RedHOT: A Corpus of Annotated Medical Questions, Experiences, and Claims on Social Media

Somin Wadhwa[†] **Vivek Khetan**[◇] **Silvio Amir**[†] **Byron C. Wallace**[†]
Northeastern University[†] Accenture AI Labs[◇]
{wadhwa.s,s.amir,b.wallace}@northeastern.edu
vivek.a.khetan@accenture.com

Abstract

We present **Red**dit **H**ealth **O**nline **T**alk (**RedHOT**), a corpus of 22,000 richly annotated social media posts from Reddit spanning 24 health conditions. Annotations include demarcations of spans corresponding to medical claims, personal experiences, and questions. We collect additional granular annotations on identified claims. Specifically, we mark snippets that describe patient **P**opulations, **I**nterventions, and **O**utcomes (PIO elements) within these. Using this corpus, we introduce the task of retrieving trustworthy evidence relevant to a given claim made on social media. We propose a new method to automatically derive (noisy) supervision for this task which we use to train a dense retrieval model; this outperforms baseline models. Manual evaluation of retrieval results performed by medical doctors indicate that while our system performance is promising, there is considerable room for improvement. We release all annotations collected (and scripts to assemble the dataset), and all code necessary to reproduce the results in this paper at: https://sominw.com/redhot.

1 Introduction

Social media platforms such as Reddit provide individuals places to discuss (potentially rare) medical conditions that affect them. This allows people to communicate with others who share in their condition, exchanging information about symptom trajectories, personal experiences, and treatment options. Such communities can provide support (Biyani et al., 2014) and access to information about rare conditions which may otherwise be difficult to find (Glenn, 2015).

However, the largely unvetted nature of social media platforms make them vulnerable to *mis* and *disinformation* (Swire-Thompson and Lazer, 2019). An illustrative and timely example is the idea that consuming bleach might be a viable treatment for COVID-19,[1] which quickly gained traction on social media. All misinformation can be dangerous, but *medical* misinformation poses unique risks to public health, especially as individuals increasingly turn to social media to inform personal health decisions (Nobles et al., 2018; Barua et al., 2020).

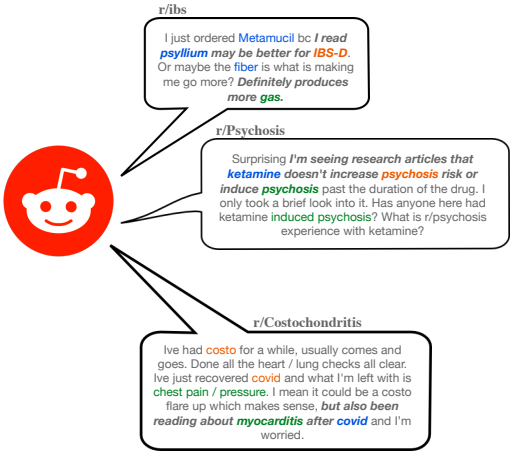

Figure 1: Examples of health-related Reddit posts annotated for populations, interventions, and outcomes.

In this paper, we introduce **RedHOT**: an annotated dataset of health-related claims, questions, and personal experiences posted to Reddit. This dataset can support development of a wide range of models for processing health-related posts from social media. Unlike existing health-related social media corpora, **RedHOT**: (a) Covers a broad range of health topics (e.g., not just COVID-19), and, (b) Comprises "natural" claims collected from real health-related fora (along with annotated questions and personal experiences). Furthermore, we have collected granular annotations on claims, demarcating descriptions of the Population (e.g., diabetics), Interventions, and Outcomes, i.e., the *PIO* elements (Richardson et al., 1995). Such annotations may permit useful downstream processing: For exam-

[1] https://www.theguardian.com/world/2020/sep/19/bleach-miracle-cure-amazon-covid

Findings of the Association for Computational Linguistics: EACL 2023, pages 809–827
May 2-6, 2023

ple, in this work we use them to facilitate retrieval of evidence relevant to a claim.

Specifically, we develop and evaluate a pipeline to automatically identify and contextualize health-related claims on social media, as we anticipate that such a tool might be useful for moderators keen to keep their communities free of potentially harmful misinformation. With this use-case in mind, we propose methods for automatically retrieving *trustworthy* published scientific evidence relevant to a given claim made on social media, which may in aggregate support or debunk a particular claim.

The contributions of this work are summarized as follows. First, we introduce **RedHOT**: A new dataset comprising 22,000 health-related Reddit posts across 24 medical conditions annotated for claims, questions, and personal experiences. Claims are additionally annotated with PIO elements. Second, we introduce the task of identifying health-related claims on social media, extracting the associated PIO elements, and then retrieving relevant and trustworthy evidence to support or refute such claims. Third, we propose **RedHOT**-DER, a Dense Evidence Retriever trained with heuristically derived supervision to retrieve medical literature relevant to health-related claims made on social media. We evaluate baseline models for the first two steps on the **RedHOT** dataset and assess the retrieval step with relevance judgments collected from domain experts (medical doctors).

The Reddit posts we have collected are public and typically made under anonymous pseudonyms, but nonetheless these are health-related comments and so inherently sensitive. To respect this, we (a) notified all users in the dataset of their (potential) inclusion in this corpus, and provided opportunity to opt-out, and, (b) we do not release the data directly, but rather a script to download annotated comments, so that individuals may choose to remove their comments in the future. Furthermore, we consulted with our Institutional Review Board (IRB) and confirmed that the initial collection and annotation of such data does not constitute human subjects research. However, EACL reviewers rightly pointed out that certain *uses* of this data may be sensitive. Therefore, to access the collected dataset we require researchers to self-attest that they have obtained prior *approval* from their own IRB regarding their intended use of the corpus.

2 The RedHOT Dataset

We have collected and manually annotated health related posts from Reddit to support development of language technologies which might, e.g., flag potentially problematic claims for moderation. Reddit is a social media platform that allows users to create their own communities (*subreddits*) focused on specific topics. Subreddits are often about niche topics, and this permits in-depth discussion catering to a long tail of interests and experiences. Notably, subreddits exist for most common (and many rare) medical conditions; we can therefore sample posts from such communities for annotation.

2.1 Data Annotation

We decomposed data annotation into two stages, performed in sequence. In the first, workers are asked to demarcate spans of text corresponding to a `Claim`, `Personal Experience`, or `Question`. We characterize these classes as follows (we provide detailed annotation instructions in Appendix A):

`Claim` suggests (explicitly or implicitly) a causal relationship between an Intervention and an Outcome (e.g., " *I* completely cured my *O*"). Operationally, we are interested in identifying statements that might reasonably be interpreted by the reader as implying a causal link between an intervention and outcome, as this may in turn influence their perception regarding the efficacy of an intervention for a particular condition and/or outcome (i.e., relationship between an *I* and *O*).

`Question` poses a direct question, e.g., "Is this normal?"; "Should I increase my dosage?".

`Personal Experience` describes an individual's experience, for instance the trajectory of their condition, or experiences with specific interventions.

This is a *multi-label* scheme: Spans can (and often do) belong to more than one of the above categories. For example, personal experiences can often be read as implying a causal relationship. Consider this example: "My doctor put me on *I* for my *P*, and I am no longer experiencing *O*". This describes an individual treatment history, but could also be read as implying that *I* is a viable treatment for *P* (and specifically for the outcome *O*). Therefore, we would mark this as both a `Claim` and a `Personal Experience`. By contrast, a general statement asserting a causal relationship outside of any personal context like "*I* can cure *O*" is what

Reddit post	Span labels	PIO elements from claims
I've seen a bunch of posts on here from people who say that glycopyrrolate suddenly isn't working anymore for hyperhidrosis. I'm one of those person who has been facing this for a while now. Just wondering if anyone fixed it? Can't really ask my GP about it since he didn't even know the meds existed. He just prescribed them for me when I asked for it	`Claim:` I've seen a bunch of posts on here from people who say that glycopyrrolate suddenly isn't working anymore for Hyperhidrosis `Question:` Just wondering if anyone fixed it?	**P** hyperhidrosis **I** glycopyrrolate
so i recently read that adderall can trigger a psychotic break & i was prescribed adderall years ago for my adhd but now i just have constant hallucination episodes. anyone else experience adderall induced psychosis?	`Claim:` so i recently read that adderall can trigger a psychotic break `Personal Experience:` i was prescribed adderall years ago for my adhd but now i just have constant hallucination episodes `Question:` anyone else experience adderall induced psychosis?	**P** adhd **I** adderall **O** hallucinations
I've had costochondritis for a while, usually comes and goes. Done all the heart/lung checks all clear. I've just recovered covid and what I'm left with is chest pain/pressure. I mean it could be a costo flare up which makes sense, but also *been reading about myocarditis after covid* and I'm worried, how can I tell which is which?	`Claim:` been reading about myocarditis after covid `Personal Experience:` I'm left with is chest pain/pressure `Question:` how can I tell which is which?	**P** costochondritis **I** covid **O** myocarditis, chest-pain

Table 1: Example annotations, which include: extracted spans (phase 1), and spans describing Populations, Interventions, and Outcomes — PIO elements — within them (phase 2). We collect the latter only for claims.

we will refer to as a "pure claim", meaning it exclusively belongs to the `Claim` category.

