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

Scope of Pre-trained Language Models for Detecting Conflicting Health Information

Joseph Gatto, Madhusudan Basak, Sarah Masud Preum

Department of Computer Science, Dartmouth College

{joseph.m.gatto.gr, madhusudan.basak.gr, sarah.masud.preum}@dartmouth.edu

Abstract

An increasing number of people now rely on online platforms to meet their health information needs. Thus identifying inconsistent or conflicting textual health information has become a safety-critical task. Health advice data poses a unique challenge where information that is accurate in the context of one diagnosis can be conflicting in the context of another. For example, people suffering from diabetes and hypertension often receive conflicting health advice on diet. This motivates the need for technologies which can provide **contextualized, user-specific health advice**. A crucial step towards contextualized advice is the ability to compare health advice statements and detect *if and how they are conflicting*. This is the task of health conflict detection (HCD). Given two pieces of health advice, the goal of HCD is to detect and categorize the type of conflict. It is a challenging task, as (i) automatically identifying and categorizing conflicts requires a deeper understanding of the semantics of the text, and (ii) the amount of available data is quite limited.

In this study, we are the first to explore HCD in the context of pre-trained language models. We find that DeBERTa-v3 performs best with a mean F1 score of 0.68 across all experiments. We additionally investigate the challenges posed by different conflict types and how synthetic data improves a model's understanding of conflict-specific semantics. Finally, we highlight the difficulty in collecting real health conflicts and propose a human-in-the-loop synthetic data augmentation approach to expand existing HCD datasets. Our HCD training dataset is over 2x bigger than the existing HCD dataset and is made publicly available on Github.

Introduction

In recent years, quick and easy access to online health information has changed the way people eat, exercise, and interact with medical professionals (Bujnowska-Fedak and Wegierek 2020). As we continue to live through a global pandemic, people are increasingly relying on online resources for health advice (Loomba et al. 2021). However, a significant limitation of online health platforms is their inability to provide consistent, contextual health advice. For instance, while many information sources promote taking daily aspirin after a heart attack, this advice is dangerous to those on warfarin, as combining the medicines can have life-threatening side-effects (See example 5 in Table 1). Moreover, as new clinical evidence emerges, some health advice will inevitably become outdated, generating conflicts with recent findings. Such situations motivate the need for automatic detection of conflicting health information across multiple relevant sources to facilitate safer interactions with online health platforms. This task is referred to as **health conflict detection** (HCD).

Given two pieces of health advice statements, the goal of HCD is to both identify the presence of a conflict *and* to recognize what type of conflict occurs [1]. Different types of conflicts are presented in Table 1. Recognizing the conflict type will aid the user in interpreting the reason for a conflict correctly. For example, consider sample 5 in Table 1. Simply expressing that a conflict is present does not guide the user on how to react to the conflicting information on aspirin. Explicitly stating the conflict type communicates how one should be careful regarding *if* (conditional) and *when* (temporal) aspirin should be taken. With more people relying on online health websites and the surge of a global infodemic (Buchanan 2020), detecting inconsistent/conflicting health information at large is even more critical now, as widespread health misinformation is a risk to public health and extremely expensive to combat (Cornish 2020).

Conflicting health information appears in many forms across the internet. HCD was first introduced in (Preum et al. 2017a,b) where the authors introduce a novel HCD dataset using data extracted from several health apps and medical information websites. More recently, the need for conflict detection on other sources, including social media, has presented itself. Consider the sub-reddit r/AskDocs, a forum with 380k+ members where anyone can ask questions to a verified medical professional for free. In Table 9 we highlight conflicting pairs of advice statements from registered nurses and licensed physicians identified on r/AskDocs to illustrate the need for more reliable and consistent social interaction on health.

[1]HCD does not aim at discounting a source of information as health conflicts can occur between a pair of advice even when both advice statements are correct (Carpenter et al. 2016). Also, conflict resolution is beyond the scope of HCD as it requires the expertise of a professional healthcare provider and is often subjective to an individual.

	Advice 1	Advice 2	Conflict Type
1	Limit intake of sweetened drinks, **snacks** and desserts by eating them less often and in smaller amounts.	Many fruits need little preparation to become a healthy part of a meal or **snack**.	Direct
2	Limit **liquids** before bed.	To prevent dehydration, drink plenty of **fluids** unless your doctor directs you otherwise.	Temporal
3	Pasteurized dairy products, **milk**, cheese, yogurt, spinach are rich in calcium	Dilute full-fat **milk** in your cereal with water. This reduces the absorption of sugar and decreases fat intake.	Sub-typical
4	Eat dark green vegetables such as spinach, **collard greens, kale and broccoli**	Avoid sudden increase of **cruciferous vegetables** if you are on Coumadin. They may affect your treatment and dosage	Conditional
5	**Aspirin** helps prevent future blood clots and decreases the risk of death after a heart attack.	**Aspirin**, aspirin-like drugs (salicylates), and nonsteroidal anti-inflammatory drugs ... may have effects similar to warfarin. These drugs may increase the risk of bleeding problems if taken during treatment with warfarin.	Conditional, Temporal

Table 1: Example health conflict pairs from the HCD dataset. Each sample contains a pair of health advice regarding a common topic. The conflict label is annotated with respect to the common topic between pieces of advice. Conflict topics are bold in the examples above.

In this study, we build on top of the initial HCD work proposed in (Preum et al. 2017a,b) by focusing on a scaleable approach to HCD data expansion. (Preum et al. 2017a,b) use a rule based solution to detect health conflicts. Although the proposed solution results in high recall the following issues limit its applicability to detect conflicting health information at large. First, rule base solutions are brittle and harder to generalize. Secondly, their solution requires a prohibitively expensive amount of annotation for each piece of advice and is hard to scale to other data. So, the next step in the advancement of HCD is to develop an end-to-end system which does not require granular levels of annotation and can work in a production environment on free-text inputs. The recent advances in transfer learning inspire us to investigate the scope of pre-trained language models to detect health conflicts automatically. Successful implementation of a transfer learning based solution can effectively address this critical challenge of automatically detecting inconsistent/conflicting health information at large.

One of the main challenges with using pre-trained language models for HCD is that they require sufficient labeled data for fine-tuning while the HCD task is inherently resource-constrained. For example, given a set of 500 health advice statements, there exists $(500)^2 = 250000$ candidate advice pairs from which a very small set will be conflicting — making data collection extremely difficult. Given the limited number of health advice pairs presented in (Preum et al. 2017a,b), this study explores methods of *synthetic conflict augmentation* towards making HCD suitable for use in a transfer learning framework. After expanding the original HCD training set by 2x, we explore the task in the context of pre-trained language models. Finally, to verify prediction reliability, we test if transformer models value the same information as human annotators when asked to identify the content most important to the prediction of a conflict. In summary, we considering the following research questions:

RQ1: How Effective Pre-trained Language Models Are in Detecting Conflicting Health Information? We provide thorough analysis of five different language models on the HCD task, investigating the challenges associated with understanding different conflict types. We additionally contrast our work with a related task, Natural Language Inference (NLI), which aims to classify pairwise inputs as entailed, contradictory, or neutral (Williams, Nangia, and Bowman 2018). We investigate the relationship between notions of conflict and contradiction by applying a pre-trained NLI classifier to the HCD task.

