data (Negi, Daudert, and Buitelaar 2019). Similar works on general advice classification such as Govindarajan et al. (2020) have explored identifying advice on Reddit. Although these works are formulated similar to HAC, they don't focus on health advice. Future works may explore the impact of cross-domain training using data from these related works to improve the performance of health advice classifiers.

Data Collection & Annotation

Collecting New Health Advice Statements Data collection for HealthE began by scraping over 30,000 candidate advice statements from four online health sources: WebMD, Medline Plus, CDC, and Covid Protocols. The scraping process took place between Aug 15, 2021 and September 10, 2021. A random sample of 4,500 candidates were then chosen for annotation by three human annotators. Samples were classified by human annotators as being "Advice" or "Not Advice". 2,244 out of the 4,500 statements were identified as "Advice" by at least two annotators. This annotation process resulted in a Fleiss' Kappa coefficient of 0.71, indicating substantial inter-annotator agreement (McHugh 2012). All samples annotated as advice received annotation for health entity recognition — the details of which are described later in this section.

Leveraging Existing Advice Data In addition to the 2,244 advice samples collected for this study, we include 1,156 samples from the Preclude dataset (Preum et al. 2017a). Preclude is a public dataset of health advice statements that primarily relate to food, exercise, lifestyle, and over-the-counter drugs. 790 of these advice statements came from 8 different mobile health apps and 366 of them came from WebMD, Yahoo! Health, MayoClinic, and Healthline. The intended use of Preclude data is to detect conflicting medical advice between two texts. Preclude had 3 human annotators annotate entities in health advice statements and their corresponding polarity. In this work, we utilize the advice statements and extracted entities provided by Preclude, with each entity manually mapped to one of our six classes by the authors.

Dataset Description The resulting HealthE dataset contains 3,400 health advice statements. Each sample is annotated for the following entity types: (i) Food / Nutrient, (ii) Disease / Condition, (iii) Medicine / Supplement / Treatment, (iv) Exercise, (v) Vitals / Physiological Status, and (vi) Other. A sample with annotation from all classes is highlighted in Figure 1. Annotation details and guidelines are described in the following section.

Health Entity Annotation Scheme

The HealthE label space is comprised of six entity classes related to the modeling of health advice. In this section, 1) We provide the definition of the entity class provided to annotators and 2) We highlight one sample from HealthE which is representative of each class. In each sample, we bold any entities with the given class type. We also note that we only bold the entity of the class being defined for illustration — other entities may be annotated in the actual dataset.

Food / Nutrient (FOOD) Entities that refer to specific food items or nutrients. Additionally, any mentions of abstract food classes (e.g. low-fiber foods).

- `Finding` **`healthy food`** `choices on the road can be an adventure. Don't fill`

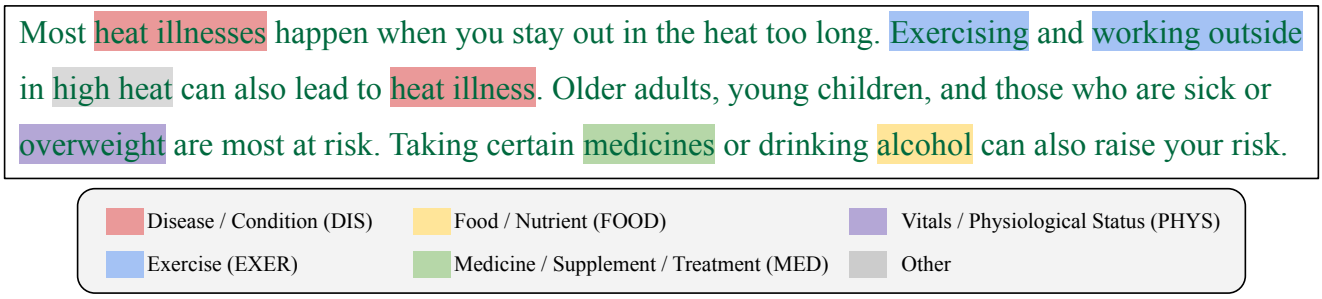

Figure 1: Sample from the HealthE dataset with entities labeled from each class in the HealthE label space.

up on **low-fiber foods** at fast food chains, rest stops, or airports. Instead, pack a few **high-fiber snacks** for your trip to help keep you regular. Good choices include **whole grain crackers, dried or fresh fruit, fresh vegetables,** or **whole grain cereals.**

Medicine / Supplement / Treatment (MED) Any entity referring to a medicine, supplement, or medical treatment.