In the second stage, workers are asked to further annotate "pure claim" instances by marking spans within them that correspond to the Populations, Interventions/Comparators,[2] Outcomes (the PIO elements) associated with the claim.

2.2 Crowdsourcing Annotations

We hired crowdworkers to perform the above annotation tasks on Amazon Mechanical Turk (AMT).[3] To estimate required annotation time and determine fair pay rates, we ran an internal pilot with two PhD students (both broadly familiar with this research area) on 100 samples.[4] To gauge quality and recruit workers from AMT, we ran two pilot experiments in which we collected sentence-level annotations on posts sampled from three medical populations (i.e., subreddits), comprising ∼6,000 posts in all.

We required all workers have an overall job approval rate of ≥90%. Based on an initial set of AMT annotations we re-hired only workers who reliably followed annotation instructions (details in Appendix A), and we actively recruited the top workers to continue on with increased pay. We obtained annotations from at least three workers for each post, allowing for robust inference of reference labels. Recruited workers were also paid periodic bonuses (equivalent to two hours of pay) based on the quality of their annotated samples.

	Fliess κ	P	R	F1
Questions	0.86	0.85	0.82	0.84
Claims	0.69	0.63	0.53	0.58
Experiences	0.71	0.78	0.69	0.73
POP	0.92	0.94	0.91	0.92
INT	0.74	0.76	0.70	0.73
OUT	0.78	0.73	0.68	0.70

Table 2: Token-wise label agreement among experts measured by Fleiss κ on a subset of data. We further compute precision, recall, and F1 scores for "aggregated" labels by evaluating them against unioned "in-house" expert labels.

2.3 Quality Validation

To evaluate annotation quality we calculate token-wise label agreement between annotators, and amongst ourselves. We emphasize here that token-level κ for sequences is quite strict and disagreements often reflect *where* annotators decide to mark

[2]This is the standard PICO framework, but we collapse Interventions and Comparators into the Intervention category, as the distinction is arbitrary.

[3]We consulted with an Institutional Review Board (IRB) to confirm that this annotation work did not constitute human subjects research.

[4]Based on the estimate from our pilot experiments, payrate for AMT workers was fixed to US $9 per hour for stage-1 annotations and US $11 per hour for stage-2 annotations, irrespective of geographic location.

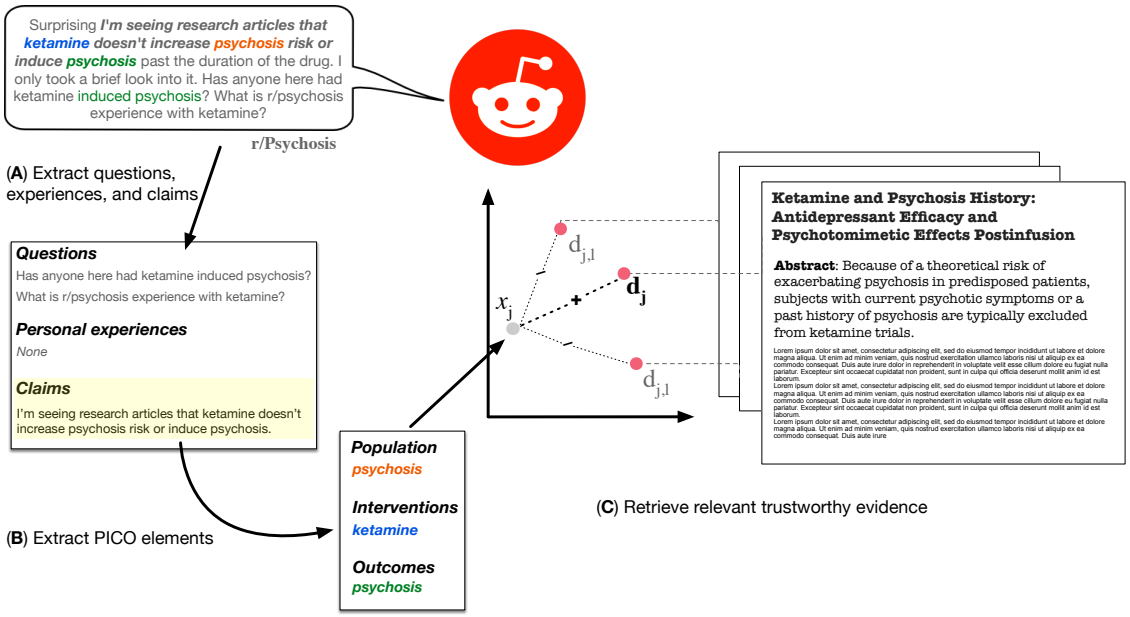

Figure 2: Examples portraying potential use cases of our corpus. We showcase three distinct tasks, to be performed in sequence. The first (A) entails extracting spans corresponding to claims (highlighted in **bold**) from a given Reddit post. The second step (B) is to identify the PICO elements associated with each claim. In the final step (C), we use the outputs of the first two models with the original post to obtain a dense representation, enabling us to retrieve relevant evidence from a large dataset of trusted medical evidence (e.g., PubMed).

span boundaries. Despite this, for the first stage agreement (Fleiss κ) on labeled questions, experiences, and claims was 0.62, and for the second stage 0.55. We consider this *moderately strong* agreement, in line with agreement reported for related annotation tasks in the literature (Nye et al., 2018; Deléger et al., 2012). To quantify this and further gauge the quality of collected annotations, we run a few additional analyses.

As previously stated, prior to collecting annotations on Amazon MTurk, we (the authors) annotated a subset of data (100 samples/stage) internally to assess task difficulty and to estimate the time required for annotation. As an additional quality check, we use these annotations to calculate token-wise label agreement. Table 2 reports the results; while there remains some discrepancy owing to the inherent complexity of the task, there is higher agreement between the us than between workers.

Each of these samples was also annotated by three workers. We aggregate these labels using majority-vote and compute token-wise precision-recall of these aggregated labels against the reference "in-house" labels (Table 2). We report the same metrics per annotator evaluated against aggregated MTurk labels in Table 9 (Appendix B). Despite moderate agreement between annotators, aggregated labels agree comparatively well with the "expert" consensus, indicating that while individual worker annotations are somewhat noisy, aggregated annotations are reasonably robust.

2.4 Dataset Details

Table 1 provides illustrative samples from **RedHOT** and Table 8 provides some descriptive statistics along with examples of included health populations. We broadly characterize populations (conditions) as *Very Common*, *Common* or *Rare*, and sought a mix of these. This was not the only attribute that informed which conditions we selected for inclusion in our dataset, however. For example, we wanted a mix of populations with respect to volume of online activity (e.g., the Diabetes subreddit has over $60k$ active visitors; Lupus has $8k$). We also wanted to include both chronic and treatable conditions (e.g., Narcolepsy is a rare and chronic condition, while Gout is common and treatable), and mental and physical disorders (e.g., ADHD, Rheumatoid Arthritis). Another consideration was whether a condition can be self-diagnosed or requires professional assessment (e.g., Bulimia is usually self-diagnosable but can potentially be life-threatening; Gastroparesis is chronic but requires a professional medical diagnosis).

The number of *claims* across different categories of health populations are far outnumbered by *ques-*

tions (∼10x) and *experiences* (∼13x). The average post length is ∼117 tokens while the average length of a *claim* within a post is ∼20 tokens. *Questions* and *experiences* have average lengths of ∼11 and ∼27 tokens, respectively. We provide per condition statistics in Appendix B.

