Extracting Medical Entities from Social Media

Sanja Šćepanović
Nokia Bell Labs
Cambridge, UK
sanja.scepanovic@nokia-bell-labs.com

Enrique Martín-López
Nokia Bell Labs
Cambridge, UK
enrique.martin-lopez@nokia-bell-labs.com

Daniele Quercia
Nokia Bell Labs
Cambridge, UK
quercia@cantab.net

Khan Baykaner
Nokia Bell Labs
Cambridge, UK
khan.baykaner@nokia-bell-labs.com

ABSTRACT

Accurately extracting medical entities from social media is challenging because people use informal language with different expressions for the same concept, and they also make spelling mistakes. Previous work either focused on specific diseases (e.g., depression) or drugs (e.g., opioids) or, if working with a wide-set of medical entities, only tackled individual and small-scale benchmark datasets (e.g., AskaPatient). In this work, we first demonstrated how to accurately extract a wide variety of medical entities such as symptoms, diseases, and drug names on three benchmark datasets from varied social media sources, and then also validated this approach on a large-scale Reddit dataset.

We first implemented a deep-learning method using contextual embeddings that upon two existing benchmark datasets, one containing annotated AskaPatient posts (CADEC) and the other containing annotated tweets (Micromed), outperformed existing state-of-the-art methods. Second, we created an additional benchmark dataset by annotating medical entities in 2K Reddit posts (made publicly available under the name of MedRed) and showed that our method also performs well on this new dataset.

Finally, to validate the accuracy of our method on a large scale, we applied the model pre-trained on MedRed to half a million Reddit posts. The posts came from disease-specific subreddits so we could categorise them into 18 diseases based on the subreddit. We then trained a machine-learning classifier to predict the post's category solely from the extracted medical entities. The average F1 score across categories was .87. These results open up new cost-effective opportunities for modeling, tracking and even predicting health behavior at scale.

CCS CONCEPTS

• **Applied computing** → **Health informatics**; • **Computing methodologies** → **Natural language processing**.

ACM CHIL '20, April 2–4, 2020, Toronto, ON, Canada

ACM ISBN 978-1-4503-7046-2/20/04…$15.00
https://doi.org/10.1145/3368555.3384467

KEYWORDS

health discussions mining, deep learning, Reddit

ACM Reference Format:
Sanja Šćepanović, Enrique Martín-López, Daniele Quercia, and Khan Baykaner. 2020. Extracting Medical Entities from Social Media. In *ACM Conference on Health, Inference, and Learning (ACM CHIL '20), April 2–4, 2020, Toronto, ON, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3368555.3384467

1 INTRODUCTION

An increasing number of people uses online forums to discuss their health. Such forums range from those where patients ask medical experts for advice (The Body, Health24), to those where they talk with each other (AskaPatient, MedHelp), or those where they talk with the general public (Reddit). In them, people tend to discuss the diseases and symptoms they experience as well as the different medications and remedies that they have found helpful. Not only patients but also health professionals [21, 24, 39, 56, 60] and health institutions [14, 21, 22, 31, 68] are increasingly using social media. Moorhead et al. [44]'s review has indeed shown that social media has offered a source of information that is alternative to traditional patient interviews, clinical reports, and electronic health records.

One of the most widely applied tools for extracting medical entities from formal medical texts such as electronic health records is MetaMap.[1] This uses a knowledge-intensive (symbolic) approach based on the Unified Medical Language System (UMLS). Recent studies have shown that MetaMap does not perform as well on social media data as it does on formal medical text [11, 20, 69].

As a result, as we shall see in Section 2, research focus has shifted from traditional tools like MetaMap to statistical NLP techniques, including deep learning. These techniques promise to overcome some of the weaknesses of a symbolic approach by better accounting for two aspects that are key for accurately analyzing social media text: context, and subtle patterns of informal expressions. However, these techniques require expertly annotated datasets for training, which are scarce, limited in size, and not always publicly available. As a result, no study has investigated the applicability of medical entity extraction models beyond single, small datasets such as CADEC [28] (1250 annotated posts from AskaPatient), or Micromed [27] (1300 annotated tweets).

The goal of this work was to build a framework that *accurately extracts a variety of medical entities from different social media sites,*

[1] https://metamap.nlm.nih.gov/

and to demonstrate its applicability on a large-scale. To meet this goal, we made three main contributions:

(1) We designed a deep learning-based framework using contextual embeddings that accurately extracts a wide variety of medical entities such as diseases, symptoms, and drug names (Section 3). This method outperformed the best published performances [67, 73] on the two benchmark datasets (CADEC and Micromed) achieving an F1 score of .82 on AskaPatient posts and .72 on tweets (Section 5).
(2) To complement the existing benchmark datasets, we crowdsourced annotation of 1977 Reddit posts in terms of diseases, symptoms, and drug names (Section 4.4). This dataset is called *MedRed* and is now publicly available[2]. We evaluated our method on the MedRed dataset (Section 5) and it achieved an F1 score of .73.
(3) Finally, we validated our method on half a million Reddit posts categorized based on the disease-specific subreddits in which they were posted. We used our method to extract medical entities, and, for each post, we predicted the post's category (i.e., disease) solely based on the extracted entities (Section 6). Given the widely accepted use of lexicons in the social media literature, as a baseline, we created a lexicon (Dis-LIWC) containing symptoms and drug names related to the 18 diseases present in the Reddit posts. We then extracted medical entities from the Reddit posts using Dis-LIWC, as well as using MetaMap. Across the 18 diseases, the classification average F1 score using extracted entities by our method was .87 as opposed to .61 by MetaMap's and .45 by Dis-LIWC's entities.

2 RELATED WORK

We review both the existing methods for extracting medical entities from social media (Section 2.1) and their applications (Section 2.2).

2.1 Medical Entity Extraction

Initial approaches to medical entity extraction from social media were *keyword-based* [20, 34, 39, 58], which were then followed by those based on *domain-specific lexicons* [20, 47, 50, 69, 71]. While these approaches can work reasonably well on formal medical texts, they have well-known limitations when applied to social media data: they fail to capture the semantic heterogeneity of user's expressions, and to adapt to the variability of informal language, and spelling mistakes [10, 51].

Because of that, machine learning methods such as Conditional Random Fields (CRFs) have been increasinlgy applied to mining social media text [20, 35, 46, 72]. More recently, however, deep-learning methods such as Recurrent Neural Networks (RNNs) have gained popularity over CRFs [61, 63] and have become the go-to technique for extracting medical entites from social media [67]. Xia et al. [70] proposed an RNN model augmented with embeddings trained on a medical corpus, and evaluated it on a dataset of around 6K posts, which is not publicly available.

We compared our work to two approaches that have shown the best published results on the AskaPatient Dataset (CADEC) [67] and on the Twitter Dataset (Micromed) [73], which are two publicly-available datasets:

(1) The tweets from Micromed were mined by Yepes and MacKinlay [73]'s Long Short-Term Memory (LSTM) RNN.
(2) The AskaPatient posts from CADEC were mined by Tutubalina and Nikolenko [67]'s set of techniques which included LSTM, Gated Recurrent Units (GRU), and Convolutional Neural Network (CNN) units.

2.2 Social Media Applications

Health mining has been applied to not only health forums but also to the social media sites of Twitter and Reddit [23, 45, 51].

Twitter. Sarker et al. [57] developed a supervised classifier to detect tweets mentioning drug abuse and Karisani and Agichtein [29] to detect tweets with any personal health mentions. More generally, MacKinlay et al. [40] applied a method to extract medical entities from Twitter and then studied co-occurrences of different symptom mentions with an ibuprofen medicine called Advil. In a similar way, Yepes et al. [74] applied topic modeling to track symptoms related to fatigue from a city's geolocated tweets over one year. All these studies were descriptive in nature and, as such, they did not focus on the technical challenges related to throughly mining a diverse set of medical entities.