RQ2: Can Synthetic Health Conflicts Aid in Predicting Real Health Conflicts?

While synthetic conflicts are shown to be linguistically and semantically valid, they are often factually incorrect and contain distributional drift in terms of style. So, we explore the capability of synthetic conflicts to aid in predicting real conflicts.

RQ3: Do Language Models Agree with Humans When Identifying Features Most Influential to Conflict Detection? The interpretability of AI systems is crucial to the integration of machine learning in healthcare systems (Tjoa and Guan 2019). We use a popular model interpretability tool, Captum (Kokhlikyan et al. 2020), to identify the input tokens our top-performing language model attributes as being most important to the prediction of a given output. In addition, we explore if language models pay attention to known conflict indicators such as the conflict topic and label-specific semantic phrases (LSSPs) (e.g. temporal or conditional clauses that result in a temporal or conditional conflict).

Dataset	Train	Test
Real	1398	608
Synthetic	1427	470

Table 2: Number of samples in each split of the HCD dataset.

Problem Formulation

Inspired by the definition presented in (Preum et al. 2017a), we define the input to a health conflict detection (HCD) task as a pair of health advice, (W_1, W_2) each with a common topic or object of an advice, t. HCD is a multiclass, multi-label classification problem. Each advice pair contains one or more of the following labels.

Direct Conflict: Occurs when the topic of a health advice pair has opposite polarity. If a piece of advice W_i suggests the user take action a, a direct conflict occurs when another piece of advice W_j suggests the user against taking action $\neg a$. In Table 1, our direct conflict example has topic **snack**. We see how advising fruit as a healthy snack directly conflicts with the suggestion to limit sweet snacks.

Temporal Conflict: Occurs when a health advice pair disagrees about *when* to take some action a. In Table 1, our temporal example agrees one should drink liquids, but disagrees as to when consumption should occur.

Sub-Typical Conflict: Occurs when the topic in reference to the suggested action disagree in type. In Table 1, we see that Advice 1 supports milk consumption while Advice 2 speaks negatively only in regards to *full-fat* milks.

Conditional Conflict: Occurs when a conflict's occurrence depends on some condition. In Table 1, our conditional example displays how Advice 1 and Advice 2 are only conflicting if the user is taking Coumadin.

Non-Conflict: Occurs when two pieces of advice are non-conflicting.

We note that this list of conflict types is non-exhaustive, as we do not explore the quantitative and cumulative conflicts defined in (Preum et al. 2017b), due to limited available annotation. Such conflicts may be explored in future works.

Dataset

The following section is organized as follows. First, we provide an overview of existing HCD data. We then discuss our attempts at performing data augmentation on existing HCD samples. Next, we discuss how new data was collected. Finally, we detail our human-in-the-loop augmentation approach, which alleviates problems faced during our attempts at data generation and augmentation. Additional dataset details are included at the end of the section.

Real Health Advice Dataset from *PreCluDe* The PreCluDe dataset contains 3285 pieces of advice statements from authentic health websites, health apps, and drug usage guidelines. The data covers general health topics, 34 chronic diseases, and their relevant prescription medications collected from authentic and reliable health websites and popular health apps. The list of diseases and medications are selected based on a set of real prescriptions for multimorbidity. These advice statements can be single sentence or multi-sentence. Each of these advice statements were annotated for topics of advice (e.g., advice on a food item, drug-drug interaction, drug-food interaction, exercise). In addition each advice was annotated based on the polarity of the advice with respect to each topic. Then potential pairs of conflicting advice statements were generated based on common topics. At least three reviewers then annotated each of these potential pairs of conflicting advice statements to determine whether they are conflicting and the type of potential conflicts (if any). The details of this annotation process are described in multiple published literature (Preum et al. 2017c, 2018). The PreCluDe dataset is the source of the 2006 *real* conflicting advice pairs utilized in this study.

Conflict Type	Augmentation Strategy	F1
Direct	None	0.67
	Paraphrase	0.66
	Text Generation	0.62
	Back Translation	0.65
	Aggregated Set	0.63
Sub-Typical	None	0.15
	Paraphrase	0.08
	Text Generation	0.19
	Back Translation	0.09
	Aggregated Set	0.13
Conditional	None	0.40
	Paraphrase	0.25
	Text Generation	0.25
	Back Translation	0.14
	Aggregated Set	0.38
Temporal	None	0.69
	Paraphrase	0.70
	Text Generation	0.57
	Back Translation	0.59
	Aggregated Set	0.69

Table 3: The F1 score of DeBERTa-v3 when trained using additional data using various popular augmentation strategies. These results highlight the complexity involved in augmentation of pair-wise multi-sentence inputs and the need for more advanced augmentation algorithms.

Synthetic Dataset Generation Using Human-in-the-Loop Data Augmentation In an attempt to expand the size of existing HCD data, we first explored various automated methods of textual augmentation. The result of these experiments can be found in Table 3. Our first experiment employed PEGASUS (Zhang et al. 2019) for paraphrase generation. Given some Advice 1, we paraphrase its corresponding Advice 2 to generate a new sample. Unfortunately, this method was prone to removing or altering the conflict topic, producing incorrect samples. We additionally tried a similar approach using Back Translation between German and English, but little textual diversity was observed. We also

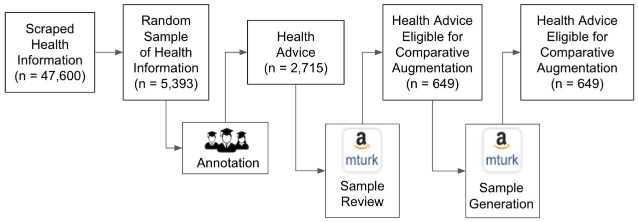

Figure 1: Data flow diagram depicting the data collection process.

explored a text generation approach using GPT-Neo (Black et al. 2021), where we employed in-context learning to generate new samples, with prompt structure inspired by GPT3-MIX (Yoo et al. 2021). Specifically, given a prompt and a small set of example advice pairs, GPT was asked to generate Advice 2 for a given Advice 1. Unfortunately, this method failed to produce coherent outputs. In each experiment, we double the dataset size (thus generating n = 1398 samples) using the described augmentation methods. We also report a result denoted Aggregated Set, an evenly distributed mix of all three augmentation approaches.