3 Tasks and Evaluation

RedHOT may support a range of tasks related to processing health-related social media posts. Here we focus on an important, timely task: Identifying medical claims on social media, and then retrieving relevant and trustworthy evidence that may support or refute them. Methods for this task could aid content moderation on health-related forums, by providing an efficient means to (in)validate claims. More generally, such methods may permit meaningful "fact checking" of health-related claims by providing relevant contextualizing evidence.

We outline a three-step approach for this task. (1) Identify spans/sentences corresponding to *pure* claims. (2) Extract from these specific PICO elements. (3) Retrieve clinical literature — specifically, reports of RCTs — relevant to the claim, i.e., the extracted PIO elements. We limit our focus to the problem of evidence retrieval here; future work might consider the subsequent step of automated claim validation on the basis on this.

Below we assess components for each of these steps. For the span and PIO extraction steps (1 and 2), we evaluate models retrospectively under standard classification metrics (i.e. precision, recall, and F1 scores) using fixed train, development, and test sets which we will distribute with **RedHOT**. The final step (3) requires relevance judgments to evaluate model performance; for this we enlisted medical doctors (Section 3.4).

3.1 Identifying Claims, Experiences, Questions, and PIO Elements

We treat the first two steps as sequence tagging tasks for which we evaluate two types of models: A simple linear-chain Conditional Random Field (CRF; Lafferty et al. 2001), and Transformer-based models (Vaswani et al., 2017) — specifically BERT variants (Devlin et al., 2019; Liu et al., 2019).[5] The features for the CRF we use are: Indicators of next, previous, and current words; Part-of-speech tags,[6] and; Indicators encoding if sentences contain digits, uppercase letters, and/or measurement units. BERT variants yield contextualized representations of input tokens, which we then use to predict labels (i.e., Claims, Experiences and Questions) by adding a linear layer on top of the encoder outputs. PIO elements are extracted using a concatenated input of the original Reddit post and an identified claim.

3.2 Evidence Retrieval

For the retrieval task we assume the model is given: (i) The original Reddit post and a claim; (ii) PIO elements associated with that claim, and; (iii) A *large* set of candidate articles featuring trustworthy evidence to rank. We use ∼800,000 abstracts from *Trialstreamer*[7] (Marshall et al., 2020), a continuously updated database of reports of randomized controlled trials (RCTs). RCTs are appropriate here because of our focus on causal claims — results from randomized trials are the most reliable means of evaluating such assertions (Meldrum, 2000).

3.2.1 Task Formulation

Formally, we represent a single input instance as $(p, c_j, \text{pop}_j, \text{int}_j, \text{out}_j)$ where p is a post comprising n sentences, c_j is the jth claim, and $\text{pop}_j, \text{int}_j, \text{out}_j$ are the *sets* of populations, interventions, outcomes associated with claim j.

The model is tasked with finding relevant abstracts from the candidate set $\mathcal{A}$,which comprises abstracts from published clinical trial reports. This is particularly challenging because a large number of candidates can mention the same set of PIO entities (i.e., investigate the same interventions and/or outcomes), but in a context unrelated to the claim being made in the social media post. This may be especially problematic for retrieval methods based primarily on string overlap measures. We therefore propose a learning based approach. This requires supervision; we next describe our approach to deriving this automatically.

3.2.2 Pseudo Training Data

Supervised neural retrieval models require annotations indicating the relevance of instances (here, published evidence) to inputs (claims on social media). We do not have such judgments, and so instead derive "pseudo" training data automatically.

We started with ∼800,000 abstracts of medical RCTs in Trialstreamer. We then used Reddit

[5] We also explored t5 (Raffel et al., 2020) with middling results, which we report in the Appendix.

[6] Extracted with SciSpacy (https://allenai.github.io/scispacy/).

[7] https://trialstreamer.ieai.robotreviewer.net/

	P	R	F1	$F1_{POP}$	$F1_{INT}$	$F1_{OUT}$
BERT (Devlin et al., 2019)	43.88	36.13	39.62	41.77	44.68	33.05
BioRedditBERT (Basaldella et al., 2020)	44.44	36.55	40.12	41.92	44.31	34.61
biomedRoBERTa (Gururangan et al., 2020)	38.80	21.48	27.66	30.54	28.13	24.54
RoBERTa (Liu et al., 2019)	**47.45**	**39.27**	**42.97**	**46.09**	**45.99**	**36.38**
t5-small (Raffel et al., 2020)	41.49	38.55	39.97	39.61	45.02	32.41

Table 3: Results on the test set for the token-level PICO tagging task.

	Claims			Experiences			Questions		
	F1	P	R	F1	P	R	F1	P	R
CRF (Lafferty et al., 2001)	33.87	35.61	32.29	40.08	40.52	39.64	86.89	85.55	88.27
BERT (Devlin et al., 2019)	**52.63**	58.82	47.61	56.68	59.46	54.33	92.39	88.76	96.34
RoBERTa (Liu et al., 2019)	47.05	61.53	**38.09**	56.81	57.11	56.52	93.06	89.01	98.34
BioRedditBERT (Basaldella et al., 2020)	45.16	**70.92**	33.29	**59.51**	**62.49**	**58.92**	**93.61**	**89.29**	**98.37**

Table 4: Results on the test-set of span-classification to identify pure claims, questions, and experiences.

posts containing pure claims as *templates* to create pseudo matches between medical claims and abstracts. Specifically, we substituted annotated PIO elements in claims made within Reddit posts with PIO elements sampled from Trialstreamer abstracts. (Trialstreamer includes PICO elements automatically extracted from all articles that it indexes.) This yields pairs of (a) naturally occurring claims (with their PIO spans replaced) and (b) RCT abstracts that are relevant to said claims by construction. We provide examples of this pseudo matching in Appendix D. We generated a total of 85,000 examples of `(pseudo claims, evidence abstract)` one-to-many pairs to be used to train a neural retrieval model (described below). The generated examples may be noisy, but hopefully sufficient to train a model to retrieve medical abstracts relevant to health-related claims made on social media.

3.2.3 RedHOT Dense Evidence Retriever

We train a neural retrieval model on the **RedHOT** corpus, using a setup similar to DPR (Karpukhin et al., 2020). We first assemble a collection of m RCT abstracts to create an evidence corpus, $\mathcal{A} = \{d_1, d_2, ..., d_m\}$. There are hundreds of thousands of RCTs, so we need an *efficient* retriever that can select a small set of relevant abstracts. Formally, a retrieval operation $\{R\colon (x_j, \mathcal{A}) \rightarrow \mathcal{A}_\mathcal{F}\}$ accepts an input contextualizing string x_j and a corpus of evidence $\mathcal{A}$, and returns a *much smaller* filtered set $\mathcal{A}_\mathcal{F} \subset \mathcal{A}$, where $|\mathcal{A}_\mathcal{F}| = k$.

We form an input context string x_j for a claim j made within a post p by concatenating the post, claim, and PIO elements extracted from the claim: $x_j = [p \oplus c_j \oplus \text{pop}_j \oplus \text{int}_j \oplus \text{out}_j]$, where $\oplus$ denotes concatenation with `[SEP]` tokens. We define two dense neural encoders (E_C, E_D; both initialized with RoBERTa-base) to project the context string x_j, and evidence (abstracts) from $\mathcal{A}$ to fixed 768 dimensional vectors. Similarity between the context string and evidence abstract is defined using the dot product of their vectors, $\phi(x_j, d_l) = E_C(x_j)^T E_D(d_l)$.

We train the model to minimize the negative log-likelihood of the positive evidence such that it pushes the context string vector x_j close to the representation of relevant evidence d_j^+, and away from b irrelevent abstracts ($d_{j1}^-, d_{j2}^-, ...d_{jb}^-$) in the same mini-batch[8] ("in-batch negative sampling"):

$$\mathcal{L} = \frac{\exp \phi(x_j, d_j^+)}{\exp \phi(x_j, d_j^+) + \sum_{l=1}^{b} \exp \phi(x_j, d_{jl}^-)}$$

In-batch negative sampling has been shown to be effective for dual-encoder training (Henderson et al., 2017; Gillick et al., 2019). Here, all samples in a minibatch are taken from the same *population* (condition) set, e.g., a mini-batch with a sample containing a claim about `diabetes` will have negative evidence abstracts that are also related to `diabetes`.