Reddit. Park et al. [48] compared three Reddit communities related to mental health (r/Depression, r/Anxiety, and r/PTSD) and found that they had four common themes: sharing of positive emotions, gratitude for receiving emotional support, sleep problems, and work-related issues. Choudhury and De [9] characterized mental health discourse on Reddit by using a combination of keyword-based search and LIWC lexicon. Park and Conway [47] tracked conversations around four topics of public interest over eight years: Ebola, electronic cigarettes, influenza, and marijuana. As one expects, discussions significantly increased in response to exogenous events such as the first case of Ebola being diagnosed, and the strain of H1N1 influenza virus being identified for the first time. Gkotsis et al. [18] predicted the disease associated with a given Reddit post from its entire content, not just from the medical entities present in it – as we restrict ourselves to do. Gaur et al. [17] predicted a Reddit user's suicide risk based on his/her posts. Finally, by geo-referencing Reddit users, Balsamo et al. [7] estimated the prevalence of opiate consumption across US states.

To sum up this literature review, we see that there are two research approaches: one focused on designing state-of-the-art machine learning solutions, which end up being applied to carefully labeled yet limited datasets; and the other focused on large-scale social media data but within limited use-cases. There has been little work in combining the two approaches.

3 METHODS

To address that research gap, we designed and made publicly available[3] a framework (Figure 1) for mining health discussions based on deep learning and contextual embeddings that serves two functions: to extract medical entities from social media text, and to predict the discussed disease.

For the entity extraction module (the left rectangle in Figure 1), we employed the *BiLSTM-CRF sequence labeling architecture* in combination with *contextual embeddings*.

[2]http://goodcitylife.org/Humane-AI

[3]http://goodcitylife.org/Humane-AI

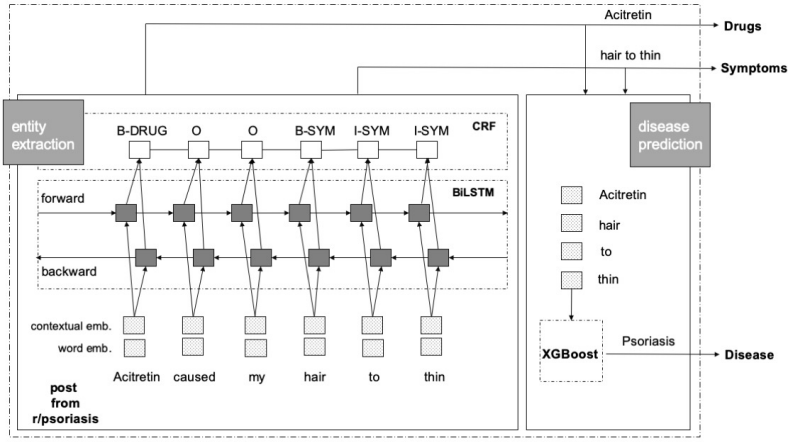

Figure 1: Our framework processing the sentence "*Acitretin caused my hair to thin*", which was taken from the r/psoriasis subreddit. The framework has two modules. The *entity extraction module* has, in turn, two layers: the BiLSTM layer with its LSTM units represented as dark squares, and the CRF layer with its units represented as white squares. The contextual and word embeddings used for input are represented as dotted squares. In the CRF layer, the extracted symptoms, diseases, and drug names are shown: each word is marked as DRUG/SYM, if it is a part of a drug/symptom entity, or as O if it is none of the two. The second module – the *disease prediction module* – uses an XGBoost classifier which takes the extracted symptoms and drug names as input and predicts the corresponding disease.

BiLSTM-CRF. This architecture was introduced by Huang et al. [26], has been repeatedly shown to accurately extract entities [2, 54, 62], and consists of two layers. The first layer is a BiLSMT network (the dashed rectangle in Figure 1), which stands for Bi-directional LSTM. It is called "bi-directional" because there are two sets of LSTM units (the small dark squares in Figure 1) through which "network activations" flow in the two directions, forward and backward. The outputs of the BiLSTM are then passed to the second layer: the CRF layer (enclosed in the other dashed rectangle). The predictions of this layer (the white squares) represent the output of the entity extraction module. To extract the medical entities of symptoms and drug names, BiLSTM-CRF needs to be trained with *labeled data*. To this end, we used the IoB tagging scheme [55, 65]. Each word in this scheme is labeled as being either a symptom (SYM), a drug (DRUG), or none of the two (O); and its position is marked as being at the beginning of an entity (B-) or not (I-). For example, in the sentence in Figure 1, the word "*Acitretin*" is labeled as B-DRUG, the three words "*hair to thin*" as B-SYM, I-SYM, I-SYM, while the remaining words are labeled as O.

Having labeled data, the entity extraction module goes through the two typical phases of training and testing. During the *training* phase, we associated each labeled word with an embedding representation (the dotted square), which is a vector that reflects the word's position in a semantic space. The resulting (embedding, label) pairs are sequentially fed into the BiLSTM, updating its network's weights. Finally, the CRF layer further improved the associations concerning all (embeddings, label) pairs.

During the *testing* phase (i.e., a phase in which the module has to extract the entities from each testing set's sentence), we associated each word in a sentence in the test set with its embedding, and fed the embedding through the previously learned weights. This resulted in a sequence of labels (outputted by the white squares in the figure, that is, by the nodes in the CRF's graph): the labels marked with either SYM or DRUG represent the extracted both symptoms (including diseases) and drug names. If we were to limit ourselves to a unidirectional LSTM, the network would have predicted the word's label only based on its *preceding* words. Instead, by using a bidirectional LSTM, the network predicted the word's label based on the entire sentence. In our running example, the word *hair* taken in isolation might indicate different things, including a body part or a symptom. By contrast, in a bidirectional LSTM, the word *hair* is taken in context (in the context of the words *to thin*) and is, as such, understood to be part of a symptom.

Contextual embeddings. The previous associations between words and embeddings can be done with different types of embeddings. The most commonly used embeddings are Global Vectors for Word Representation (GloVe) [53] and Distributed Representations of Words (word2vec) [43]. However, these do not take into account a word's context. The word 'pain', for example, could be a symptom (e.g., *'I felt **pain** all over my body'*) or could be used figuratively (e.g., *'He was such a **pain** to deal with'*). To account for context, the research has recently moved from word-based to *contextual embeddings*. We tested four types of such embeddings (Section 5.6):

Embeddings from Language Models (ELMo). ELMo embeddings are built by assigning to each token a representation obtained by training a BiLSTM on a large text corpus. Peters et al. [54] have shown that the lower-level LSTM states capture aspects of word syntax, while the higher-level states capture semantics.

Flair Embeddings. These embeddings are built by learning to predict the next character in a sequence of characters. Akbik et al. [3] have showed that in such a way, linguistic concepts such as words, sentences, and even sentiment are automatically internalized. There

is also in improved version of this embeddings called *Pooled Flair Embeddings*.

Bidirectional Encoder Representations from Transformers (BERT) embeddings. BERT is pre-trained on two unsupervised tasks. In the first task, some percentage of the input tokens is masked at random, and then the model is trained to predict the masked tokens. In the second task, given two sentences, the model is trained to predict if one follows the other one in a piece of text. Hence, BERT is a general NLP model [12]. Liu et al. [38] have replicated the BERT model with different parameters and design choices producing an enhanced version called *Robustly Optimized BERT Pretraining Approach (RoBERTa) embeddings*.

Specialized medical BERT embeddings. By pre-training a BERT model on large biomedical corpora instead of general corpora, Lee et al. [36] created BioBERT. Using BioBERT significantly outperformed state-of-the-art on three representative biomedical text mining tasks, including biomedical entity extraction. Alsentzer et al. [4] have pre-trained a BERT model on yet another type of corpora – clinical notes. They have showed that ClinicalBERT outperforms both BERT and BioBERT on the specific tasks, such as hospital readmission prediction based on clinical notes or on discharge summaries.

Given the small sizes of our training datasets (Section 4), we used the contextual embeddings as the input to the Bi-LSTM-CRF architecture without training them any further.