- Certain medical conditions that increase bone breakdown, including kidney disease, Cushing's syndrome, and an overactive thyroid or parathyroid, can also lead to osteoporosis. **Glucocorticoids,** also known as **steroids,** also increase bone loss. **Anti-seizure drugs** and long-term immobility because of paralysis or illness can also cause bone loss.

Disease / Condition (DIS) Entities which refer to diseases or medical conditions. Additionally, entities which may represent addictions such as "smoking".

- Has your doctor said you have **high cholesterol**? Then you know you need to change your diet and lifestyle to lower cholesterol and your chance of getting **heart disease.** Even if you get a prescription for a cholesterol drug to help, you'll still need to change your diet and become more active for heart health. Start with these steps.

Exercise (EXER) Entities that mention explicit or implicit forms of exercise

- Once you can do **stretching** and **strengthening exercises** without pain, you can gradually begin **running** or **cycling** again. Slowly build up distance and speed.

Vitals / Physiological Status (PHYS) Entities that are the name of a vital or physiological status. Specifically, this class encompasses things like vitals, organs, physical characteristics, symptoms, behaviors, and general state of health/well-being.

- A class of medications known as 5-alpha reductase inhibitors help stop the prostate from growing or even shrink it in some men. They lower the production of **dht,** a hormone involved in prostate growth. However, these medications can also lower **sex drive** and cause erectile dysfunction. And it can take 6 months or more to feel the benefits.

Other (OTH) The Other class encapsulates all entities that were not appropriate for the other five categories but are still deemed crucial to the modeling of a given health advice statement. This category exists in order to be exhaustive and inclusive of medically relevant objects or phrases that do not fit into the other five categories.

- Getting enough quality slumber may lower your pain and fatigue. Limit caffeine and alcohol and avoid tobacco. Eat your last meal of the day several hours before you go to sleep. Keep your bedroom comfortable and free of **electronics.**

Health Entity Annotation Process

A team of computational health researchers completed the HER annotation tasks. Each annotator frequently utilizes online resources for health information seeking and was trained on the annotation tasks. The team consists of 14 annotators in total. Each advice statement was annotated by one annotator. The annotation was then independently reviewed by at least two different annotators before a sample was to be included in HealthE. Thus, three annotators independently confirmed each annotation. This method is in contrast to having multiple annotators do the entity extraction independently, which had various drawbacks, including annotation cost and misleading annotation mismatches. To illustrate the latter, consider the sentence, "Make sure to eat

Model	DIS	MED	FOOD	EXER	PHYS	OTH	AVG
SciSpacy	0.46	0.38	-	-	-	-	0.42
BioBERT	0.23	0.29	-	-	-	-	0.26
ACM	0.34	0.47	-	-	0.02	-	0.28
cTAKES	0.36	0.42	-	-	0.03	-	0.27
GHNLP	0.40	0.56	-	-	0.11	-	0.18
Distil-BERT	0.70	0.65	0.74	0.64	0.62	0.00	0.56
BERT	0.72	0.67	0.74	0.66	0.63	0.03	0.58
DeBERTa-v3	0.74	0.69	0.78	0.67	0.66	0.02	0.59

Table 2: Experimental results of HER task on HealthE dataset. Off-the-shelf HER models received no additional training and performance is reported for the classes which we can map to the HealthE label space, along with macro F1. Our end-to-end baselines display performance on HealthE with end-to-end fine-tuning.