For test examples, we rank all evidence (abstracts in Trialstreamer) according to their similarity to the context string. To do this efficiently, we induce representations of all the abstracts in the Trialstreamer database using the evidence encoder and index these using the Facebook AI Similarity Search library (Johnson et al., 2021).[9]

[8] We set the size of the mini-batch to 100.

[9] FAISS: Open-source library for efficient similarity search

k	MRR @k					Precision @k				
	1	5	10	50	100	1	5	10	50	100
random	0.00	0.003	0.02	0.02	0.02	0.00	0.02	0.00	1.10	2.80
BM25	5.34	7.98	9.86	14.36	16.70	5.34	10.40	14.45	26.20	33.14
DPR (Karpukhin et al., 2020)	8.07	10.96	11.89	12.20	13.77	8.07	16.50	23.58	31.98	36.87
*(trained on the **RedHOT** pseudo training set)*										
RedHOT-DER (BERT-based)	39.14	47.99	49.3	50.28	50.35	39.14	62.55	72.64	83.73	91.74
RedHOT-DER (RoBERTa-based)	**45.93**	**54.60**	**55.90**	**56.73**	**56.78**	**45.93**	**69.90**	**78.81**	**94.73**	**98.06**

Table 5: Results of evidence retrieval baselines evaluated on pseudo test data.

3.2.4 Baseline Models

BM25 A standard Bag-of-Words method for IR (Robertson et al., 1995). We form queries by concatenating the Reddit post with a single claim and its corresponding PIO frames. We used a publicly available BM25 implementation from the Rank-BM25 library.[10]

Dense Passage Retrieval (DPR) is a dense retrieval model trained to retrieve *relevant* context spans ("paragraphs") in an open domain question-answering setting (Karpukhin et al., 2020). In general, such models map **queries** and **candidates** to embeddings, and then rank candidates with respect to a similarity measure (e.g., dot product) taken between these. While originally designed for open-domain question answering, use of DPR-inspired models has been extended to general retrieval tasks (Thai et al., 2022a). We use a DPR context encoder trained on Natural Questions (Kwiatkowski et al., 2019) with dot product similarity.[11]

3.3 Results

We evaluate models for the tasks of identifying claims, experiences, and questions and extracting PIO elements using precision, recall, and F1 scores. We report results per class for the first task in Table 4. `BioRedditBERT` (Basaldella et al., 2020) — a BERT model initialized from `BioBERT` (Lee et al., 2019) and further pre-trained on health-related Reddit posts — fares best here. We report results for the second task (PIO tagging) in Table 3.[12] Here `RoBERTa` (Liu et al., 2019) modestly outperforms `BioRedditBERT` (Basaldella et al., 2020).

Models for the retrieval task rank evidence candidates for each input (post, claim, PIO frame). We therefore use standard ranking metrics for evaluation, including mean reciprocal rank, and precision@k (for $k = 1, 5, 10, 50, 100$). Baseline results are reported in Table 5. We emphasize that these results are with respect to pseudo annotated data, effectively providing an unfair advantage to **RedHOT**-DER, given that this was optimized on data from this distribution. We report results with respect to manual relevance judgments provided by experts in Section 3.4.

As we might expect, the pre-trained neural DPR model outperforms the naive string matching BM25 method. Furthermore, as anticipated, explicitly training for evidence retrieval confers pronounced advantages: **RedHOT**-DER fares ∼8x better than BM25 and ∼5x better than "off-the-shelf" pre-trained DPR (Karpukhin et al., 2020) with respect to retrieving relevant evidence (precision@1) corresponding to medical claims. Again, this is not particularly surprising given that we are evaluating models with respect to the pseudo annotations with which **RedHOT**-DER was trained (because we do not otherwise have access to explicit relevance judgments). Therefore, we next present results from more meaningful manual relevance evaluations performed by domain experts.

3.4 Expert Manual Relevance Judgments

We evaluated models in terms of retrieving evidence relevant to *naturally occurring* medical claims, as opposed to the *pseudo* data derived for training. We hired three domain experts (medical doctors) on the Upwork platform.[13] Providing hundreds of retrieved medical abstracts per claim to a human evaluator for assessment is infeasible, so

and clustering of dense vectors; https://ai.facebook.com/tools/faiss/.

[10] https://github.com/dorianbrown/rank_bm25

[11] https://huggingface.co/facebook/dpr-ctx-encoder-single-nq-base

[12] Results from additional experiments using other model variants are reported in Appendix C.

[13] Upwork (https://www.upwork.com/) allows clients to interview, hire and work with freelancers. All of our evaluators had medical degrees and were hired at wages ranging from $15 to $20 per hour for a minimum of 15 hours.

Cumulative # of relevant abstracts @k				
k	1	3	5	10
Pre-trained DPR (Karpukhin et al., 2020)				
Relevant	6	16	29	58
Somewhat relevant	14	39	66	135
Irrelevant	80	245	405	807
***RedHOT*-DER trained on pseudo data**				
Relevant	18	62	101	201
Somewhat relevant	17	49	87	193
Irrelevant	65	189	312	606

Table 6: Results from manual (domain expert) evaluations for DPR and our pseudo-supervised DER model.

we instead provided evaluators with 10 retrieved abstracts each for 100 individual claims, retrieved using the pretrained DPR (Karpukhin et al., 2020) model and our **RedHOT**-DER trained on pseudo data. (We compared the proposed distantly supervised model to DPR because it is the strongest baseline we evaluated in preliminary experiments.)

We asked evaluators to categorize each retrieved abstract as: (1) `Relevant`; (2) `Somewhat Relevant`, or; (3) `Irrelevant` to the corresponding claim. An abstract was to be considered `Relevant` if and only if it (1) contained to the same **P**, **I**, and **O** elements mentioned in the original Reddit post, **and** (2) provided information to support or refute the claim in question. An abstract might be deemed `Somewhat Relevant` if it contains a **P**, **I**, and **O** set in line with the given claim, but does not provide any information relating these elements. We provide examples in the Appendix D.

Human evaluators achieve strong agreement: All three evaluators chose the same relevance label 71.33% of the time, while they all chose a different label only in 1.29% of the total instances. They also show substantial agreement in terms of Fleiss κ (0.71). We derive final relevance labels by majority vote. Comparing results from Table 5 and Table 6, at $k = 1$ we see similar values of precision in the manually annotated data and pseudo test data. However, for higher values of k large differences emerge, indicating considerable room for improvement. Compared to the pre-trained DPR model, at $k = 1$ **RedHOT**-DER retrieves a substantially larger fraction of relevant evidence abstracts (3x). At higher k, we also observe a large reduction in the number of *irrelevant* abstracts retrieved (e.g., at $k = 10$, the number of irrelevant abstracts decreases by $\sim 30\%$). We believe this highlights the value of our proposed distant supervision scheme.

4 Related Work

Claim validation via evidence retrieval Past work has typically treated (open domain) claim validation as a two-step process in which one retrieves evidence relevant to a given claim, and then makes a prediction regarding claim validity on the basis of this. Information retrieval (IR) models are usually used in the first step to rank order documents based on relevance to a given claim (Thorne et al., 2018; Wadden et al., 2020; Thai et al., 2022b; Hanselowski et al., 2018; Samarinas et al., 2021; Saeed et al., 2021). The next step is usually to characterize *retrieved* evidence as `supporting`, `refuting`, or `not providing enough information` (although this latter category is not always included). Evidence might be individually characterized (Pradeep et al., 2021), or aggregated to make a single prediction about the veracity of the claim (Sarrouti et al., 2021).

Scientific claim verification Beyond "general domain" verification, there have been efforts focused specifically on vetting *scientific* claims. `SciFact` (Wadden et al., 2020) largely follows the typical fact verification setup outlined above (but for scientific claims). Subsequent efforts have focused specifically on verifying claims related to COVID-19 (Saakyan et al., 2021). The `evidence inference` task (Lehman et al., 2019; DeYoung et al., 2020) entails inferring whether a given trial report supports a significant effect concerning a specific intervention, comparator, and outcome.