New Online Health Advice Dataset To expand upon the existing HCD dataset with additional real-world samples, we scraped 4 popular health information websites, including WebMD[2], MedLine[3], Covid protocols[4], and CDC[5]. We collected a total of 47,600 candidate advice statements from a wide variety of health topics. Three annotators were then tasked with annotating a random sample of 5393 advice statements, each answering the following questions: 1) Does this statement contain health advice? 2) What are the topics of the advice? 3) What is the polarity of each topic in a piece of health advice? This allowed us to generate potential conflicting advice pairs by considering samples with common topics containing opposite polarities.

The annotators found 2715 samples that were in fact health advice and thus eligible for inclusion in a conflicting advice pair. Unfortunately, very few useful organic conflicts were found when pairing together samples with opposite topic polarities. This occurred because 1) there is a wide variety of topic entities annotated, leaving few opportunities for sample pairing, 2) many topics have inherent static polarity (e.g. cancer is always negative), which makes finding a conflict for such samples difficult. We thus concluded that automatic conflict generation using randomly sampled, real-world text was not a feasible strategy to *significantly* increase the dataset size.

Human-in-the-Loop Data Augmentation The failure of the two approaches above to generate data inspired us to take a human-in-the-loop approach for data augmentation using mTurk. This approach allows us to generate training samples that encode the relevant linguistic phenomena associated with each conflict type, even if the style and prose of human-written samples do not perfectly reflect real-world data. Although the synthetic data generated in this process is often not clinically accurate, they help capture the linguistic phenomena of conflicting textual information. In this study, we recruit 88 master-qualified mTurk workers. Additionally, we only recruited US-based annotators with a minimum 90% acceptance rating. To generate a synthetic advice conflict, an mTurk worker (Turker) is provided with a real health advice statement, the topic of the advice, and instructions on how to write samples for each conflict type. For example, a Turker may be prompted with the advice statement "Drink less alcohol." If asked to write a conditional conflict, a valid synthetic sample would be "Drink less alcohol *if you suffer from heart disease*". Full details of the data collection process are depicted in Figure 1.

Of the 2715 newly collected health advice statements deemed useful, an additional round of human review extracted 649 advice statements most suitable for augmentation. Each sample was annotated for each conflict type, and then **verified by at least one research assistant at the authors institution.** The resulting dataset after sample verification contains 1897 unique advice statements from Amazon's mTurk platform. Finally, we display the distribution of health domains in the synthetic dataset in Table 4.

Advice Topic	Number of Samples
Adhd	19
Allergies	52
Alzheimers	59
Arthritis	115
Asthma	73
Diabetes	51
Druginfo	44
Health topics	32
Hypertension	86
Pregnancy	92
Other	26

Table 4: Distribution of health advice domains in the newly collected data.

Dataset Details In total, there are 2006 real annotated advice pairs, and 1897 synthetic pairs. Our training set contains 2825 samples, comprised of a mix of real and synthetic advice pairs. We evaluate our model on two test sets: One with entirely real-world data, and another which contains only synthetic advice pairs. The full data distribution is shown in Table 2. The ratio of single-sentence advice to multi-sentence advice is 3.07. The length of a multi-sentence advice can vary from 2 to 11 sentences. On average, one health advice is 1.4 sentences long, containing close to 25 words. Given that HCD is a pairwise task, each input thus requires modeling of about 2.8 sentences, demanding long-range dependency modeling.

[2]https://www.webmd.com/a-to-z-guides/common-topics

[3]https://medlineplus.gov/healthtopics.html

[4]https://covidprotocols.org/en/chapters/home-and-outpatient-management/

[5]https://www.cdc.gov/DiseasesConditions/

Methods

RQ1: How Effective Pre-trained Language Models Are in Detecting Conflicting Health Information?

To investigate the performance of pre-trained language models on the HCD task, we aim first to quantify performance when fine-tuning transformer-based models for HCD. We fine-tune 5 different models for each architecture in our experiments. Four out of these five models are 1-vs-all classifiers for each conflict type, with the fifth being a multi-label classifier. We then report results on two different test sets, each containing exclusively real or synthetic data. These experiments will identify (i) which conflict-types are most and least challenging to detect and (ii) allow us to quantify the difficulty level in predicting real vs synthetic test samples.

We additionally explore the relationship between conflict detection and the relevant NLP task of Natural Language Inference, which includes *contradiction detection*. To do so, we apply a DeBERTa-v3 model fine-tuned on the MNLI dataset (Williams, Nangia, and Bowman 2018) and consider contradictory predictions to be conflicts in our 1-vs-all HCD setting. This experiment aims to highlight the differences between conflicts and contradictions and motivate the need for HCD-specific algorithms. We additionally experiment with intermediate fine-tuning (Chang and Lu 2021) on MNLI, before fine-tuning on HCD. This experiment aims to explore if leveraging the subtle relationships between MNLI and HCD provides any boost in predictive performance.

RQ2: Are Synthetic Health Conflicts Able to Aid in the Prediction of Real Health Conflicts?

We have thus far motivated the need for better health conflict detection solutions, which will require more data to train reliable end-to-end models. Thus, RQ2 measures the effectiveness of the human-in-the-loop synthetic data augmentation technique. This experiment evaluates the predictive power of synthetic training samples on real-world health conflicts. To evaluate, we train 3 separate classifiers on each conflict type, using real-only, synthetic-only, and real+synthetic training data respectively.

RQ3: Do Language Models Agree with Humans When Identifying Features Most Influential to Conflict Detection?

With the long-term goal of deploying a scaleable HCD system to alert health information consumers of potential conflicts, we aim to explore the reliability of transformer-based models by taking a low-level look at their predictions on the HCD task. To do so, we use the Layer Integrated Gradients algorithm (Sundararajan, Taly, and Yan 2017) provided by Captum (Kokhlikyan et al. 2020), to quantify the attribution of each input token to the prediction of the conflict label.

Specifically, Captum allows us to produce a numeric attribution score for each input token. This enables visualization of token importance by toggling the opacity of highlighted tokens based on the magnitude of their attributions. An example Captum output for each conflict type is shown in Figure 2. Tokens highlighted in Green contribute positively to the prediction of the positive class. Tokens highlighted in red contribute positively towards the prediction of the negative class. In this experiment, we employ 3 human annotators experienced in health conflict detection to answer the following two questions about the tokens Captum identifies as most important to the prediction of the positive class.