more fresh vegetables." If annotator 1 marked "fresh vegetables" as FOOD and annotator 2 marked only "vegetables," the annotation for this entity may be excluded from the dataset if we only considered entities with complete annotator intersection. Thus, we employ an extract-then-check method to avoid this issue. To verify the reliability of the annotation, we sampled 50 advice statements and had two authors independently label them using HealthE guidelines. We report the inter-annotator agreement (IAA) using standard classification metrics, where one annotator is treated as the prediction and the other the ground truth label (as done in similar works such as (Deleger et al. 2012)). The weighted average F1 score between both annotators is 0.68. This is a significant IAA for NER as no credit is given for partial entity matches. F1 scores for each category are DIS (0.71), EXER (0.75), FOOD (0.84), MED (0.61), OTH (0.30), and PHYS (0.61). A common error found in IAA analysis was occasional subjectivity in annotating DIS vs PHYS. For example, in one instance annotators disagreed on if "elevated blood pressure" was a DIS or PHYS — an entity annotation which is dependent on context. We note that this method for analyzing label quality, i.e. calculating subsequent IAA for a sample of documents, is commonly performed to verify IAA of NER datasets (Bamman, Lewke, and Mansoor 2019; Derczynski, Bontcheva, and Roberts 2016).

Data Characterization

Within HealthE's 3,400 health advice statements we find 13,989 labeled entity instances. The distribution of class labels is given in Figure 2. The MED class is found most often in the dataset, accounting for 28% of all labeled entities. The EXER class and OTH class are the least frequent labels in HealthE, representing just 5% and 3% of entity labels, respectively.

Of the 13,989 labeled entity instances in HealthE, 5,418 entities are unique (i.e. 8,571 labeled health entities are repeated more than once across the 3,400 health advice statements). In addition to having the most total entities, the MED class has the highest number of unique entities. This is in part due to the nature of medication brands appearing in various forms e.g., the entities "ibuprofen," "advil," and "motrin" all refer to the same medication. We also note that while the EXER class has more total labeled entity instances than the OTH class, it also encompasses far fewer unique en-

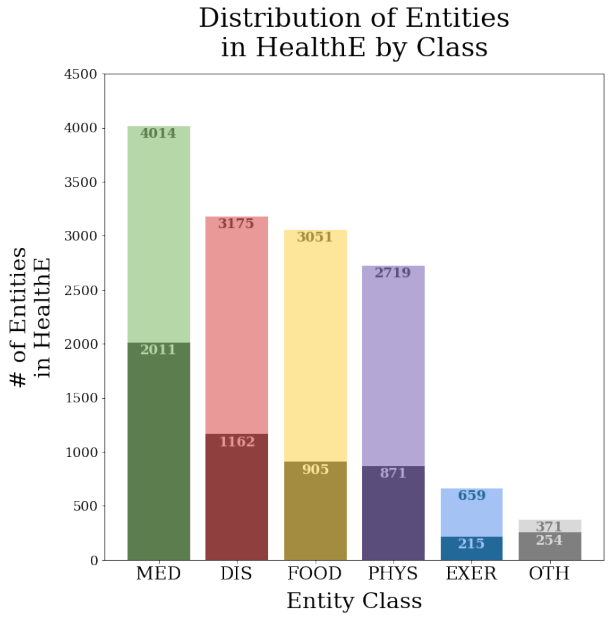

Figure 2: Distribution of entities by class label in HealthE. The light bars represent the total number of labeled entity instances in each class, and the dark bars represent the total number of unique entity instances in each class. For example, there are 3,051 total and 905 unique food labels across the 3,400 HealthE samples.

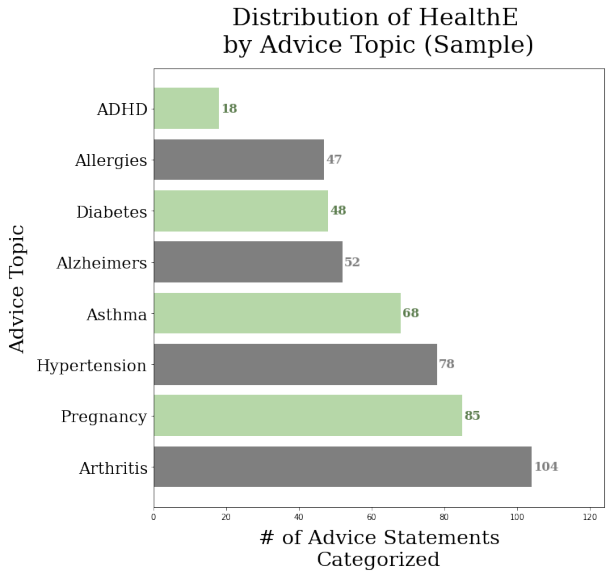

Figure 3: The distribution of advice topics across a sample of 500 advice statements in HealthE. Sample topics were determined by reviewing the source website/application from which a sample was scraped.

tities than the OTH class. This is partially because 20% of entities labeled as the EXER class are simply the word "exercise," as in the advice statement "you can **exercise** anywhere, any time."