Disease Prediction Module. For each social media post, the second module (the rectangle on the right in Figure 1) then received the entities extracted from the post by the first module (more specifically, their stacked word embeddings) in input, and predicted the most likely disease discussed in the post. Its core component is a classifier based on an ensemble of decision trees with extreme gradient boosting (XGBoost) [16]. XGBoost has been found to consistently provide good performance by iteratively combining the results of a set of classification and regression trees into a single strong learner [8].

4 DATASETS

To study the general applicability of our method, we needed to evaluate it upon a variety of datasets. In addition to the two benchmark datasets from AskaPatient and Twitter (Sections 4.1 and 4.2), we categorized half a million Reddit posts in terms of diseases (Section 4.3), and annotated symptom and drug entities in a subset of those posts (Section 4.4).

4.1 AskaPatient Dataset (CADEC)

CADEC has become the standard corpus for medical entity extraction from social media, and using it made it possible to compare our results with those in previous studies. It consists of over 1K annotated posts from AskaPatient[4] (Table 1). In this forum, patients discuss their own experiences concerning a variety of health-related aspects. In the corpus, mentions of *adverse drug reactions* (6318 entities), *symptoms* (275), *clinical findings* (435), *diseases* (283), and of *drug names* (1800) are annotated [28].

[4]https://www.askapatient.com

Table 1: Statistics for the Reddit Medical Entities (MedRed) dataset in comparison with the AskaPatient (CADEC) [28] and Micromed [27] datasets that was used as a reference for creating MedRed.

	MedRed	CADEC	Micromed
# posts	1977	1321	734
# sent.	8794	7632	1027
# words	147,915	101,486	15,690
time span	Jan-Jun 2017	Jan 2001-Sep 2013	May 2014
comm.	18 subreddits	lipitor, diclofenac	Twitter
entities	4485	9111	757

4.2 Twitter Dataset (Micromed)

Another existing yet less adopted dataset is Micromed. Jimeno-Yepes et al. [27] annotated 1300 tweets in terms of *symptoms* (764 entities), *diseases* (253), and *pharmacologic substances* (233). Since words (including health-related ones) might be figuratively used on social media [40], Micromed also comes with a flag indicating whether each extracted word (e.g., pain) is actually a medical entity (e.g., *'I felt **pain** all over my body'*) or not (e.g., *'He was such a **pain** to deal with'*). Since the dataset makes only the tweet identifiers available (and not the actual tweets), we crawled those identifiers, and, at the time of writing, 734 out of the original 1300 tweets were still available.

4.3 Full Reddit Dataset (Disease Subreddits)

Given the specificity of the AskaPatient dataset (which only covered two drugs) and the limited size of the Twitter dataset, we turned to Reddit for further data collection. Reddit is a popular social platform for news and entertainment where users discuss a variety of topics. According to the official statistics from the site,[5] Reddit has over 330M average monthly active users, over 138K active communities (subreddits), and the majority of its users are from the US, Canada, and the UK. The discussions are grouped into subreddits (subgroups) by subject (e.g., depression). Subreddits consist of discussion threads that are initiated by a user's post, which is generally followed by comments/answers from other users.

Users discuss a wide variety of topics, including health matters, and share their own experiences, ask for advice, and learn from others. There is an entire 'health' category on Reddit, and it contains around 40 subreddits in total. Out of these subreddits, we selected the ones that met two main criteria:

(1) Were active, i.e., those having an average of at least hundred posts per month. This first criterion made it possible to select diseases of general interest.
(2) Focused on a specific disease, such as *r/depression* and *r/kidneyStones*. This second criterion creates an unequivocal relationship between subreddit and disease and, as such, allowed us to implicitly annotate the Reddit posts depending on which subreddits they appeared.

By applying both criteria, we were left with 18 subreddits, from which we then downloaded the posts for the first six months of 2017 (Table 2).

[5]https://www.redditinc.com/press

Table 2: Reddit Dataset (labeled for Diseases): total number of posts in each subreddit during January-June 2017.

subreddit	disease name	posts
r/bpd	Borderline personality disorder	48000
r/cfs	Chronic fatigue syndrome	10711
r/crohnsdisease	Crohn's disease	30774
r/dementia	Dementia	1979
r/depression	Depression	286968
r/diabetes	Diabetes mellitus	54285
r/dysautonomia	Disorder of autonomic nervous system	1655
r/gastroparesis	Gastroparesis syndrome	679
r/hypothyroidism	Hypothyroidism	7990
r/ibs	Irritable bowel syndrome	19497
r/interstitialcystitis	Chronic interstitial cystitis	1851
r/kidneystones	Kidney disease	1301
r/menieres	Menieres disease	613
r/multiplesclerosis	Multiple sclerosis	12996
r/parkinsons	Parkinson's disease	703
r/psoriasis	Psoriasis	5734
r/rheumatoid	Rheumatoid arthritis	3736
r/sleepapnea	Sleep apnea	7486
total		496958

4.4 Medical Entities in Reddit (MedRed) Dataset

For 1980 posts from Reddit (110 randomly sampled posts from each of the 18 subreddits), we set up a Mechanical Turk (MT) experiment to annotate medical entities in them. Since CADEC is the most commonly used benchmark dataset, which is annotated by experts, we used some of its posts it to ensure the quality of our annotations. Each MT task consisted of six posts to be labeled: four were posts randomly selected from the 1980 Reddit posts; one was a *'control post'* carefully sampled from CADEC to resemble a typical Reddit post; and the final one was a manually created *'trap post'* containing exactly a one symptom and a one drug name. The positions of the different types of posts were randomized in each task.

We instructed workers to extract entities of two types: *i)* symptom/disease, and *ii)* drug names. Instructions for what constituted a relevant entity were similar to those by Karimi et al. when creating the CADEC dataset.

To ensure high quality annotations:

(1) A task could be performed only by workers with an approval rate above 95%.
(2) A task result was accepted only if its *trap post's* symptom and drug name were both correctly identified (which discarded 21% of the responses).
(3) Each post was labeled by at least 10 different workers.
(4) Each post was annotated only with the entities that were extracted by at least two workers independently, which was found to be a good agreement number by previous work [30].

Finally, we used the *control post* from CADEC out of the 6 posts in each task to compute the pair-wise agreement of the annotations, as follows:

$$Agr(i,j) = \frac{match(A_i, A_j)}{max(n_{A_i}, n_{A_j})},$$

Table 3: Number of labeled entities in the MedRed dataset.

subreddit	drugs	symptoms	all
r/bpd	6	152	158
r/cfs	33	226	259
r/crohnsdisease	51	134	185
r/dementia	10	184	194
r/depression	11	65	76
r/diabetes	46	93	139
r/dysautonomia	54	333	387
r/gastroparesis	69	251	320
r/hypothyroidism	76	200	276
r/ibs	39	135	174
r/interstitialcystitis	84	252	336
r/kidneystones	55	223	278
r/menieres	43	306	349
r/multiplesclerosis	44	161	205
r/parkinsons	76	259	335
r/psoriasis	93	148	241
r/rheumatoid	143	236	379
r/sleepapnea	41	158	199
corpus	974	3511	4485

where A_i is the list of medical entities extracted by the MT workers, A_j is the list of medical entities extracted by the CADEC experts (control posts), n_{A_i} is the number of medical entities in A_i, n_{A_j} is the number of medical entities in A_j, and $match(A_i, A_j)$ is the number of medical entities that were extracted by both the MT workers and the experts. The average pair-wise agreement in its strict form, i.e., when allowing only for exact matches, was .62 for symptoms, and .75 for drugs, while in its relaxed form, i.e., when allowing entities to overlap (e.g., 'pain' to overlap with 'strong pain'), was .77 for symptoms, and .83 for drug names. These scores are comparable to those previously obtained for CADEC expert annotations [28], hence confirming the quality of the MedRed annotations.