Crowd-sourcing annotation of scientific and medical texts We have relied on crowdworkers to annotate the instances comprising **RedHOT**. This is in keeping with a body of work that has shown crowdworkers capable of annotating health-related texts, even when these are technical (Drutsa et al., 2021). For example, several past efforts have crowdsourced annotation of texts drawn from PubMed, e.g. for mentions of diseases (Nye et al., 2018; Good et al., 2014). More recently, Bogensperger et al. (2021) crowdsourced a dataset of drug mentions (a type of *intervention*) on the darknet. Khetan et al. (2022) crowdsourced annotations of electronic health records to identify causal relations between medical entities. Similarly, there is a body of work relying on crowdsourcing to accom-

plish a diverse set of domain-specific non-medical NLP tasks (Sukhareva et al., 2016; Fromreide et al., 2014; Bhardwaj et al., 2019; Gardner et al., 2020).

Health-related Reddit corpora Past work has also built corpora of health-related Reddit posts. For example, Cohan et al. (2018) assembled a dataset of Reddit posts made by individuals who self-reported one of nine mental health diagnoses of interest. Building on this work, Jiang et al. (2020) introduced a dataset of Reddit posts to evaluate models for automatically detecting psychiatric disorders.

5 Conclusions

We presented **RedHOT**: a new, publicly available dataset comprising of about 22,000 richly annotated Reddit posts extracted from 24 medical condition-based communities ("subreddits"). This dataset meets a need for corpora that can facilitate development of language technologies for processing health-related social media posts.

We evaluated baseline models for categorizing posts as containing claims, personal experiences, and/or questions. Focusing on claims, we then proposed and evaluated models for extracting descriptions of populations, interventions, and outcomes, and then using such snippets to inform retrieval of trustworthy (published) evidence relevant to a given claim. To this end, we introduced a heuristic supervision strategy, and found that this outperformed pre-trained retrieval models.

Limitations

We have introduced a new annotated dataset of medical questions, experiences, and claims across a range of health populations from social media. We showed that this data can be used to train models potentially useful for downstream applications, e.g., by facilitating content moderation. However, there are important limitations to this work, specifically with respect to the raw data we sampled and the annotations on this that we have collected.

First, the dataset we have annotated is inherently limited. While we have tried to select a diverse set of health populations (i.e., subreddits), these nonetheless constitute a small sample of the diverse set of existing health conditions. Moreover, our selection has led to a corpus comprising nearly entirely of English-language posts, which is a clear limitation.

We relied on non-expert (layperson) workers from Amazon Mechanical Turk (AMT) to carry out the bulk of annotation work. While we took steps to try and ensure annotation quality (described in Section 2), we nonetheless acknowledge that these annotations will contain noise. This is especially true given that AMT workers are not medical-experts and ultimately do not have (nor are they expected to have) sufficient knowledge of different kinds of medical terms appearing in the dataset (e.g., SSRIs' stand for *selective serotonin reuptake inhibitor* and is a common form of intervention which may lead to outcomes like dizziness, anxiety, and/or insomnia, but many laypeople might simply be unaware of ordinary meaning of complicated medical terms leading them to *not* matching all or part of such terms to their respective labels).

In Section 3.2.2, we describe how we obtained *pseudo* training labels to build a supervised dense retriever. To generate this data, several natural language claims get reused with substitute set of populations/interventions/outcomes. This heuristic may induce certain biases (as evident from Table 6 and Table 5). An ideal way to train a dense retriever here would be to collect positive annotation labels for *every* claim in our dataset. Collecting such supervision at scale sufficient for model training would be expensive, given that one would strongly prefer expert (medical doctor) annotations concerning the factual accuracy of claims.

Ethics Statement

This work has the potential to contribute to human well-being by supporting development of language technologies for processing health-related social media posts. Such models might in turn provide insights about patient experiences and viewpoints in general, and more specifically may help community moderators identify and remove posts containing medical misinformation.

Realizing these potentially positive contributions requires annotated data with which to train relevant models; such data is the main contribution on offer in this work. However, releasing an annotated corpus of health-related social media posts raises concerns regarding individual privacy. The Reddit posts we have assembled and collected annotations were posted publicly on the Internet (almost always under pseudonyms), but nonetheless we have taken steps to ensure that individuals can choose not to be represented in this dataset.

Specifically, we sent a message to every user in the **RedHOT** explaining our intent to construct and

release this dataset and offering the option to "opt out". In addition, although this is not required by Reddit, we have decided not to release the collected *posts* directly. Instead we release a script that will download the posts comprising our data on-demand and align these with the collected annotations. This means that if a user chooses to delete their post(s) from Reddit, they will also effectively be removed from our dataset. Further, we require anyone accessing this data to self-certify that they have obtain prior approval from their own IRB concerning the use-cases of their research.

Acknowledgements

This work was supported in part by the National Science Foundation (NSF) CAREER award 1750978.

References

Zapan Barua, Sajib Barua, Salma Aktar, Najma Kabir, and Mingze Li. 2020. Effects of misinformation on covid-19 individual responses and recommendations for resilience of disastrous consequences of misinformation. *Progress in Disaster Science*, 8:100119.

Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. COMETA: A corpus for medical entity linking in the social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3122–3137, Online. Association for Computational Linguistics.

Sangnie Bhardwaj, Samarth Aggarwal, and Mausam Mausam. 2019. CaRB: A crowdsourced benchmark for open IE. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6262–6267, Hong Kong, China. Association for Computational Linguistics.

Prakhar Biyani, Cornelia Caragea, Prasenjit Mitra, and John Yen. 2014. Identifying emotional and informational support in online health communities. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*, pages 827–836.

Johannes Bogensperger, Sven Schlarb, Allan Hanbury, and Gábor Recski. 2021. DreamDrug - a crowdsourced NER dataset for detecting drugs in darknet markets. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 137–157, Online. Association for Computational Linguistics.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1485–1497, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Louise Deléger, Qi Li, Todd Lingren, Megan Kaiser, Katalin Molnár, Laura Stoutenborough, Michal Kouril, Keith A. Marsolo, and Imre Solti. 2012. Building gold standard corpora for medical natural language processing tasks. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2012:144–53.

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 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Eric Lehman, Ben Nye, Iain J Marshall, and Byron C Wallace. 2020. Evidence inference 2.0: More data, better models. *arXiv preprint arXiv:2005.04177*.

Alexey Drutsa, Dmitry Ustalov, Valentina Fedorova, Olga Megorskaya, and Daria Baidakova. 2021. Crowdsourcing natural language data at scale: A hands-on tutorial. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*, pages 25–30, Online. Association for Computational Linguistics.

Hege Fromreide, Dirk Hovy, and Anders Søgaard. 2014. Crowdsourcing and annotating NER for Twitter #drift. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2544–2547, Reykjavik, Iceland. European Language Resources Association (ELRA).

Rachel Gardner, Maya Varma, Clare Zhu, and Ranjay Krishna. 2020. Determining question-answer plausibility in crowdsourced datasets using multi-task learning. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 22–27, Online. Association for Computational Linguistics.

Daniel Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldridge, Eugene Ie, and Diego Garcia-Olano. 2019. Learning dense representations for entity retrieval. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 528–537, Hong Kong, China. Association for Computational Linguistics.

Adriana D Glenn. 2015. Using online health communication to manage chronic sorrow: mothers of children with rare diseases speak. *Journal of Pediatric Nursing*, 30(1):17–24.

Benjamin M Good, Max Nanis, Chunlei Wu, and Andrew I Su. 2014. Microtask crowdsourcing for disease mention annotation in pubmed abstracts. In *Pacific Symposium on Biocomputing Co-Chairs*, pages 282–293. World Scientific.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. UKP-athene: Multi-sentence textual entailment for claim verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108, Brussels, Belgium. Association for Computational Linguistics.

Matthew L. Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *CoRR*, abs/1705.00652.

Zhengping Jiang, Sarah Ita Levitan, Jonathan Zomick, and Julia Hirschberg. 2020. Detection of mental health from Reddit via deep contextualized representations. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 147–156, Online. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Vivek Khetan, Md Imbesat Rizvi, Jessica Huber, Paige Bartusiak, Bogdan Sacaleanu, and Andrew Fano. 2022. MIMICause: Representation and automatic extraction of causal relation types from clinical notes. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 764–773, Dublin, Ireland. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

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*, 36(4):1234–1240.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. 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)*, pages 3705–3717, Minneapolis, Minnesota. Association for Computational Linguistics.

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*, abs/1907.11692.