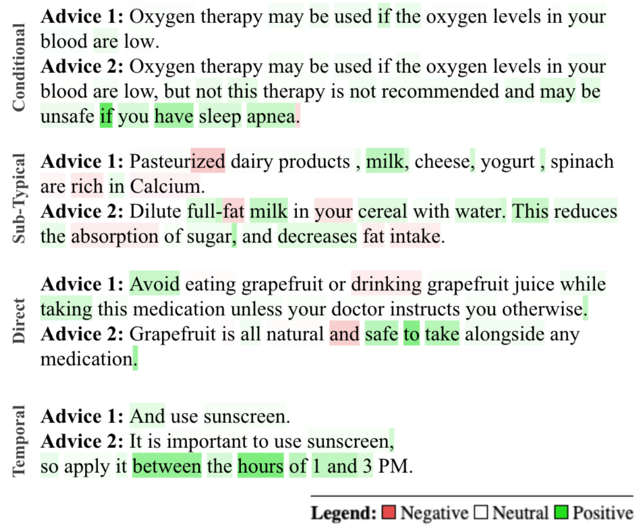

Figure 2: Visualizing sample outputs of token importance attributions from the Captum library.

1. Does the model identify the topic as important to classification?
2. Does the model identify the correct Label-Specific Semantic Phrase (LSSP) as important to classification? Where LSSPs are defined as the topic-altering tokens which generate a specific conflict.

In other words, the annotator is asked to look at a given sample, gather their understanding of the correct Topic and LSSP, and confirm their intuitions align with the transformer model by observing which tokens DeBERTa-v3 highlights in green. Given the importance of model interpretability for health data (Vellido 2020; Dai, Sun, and Wang 2020), this experiment will verify if transformers are making judgements which align with human intuitions. In this experiment, each annotator looks at 10 samples from each class, from both real and sythetic test sets (resulting in 80 samples in total). All samples were selected randomly from the set of correct predictions made by the DeBERTa-v3 model. Each annotator provides a binary response (Yes/No) for each question, and the average score is reported. The Fleiss Kappa (Fleiss 1971) amongst the three annotators for all 160 annotations (80 Topic + 80 LSSP annotations) is 0.58.

Evaluation Setting

To ensure results are robust to random weight initializations, we run each experiment 3 times, each with a different random seed. Each reported result is the macro average of all experiments. For each individual experiment, we compute the F1 score of the positive class in a 1-vs-All setting or the

Conflict Type	**NLI**$_{OTS}$	**NLI**$_{FT}$	DeBERTa-V3
Direct	0.53	0.69	0.75
Sub-Typical	0.12	0.29	0.36
Conditional	0.08	0.69	0.53
Temporal	0.05	0.81	0.70
Average	0.19	0.62	0.58

Table 5: Experiments exploring the relationship of HCD and NLI. NLI$_{OTS}$ identifies how well an off-the-shelf (OTS) NLI model predicts conflicts. NLI$_{FT}$ highlights the impact of intermediate fine-tuning (FT) DeBERTa-V3 on the MNLI task. i.e. this model is first fine-tuned to MNLI, then fine-tuned to HCD. DeBERTa-V3 is the reference HCD score with no NLI training. All results are mean F1 score over 3 experimental trials.

weighted F1 in the multi-label setting. All metrics are computed using the scikit-learn package (Pedregosa et al. 2011). We note that we are unable to perform cross-validation due to the fact that a given advice statement may be paired with multiple other advice statements. Thus, performing cross validation runs the risk of input memorization as has been observed in similar pairwise inference tasks (Gururangan et al. 2018) (Herlihy and Rudinger 2021). This issue additionally inhibits our ability to construct a separate validation set. For example, we cannot simply take 200 samples out of the real test set and create a validation set as many pieces of advice are used in multiple conflicting advice pairs — which would cause data leakage. Thus, hyperparameter optimization was considered out-of-scope of this work but should be explored in future works.

Experiments using static embedding methods, namely Sentence-BERT (SBERT) (Reimers and Gurevych 2019) and GloVE (Pennington, Socher, and Manning 2014) each function by first computing a single embedding vector for each piece of advice. Next, the embeddings for Advice 1 and Advice 2 are concatenated and fed into a linear classification head using scikit-learn. In these experiments, only the classification head undergoes training. Each dynamic contextual embedding model (BERT, Bio+Clinical BERT, RoBERTa, DeBERTa-v3) is fine-tuned using the Transformers library [6]. We use the pair-wise classification paradigm outlined in (Devlin et al. 2018) for a fixed 5-epochs with the AdamW (Loshchilov and Hutter 2017) optimizer set with learning rate = $2e-5$ and a weight decay = 0.01 to combat overfitting. We use the standard Cross Entropy loss for each dynamic training task, where in the 1-vs-All experiments the loss is weighted using the distribution of the training labels to account for class imbalance.

Results

RQ1: How Effective Pre-trained Language Models Are in Detecting Conflicting Health Information?

Model Analysis We report the performance of 5 different transformer-based classifiers in Table 6 as well as a random guess, lexical feature, and word embedding-based baseline. The lexical model, which uses TFIDF (Manning 2009) for feature generation with a class-weighted Random Forest classifier (Breiman 2001), under-performs random guess on 90% of experiments across all conflict types. This is expected as TFIDF generates features based on word count and frequency, ignoring the fact that the input text is comprised of two disjoint advice statements which need to be *compared*. Additionally, we note that the Random Forest model was unable to overcome the severe class imbalance with TFIDF features, often over-predicting the non-conflicting class for all labels.

The static embedding methods, SBERT and GloVE, provide an improvement over our lexical baseline while outperforming random guess in most experiments. However, static embedding methods are not constructed to perform pair-wise comparison of longer texts. Thus, all textual comparison must happen in the classification head, which is shown to be inferior to the dynamic embedding approaches shown in Table 6.

Transformer-based models such as BERT provide significant improvement over both TFIDF and static embeddings, as their contextual embedding approach permits direct comparison of two texts. The BERT baseline, however, is shown to be inferior to more modern transformer models such as RoBERTa and DeBERTa-v3, both of which outperform BERT on 80% of experiments. From our experiments using Bio+Clinical BERT (B+C BERT), we show that BERT's performance issues are not due to BERT's lack of domain specific training data, as B+C BERT under-performs or matches BERT's performance on 70% of experiments. This again follows intuition as, while this task is called health conflict detection, the addition of knowledge from sources like biomedical corpora and clinical notes are intuitively not relevant to this task. Health advice contains common language alongside occasional references to medications or physiological effects (as shown in Table 1).

Across all experiments for which scores were higher than random guess, results on the synthetic test set outperform results on real health conflicts — often by a significant margin. This confirms our understanding of synthetic data outlined in the dataset description - where we hypothesize that non-organic conflict data may be easier to understand as they often contain simpler and less diverse sentence semantics, using phrases specific to label definitions (as discussed in the *Dataset* Section).