MedRed is a new benchmark dataset for medical entity extraction openly available to other researchers [59][6].

5 EVALUATION

The main goal of our evaluation was to assess whether our method performed competitively on the datasets from diverse sources.

5.1 Evaluation Metrics

Our evaluation metric is $F1$ score, $F1 = 2P \cdot R/(P+R)$, i.e., the harmonic mean of precision P and recall R, where:

$$P = \frac{\text{\#correctly classified medical entities}}{\text{\#total entities classified as being medical}},$$

and

$$R = \frac{\text{\#correctly classified medical entities}}{\text{\#total medical entities}}.$$

To be conservative, we counted as "correctly classified" only the entities that were *exactly* matching the ground truth labels. This means that we used the strict F1-score, as opposed to its relaxed version that is sometimes used when reporting entity extraction results. Also, given that our data comes with class imbalance (i.e.,

[6] http://goodcitylife.org/Humane-AI

text tokens do not correspond equally to symptoms, drugs, or non-medical entities), we corrected for that by computing P and R using micro-averages [6].

5.2 MetaMap

MetaMap is a well-established tool for extracting medical concepts from text using symbolic NLP and computational-linguistic techniques [5], and has become a de-facto baseline method for NLP studies related to health [66]. MetaMap performs entity extraction by following a knowledge-intensive rule-based approach with the UMLS Metathesaurus as its knowledge source. As a result, when processing a sentence, it returns a list of tokens that correspond to the medical entities it finds in the sentence. These entities are either drug names – which we defined as MetaMap's categories[7] of *Antibiotic*, *Clinical Drug* and *Pharmacologic Substance* – or symptoms – which were under *Disease or Syndrome*, *Finding*, and *Sign or Symptom*. Apart from this type of post-processing, we also limited our results to two vocabulary sources: SNOMEDCT_US for symptoms, and RxNorm for drugs.

5.3 TaggerOne

TaggerOne is a machine learning tool using semi-Markov models to jointly perform two tasks: entity extraction and entity normalization. The tool does so using a medical lexicon [33]. However, since we had training data for entity extraction but not for normalization, we could not train TaggerOne on our data. We therefore adopted the version of it previously trained on the biomedical corpora "BioCreative V CDR corpus" [37] as one of our baselines, and we could do so because its extracted medical entities are similar to ours.

5.4 Previous Deep Learning Methods

Deep learning (DL) models have increasingly become the state-of-the-art solution for medical entity extraction. Our approach was evaluated against two existing approaches with best results on the respective datasets:

CADEC DL[67]. Tutubalina and Nikolenko [67] applied BiLSTM-CRF using the specialized word embeddings HealthVec [41] on the AskaPatient Dataset (CADEC). Hence, we refer to their method as CADEC DL.

Micromed DL[73]. Yepes and MacKinlay [73] proposed an LSTM RNN whose outputs were passed to a linear classifier trained using a multiclass hinge loss. We refer to their method as Micromed DL.

Given the unavailability of the source code, we took the best performance results for these two approaches from the corresponding publications [67, 73], ensuring a fair comparative analysis.

5.5 Implementation and Training Setup

For implementation of the BiLSTM-CRF model, we used Python with Flair library [1] and Pytorch backend [49]. For each of the embeddings under consideration, we used their language models from the respective open source repositories. The network was set up with following parameters: 256 hidden units, learning rate starting from 0.1 and being gradually halved every time when there is no improvement after 3 epochs, batch size of 4, and we trained using both the training and development sets. Training was done on a single GeForce GTX 1080 GPU for a maximum of 200 epochs or before the learning rate become too small ($\leq .0001$).

[7] https://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml

Table 4: F1 scores (P/R) for our method when using different embeddings to extract entities from the three datasets.

Embeddings	AskaPatient (CADEC)	Twitter (Micromed)	Reddit (MedRed)
	Individual contextual emb.		
ELMo	.80 (.79/.80)	.69 (.69/.69)	.70 (.69/.71)
Flair	.79 (.79/.78)	.62 (.66/.58)	.69 (.68/.69)
Pooled Flair	.80 (.81/.79)	.63 (.66/.61)	.70 (.77/.64)
BERT	.80 (.79/.82)	**.70 (.66/.75)**	.70 (.74/.66)
RoBERTa	**.81 (.81/.82)**	.67 (.65/.69)	**.73 (.77/.69)**
BioBERT	.80 (.79/.81)	.59 (.52/.70)	.66 (.71/.61)
Clinical BERT	.79 (.80/.78)	.66 (.62/.70)	.64 (.72/.58)
	Combined contextual and word emb.		
RoBERTa + GloVe	**.82 (.81/.82)**	**.72 (.69/.74)**	**.73 (.75/.70)**

5.6 Results

We first tested different versions of our framework that used different contextual embeddings (without any word embedding), and did so on each of the three datasets (the three columns in Table 4). The differences in scores were subtle in the case of AskaPatient, but notable in the case of Twitter and Reddit. Across the three datasets, the specialized embeddings for the medical domain (i.e., BioBERT and Clinical BERT) were outperformed by two of the general embeddings (RoBERTA yielded the best results on AskaPatient and Reddit, while BERT did so on Twitter). That is mainly because specialized embeddings capture formal medical language, while health expressions on social media are of more informal nature. By then stacking RoBERTa with GloVe word embeddings [53], our framework yielded the best performance on all of the three datasets. Therefore, in what follows, we report results for the version that used the combination of RoBERTA and GloVE embeddings (Figure 2).

AskaPatient (CADEC). The AskaPatient (CADEC) dataset was split into train (60%), dev (20%), and test (20%) sets, a split used by previous work [42]. By considering the F1 scores for the AskaPatient (CADEC) dataset (Figure 2), we see that on symptoms our method has the F1 of .78 and, as such, outperformed MetaMap (.19) and TaggerOne (.47), and performed better than CADEC DL (.71), although the latter was restricted to extracting ADRs rather than symptoms or diseases. This latter result suggests that the use of the two types of embeddings - GloVe for words, and RoBERTa for context - made a substantial difference. Finally, in terms of drug extraction, we could compare our method only to MetaMap (which was greatly outperformed), as CADEC DL did not report any result about it (hence the empty slots in Figure 2).

Twitter (Micromed). The Twitter Micromed dataset was split into train (50%), dev (25%), and test (25%) sets (we kept at least 25% of the dataset for validation and testing because it is small).

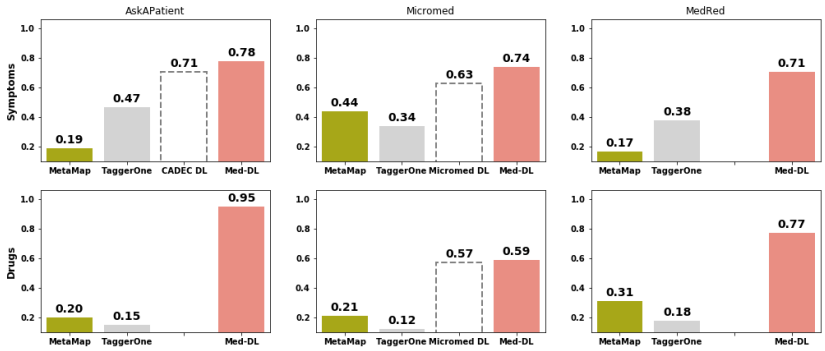

Figure 2: Evaluation of our method versus baselines in extracting medical entities on the three datasets. The F1 scores are shown separately for symptom and drug entities. Dashed lines for Micromed DL and CADEC DL signify that these results are not calculated by us but come from the respective publications ([67],[73]), and empty bars indicate that the results were not available from these publications. For Micromed DL in the case of symptoms, we took the weighted average of its scores reported separately for diseases and symptoms. CADEC DL extracted only ADRs.