To demonstrate the distribution of health topics mentioned in the advice statements from the OHC's represented in HealthE, a random sample of 500 advice statements was taken from HealthE and each advice statement was linked back to it's source website and given a general advice topic label by an annotator. We display the distribution of these advice topics in Figure 3. While the sample contained a range of topics, some topics like arthritis and hypertension were more frequent than others like ADHD.

Experiments

In this section, we introduce the model architectures used to establish baseline results on the HealthE dataset. First, we present 5 end-to-end (E2E) baselines for HAC and 3 for HER. Then, we present five off-the-shelf (OTS) HER models trained on technical texts.

End-to-End Baselines

We benchmark two traditional NLP models and three transformer models (Vaswani et al. 2017) on HealthE in the context of end-to-end binary classification (HAC). (i) TFIDF+RF (Manning 2009) a common statistical measure paired with a Random Forest (RF) classifier, (ii) GloVe (Pennington, Socher, and Manning 2014) a standard word embedding method (also paired with RF), (iii) DistilBERT (Sanh et al. 2019) a light-weight transformer model, (iv) BERT (Devlin et al. 2019) the popular large-parameter transformer model, and (v) DeBERTa-v3 (He, Gao, and Chen 2021) a more recent adaptation of the transformer with various improvements in pre-training strategy. The first two models generate embeddings for each statement and are used as input for an RF classifier to give a good baseline for text classification. The final three models represent a popular yet diverse set of accessible transformers which reflect the state-of-the-art in large language model fine-tuning. These transformer models are also used as baselines for end-to-end token classification (HER).

Existing HER Baselines

On the HER task, we investigate the capacity of five popular HER models to classify the HealthE dataset. Each model is slightly different but shares the common goal of extracting medical/health entity information from medical texts. Specifically, these models are known to work well on modeling clinical notes and biomedical research publications. We note, however, that these models are commonly applied out of their source domain such as on social media data (Dey et al. 2021; Bai and Zhou 2020; Jeon et al. 2020) and are thus highly relevant to this study. We briefly describe each model below:

- **SciSpacy** (Neumann et al. 2019): This is an NER model trained on the BC5CDR corpus (Li et al. 2016) which contains annotations for chemicals and diseases across 1500 PubMed articles. SciSpacy achieves a reported 84.53 F1 score on the task of recognizing chemical and disease entities.
- **BioBERT** (Lee et al. 2020): This model is pre-trained on both the NCBI Disease (Doğan, Leaman, and Lu 2014) and BC5CDR corpus and is fine-tuned for the recognition of chemical and disease entities. BioBERT achieves state-of-the-art performance on a variety of biomedical text mining tasks, with a greater than 0.90 F1 score on both the NCBI and BC5CDR test sets.
- **Amazon Comprehend Medical (ACM)**[6]: This service provided by Amazon claims to "understand and extract health data from medical text, such as prescriptions, procedures, or diagnoses". Various Health-NLP studies have investigated ACM as a medical information extractor (Sarabadani et al. 2022; Shah-Mohammadi, Cui, and Finkelstein 2021; Li et al. 2022).
- **Apache cTAKES** (Savova et al. 2010): cTAKES is designed to extract information from free-text clinical notes. Trained on datasets derived from Mayo Clinic electronic medical records, cTAKES reports an NER evaluation F1 score of 82.40. We use the default clinical pipeline which labels 5 entity classes.
- **Google Healthcare Natural Language API (GHNLP)** [7]: An off-the-shelf tool provided by Google for the processing of medical texts. While this tool is more recent than Amazon's and has been studied less by the NLP community, it is of interest as it has by far the most diverse label space of our HER baselines, with 28 entity

[6]https://aws.amazon.com/comprehend/medical/
[7]https://cloud.google.com/healthcare-api/

Model	F1	Precision	Recall
TFIDF+RF	0.82	0.91	0.76
GloVe	0.80	0.87	0.75
Distil-BERT	0.81	0.81	0.81
BERT	0.82	0.82	0.82
DeBERTa-v3	0.84	0.84	0.83

Table 3: Mean Health Advice Classification performance over 5-fold cross validation.

classes [8].