Label Specific Performance Analysis

Direct Conflicts: Our results show there is a variance in the data distributions between real and synthetic direct conflicts, as performance on their respective test sets has the highest discrepancy across all conflict types. We additionally note the *significant* improvement provided by DeBERTa-v3 on direct conflict understanding. We believe the DeBERTa-v3 architecture is well-suited for this conflict type for the following reasons. Direct conflicts, by definition, will not contain overt textual clues in a single advice statement which facilitate better label prediction. This in contrast to temporal conflicts, for example, where a statement such as "be-

[6] https://github.com/huggingface/transformers

	Direct		Sub-Typical		Conditional		Temporal		Multilabel		
	F_r	F_s	F_r	F_s	F_r	F_s	F_r	F_s	F_r	F_s	***Avg***
Random Guess	0.39	0.28	0.24	0.25	0.40	0.33	0.24	0.30	0.35	0.32	0.31
TFIDF+RF	0.07	0.36	0.00	0.00	0.00	0.25	0.21	0.36	0.03	0.18	0.14
GloVE	0.51	0.34	0.21	0.19	0.58	0.46	0.40	0.64	0.42	0.43	0.41
SBERT	0.34	0.55	0.17	0.25	0.31	0.50	0.49	0.72	0.37	0.51	0.42
BERT	0.38	0.73	0.28	0.18	0.44	**0.86**	0.53	**0.91**	0.40	0.70	0.54
B+C BERT	0.36	0.70	0.24	0.23	0.45	0.84	0.66	0.89	0.34	0.70	0.54
RoBERTa	0.51	0.77	0.30	0.39	**0.58**	0.85	**0.72**	0.90	**0.56**	**0.79**	0.63
DeBERTa-v3	**0.75**	**0.85**	**0.36**	**0.71**	0.53	0.83	0.70	**0.91**	0.51	0.71	**0.68**

Table 6: Average F1 score across 3 experimental trials using all training data to predict each conflict type on the real (F_r) and synthetic (F_s) test sets respectively.

fore bed" found in Advice 2 is a strong indicator that there *may* be a temporal conflict independent of Advice 1. Thus, in order to understand direct conflicts, long-range dependency modeling is required, where a common topic must be identified in both advice statements as well as their relative polarities. DeBERTa-v3 improves upon BERT and RoBERTa in a variety of ways that facilitate long-range dependency modeling. Specifically, DeBERTa-v3 uses (i) disentangled position embeddings, (ii) global position information, and (iii) replaced token detection pre-training, which has been shown to be a more efficient pre-training objective than those used by BERT and RoBERTa (He, Gao, and Chen 2021a). We find evidence of DeBERTa-v3's improved ability to reason about pair-wise inputs in their paper where DeBERTa-v3 outperforms BERT on the pairwise Recognizing Textual Entailment task by an astounding 20.3 accuracy points (He, Gao, and Chen 2021a). We visualize the relationship between input length and F1 score in Figure 3, where we observe how BERT and RoBERTa are much more sensitive to token length than DeBERTa-v3 is for direct conflicts.

A common error made when predicting Direct Conflicts was the over-prediction of other conflict samples which contained opposite topic polarities alongside conditional or temporal constraints. For example, if Advice 1 has positive sentiment towards coffee yet Advice 2 expresses negative sentiment towards coffee *only for pregnant women*, false positives tend to occur as the model is unable to understand that the opposing sentiments are tied to a conditional constraint.

Sub-Typical Conflicts: Our experimental results find that sub-typical conflicts are the most challenging to predict. Similar to direct conflicts, sub-typical conflicts by definition have less-overt textual clues than conditional and temporal conflicts. However, unlike direct conflicts, which rely on understandings of polarity with respect to each topic, subtypical conflicts often rely on world-knowledge for proper understanding. For example, the advice pair ("Don't eat cheese", "Only some cheeses are bad for you, like Goat cheese") contains a sub-typical conflict regarding cheese type. It may be easy for a pre-trained language model to understand the topic (cheese) and sentence polarities (negative, negative), but challenging to know that "goat" is a type of cheese, and that the input isn't referring to the animal itself, but rather altering the topic-type.

Temporal and Conditional Conflicts: Pre-trained language models find the most success in the prediction of conditional and temporal conflicts. This is particularly true of the synthetic test sets, where mTurk workers could add *if* or similar conjunctions or temporal clauses to the end of the original advice statements to generate conditional or temporal conflicts, respectively. Thus such samples are less challenging for the model to classify correctly. As previously stated, these conflict types contain more consistent textual patterns that make their detection easier. Interestingly, these two conflict types are the only experiments where DeBERTa-v3 does not outperform other pre-trained language models. This is likely due to the lesser dependency on long-range dependency modeling required by these two conflict types.

With these two conflict types, the presence of label-specific semantic phrases in non-conflicting samples can cause false positives. Additionally, our conditional classifier was stumped by samples such as "Don't exercise on an empty stomach" where the condition is expressed in an implicit manner. That is to say, the model may have predicted it correctly had it been phrased "Don't exercise if you have an empty stomach". Such problems can be mitigated by collecting more diverse samples in future works.

Relationship to Natural Language Inference: Table 5 highlights the results of our experiments exploring the relationship between NLI and HCD. The $\textbf{NLI}_{OTS}$ experiment tests how well an off-the-shelf (OTS) NLI model detects health conflicts. Our results show extremely poor performance from $\textbf{NLI}_{OTS}$ on most conflict types, with moderate performance on Direct Conflicts. This result was expected as direct conflicts are semantically the most similar to NLI contradictions. However, nuanced conflict types like conditional, temporal and sub-typical have no relationship to the NLI output space, making this a near impossible task for an $\textbf{NLI}_{OTS}$ model. For example, consider the instructions given to MNLI annotators when writing con-

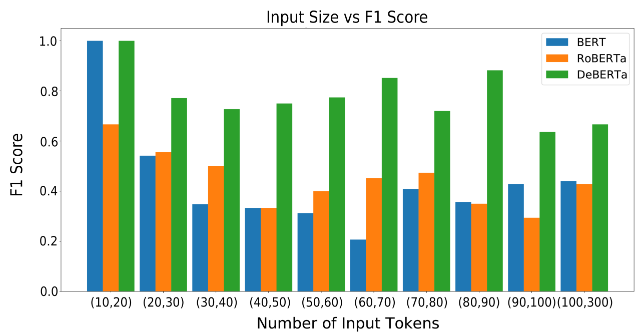

Figure 3: Plot visualizing the relationship between token length and model performance. The x-axis represents token number bins and the y-axis shows the corresponding F1 score on that data subset. Results were generated using all training data in a one-vs-all setting on Direct Conflicts.

tradictions: Given a sentence 1, write a sentence 2 which is "definitely incorrect about the event or situation" in sentence 1. In the case where Advice 1 is *Don't consume alcohol* and Advice 2 is *Don't drink alcohol if you're pregnant*, we have a conditional conflict which is not at all related to the requirements for an MNLI contradiction.