By considering the F1 scores for the Twitter (Micromed) dataset (Figure 2), we see that for symptoms our method had the F1 score .74 and, as such, outperformed MetaMap (.44) and TaggerOne (.34), and performed better than MicroMed DL (.63). On drugs, our method's performance is slightly superior to Micromed DL (.59 versus .57). Interestingly, MetaMap performed considerably better on Twitter than on the more specialized platform of AskaPatient. That is partly because the Micromed dataset was originally built by searching tweets for UMLS terms[27], and MetaMap is based on UMLS.

Reddit (MedRed). The MedRed dataset was split into train (50%), dev (25%), and test (25%) sets (again, we kept at least 25% of the dataset for validation and testing because also this dataset is small). By considering the F1 scores on the MedRed dataset (Figure 2), we see that our method has the score of .71 on symptoms, and .77 on drugs, outperforming MetaMap/TaggerOne (with the F1 scores of .17/, 38 and .31/.18, respectively).

6 VALIDATION: PREDICTING DISEASES

The previous results showed that our method performs well on three datasets, each from a different platform. We have to concede, however, that these results are not conclusive as they are produced on sets that are richly labeled but limited in size. Given the availability of the larger set of Reddit posts categorized into 18 diseases (Section 4.3), we turned to a final prediction task: that of predicting a Reddit post's disease from the set of medical entities it contains. We expected that if the entities extracted by our method are accurate, then the classifier would be able to tell apart posts from different categories, as the symptoms and drugs associated with the 18 diseases are different. In practice, predicting the disease only from a few medical entities a post is bound to contain instead of using its entire textual content (as previous work has largely done) has the benefit that the prediction results do not tend to suffer from spurious correlations, making them more robust to exogenous events and potentially more generalizable.

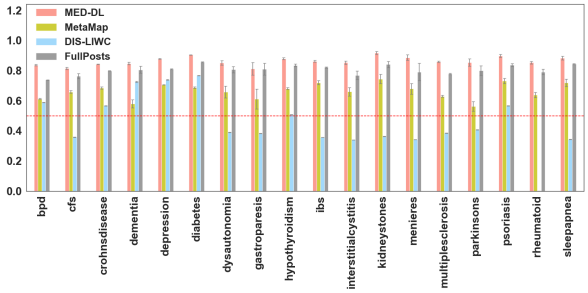

Figure 3: F1 scores for the 18 disease-specific binary classifiers, which predicted the diseases associated with Reddit posts purely based on the posts' medical entities, and also based on the full content of the posts. The error bars show the standard deviations for the 5-fold cross validation. Red line represents random classifier baseline.

6.1 Classification Setup

We selected the best performing among the deep learning models (i.e., using RoBERTa and GloVe).

We used MetaMap, and our selected method to extract medical entities from Reddit posts (generating two separate sets of medical entities), and kept only the posts in which at least one medical entity was found. These posts were then arranged in 18 balanced datasets, one for each of the 18 diseases. Each disease's set contained a subset of positive examples (all the posts related to the disease), and a subset of negative examples (posts randomly sampled from the remaining 17 disease sets).

We then trained 36 binary XGBoost classifiers (with $n = 1000$ estimators and a maximum tree depth of 4): for each of the 18 diseases, we trained three classifiers, one relying on the medical entities extracted by MetaMap, and the other relying on the medical entities extracted by our method.

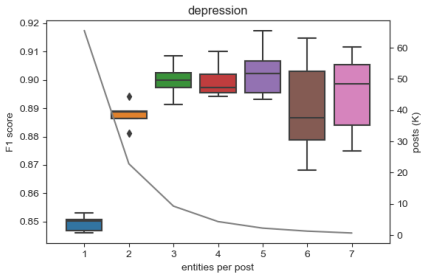

Figure 4: The ability to predict whether a Reddit post is about depression (F1 score, y-axis on the left-hand side) increases with the number of medical entities the post mentions. The gray line indicates the number of posts (number of K posts, y-axis on the right-hand side) containing a given number of entities.

6.2 Classification Metrics

Training and testing was done using a 5-fold cross validation, and the accuracy was measured using $F1$ score. This is the harmonic mean of the classifier's precision P and recall R:

$$P = \frac{\text{\#correctly classified disease posts}}{\text{\#total posts classified as disease-related}},$$

and

$$R = \frac{\text{\#correctly classified disease posts}}{\text{\#total disease-related posts}}.$$

6.3 Dictionary-based Classifier (DIS-LIWC)

Since dictionary-based approaches (e.g., matching words from the LIWC - Linguistic Inquiry and Word Count - dictionary [52]) have been widely used in social media studies and have shown good performance, we added one such an approach as a baseline. We created Dis-LIWC (Disease Linguistic Inquiry and Word Count):[8] a set of 1493 words reflecting symptoms and drug names, and organized under 18 disease categories. We did so by collecting for each disease:

Formal medical expressions for symptoms. We gathered these expressions from over 100K symptom-disease pairs from the Human Disease Network (HDN) dataset [19], which consists of pairs frequently co-mentioned in publications indexed by PubMed.

Colloquial expressions for symptoms. We manually wrote down all the symptoms that appeared in the disease's main description pages on MedScape[9], WebMed[10], and Wikipedia.

Drug names. We crawled drug names and corresponding diseases from the whole DrugBank[11] database, resulting in 100+ names in total.

6.4 Classifier based on Full Posts (FullPosts)

One might wonder to which extent a classifier taking the entire textual content of a post (not only the post's medical entities) as input would be able to predict the post's disease. To ascertain that, we created the FullPosts baseline. This encoded all the posts with less than 512 generic tokens (not necessarily medical ones) into a Roberta+Glove document embedding, and resulted into 18 balanced datasets and binary classifiers in a way similar to the setup in Section 6.1. We expect FullPosts to return accurate predictions (as it works upon a variety of entities, not only medical ones), but we also expect it to be less principled and, as such, less generalizable (e.g., it might be trained to associate the word "soup" with influenza).

6.5 Results

From the results (Figure 3), we see that, on input of the medical entities extracted by our method, one can reliably predict all the 18 diseases: for all of them, F1 scores were higher than .80. These scores are even slightly higher than the FullPosts' [38], which used the full text contained in the Reddit posts. On input of the medical entities extracted by MetaMap, on the other hand, one can still predict the majority of diseases, yet entities by our method are always associated with a 15% up to 20% increase in prediction accuracy. By contrast, on input of the entities extracted by Dis-LIWC, most diseases cannot be identified. All these results suggest that:

(1) The expected behavior of Dis-LIWC is to extract all entities matching its dictionary content. However, such extractions turned out to have less predictive power than the ones from our method, which are not based on a specialized vocabulary.
(2) A high level of accuracy is required for predicting diseases, as the prediction was done at the post level, likely from just a few medical entities. That is because, as one expects, the ability to predict a post's disease increases with the number of entities found in the post (Figure 4).

6.6 Factors Affecting Disease Prediction

F1 scores vary across diseases (Figure 3): there is nearly a 10% difference between the highest scoring disease (kidney stones) and the lowest scoring one (gastroparesis). One might now wonder which factors explain that variability.

Size of training data. One might expect that the more training data for a disease, the easier it is to predict. We correlated a disease's F1 with the logarithm of the number of posts associated with it. There is no correlation ($r = 0.006$, $p > .98$). Indeed, even for the disease with the lowest number of posts (r/menieres with 613 posts), we have an F1 score of .89.

Input size from 1 to a set of n posts. To assess the impact of the size of the input on disease prediction, we also trained 7 different versions of the 18 binary disease classifiers with increasing sizes of the input unit, from $n = 1$ to $n = 7$ posts. The n posts in each training/test unit were randomly sampled without replacement. It is indeed easier to predict a disease on input of 2 posts rather than 1 (left panel in Figure 5), and with only $n = 4$ posts, all the classifiers exceeded the $F1$ score of .90.