Given that the above models do not share a common label space with HealthE, we mapped each model's outputs to our label space. The mapping was performed carefully by the authors through the following process. 1) We map the entity class from a model to HealthE by using it's class definition 2) We empirically verify that the mapping is accurate via annotation inspection. For example, GHNLP tags some entities as "Body Measurement", which by definition aligns with our class "Vitals / Physiological Status". We then sampled the entities tagged by GHNLP in HealthE as Body Measurement to confirm this mapping is correct. Tokens annotated with a label which could *not* be mapped to our label space were treated as if they received no prediction during our evaluation (e.g. GHNLP tags tokens like "worse" and "chronic" as SEVERITY, which has no mapping to the HealthE label space).

Experimental Setting

For the transformer E2E baselines, we run our experiments using Huggingface (Wolf et al. 2019). We report the mean per-class F1 across a 5-fold cross-validation on the HealthE dataset. In each run, we fine-tune the model for 5 epochs with a learning rate of 2e-5, and weight decay of 0.01. HER evaluation metrics are computed with SeqEval (Nakayama 2018).

Results

Health Advice Classification All baseline models perform well on HAC as shown in Table 3. We find the best performance from DeBERTa-v3, however light-weight models such as TFIDF+RF also achieve a high F1 score. We find there are only minor differences in F1 score between the classical NLP methods and modern transformer-based approaches. The ability of simple statistical models like TFIDF to perform similarly to transformer models on HAC may in part be attributed to the presence of overt advice-specific linguistic markers in the data. For example, advice statements often contain second-person pronouns (e.g., "If you are taking...") or imperative sentences. Such markers may be useful for suggestion mining. Similarly, a hint to the model that a sample is not advice might be the lack of imperative language. Many non-advice samples in HealthE are statements of fact or lists of health concepts.

[8]https://cloud.google.com/healthcare-api/docs/concepts/nlp

Health Entity Recognition Results for both the OTS and end-to-end models are shown in Table 2. As mentioned throughout the paper, OTS HER models are not trained to predict most of the HealthE label space. However, we find consistent poor performance from OTS models on classes it is trained to predict, namely Disease and Medication mentions. We identify reasons for poor performance below:

- **Modeling Generic Expressions:** OTS models don't always mark abstract terms such as "medicine" or "pills" as a health entity. However, if you consider a sample such as "*ask your provider if you should stop taking any medicines that could thin your blood*", identifying medicines as a medical entity is important for representing the text computationally.
- **Niche Medical Treatments:** OTS models can struggle with unfamiliar medical treatments. For example, when presented with the following sample: "*people whose immune system isnt fully functioning should ask their doctor before trying nasal irrigation because they are at greater risk for infections*", most of the OTS models were unable to identify "nasal irrigation" as a medical treatment.
- **Correct Entity, Wrong Label:** Most models mark entities such as "headache", "fatigue", and "nausea" as *DISEASE*. However, the context of the advice data often refers to these mentions as temporary physiological status, symptoms of on ongoing condition, or side effects of a treatment. Thus, errors occur due to the failure of the models to capture the context of the annotated health advice.
- **Others**: Other reasons for poor performance include stylistic difference in language and presence of complex, multi-token entities in the advice data, such as "extreme heat and cold", "chronic lack of sleep", or "buildup of abnormal proteins".

Our E2E HER models perform reasonably well across all classes except Other. The other class is particularly difficult to predict as it occurs extremely rarely and contains a high percentage of unique entity labels (as shown in Figure 2). DeBERTa-v3 shows the best performance, with a macro F1 score of 0.59 across all classes. All models were able to predict FOOD and DIS with greater than 70% F1 score. Performance dwindled on MED and PHYS, but future works may benefit from integrating medical knowledge bases with HealthE to refine the predictions of such classes.