Iain J Marshall, Benjamin Nye, Joël Kuiper, Anna Noel-Storr, Rachel Marshall, Rory Maclean, Frank Soboczenski, Ani Nenkova, James Thomas, and Byron C Wallace. 2020. Trialstreamer: A living, automatically updated database of clinical trial reports. *Journal of the American Medical Informatics Association*, 27(12):1903–1912.

Marcia L. Meldrum. 2000. A brief history of the randomized controlled trial: From oranges and lemons to the gold standard. *Hematology/Oncology Clinics of North America*, 14(4):745–760.

Alicia L Nobles, Caitlin N Dreisbach, Jessica Keim-Malpass, and Laura E Barnes. 2018. " is this an std? please help!": Online information seeking for sexually transmitted diseases on reddit. In *Twelfth International AAAI Conference on Web and Social Media*.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Melbourne, Australia. Association for Computational Linguistics.

Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. 2021. Scientific claim verification with VerT5erini. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pages 94–103, online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

W Scott Richardson, Mark C Wilson, Jim Nishikawa, Robert S Hayward, et al. 1995. The well-built clinical question: a key to evidence-based decisions. *Acp j club*, 123(3):A12–A13.

Stephen Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. 1995. Okapi at trec-3. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 109–126. Gaithersburg, MD: NIST.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. Covid-fact: Fact extraction and verification of real-world claims on covid-19 pandemic. *arXiv preprint arXiv:2106.03794*.

Mohammed Saeed, Giulio Alfarano, Khai Nguyen, Duc Pham, Raphael Troncy, and Paolo Papotti. 2021. Neural re-rankers for evidence retrieval in the FEVEROUS task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 108–112, Dominican Republic. Association for Computational Linguistics.

Chris Samarinas, Wynne Hsu, and Mong Li Lee. 2021. Improving evidence retrieval for automated explainable fact-checking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 84–91, Online. Association for Computational Linguistics.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maria Sukhareva, Judith Eckle-Kohler, Ivan Habernal, and Iryna Gurevych. 2016. Crowdsourcing a large dataset of domain-specific context-sensitive semantic verb relations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2131–2137, Portorož, Slovenia. European Language Resources Association (ELRA).

Briony Swire-Thompson and David Lazer. 2019. Public health and online misinformation: challenges and recommendations. *Annual review of public health*, 41:433–451.

Katherine Thai, Yapei Chang, Kalpesh Krishna, and Mohit Iyyer. 2022a. Relic: Retrieving evidence for literary claims. *arXiv preprint arXiv:2203.10053*.

Katherine Thai, Yapei Chang, Kalpesh Krishna, and Mohit Iyyer. 2022b. RELiC: Retrieving evidence for literary claims. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7500–7518, Dublin, Ireland. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Appendix for "RedHOT: A Corpus of Annotated Medical Questions, Experiences, and Claims on Social Media"

A Data Collection

A.1 Sampling from Reddit

We retrieved the *newest* 1,000 posts from the respective subreddits using the Reddit PRAW[14] API. While we could have relied on alternative sampling strategies — e.g., ranking posts according to "hot" or "best" under Reddit's metrics — retrieving the newest posts yields an unfiltered snapshot of the full variety of posts made to social media. We also considered performing completely uniform sampling over all posts ever made to a given forum, but the Reddit API limits callers to retrieving 1000 posts for any search criteria; this practically precludes uniform sampling across all time periods.

Preprocessing We identified and removed all non-English text post extraction.[15] Reddit allows its users to post media content (images/videos) in addition to text, and such imagery can be explicit or disturbing. Therefore, we only retained posts that did not contain any media content.

A.2 Annotations on Amazon Mechanical Turk

Amazon Mechanical Turk (AMT) is a popular platform for recruiting non-expert workers to perform "micro-tasks" (here, annotation). We initially recruited workers by collecting annotations on relatively simple examples for which we already had ground truth labels. We provided AMT workers with a comprehensive set of instructions including (templated) examples of the respective categories. For instance:

- **Questions:** Does this work?; Are X, Y, Z symptoms normal for Condition A?; Will increasing my dosage of X help Y in any way?
- **Personal Experiences:** I was diagnosed with X and have since experienced symptoms Y,Z.; I took X and it seemed to help.; My mother took Y and it helped improved her Z
- **Claim:** My doctor told me that X should help with Y; Since increasing dosage of Z, my X levels have normalized (also an example of personal-experience); I heard from multiple people that A helps with C; I read online that X & Y are directly causing Z; I heard from my cousin that X helps control Z

For additional context we provided workers with the "Topic", i.e., the subreddit from which the post being annotated was sampled. For example, if the topic was "Diabetes", the piece of text will (presumably) be about diabetes, its treatments, individual experiences with the condition, and so on. We highlight the stage-1 annotation interface in Figure 3. The complete set of instructions we provided to AMT workers are available at https://anonymous.4open.science/r/med_val-64C2/stg1_instructions.pdf.

We retained all qualified AMT workers from stage-1 to carry out additional annotations for us in stage-2, with a higher pay rate. The objective here was to recruit people who had established a working understanding of the data, and would presumably be proficient as a result. Similar to stage-1, we provided workers with a comprehensive set of instructions containing (templated) examples to give a sense of what might be qualify as PIO elements:

- **Populations** coronavirus, asthma, narcoleptic, diabetic, children, young, women etc.
- **Interventions** diet, aspirin, allopurinol, insulin, exercise, botox etc
- **Outcomes** depression, sweating, anxiety, pain, flares, covid etc

Interface used for stage-2 annotations is provided in Figure 4. Complete set of stage-2 instructions provided to AMT workers are available at https://anonymous.4open.science/r/med_val-64C2/stg2_instructions.pdf.

B Dataset Summary

Table 7 provides descriptive statistics for all patient populations (that is, subreddits) included in our dataset. Dysthymia has the highest number of posts included in our corpus while Ankylosing Spondylitis has the lowest (due to data filtering described above). There is substantial variation in the length of the posts written under different subreddits (e.g., in r/ADHD the average post is ∼222 tokens, while in r/Lupus it's only ∼93 tokens long). Similarly, there are variations in the number of questions, claims, and experiences across populations. We used subscriber count as a proxy for

[14] https://github.com/praw-dev/praw

[15] Langid is a python tool that allows filtering data by language: https://github.com/saffsd/langid.py.

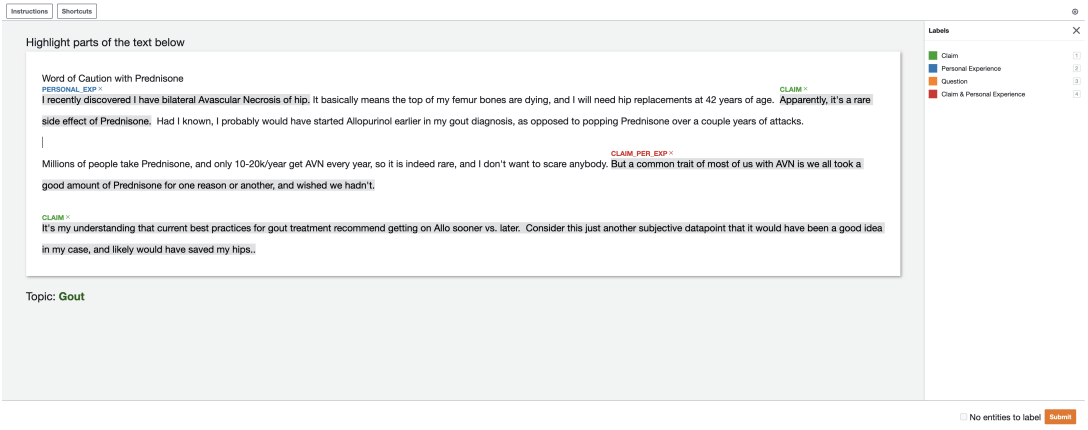

Figure 3: Stage-1 annotations interface for demarcation of spans associated with *questions*, *experiences*, and *claims*.