Number of symptoms/drugs for each disease. Our classification is done on input of medical entities, including symptoms and drug names. One might expect that the higher the number of drug names or symptoms associated with a disease, the more uniquely classifiable a disease. To test that, we took the number of symptoms and the number of drug names associated with each

[8] Dis-LIWC is publicly available under http://goodcitylife.org/Humane-AI
[9] https://www.medscape.com
[10] https://www.webmd.com
[11] https://www.drugbank.ca

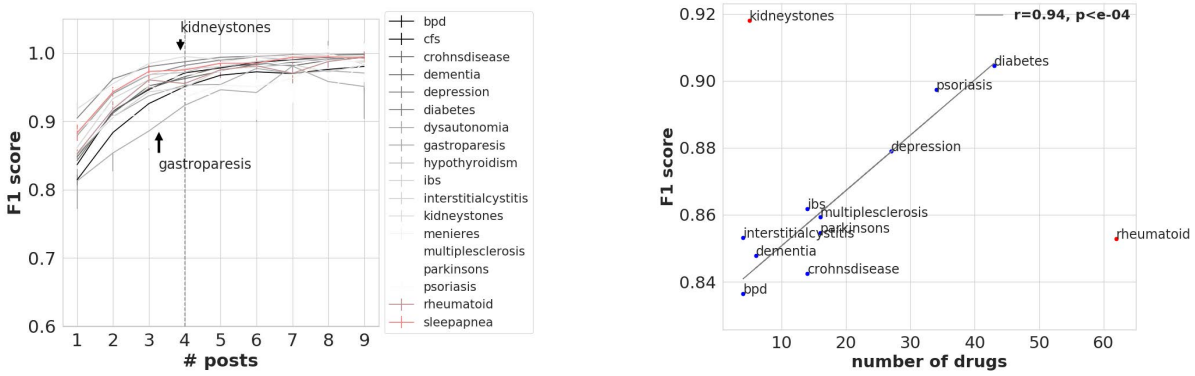

Figure 5: The ability to predict diseases (F1 scores) depending on: *(left)* the number of posts as a unit of input; and *(right)* the extent to which a disease is associated with a large number of drugs.

disease (from our DIS-LIWC in Section 6.3), and computed their correlations with each disease's $F1$ score. As we hypothesized, the higher the number of symptoms a disease is associated with, the easier to predict ($r = .59$, $p < 0.01$); in a similar way, the higher the number of drug names a disease is associated with[12], the easier to predict (right panel in Figure 5), with a correlation as high as $r = .94$ ($p < .0001$). The corollary of this result is that common diseases (which tend to be treated by a large number of drugs)[13] are easier to be detected on social media than less common ones. More interestingly, there are two exceptions to our hypothesis (the two red dots in the right panel of Figure 5):

(1) Rheumatoid arthritis is treated by many drugs but is relatively difficult to classify. This condition is known to be difficult to diagnose because its symptoms are associated with a variety of other diseases.[14]
(2) Kidney stones is treated by only a few drugs but is relatively easy to classify. That is because it pertains to one body organ, and comes with well-defined symptoms and specific drugs. This sets kidney stones apart from any other condition.

Shared symptoms between diseases. To test which diseases are harder to tell apart from each other, in addition to having a binary classifier for each of the 18 diseases, we had a unique multi-class classifier for the 18 diseases. To create a balanced dataset, for each subreddit, we randomly took the number of posts equal to that of the smallest subreddit. The confusion matrix in Figure 6 reports the F1 score for a disease on each row (the random baseline in the multi-class case has an F1 score of .06), its diagonal corresponds to when disease i is correctly predicted, and its (i, j) element reports the number of disease j posts wrongly labeled as i. Consider the Crohn's (inflammatory bowel) disease, which has the lowest F1 score of .44, which is still far higher than the baseline's score of .06. Most of the posts wrongly labeled as being Crohn's were instead either about:

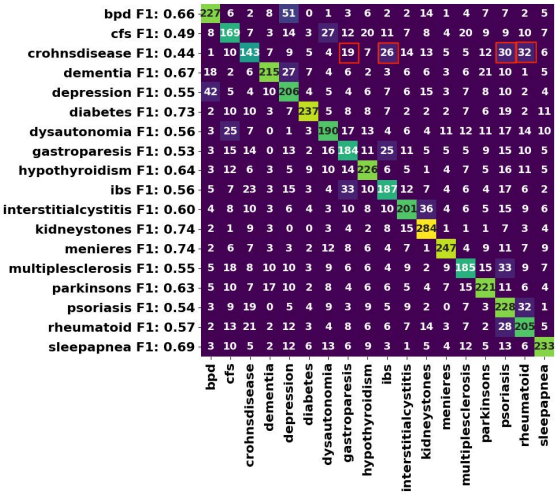

Figure 6: Confusion matrix for the multi-class classifier predicting the 18 diseases associated with Reddit posts purely based on the medical entities extracted by our method.

two other bowel diseases - "inflammatory bowel syndrome" (26 posts) or "gastroparesis" (19 posts); a disease it shares fundamental biological mechanisms with called "psoriasis" (30 posts) [15]; or "rheumatoid arthritis" (32 posts), which is a complication of Crohn's outside the digestive tract. Given this result, we hypothesized that two diseases are hard to tell apart from each other, if they tend to share symptoms. To test that, we computed the number of shared symptoms for each disease pair $(D1, D2)$: $J_{D1,D2} = \frac{S_{D1} \cap S_{D2}}{S_{D1} \cup S_{D2}}$, where (S_{D1}, S_{D2}) are the symptom sets for the two diseases (computed from the Dis-LIWC symptom list in Section 6.3), and J is the Jaccard index (similarity index) of the two sets. As we hypothesized, there is a statistically significant and positive correlation ($r = .31$) between

[12]When associating drug names with diseases in Section 6.3, we could find drug names for 12 diseases out of the 18 and, as such, Figure 5*(right)* will show the results for these 12 diseases.
[13]Pharmaceutical companies target profitable diseases https://www.focusforhealth.org/big-pharma-creates-diseases-medications-big-business
[14]https://www.nhs.uk/conditions/rheumatoid-arthritis/diagnosis

J (similarity between two diseases in terms of the number of shared symptoms) and misclassification.

7 DISCUSSION

Applying medical entity extraction to social media is more challenging than applying it to electronic health records. This is due to the unconventional ways in which users express themselves, possible misspellings, and use of internet slang [63]. The demonstrated deep learning's ability to extract symptoms in this social media context, which is arguably more abundant in data and also more accessible, has a variety of theoretical and practical implications. Before being able to realize those implications though, researchers have to tackle two main limitations.

Medical annotations are costly. Machine learning algorithms need data and, in the particular case of health applications, it is costly and hard to generate annotations: employing expert annotators is costly, and crowd workers might not be the best candidates for highly-specialized medical annotations. New crowdsourcing approaches that minimize annotation costs without compromising quality are needed. To this end, the extent to which gamification techniques could be introduced in the healthcare sector should be explored in the future [13].

The quirkiness of social media. There is still plenty of room for improving accuracy. To begin with, our results suggest that figurative use of language is present on social media platforms, and our method should be further improved to deal with that. Also, this work has not dealt with spam or malicious content, and the integration of techniques that filter content based, for example, on topical coherence or on trustworthiness of content creators might well enhance performance [64].

Theoretical implications. We have identified immediate theoretical implications in two medical fields:

(1) *Phenotypic human disease networks.* In this type of networks, nodes are diseases, and link weights reflect the extent to which the corresponding disease pairs share the same symptoms. New ways of extracting symptoms from social media such as the one presented here would result in the creation of new phenotypic human disease networks, perhaps opening up a new field which could be called '*social* phenotypic disease networks'. With multiple phenotypic networks and the methodologies developed by the 'complex networks' research community at hand, researchers might be able to take a 'fresh look' at the complex interplay of symptoms and diseases.
(2) *Genetic studies.* Such studies have been increasingly able to characterize diseases with common genetic associations. Further characterization can be based on the symptoms shared by diseases. Two recent papers achieved this by mining abstracts of medical research publications [25, 75]. In so doing, they discovered that indeed *"symptom-based similarity of diseases correlates strongly with the number of shared genetic associations and the extent to which their associated proteins interact"*. The deep learning's ability of extracting symptoms from social media might further contribute to genetic studies.