The $\mathbf{NLI}_{FT}$ experiment highlights how NLI can help to infer some conflict types as an intermediate fine-tuning task. Specifically, Temporal and Conditional conflicts benefited significantly from this pre-training strategy. However, given the previous discussion on the difference in definitions between conflict and contradiction, we suppose the performance increase is due to a combination of the following reasons. 1) The introduction of DeBERTa-v3 to a pairwise inference task. 2) These conflict types have overt textual clues, making them easier to classify. In other words, we can classify more of the simpler sample types with additional pairwise inference pre-training. However, this did not affect Direct and Sub-Typical conflicts.

RQ2: Are Synthetic Health Conflicts Able to Aid in the Prediction of Real Health Conflicts?

A core goal of this paper is to expand the size of the original HCD datasets presented in a scaleable manner as collecting real data on conflicting advice is prohibitively expensive (Preum et al. 2017a,b). We run experiments using different combinations of training data to quantify the contribution of synthetic samples on the real-world test set. Table 8 displays the results of this experiment. The synthetic data helped all transformer based models to classify conditional and temporal conflicts. This validates our hypothesis that, even when there exists severe stylistic drift [7] between real and synthetic data, models can benefit from examples of LSSP's, regardless of their context. All models additionally benefit from synthetic sub-typical conflicts. In fact, **each model found synthetic data more useful than real data when predicting sub-typical conflicts.**

[7] Additional discussion of stylistic drift can be found in the limitations section.

Conflict Type	Topic	LSSP
Direct	0.18	0.75
Sub-Typical	0.66	0.63
Conditional	0.45	0.95
Temporal	0.2	0.86

Table 7: Experimental results from RQ3. Scores in the column Topic confirm if DeBERTa identifies the conflict topic as important. Scores in the column LSSP confirm if DeBERTa recognizes tokens that are relevant to the label type.

However, the synthetic data was distracting when predicting direct conflicts for BERT and RoBERTa, while DeBERTa-v3 found synthetic data useful. This can be explained by DeBERTa-v3's superior long-range dependency modeling (as discussed in RQ1). Consider that the mean token lengths for a synthetic direct conflict and real direct conflict are 55 tokens and 86 tokens, respectively. In terms of sentence length, this on average translates to real direct conflicts being 1 full sentence longer than synthetic conflicts. Thus, synthetic direct conflicts are significantly shorter and perhaps not helpful in aiding BERT and RoBERTa to resolve the long-range dependency in real world data. We find that DeBERTa-v3 is, in general, able to best utilize synthetic samples as they improve DeBERTa-v3's performance on real test data across all conflict types.

RQ3: Do Language Models Agree with Humans When Identifying Features Most Influential to Conflict Detection?

Table 7 displays the results of our experiment in which three human annotators are asked to confirm if DeBERTa-v3 identifies what tokens are most important to understand a conflict. For all experiments, the **Topic** score displayed in Table 7 is defined the same — the human annotator is asked to confirm if the model has attributed a high score to the conflict topic. LSSP's however are defined per conflict type. In the following sub-sections, we define the LSSP for each conflict type and analyze the results.

Direct Conflicts: In this experiment, LSSP points were awarded to samples where DeBERTA-v3 identified relevant polarity tokens associated with each conflict topic. In Figure 2, we see that the topic *Grapefruit* is not emphasized — which is surprising given the importance of conflict understanding in humans. However, the model correctly attributes "Avoid" and "safe to take", the two opposing polarity phrases, as being more important to classification. On average, Topics were attributed highly in 18% of samples, while LSSP's were attributed highly in 75% of examples.

Sub-Typical Conflicts: In this experiment, the model got points for LSSP identification if tokens which altered the type of the conflict topic were given high attribution scores. Our results found that 66% of samples identified Topic and 63% of samples identified LSSP's.

In the sub-typical example provided in Figure 2, the topic "Milk" is identified in both Advice 1 and 2, while the topic

Conflict Type	**Training Set**	B_{F1}	R_{F1}	D_{F1}
Direct	Real	**0.49**	**0.62**	0.67
Direct	Synthetic.	0.07	0.20	0.24
Direct	Real + Synthetic	0.38	0.51	**0.75**
Sub-Typical	Real	0.18	0.13	0.15
Sub-Typical	Synthetic	**0.31**	0.35	0.33
Sub-Typical	Real + Synthetic	0.28	**0.39**	**0.36**
Conditional	Real	0.24	0.42	0.40
Conditional	Synthetic	0.19	0.27	0.26
Conditional	Real + Synthetic	**0.44**	**0.58**	**0.53**
Temporal	Real.	0.51	0.65	0.69
Temporal	Synthetic	0.44	0.52	0.57
Temporal	Real + Synthetic	**0.53**	**0.72**	**0.70**

Table 8: The F1 score of BERT, RoBERTa, and DeBERTa (B_{F1}, R_{F1}, D_{F1} respectively) on the HCD real test set using real-only, synthetic-only, and all training data sets.

altering tokens “full -” are highlighted in Advice 2. Given the need to identify topic to understand sub-typical conflicts, it is expected that the topic attribution rate is high for this conflict type.

Conditional and Temporal Conflicts In these two experiments, LSSP attribution was credited when the model highlighted either conditional or temporal clauses respectively. For example, in Figure 2, the conditional example gives strong attribution to the statement “if you have sleep apnea”. In the temporal example, we also observe strong attribution given to “between the hours of 1 and 3 pm”. LSSP’s for these two label types are shown to be easy for the model to understand as conditional and temporal tokens are identified as LSSP’s 95% and 86% of the time respectively. However, topic attribution scores for these two conflict types are low, which is expected given that the model can rely on the overt label-specific semantic phrases found in conditional and temporal samples.

Related Work

Pre-trained Language Models

From their inception when introduced in (Vaswani et al. 2017), the use of the Transformer architecture has evolved greatly over the years. (Devlin et al. 2018) released BERT, a transformer-based language model pre-trained on a 3.4 billion word corpus which produced state-of-the-art results on popular NLP benchmarks. Modern adaptations of the BERT pre-training paradigm such as RoBERTa (Liu et al. 2019) and DeBERTa-v3 (He, Gao, and Chen 2021b) improve upon BERT by optimizing the pre-training tasks used to generate high-quality contextual token representations. We explore all three transformer models in this study.

Additionally, we investigate the benefits of medical training data on HCD via Bio+Clinical BERT (Alsentzer et al. 2019), which is the BERT framework described above, but instead of being pre-trained on common texts, it is trained on biomedical corpora and clinical notes. Finally, we evaluate the performance of static sentence embeddings using SBERT (Reimers and Gurevych 2019). This experiment provides a highly regularized transformer baseline where only the classification head is fine-tuned for HCD.

Natural Language Inference

The task of Natural Language Inference (NLI) is a multi-class pairwise classification problem for which, given a premise and a hypothesis, the model is tasked with predicting if the premise, Entails, Contradicts, or is Neutral towards the hypothesis. NLI data sets such as MNLI (Williams, Nangia, and Bowman 2018) are commonly used to evaluate a language model’s ability to reason about complex language understanding tasks. NLI and HCD both depend on the compare and contrast of two disjoint texts. Additionally, it is reasonable for one to posit that there is a relationship between the notion of a pairwise contradiction and conflict. Thus in this study, we explore the relationship between notions of *conflict* and *contradiction* in effort to identify if NLI can be useful in the prediction of conflicting health information.