Broader Impacts

Novel Public Dataset with Reliable Annotation

We release a public dataset on health advice that contains labels for health advice classification (HAC) and health entity recognition (HER). Our work on HAC can help facilitate the detection of health advice statements in downstream Health NLP applications as well as enhance suggestion mining for the healthcare domain. Our HER data has a novel label space and brings annotated data from online health communities into the realm of HER. Additionally, our annotation policies ensure high-quality data reducing errors that can occur via

the more traditional alternative, i.e., crowd-sourced annotation. This is achieved by using multiple annotators per sample and training the annotators before annotation to extract health entities.

Expand the Capacity of HER-Dependent NLP

The ability to automatically identify and classify health entities in natural language has improved the performance of various systems, including misinformation detection, medical dialogue modeling, and health conflict detection. An example of a system that utilizes recognized health entities is DETERRENT, which propagates extracted textual entities through a medical knowledge graph to assess whether a healthcare statement is true (Cui et al. 2020). HealthE provides an additional set of annotations/labels that may allow future Health NLP systems, particularly those like DETERRENT that aim to model health advice, to build more useful computational text representations.

Representation of OHC data in Health NLP

As discussed throughout the paper, most large annotated HER datasets are in the domain of biomedical research literature or social media data. Representation of data from OHCs like WebMD and HealthLine is crucial to future works which aim to build patient-centric health management tools which rely on official online sources. For example, Preum et al. (2017a) aims to detect conflicting health advice statements from official websites and health apps. This is a safety-critical task as someone with a pre-existing condition may not know that following advice from an OHC could have fatal consequences (e.g., adverse drug interaction with blood thinners during pregnancy). Similar works building patient-centric tools may too benefit from improved HER via HealthE.

FAIR Data Principles

To ensure FAIR principles for the released HealthE dataset and non-advice samples, the authors provide the following information (Hagstrom 2014). The released datasets are **findable** in that they are hosted on the data publishing service Zenodo with the unique DOI [9]. The released datasets are **accessible** in that they are free to access. They are **interoperable** in that they are formatted as plain text samples paired with simply-structured labels and entity tags in a Python-readable format. Finally, the released datasets are **re-usable** as they are paired with a README file explaining correct data usage, as well as an accessible data usage license.

Discussion

From trusted online sources like WebMD and HealthLine to unverified sources like social media and blogs, online health information has greatly influenced how people make personal healthcare decisions. Identifying actionable health advice and computationally modeling its properties are essential to future Health-NLP tools, which aim to help manage public health information. From the perspective of personal health management, HealthE promotes the creation of more advanced patient-centric healthcare applications. Such technology is not possible without a low-level understanding of textual health advice. When considering large-scale public health monitoring, being able to flag which texts contain health advice (similar to how *Misinformation systems* flag claims as being "check-worthy") is crucial for disrupting the spread of misinformation and disinformation. Additionally, improving the quality of health entity recognizers has vast implications for downstream tasks which leverage knowledge bases or graphical text representations (Roy and Pan 2021), among others.

Ethical Impact

This work only involves humans at the data annotation level. Since the data is collected from OHCs, there is no danger to any human subject in this study concerning the data in HealthE. Given that data from OHCs are layperson-targeted with a general audience in mind, underrepresented communities may not be well represented in this dataset. In other words, OHCs may not formulate advice statements targeting smaller sub-populations of users with diverse factors of social determinants of health. Thus, future applications that leverage HealthE to model patient data or specific types of advice data should consider the bias toward a more generic audience.

References

Bai, Y.; and Zhou, X. 2020. Automatic Detecting for Health-related Twitter Data with BioBERT. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, 63–69. Barcelona, Spain (Online): Association for Computational Linguistics.

Bamman, D.; Lewke, O.; and Mansoor, A. 2019. An annotated dataset of coreference in English literature. *arXiv preprint arXiv:1912.01140*.

Batbaatar, E.; and Ryu, K. H. 2019. Ontology-based healthcare named entity recognition from twitter messages using a recurrent neural network approach. *International journal of environmental research and public health*, 16(19): 3628.