Practical Implications. Our method was evaluated for 18 diseases but it could be applied to any disease. That is because the specific type of embeddings makes it possible to recognize mentions that were not present in the training data. To see how, let us name a few examples out of the many we encountered: '*medium-frequency sensorineural hearing loss', 'thumping or heartbeat like sound in my ears', 'sugars aren't dropping', 'dramatic swings in his blood glucose', 'shoulder was too high on my desk until my arm started going numb'*, but also the abbreviations, such as *'DKA'* (for diabetic ketoacidosis), and *'OCD'* for (obsessive-compulsive disorder). Given its generalizability, our method could be used for:

(1) *Pharmacovigilance.* Pharmacovigilance requires the ability to identify a specific set of symptoms called 'Adverse Drug Reactions' (ADRs). Since our method is able to extract symptoms at the fine-grained level of post, it might well be used for pharmacovigilance, and its application to social media might well result in the discovery of unknown ADRs.
(2) *Tracking diseases in time and space.* Considerable research has shown that social media can be used to track diseases over time and across geographies by simply mining posts which come with locations and timestamps [47, 57]. Here we have proposed a new way of identifying diseases – one that is based on extracting symptoms and, upon them, predicting the disease being discussed. This way of identifying diseases is principled: as opposed to existing approaches, by constraining the extracted words to medical entities, it does not suffer from spurious associations between non-medical words and diseases. This was the main reason Google terminated its 'flu trends' project, in which flu outbreaks were tracked from people's searches – by considering any word useful for prediction, the system ended up failing to be robust to exogenous events [32].
(3) *Drug re-purposing.* By mining symptoms from social media posts that mention specific drug names (e.g., from drug review forums), pharmaceutical companies might discover potential candidates for what in the industry is called 'drug re-purposing', that is, determining which additional diseases/symptoms could be treated by drugs that are currently prescribed for other conditions.

8 CONCLUSION

We presented a widely applicable deep learning framework for reliably extracting medical entities such as symptoms and drug names, and for accurately predicting diseases purely from the extracted medical entities. By evaluating it on three datasets originating from AskaPatient, Twitter, and Reddit, we showed that it consistently outperformed baseline and state-of-the-art approaches, showing generalizable results. In the future, more research should go into: new data collection efforts, including the design of novel crowdsourcing solutions tailored to the highly-specialized health domain; text mining techniques that are able to deal with figurative uses of language; and social media mining techniques that are able to filter malicious and inaccurate content in robust ways. All this work might well enable large-scale health tracking applications, lead to the construction of new phenotypic human disease networks, and even impact genetic studies.

REFERENCES

[1] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*. 54–59.

[2] Alan Akbik, Tanja Bergmann, and Roland Vollgraf. 2019. Pooled contextualized embeddings for Named Entity Recognition. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*. 724–728.

[3] Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the International Conference on Computational Linguistics of the Association for Computational Linguistics*. 1638–1649.

[4] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the Clinical Natural Language Processing Workshop*. 72–78.

[5] Alan R Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association* 17, 3 (2010), 229–236.

[6] Vincent Van Asch. 2013. Macro-and micro-averaged evaluation measures. *Technical Report* (2013).

[7] Duilio Balsamo, Paolo Bajardi, and André Panisson. 2019. Firsthand Opiates Abuse on Social Media: Monitoring Geospatial Patterns of Interest Through a Digital Cohort. In *Proceedings of of the ACM World Wide Web Conference*. 2572–2579.

[8] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 785–794.

[9] Munmun De Choudhury and Sushovan De. 2014. Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*.

[10] Aaron M Cohen and William R Hersh. 2005. A survey of current work in biomedical text mining. *Briefings in bioinformatics* 6, 1 (2005), 57–71.

[11] Kerstin Denecke. 2014. Extracting medical concepts from medical social media with clinical NLP tools: a qualitative study. In *Proceedings of the Workshop on Building and Evaluation Resources for Health and Biomedical Text Processing*.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*. 4171–4186.

[13] Emilia Duarte, Pedro Pereira, Francisco Rebelo, and Paulo Noriega. 2014. A Review of Gamification for Health-Related Contexts. In *Proceedings of the 3rd International Conference on Design, User Experience, and Usability*. 742–753.

[14] Tara Fenwick. 2014. Social media and medical professionalism: rethinking the debate and the way forward. *Academic Medicine* 89, 10 (2014), 1331–1334.

[15] Gionata Fiorino and Paolo D Omodei. 2015. Psoriasis and Inflammatory Bowel Disease: Two Sides of the Same Coin? *Journal of Crohn's & Colitis* 9, 9 (2015), 697–698.

[16] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* (2001), 1189–1232.

[17] Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. 2019. Knowledge-aware Assessment of Severity of Suicide Risk for Early Intervention. In *Proceedings of the ACM World Wide Web Conference*. 514–525.

[18] George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim JP Hubbard, Richard JB Dobson, and Rina Dutta. 2017. Characterisation of mental health conditions in social media using Informed Deep Learning. *Scientific Reports* 7 (2017), 45141.

[19] Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. 2007. The human disease network. *Proceedings of the National Academy of Sciences* 104, 21 (2007), 8685–8690.

[20] G Gonzalez-Hernandez, A Sarker, K O'Connor, and G Savova. 2017. Capturing the Patient's Perspective: a Review of Advances in Natural Language Processing of Health-Related Text. *Yearbook of Medical Informatics* 26, 01 (2017), 214–227.

[21] Frances Griffiths, Jonathan Cave, Felicity Boardman, Justin Ren, Teresa Pawlikowska, Robin Ball, Aileen Clarke, and Alan Cohen. 2012. Social networks–the future for health care delivery. *Social Science & Medicine* 75, 12 (2012), 2233–2241.

[22] Carleen Hawn. 2009. Take two aspirin and tweet me in the morning: how Twitter, Facebook, and other social media are reshaping health care. *Health affairs* 28, 2 (2009), 361–368.

[23] Matthew Herland, Taghi M Khoshgoftaar, and Randall Wald. 2014. A review of data mining using big data in health informatics. *Journal of Big Data* 1, 1 (2014), 2.

[24] William R Hersh. 2002. Medical informatics: improving health care through information. *Jama* 288, 16 (2002), 1955–1958.

[25] Robert Hoehndorf, Paul N Schofield, and Georgios V Gkoutos. 2015. Analysis of the human diseasome using phenotype similarity between common, genetic, and infectious diseases. *Scientific Reports* 5 (2015), 10888.

[26] Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991* (2015).

[27] Antonio Jimeno-Yepes, Andrew MacKinlay, Bo Han, and Qiang Chen. 2015. Identifying Diseases, Drugs, and Symptoms in Twitter. *Studies in Health Technology and Informatics* 216 (2015), 643.

[28] Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics* 55 (2015), 73–81.

[29] Payam Karisani and Eugene Agichtein. 2018. Did you really just have a heart attack?: towards robust detection of personal health mentions in social media. In *Proceedings of the ACM World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 137–146.

[30] Nolan Lawson, Kevin Eustice, Mike Perkowitz, and Meliha Yetisgen-Yildiz. 2010. Annotating Large Email Datasets for Named Entity Recognition with Mechanical Turk. In *Proceedings of the ACM NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. 71–79.

[31] Allison J Lazard, Emily Scheinfeld, Jay M Bernhardt, Gary B Wilcox, and Melissa Suran. 2015. Detecting themes of public concern: a text mining analysis of the Centers for Disease Control and Prevention's Ebola live Twitter chat. *American Journal of Infection Control* 43, 10 (2015), 1109–1111.

[32] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The parable of Google Flu: traps in big data analysis. *Science* 343, 6176 (2014), 1203–1205.