Health Misinformation Detection

Recently there has been increasing interest in misinformation detection for specific health topics from social media data (Sager et al. 2021; ElSherief et al. 2021; Weinzierl, Hopfer, and Harabagiu 2021). ElSherief et al. focus on detecting misinformation regarding medications used for opioid use disorder (OUD) treatment from multiple social media using traditional machine learning models and BERT (ElSherief et al. 2021).

They formulate the problem as a binary classification task where a positive class refers to a post discussing a piece of misinformation challenging OUD treatment and the negative class refers to any post that was not relevant. They report that logistic regression on TF-IDF based features perform better than BERT. Weinzierl et al. focus on identifying adoption or rejection of COVID-related misinformation on Twitter to better understand the effect of exposure to misinformation on COVID-19. They formulate the problem as stance classification and develop an ensemble architecture consisting of graph attention networks and BERT (Weinzierl, Hopfer, and Harabagiu 2021). Both HCD and health misinformation detection aim to measure information quality and improve public interactions with online health texts.

Limitations and Future Work

Limitations: Although the human-in-the-loop approach using mTurk yielded reasonable conflicting health advice pairs, the resulting dataset has a couple of limitations. First, we found that mTurk workers do not generate health conflict data which stylistically reflects real-world data. Specifically, mTurk workers struggle to write diverse sub-typical conflicts and often generate repetitive conditional or temporal conflicts with no or limited stylistic and semantic variation. In general, synthetic samples have much more textual overlap than real samples, as mTurk annotators can often change/add a small number of tokens to generate a conflict.

Medical Professional A	Medical Professional B	Post Topic	Conflict Type
For clarity's sake, are you saying you took 12x500 mg over the course of 23 hours or 11 hours? ... If it's the latter it may be worth getting checked as it's double the recommended dose.	Assuming the doses were spaced out evenly then concentrations would not have reached toxic levels at any point (see calculation below if you're interested).	Fever Medication	Direct
If the left breast has been getting larger and developing those veins recently, and if you feel firmness of the left breast that is different from the right, then I would get it checked out by your PCP or GYN.	They're normal. Breasts (like eyebrows) are "sisters, not twins" - no one has perfectly symmetrical breasts.	Breast Cancer	Conditional

Table 9: Conflicting health information found on reddit.com/r/AskDocs — A social media forum where anyone can get medical advice from a professional whose medical certifications are verified by the forum moderators. We show that suggestions from medical professionals can be conflicting.

These common themes produce a stylistic drift between real and synthetic samples in our dataset. However, this is an expected limitation of mTurk-generated samples as they were only tasked with mimicking an advice conflict's linguistic phenomena, not generating realistic samples.

Another limitation of this work is that the distribution of conflicts found in this dataset will be biased towards the official data sources used to collect real advice statements. As the HCD task moves to new data domains such as social media data, different conflict types/distributions may emerge.

The proposed solution is limited to understanding the problem linguistically. In the real world, health conflicts are often subjective to user physiology, diagnosis and prognosis. Effects of health conflicts might differ in terms of degree of risk and potential severity. A user-centric conflict detection system should be personalized to avoid false alarms and cause user anxiety. This demands continuous user sensing, as information of the users' medical context is required for the personalized system to detect what is conflicting for a given user. For example, the HCD system can only flag potentially dangerous drug-to-drug interactions if it knows what medications the end-user is currently taking.

Future work: Future work on HCD should investigate better use of medical knowledge. While experimental results on Bio+Clinical BERT showed HCD does not benefit from pre-training on scientific texts and clinical notes, it may be the case that HCD requires a medical knowledge integration solution that is well-suited for non-clinical text related to public health. Medical knowledge graphs like UMLS (Bodenreider 2004) may aid BERT-like models in the understanding of rare diseases and medications not common in language model vocabularies.

Additionally, our attempt at organic HCD dataset expansion only targeted professional sources. Future works may consider non-professional sources such as social media data. However, advice extraction from social media data is itself a separate research question. Additionally, non-professional sources may provoke differing data distributions with new conflict types. Exploration of such data sources should be the subject of future works.

Broader Impacts

New dataset with high quality and reliability: We provide a new public dataset on health conflict detection. The data is reliable as using multi-stage human review helps mitigate annotation artifacts and human errors commonly found in large-scale crowd-sourced datasets. Additionally, our work on expanding the original HCD dataset uses data from various topics on various diseases and conditions. This makes HCD more health topic inclusive.

Providing consistent online health information: An accurate health conflict detection system will protect consumers of online health information by providing a cautionary signal upon detection of potential health information conflicts. In addition, such technology can reduce the number of errors in medical decisions made by patients using online medical information sources. This is particularly important for individuals who rely on online sources for decision-making due to limited access to healthcare, an increasingly common case since the onset of a global pandemic.

Reducing cognitive overload and aiding patient-provider communication: In general, health information inconsistency checking is a challenging task for most humans. It can be overwhelming for patients to memorize all of the health advice they receive and cross-reference it with all the new pieces of advice they encounter. Automatic conflict detection can help patients identify potential conflicts and consider when to raise questions to their medical providers. Such a solution can aid intelligent assistants targeting patients with multiple conditions, caregivers of individuals with cognitive challenges, or those undergoing stressful situations, where mental cross-referencing between sources becomes increasingly difficult (Preum et al. 2021; Stankovic et al. 2021). Even for the general population, HCD reduces the burden to understand and memorize all of the fine-grained details of their medical conditions/medications by automatically detecting conflicts *which pertain to a patient's specific condition.* Also, the goal of this work is to flag potential conflicts found in textual health information so an end-user can bring them to the attention of their medical provider. Thus there is no conflict resolution currently in place, as it is beyond the scope of this work.

Extending HCD task for detecting and characterizing misinformation: We additionally note that, while the non-augmented HCD samples in this study were taken from verified medical sources, the ideas proposed in this work extend to non-official sources such as text collected from Twitter, Reddit, Facebook or other social media (Sager et al. 2021; ElSherief et al. 2021; Weinzierl, Hopfer, and Harabagiu 2021). For example, one can compare health advice between official social media sources (e.g. @CDCgov on Twitter) and non-official sources aiming to spread health misinformation. This differs from other health misinformation formulations where the veracity of a claim is verified using the text itself, or an external scientific resource. In the HCD formulation, we not only better permit the direct comparison of two pieces of advice, but can predict fine-grained labels that provide better insight into *how* two advice texts differ. Thus it can help to characterize misinformation from a linguistic standpoint. Additionally, fact-checking often uses knowledge base triplets for claim verification. Under this paradigm, it may be hard to extend misinformation detection to new/niche health domains for which knowledge triplets are unavailable.