Beaunoyer, E.; Arsenault, M.; Lomanowska, A. M.; and Guitton, M. J. 2017. Understanding online health information: Evaluation, tools, and strategies. *Patient education and counseling*, 100(2): 183–189.

Bodenreider, O. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database issue): D267–70.

Calixte, R.; Rivera, A.; Oridota, O.; Beauchamp, W.; and Camacho-Rivera, M. 2020. Social and demographic patterns of health-related Internet use among adults in the United States: a secondary data analysis of the health information national trends survey. *International Journal of Environmental Research and Public Health*, 17(18): 6856.

[9] 10.5281/zenodo.7539392

Chintagunta, B.; Katariya, N.; Amatriain, X.; and Kannan, A. 2021. Medically aware gpt-3 as a data generator for medical dialogue summarization. In *Machine Learning for Healthcare Conference*, 354–372. PMLR.

Cui, L.; Seo, H.; Tabar, M.; Ma, F.; Wang, S.; and Lee, D. 2020. Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 492–502.

Dai, E.; Sun, Y.; and Wang, S. 2020. Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 853–862.

Deleger, L.; Li, Q.; Lingren, T.; Kaiser, M.; Molnar, K.; Stoutenborough, L.; Kouril, M.; Marsolo, K.; and Solti, I. 2012. Building gold standard corpora for medical natural language processing tasks. *AMIA Annu. Symp. Proc.*, 2012: 144–153.

Derczynski, L.; Bontcheva, K.; and Roberts, I. 2016. Broad Twitter Corpus: A Diverse Named Entity Recognition Resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 1169–1179. Osaka, Japan: The COLING 2016 Organizing Committee.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dey, V.; Krasniak, P.; Nguyen, M.; Lee, C.; and Ning, X. 2021. A Pipeline to Understand Emerging Illness Via Social Media Data Analysis: Case Study on Breast Implant Illness. *JMIR Med Inform*, 9(11): e29768.

Doğan, R. I.; Leaman, R.; and Lu, Z. 2014. NCBI disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47: 1–10.

Eligibility. 2020. Every state's most googled medical symptoms. https://eligibility.com/medicare/states-most-googled-medical-symptom. Accessed: 2023-01-15.

Gatto, J.; Basak, M.; and Preum, S. M. 2023. Scope of Pre-trained Language Models for Detecting Conflicting Health Information. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM), AAAI Press*.

Govindarajan, V. S.; Chen, B. T.; Warholic, R.; Li, J. J.; and Erk, K. 2020. Help! Need Advice on Identifying Advice. In *Proceedings of The 2020 Conference on Empirical Methods in Natural Language Processing*.

Gualtieri, L. N. 2009. The doctor as the second opinion and the internet as the first. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*, 2489–2498.

Hagstrom, S. 2014. The FAIR data principles. *FORCE11*.

He, P.; Gao, J.; and Chen, W. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Islamaj, R.; Leaman, R.; Kim, S.; Kwon, D.; Wei, C.-H.; Comeau, D. C.; Peng, Y.; Cissel, D.; Coss, C.; Fisher, C.; et al. 2021. NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature. *Scientific Data*, 8(1): 1–12.

Jeon, J.; Baruah, G.; Sarabadani, S.; and Palanica, A. 2020. Identification of Risk Factors and Symptoms of COVID-19: Analysis of Biomedical Literature and Social Media Data. *J Med Internet Res*, 22(10): e20509.

Jimeno-Yepes, A.; MacKinlay, A.; Han, B.; and Chen, Q. 2015. Identifying diseases, drugs, and symptoms in twitter. In *MEDINFO 2015: eHealth-enabled Health*, 643–647. IOS Press.

Kalyan, K. S.; and Sangeetha, S. 2020. Social Media Medical Concept Normalization using RoBERTa in Ontology Enriched Text Similarity Framework. In *Proceedings of Knowledgeable NLP: the First Workshop on Integrating Structured Knowledge and Neural Networks for NLP*, 21–26. Suzhou, China: Association for Computational Linguistics.

Krallinger, M.; Rabal, O.; Leitner, F.; Vazquez, M.; Salgado, D.; Lu, Z.; Leaman, R.; Lu, Y.; Ji, D.; Lowe, D. M.; et al. 2015. The CHEMDNER corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(1): 1–17.