[33] Robert Leaman and Zhiyong Lu. 2016. TaggerOne: joint named entity recognition and normalization with semi-Markov Models. *Bioinformatics* 32, 18 (2016), 2839–2846.

[34] Robert Leaman, Laura Wojtulewicz, Ryan Sullivan, Annie Skariah, Jian Yang, and Graciela Gonzalez. 2010. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. In *Proceedings of the Worshop on Biomedical Natural Language Processing*. Association for Computational Linguistics, 117–125.

[35] Hsin-Chun Lee, Yi-Yu Hsu, and Hung-Yu Kao. 2015. An enhanced CRF-based system for disease name entity recognition and normalization on BioCreative V DNER Task. In *Proceedings of the BioCreative Challenge Evaluation Workshop*. 226–233.

[36] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* (2019).

[37] Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database* 2016 (2016).

[38] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* (2019).

[39] Yingjie Lu, Yang Wu, Jingfang Liu, Jia Li, and Pengzhu Zhang. 2017. Understanding health care social media use from different stakeholder perspectives: a content analysis of an online health community. *Journal of medical Internet research* 19, 4 (2017).

[40] Andrew MacKinlay, Antonio Jimeno Yepes, and Bo Han. 2015. Identification and Analysis of Medical Entity Co-occurrences in Twitter. In *Proceedings of the ACM International Workshop on Data and Text Mining in Biomedical Informatics*. 22–22.

[41] Zulfat Miftahutdinov, Elena Tutubalina, and Alexander Tropsha. 2017. Identifying Disease-related Expressions in Reviews using Conditional Random Fields. In *Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog*, Vol. 1. 155–167.

[42] Zulfat Miftahutdinov, Elena Tutubalina, and Alexander Tropsha. 2017. Identifying Disease-related Expressions in Reviews using Conditional Random Fields. In *Proceedings of International Conference Dialog*, Vol. 1. 155–167.

[43] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *In Proceedings of the Advances in neural information processing systems conference*. 3111–3119.

[44] S Anne Moorhead, Diane E Hazlett, Laura Harrison, Jennifer K Carroll, Anthea Irwin, and Ciska Hoving. 2013. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. *Journal of medical Internet research* 15, 4 (2013).

[45] Ramona Nelson and Nancy Staggers. 2016. *Health informatics: An interprofessional approach*. Elsevier Health Sciences.

[46] Azadeh Nikfarjam, Abeed Sarker, Karen O'Connor, Rachel Ginn, and Graciela Gonzalez. 2015. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association* 22, 3 (2015), 671–681.

[47] Albert Park and Mike Conway. 2017. Tracking Health Related Discussions on Reddit for Public Health Applications. In *AMIA Annual Symposium Proceedings*,

Vol. 2017. American Medical Informatics Association, 1362.
[48] Albert Park, Mike Conway, and Annie T Chen. 2018. Examining thematic similarity, difference, and membership in three online mental health communities from Reddit: a text mining and visualization approach. *Computers in human behavior* 78 (2018), 98–112.
[49] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic Differentiation in PyTorch. In *Proceedings of the Advances in Neural Information Processing Systems Autodiff Workshop*.
[50] Michael J Paul and Mark Dredze. 2011. You are what you Tweet: Analyzing Twitter for public health. *Proceedings of the International AAAI Conference on Web and Social Media* 20 (2011), 265–272.
[51] Michael J Paul, Abeed Sarker, John S Brownstein, Azadeh Nikfarjam, Matthew Scotch, Karen L Smith, and Graciela Gonzalez. 2016. Social media mining for public health monitoring and surveillance. In *Biocomputing: Proceedings of the Pacific symposium*. 468–479.
[52] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71 (2001).
[53] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 1532–1543.
[54] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 2227–2237.
[55] Lance A Ramshaw and Mitchell P Marcus. 1995. Text chunking using transformation-based learning. CoRR. *arXiv preprint cmp-lg/9505040* 50 (1995).
[56] Frederic G Reamer. 2015. Clinical social work in a digital environment: Ethical and risk-management challenges. *Clinical Social Work Journal* 43, 2 (2015), 120–132.
[57] Abeed Sarker, Karen O'Connor, Rachel Ginn, Matthew Scotch, Karen Smith, Dan Malone, and Graciela Gonzalez. 2016. Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from Twitter. *Drug safety* 39, 3 (2016), 231–240.
[58] Daniel Scanfeld, Vanessa Scanfeld, and Elaine L Larson. 2010. Dissemination of health information through social networks: Twitter and antibiotics. *American Journal of infection control* 38, 3 (2010), 182–188.
[59] Sanja Scepanovic, Enrique Martin-Lopez, and Daniele Quercia. 2020. MedRed. https://doi.org/10.7910/DVN/8YVINU
[60] Wendy Sinclair, Moira McLoughlin, and Tony Warne. 2015. To Twitter to woo: Harnessing the power of social media (some) in nurse education to enhance the student's experience. *Nurse education in practice* 15, 6 (2015), 507–511.
[61] Gabriel Stanovsky, Daniel Gruhl, and Pablo Mendes. 2017. Recognizing mentions of adverse drug reaction in social media using knowledge-infused recurrent models. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, Vol. 1. 142–151.
[62] Jana Straková, Milan Straka, and Jan Hajic. 2019. Neural Architectures for Nested NER through Linearization. In *Proceedings of the Conference of the Association for Computational Linguistics*. 5326–5331.
[63] Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine de Marneffe, and Wei Xu. 2016. Results of the WNUT16 Named Entity Recognition shared task. In *Proceedings of the ACM Workshop on Noisy User-generated Text*. 138–144.
[64] Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. 2010. Detecting Spammers on Social Networks. In *Proceedings of the ACM Annual Computer Security Applications Conference*. 1–9.
[65] Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In *Proceedings of the ACM Conference on Natural language learning at HLT-NAACL*. 142–147.
[66] Elena Tutubalina, Zulfat Miftahutdinov, Sergey Nikolenko, and Valentin Malykh. 2018. Medical Concept Normalization in Social Media Posts with Recurrent Neural Networks. *Journal of Biomedical Informatics* (2018).
[67] Elena Tutubalina and Sergey Nikolenko. 2017. Combination of deep recurrent neural networks and conditional random fields for extracting adverse drug reactions from user reviews. *Journal of Healthcare Engineering* 2017 (2017).
[68] C Lee Ventola. 2014. Social media and health care professionals: benefits, risks, and best practices. *Pharmacy and Therapeutics* 39, 7 (2014), 491.
[69] Matthew T Wiley, Canghong Jin, Vagelis Hristidis, and Kevin M Esterling. 2014. Pharmaceutical drugs chatter on online social networks. *Journal of biomedical informatics* 49 (2014), 245–254.
[70] Long Xia, G Alan Wang, and Weiguo Fan. 2017. A Deep Learning Based Named Entity Recognition Approach for Adverse Drug Events Identification and Extraction in Health Social Media. In *Proceedings of the Springer International Conference on Smart Health*. 237–248.
[71] Christopher C Yang, Haodong Yang, Ling Jiang, and Mi Zhang. 2012. Social media mining for drug safety signal detection. In *Proceedings of the ACM International workshop on Smart health and wellbeing*. 33–40.
[72] Andrew Yates, Nazli Goharian, and Ophir Frieder. 2015. Extracting Adverse Drug Reactions from Social Media. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 15. 2460–2467.
[73] Antonio Jimeno Yepes and Andrew MacKinlay. 2016. NER for medical entities in Twitter using sequence to sequence neural networks. In *Proceedings of the Australasian Language Technology Association Workshop*. 138–142.
[74] Antonio Jimeno Yepes, Andrew MacKinlay, and Bo Han. 2015. Investigating public health surveillance using twitter. In *Proceedings of Biomedical Natural Language Processing Workshop*. 164–170.
[75] XueZhong Zhou, Jörg Menche, Albert-László Barabási, and Amitabh Sharma. 2014. Human symptoms–disease network. *Nature communications* 5 (2014), 4212.