Ethical Impact

This research involved human subjects only for data annotation, review, and prompt-based synthetic data generation. The project did not provide any intervention to any human subjects nor did it collect any user level data. So there is no risk to human subjects. The Amazon mechanical turk workers were recruited only for prompt-based synthetic data generation using formal procedures and no personal data was collected from them. The authors took careful measures to avoid annotation errors and maintain the quality and reliability of the data. Future progressions of HCD studies, to provide contextualized user-centric health information may require end-user medical data. This might raise various ethical and privacy concerns and will require careful consideration in the future.

The collected advice statements are often for the general population. Thus, the HCD dataset may not contain many advice statements directed at underrepresented populations. Future data collection rounds should focus on diversifying both medical condition coverage and target audience. User-generated medical content from social media can be a good source to collect health advice targeted to a specific sub-population, e.g. *heart disease* related advice targeted to African Americans.

Conclusion

Detecting conflicting health information makes human interaction with online health platforms safer. In this work, we introduce the HCD task in the context of pre-trained language models. We provide a detailed analysis of five different language models, examining the challenges brought about when predicting different health conflict types. Our experimental results find that the DeBERTa-v3 architecture performs better on challenging conflict types with less obvious semantic patterns. However, for simpler conflicts like conditional and temporal, BERT and RoBERTa provides comparable performance. We additionally confirm that the token attribution scores provided by a deep interpretability model, Captum, align with human judgement regarding input importance for predicting different conflict types. Finally, we show that the addition of synthetic conflict data does help with the prediction of real-world data, irregardless of differences in style and content accuracy.

References

Alsentzer, E.; Murphy, J. R.; Boag, W.; Weng, W.; Jin, D.; Naumann, T.; and McDermott, M. B. A. 2019. Publicly Available Clinical BERT Embeddings. *CoRR*, abs/1904.03323.

Black, S.; Leo, G.; Wang, P.; Leahy, C.; and Biderman, S. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. Zenodo. https://doi.org/10.5281/zenodo.5297715.

Bodenreider, O. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1): D267–D270.

Breiman, L. 2001. Random Forests. *Machine Learning*, 45(1): 5–32.

Buchanan, M. 2020. Managing the infodemic. *Nature Physics*, 16(9): 894–894.

Bujnowska-Fedak, M. M.; and Wegierek, P. 2020. The Impact of Online Health Information on Patient Health Behaviours and Making Decisions Concerning Health. *International journal of environmental research and public health*, 17(3): 880. 32023828[pmid].

Carpenter, D. M.; Geryk, L. L.; Chen, A. T.; Nagler, R. H.; Dieckmann, N. F.; and Han, P. K. 2016. Conflicting health information: a critical research need. *Health Expectations*, 19(6): 1173–1182.

Chang, T.; and Lu, C. 2021. Rethinking Why Intermediate-Task Fine-Tuning Works. *CoRR*, abs/2108.11696.

Cornish, L. 2020. Targeting false information on COVID-19: What the funding data shows. https://www.devex.com/news/targeting-false-information-on-covid-19-what-the-funding-data-shows-97536. Accessed: 2022-04-23.

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

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805.

ElSherief, M.; Sumner, S. A.; Jones, C. M.; Law, R. K.; Kacha-Ochana, A.; Shieber, L.; Cordier, L.; Holton, K.; De Choudhury, M.; et al. 2021. Characterizing and Identifying the Prevalence of Web-Based Misinformation Relating to Medication for Opioid Use Disorder: Machine Learning Approach. *Journal of medical Internet research*, 23(12): e30753.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychol. Bull.*, 76(5): 378–382.

Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S. R.; and Smith, N. A. 2018. Annotation Artifacts in Natural Language Inference Data. In *NAACL*.

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

He, P.; Gao, J.; and Chen, W. 2021b. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *CoRR*, abs/2111.09543.

Herlihy, C.; and Rudinger, R. 2021. MedNLI Is Not Immune: Natural Language Inference Artifacts in the Clinical Domain. *arXiv preprint arXiv:2106.01491*.

Kokhlikyan, N.; Miglani, V.; Martin, M.; Wang, E.; Alsallakh, B.; Reynolds, J.; Melnikov, A.; Kliushkina, N.; Araya, C.; Yan, S.; et al. 2020. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Loomba, S.; de Figueiredo, A.; Piatek, S. J.; de Graaf, K.; and Larson, H. J. 2021. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour*, 5(3): 337–348.

Loshchilov, I.; and Hutter, F. 2017. Fixing Weight Decay Regularization in Adam. *CoRR*, abs/1711.05101.

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

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct): 2825–2830.

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

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

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

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

Preum, S. M.; Munir, S.; Ma, M.; Yasar, M. S.; Stone, D. J.; Williams, R.; Alemzadeh, H.; and Stankovic, J. A. 2021. A review of cognitive assistants for healthcare: Trends, prospects, and future directions. *ACM Computing Surveys (CSUR)*, 53(6): 1–37.

Preum, S. M.; Parvez, M. R.; Chang, K.-W.; and Stankovic, J. 2018. A corpus of drug usage guidelines annotated with type of advice. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Sager, M. A.; Kashyap, A. M.; Tamminga, M.; Ravoori, S.; Callison-Burch, C.; and Lipoff, J. B. 2021. Identifying and Responding to Health Misinformation on Reddit Dermatology Forums With Artificially Intelligent Bots Using Natural Language Processing: Design and Evaluation Study. *JMIR Dermatology*, 4(2): e20975.

Stankovic, J. A.; Ma, M.; Preum, S. M.; and Alemzadeh, H. 2021. Challenges and Directions for Ambient Intelligence: A Cyber Physical Systems Perspective. In *2021 IEEE Third International Conference on Cognitive Machine Intelligence (CogMI)*, 232–241. IEEE.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. *CoRR*, abs/1703.01365.

Tjoa, E.; and Guan, C. 2019. A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. *CoRR*, abs/1907.07374.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. *CoRR*, abs/1706.03762.

Vellido, A. 2020. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural computing and applications*, 32(24): 18069–18083.

Weinzierl, M.; Hopfer, S.; and Harabagiu, S. M. 2021. Misinformation adoption or rejection in the era of covid-19. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM), AAAI Press*.

Williams, A.; Nangia, N.; and Bowman, S. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1112–1122. Association for Computational Linguistics.

Yoo, K. M.; Park, D.; Kang, J.; Lee, S.-W.; and Park, W. 2021. GPT3Mix: Leveraging Large-scale Language Models for Text Augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2225–2239. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Zhang, J.; Zhao, Y.; Saleh, M.; and Liu, P. J. 2019. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. *CoRR*, abs/1912.08777.