Kubb, C.; Foran, H. M.; et al. 2020. Online health information seeking by parents for their children: systematic review and agenda for further research. *Journal of Medical Internet Research*, 22(8): e19985.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Li, I.; Pan, J.; Goldwasser, J.; Verma, N.; Wong, W. P.; Nuzumlalı, M. Y.; Rosand, B.; Li, Y.; Zhang, M.; Chang, D.; Taylor, R. A.; Krumholz, H. M.; and Radev, D. 2022. Neural Natural Language Processing for unstructured data in electronic health records: A review. *Computer Science Review*, 46: 100511.

Li, J.; Sun, Y.; Johnson, R. J.; Sciaky, D.; Wei, C.-H.; Leaman, R.; Davis, A. P.; Mattingly, C. J.; Wiegers, T. C.; and Lu, Z. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Li, Y.; Wang, J.; and Yu, B. 2021. Detecting Health Advice in Medical Research Literature. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6018–6029. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Manning, C. D. 2009. *An introduction to information retrieval*. Cambridge university press.

Masoni, M.; Guelfi, M. R.; Conti, A.; and Gensini, G. F. 2013. Pharmacovigilance and use of online health information. *Trends in pharmacological sciences*, 34(7): 357–358.

McHugh, M. L. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3): 276–282.

Mohan, S.; and Li, D. 2019. Medmentions: A large biomedical corpus annotated with umls concepts. *arXiv preprint arXiv:1902.09476*.

Nakayama, H. 2018. seqeval: A Python framework for sequence labeling evaluation. https://github.com/chakki-works/seqeval. Accessed: 2023-01-15.

Negi, S.; Daudert, T.; and Buitelaar, P. 2019. SemEval-2019 Task 9: Suggestion Mining from Online Reviews and Forums. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, 877–887. Minneapolis, Minnesota, USA: Association for Computational Linguistics.

Neumann, M.; King, D.; Beltagy, I.; and Ammar, W. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, 319–327. Florence, Italy: Association for Computational Linguistics.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Prasad, V.; Jorgenson, J.; Ioannidis, J. P.; and Cifu, A. 2013. Observational studies often make clinical practice recommendations: an empirical evaluation of authors' attitudes. *Journal of clinical epidemiology*, 66(4): 361–366.

Preum, S. M.; Mondol, A. S.; Ma, M.; Wang, H.; and Stankovic, J. A. 2017a. Preclude: Conflict detection in textual health advice. In *2017 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 286–296.

Preum, S. M.; Mondol, A. S.; Ma, M.; Wang, H.; and Stankovic, J. A. 2017b. Preclude2: Personalized conflict detection in heterogeneous health applications. *Pervasive and Mobile Computing*, 42: 226–247.

Roy, A.; and Pan, S. 2021. Incorporating medical knowledge in BERT for clinical relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 5357–5366. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sarabadani, S.; Baruah, G.; Fossat, Y.; and Jeon, J. 2022. Longitudinal Changes of COVID-19 Symptoms in Social Media: Observational Study. *J Med Internet Res*, 24(2): e33959.

Savova, G. K.; Masanz, J. J.; Ogren, P. V.; Zheng, J.; Sohn, S.; Kipper-Schuler, K. C.; and Chute, C. G. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5): 507–513.

Shah-Mohammadi, F.; Cui, W.; and Finkelstein, J. 2021. Comparison of ACM and CLAMP for Entity Extraction in Clinical Notes. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 1989–1992.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; and Brew, J. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *CoRR*, abs/1910.03771.

Wührl, A.; and Klinger, R. 2022. Entity-based Claim Representation Improves Fact-Checking of Medical Content in Tweets. In *Proceedings of the 9th Workshop on Argument Mining*, 187–198. Online and in Gyeongju, Republic of Korea: International Conference on Computational Linguistics.

Zhou, P.; Wang, Z.; Chong, D.; Guo, Z.; Hua, Y.; Su, Z.; Teng, Z.; Wu, J.; and Yang, J. 2022. METS-CoV: A Dataset of Medical Entity and Targeted Sentiment on COVID-19 Related Tweets. *arXiv preprint arXiv:2209.13773*.