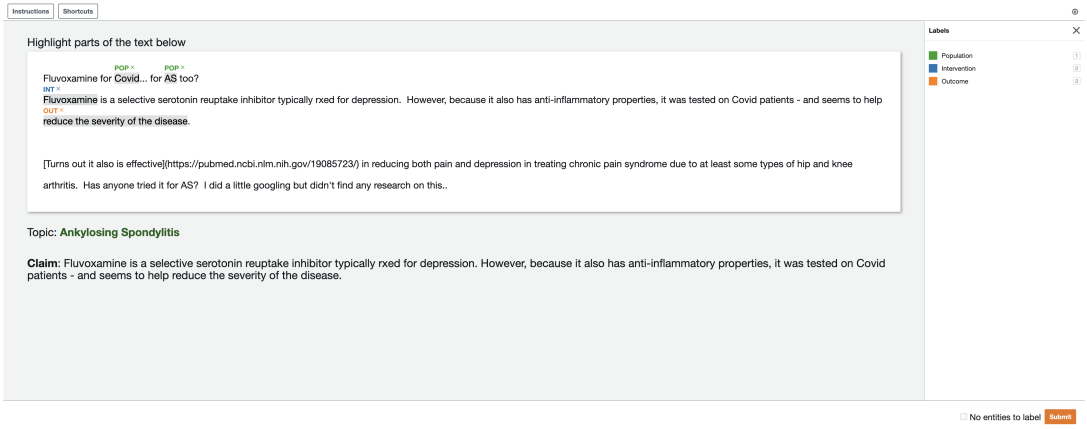

Figure 4: Stage-2 annotations interface for demarcation of PICO frames associated with a given Reddit post and a claim.

Population Type (subreddit)	# of posts in RedHOT	Avg. length of post (# tokens)	# claims	# questions	# experiences	# subscribers on Reddit
Dysthymia	999	175.42	102	989	1387	6.8k
Chronic Fatigue Syndrome	998	139.50	162	1034	1292	31.1k
IBS	998	118.70	71	987	1337	77.1k
Narcolepsy	997	148.65	121	1311	1547	18.9k
Bulimia	996	122.99	46	761	1316	32.8k
Hypothyroidism	995	125.91	111	1585	2088	35.2k
Costochondritis	995	116.97	98	1136	1488	8.8k
Hyperhidrosis	994	97.21	184	1076	1245	25k
Sinusitis	991	135.45	136	1242	1979	5.9k
Psychosis	984	122.91	53	932	933	39.8k
Thyroid Cancer	976	121.80	143	1157	1405	3.2k
Cystic Fibrosis	970	96.11	77	1001	882	7.1k
POTS	963	111.03	77	1155	1274	21.8k
Multiple Sclerosis	958	152.47	129	1081	1309	31.6k
Gout	933	128.87	154	1251	1730	14.2k
ADHD	899	222.41	141	875	1222	1.4M
Gastroparesis	861	134.91	52	909	1319	8k
Diabetes (Type I & II)	748	113.85	40	667	620	90.4k
Crohn's Disease	791	99.79	92	1026	995	43.7k
Lupus	784	93.13	96	978	972	18.2k
Rheumatoid Arthritis	759	103.08	105	1033	1010	6.4k
Epilepsy	670	165.77	37	634	1170	27.8k
GERD	650	164.12	45	669	1518	44.2k
Ankylosing Spondylitis	644	170.83	32	649	1139	12.6k

Table 7: Population-wise descriptive statistics.

		Average # *per population*			Average # *per claim*		
Population type	**# Posts**	**Questions**	**Experiences**	**Claims**	**Populations**	**Interventions**	**Outcomes**
Very Common (Dysthymia, Hypothyroidism, Gout, etc)	5467	1101.82	1654.00	114.83	0.82	2.66	3.57
Common (Chronic Fatigue Syndrome, Bulimia, Psychosis, etc)	9539	847.01	1141.72	74.27	1.05	2.95	3.22
Rare (Narcolepsy, Hyperhidrosis, Thyroid Cancer, etc)	7295	1028.50	1166.25	104.75	0.97	2.79	3.81

Table 8: For descriptive purposes we categorize conditions into: Very Common (>3 million US cases per year), Common (>200k US cases per year), and Rare (<200k US cases per year). We only include posts that do *not* contain any media (photos/videos). Number of experiences here *include* claims based on personal experiences. Diabetes is included as both Common (Type II) and a Rare (Type I) type.

	P	R	F1
Questions	0.73	0.68	0.70
Claims	0.47	0.40	0.43
Experiences	0.33	0.29	0.31
POP	0.85	0.81	0.83
INT	0.56	0.50	0.53
OUT	0.48	0.37	0.42

Table 9: Individual annotator labels evaluated against their own "aggregated" labels.

gauging how active a community on Reddit is. For instance, `r/ADHD` has 1.4M subscribers and so can be considered substantially more active than, say, `r/Psychosis`, which has 39.8K subscribers.

C Additional Results and Experimental Details

We provide results from additional BERT (Devlin et al., 2019) variants for the first task of identifying claims, questions, and experiences in Table 10. Unsurprisingly, pre-trained neural models consistently outperform linear-chain Bag-of-Words CRFs. Similarly, Table 11 provides results from BERT variants and t5-small (Raffel et al., 2020) for the second task of extracting PICO elements conditioned on the post and a given claim. For the t5 model, the target was to produce `<entity token>` followed by `<entity label>` in the same order as they appear in the input sentence (sequential linearization scheme). We evaluated the generated entities against the true sets of PICO elements for each output. While it may be possible to come up with a more optimal linearization scheme for sequence labelling, we posit that to be beyond the scope of our work.

To use dense retrieval models to rank evidence (abstracts) with respect to their relevance to a given claim we need an efficient means to index vectors for ∼800k abstracts of RCTs in the Trialstreamer database. We did so using FAISS (Johnson et al., 2021) on an Intel Xeon E5-2650 V3 CPU @2.3GHz with 512GB memory. Building an index of dense embeddings for hundreds of thousands passages is highly resource intensive and required roughly 9 hours on two NVIDIA GeForce GTX 1080Ti GPUs.

To train the dense retriever, we used standard split of train, development, and test sets (80%-10%-10%). We trained the two encoders for 40 epochs with a learning rate of 10^{-5} using the Adam optimizer, linear scheduling with warm up, and a dropout rate of 0.1. We parallelized training over multiple-GPUs; it took roughly 40 hours to train the retriever. Our best-performing retrieval model was initialized with RoBERTa-base (250M parameters). In addition to the results provided in section 3.3, we provide additional results for the retrieval task (evaluated on pseudo test set) in Table 12.

D Deriving Pseudo Training Data: Examples

Generating pseudo training data — i.e., matching reddit annotated reddit posts to "relevant" abstracts of RCTs — is an important component of our dense retrieval pipeline. In Table 13 we provide several examples of the pseudo data we generated from annotated claims. For each row we have inserted intervention and outcome elements from abstracts indexed in Trialstreamer, which makes them "relevant" by construction (while still featuring natural language as it used on social media). We showcase how stage-2 annotated (post, claim) pairs serve as templates to create pseudo claims by substituting PICO elements from an existing corpus.

In Section 3.4 we emphasize the need to evaluate retrieved evidence relevant to *naturally occurring* medical claims, as opposed to the *pseudo* data we derived for training. To this end, we hired domain experts (medical doctors) to look at the evidence abstracts from our retrieval model and assign a relevance score to each abstract (3: relevant, 2: somewhat relevant, 1: irrelevant). We provide some examples of retrieved evidence in Table 14 annotated by our experts as *relevant* (score: 3). Due to space constraints, we provide a link to the full article instead of the full abstract text.

	Claims			Experiences			Questions		
	F1	P	R	F1	P	R	F1	P	R
CRF (Lafferty et al., 2001)	33.87	35.61	32.29	40.08	40.52	39.64	86.89	85.55	88.27
BERT (Devlin et al., 2019)	**52.63**	58.82	47.61	56.68	59.46	54.33	92.39	88.76	96.34
BioRedditBERT (Basaldella et al., 2020)	45.16	**70.92**	33.29	**59.51**	**62.49**	**58.92**	**93.61**	**89.29**	**98.37**
RoBERTa (Liu et al., 2019)	47.05	61.53	**38.09**	56.81	57.11	56.52	93.06	89.01	98.34

Table 10: Additional results from the test set for the task of identifying spans of Claims, Experiences, and Questions.

	P	R	F1	$F1_{POP}$	$F1_{INT}$	$F1_{OUT}$
BERT (Devlin et al., 2019)	43.88	36.13	39.62	41.77	44.68	33.05
RoBERTa (Liu et al., 2019)	**47.45**	**39.27**	**42.97**	**46.09**	**45.99**	**36.38**
BioRedditBERT (Basaldella et al., 2020)	44.44	36.55	40.12	41.92	44.31	34.61
biomedRoBERTa (Gururangan et al., 2020)	38.80	21.48	27.66	30.54	28.13	24.54
t5-small (Raffel et al., 2020)	41.49	38.55	39.97	39.61	45.02	32.41

Table 11: Additional results from the test set for the token-level PIO labelling task.

	MRR @k					Precision @k				
k	1	5	10	50	100	1	5	10	50	100
random	0.00	0.003	0.02	0.02	0.02	0.00	0.02	0.00	1.10	2.80
BM25	5.34	7.98	9.86	14.36	16.70	5.34	10.40	14.45	26.20	33.14
DPR (Karpukhin et al., 2020)	8.07	10.96	11.89	12.20	13.77	8.07	16.50	23.58	31.98	36.87
*(trained on the **RedHOT** pseudo training set)*										
RedHOT-DER (BERT-based)	39.14	47.99	49.3	50.28	50.35	39.14	62.55	72.64	83.73	91.74
RedHOT-DER (RoBERTa-based)	**45.93**	**54.60**	**55.90**	**56.73**	**56.78**	**45.93**	**69.90**	**78.81**	**94.73**	**98.06**

Table 12: Additional results from the retrieval task (tested on the pseudo test set).

Table 13: Examples of template claims used for the creation of pseudo training labels for training a supervised evidence retrieval model.

	Original w/ PIO placeholders (Template)	w/ Substituted PIO elements (Pseudo)	Population
Claim	Global spread of [OUT] blamed on [INT]	Global spread of Gradual deterioration of renal function blamed on cyclophosphamide	Lupus
Post	Global spread of [OUT] blamed on [INT]	Global spread of Gradual deterioration of renal function blamed on cyclophosphamide	
Claim	I'll be starting [INT] soon and have heard/been told it can cause some serious side effects when first starting to take it.	I'll be starting solriamfetol treatment soon and have heard/been told it can cause some serious side effects when first starting to take it.	Narcolepsy
Post	I'll be starting [INT] soon and have heard/been told it can cause some serious side effects when first starting to take it. Because of this, I let my employer know I may have to be out for a day or two during busiest time of the year, and I'm worried I overshared.	I'll be starting solriamfetol treatment soon and have heard/been told it can cause some serious side effects when first starting to take it. Because of this, I let my employer know I may have to be out for a day or two during busiest time of the year, and I'm worried I overshared.	
Claim	I read that [OUT] could be due to [POP].	I read that hip and lumbar bone mineral density differences could be due to Ankylosing Spondylitis.	Ankylosing Spondylitis
Post	I'm 40M with [POP] and UC and my annual blood work just came back [OUT] (around 2.5). However, my other blood levels are all fine, I eat well, am relatively thin (BMI 24), exercise a lot. I read that [OUT] could be due to [POP].	I'm 40M with Ankylosing Spondylitis and UC and my annual blood work just came back hip and lumbar bone mineral density differences (around 2.5). However, my other blood levels are all fine, I eat well, am relatively thin (BMI 24), exercise a lot. I read that hip and lumbar bone mineral density differences could be due to Ankylosing Spondylitis.	
Claim	Surprising I'm seeing research articles that [INT] causes [OUT] past the duration of the drug	∗ Surprising I'm seeing research articles that quetiapine versus aripiprazole causes psychopathology, cognition, health-related quality of life, and adverse events past the duration of the drug. ◇ Surprising I'm seeing research articles that IPS causes levels of stress past the duration of the drug • Surprising I'm seeing research articles that olanzapine causes discontinuation rate past the duration of the drug	Psychosis
Post	Surprising I'm seeing research articles that [INT] causes [OUT] past the duration of the drug. I only took a brief look into it. Has anyone here had [INT] induced [OUT]? What is r/psychosis experience with [INT]?	∗ Surprising I'm seeing research articles that quetiapine versus aripiprazole causes psychopathology, cognition, health-related quality of life, and adverse events past the duration of the drug. I only took a brief look into it. Has anyone here had quetiapine versus aripiprazole induced psychopathology, cognition, health-related quality of life, and adverse events? What is r/psychosis experience with quetiapine versus aripiprazole? ◇ Surprising I'm seeing research articles that IPS causes levels of stress past the duration of the drug. I only took a brief look into it. Has anyone here had IPS induced levels of stress? What is r/psychosis experience with IPS? • Surprising I'm seeing research articles that olanzapine causes discontinuation rate past the duration of the drug. I only took a brief look into it. Has anyone here had olanzapine induced discontinuation rate? What is r/psychosis experience with olanzapine?	

Table 14: Examples of evidence abstracts (marked relevant by domain experts) retrieved by the RoBERTa-based **RedHOT-DER** model trained on pseduo data.

	Title of trial paper	Link to abstract/trial
Claim: Vitamin D may prevent autoimmune diease **Post:** Okay so... the only bloodwork for me that was pretty abnormal was vitamin D. My neurologist did bloodwork for it a year ago and it was in the 20s. He said it should be 50+ and that *Vitamin D may prevent autoimmune diease*. Are there any long term problems I should be aware about if I can, how get it to go up?	Vitamin D and marine omega 3 fatty acid supplementation and incident autoimmune disease: VITAL randomized controlled trial.	https://dx.doi.org/10.1136/bmj-2021-066452
Claim: been researching few weeks now and I recently came across POTS Syndrome. I found that it affects your heart rate so I decided to test mine while resting and then standing to see if maybe thats what it could be. **Post:** I just joined this group, so I apologize if this is not allowed. I have been researching what I feel to be abnormal symptoms Ive been dealing with the majority of my life (dizziness, nausea when standing, etc).. Anyways, Ive *been researching few weeks now and I recently came across POTS Syndrome. I found that it affects your heart rate so I decided to test mine while resting and then standing to see if maybe thats what it could be.* I took my heart rate three times while laying in bed. at 1:37am, by heart rate was 73bpm. at 1:39am, my heart rate was 74bpm. at 1:40am, my heart rate was 73 bpm again. I then stood up (right next to my bed) and proceeded to take my heart rate again. Immediately it shot up to more than double my resting heart rate at 1:41am my heart rate was 156bpm i took it again a minute later and at 1:42am my heart rate was 153bpm. Even if its not pots, just from standing up, I feel like this is not a normal bodily response for the majority of the population. Dont know how to go about getting this checked out. By the way, not sure if it matters, but I am a 19 year old girl.	Cardiovascular exercise as a treatment of postural orthostatic tachycardia syndrome: A pragmatic treatment trial.	https://dx.doi.org/10.1016/j.hrthm.2021.01.017
Claim: did some research and apparently smoking can effect bowel movements (bloating,cramping) which is what i struggle with exactly **Post:** i have been a smoker for only 3 years, and i recently had the realisation that my IBS(like) symptoms correlated to the same period of time i started smoking, i then did some research and apparently smoking can effect bowel movements (bloating,cramping) which is what i struggle with exactly, so i dont know if anyone has a similar story or if quitting smoking helped with their IBS ?	The effect of alpha-tocopherol and beta-carotene supplementation on colorectal adenomas in middle-aged male smokers.	https://www.ncbi.nlm.nih.gov/pubmed/10385137
Claim: I read of the issues it can cause the body but so much out there has it. **Post:** sugar alcohol vs sugar Just wondering what your thoughts are of sugar alcohol. I noticed a lot of sugar free foods have sugar alcohol inplace of sugar. I read of the issues it can cause the body but so much out there has it. Do you avoid sugar alcohol products or do you embrace it as a sugar alternative?	Glycemic Effects of Rebaudioside A and Erythritol in People with Glucose Intolerance.	https://dx.doi.org/10.4093/dmj.2016.40.4.283
Claim: I cant help thinking it may be related to my meds **Post:** I stopped taking Levothyroxin for about a month. Ever since I started taking it again I feel like crying after taking it in the mornings. It could be that I really dont want to go to work, but I cant help thinking it may be related to my meds. Does this happen to anyone else?	Clinical Observation of Levothyroxine Sodium Combined with Selenium in the Treatment of Patients with Chronic Lymphocytic Thyroiditis and Hypothyroidism and the Effects on Thyroid Function, Mood, and Inflammatory Factors.	https://dx.doi.org/10.1155/2021/5471281