SemEval-2023 Task 8: Causal Medical Claim Identification and Related PIO Frame Extraction from Social Media Posts

Vivek Khetan
Accenture Labs
vivek.a.khetan@accenture.com

Somin Wadhwa
Northeastern University
wadhwa.s@northeastern.edu

Byron C. Wallace
Northeastern University
b.wallace@northeastern.edu

Silvio Amir
Northeastern University
s.amir@northeastern.edu

Abstract

Identification of medical claims from user-generated text data is an onerous but essential step for various tasks including content moderation, and hypothesis generation. SemEval-2023 Task 8 is an effort towards building those capabilities and motivating further research in this direction. This paper summarizes the details and results of shared task 8 at SemEval-2023 which involved identifying causal medical claims and extracting related Populations, Interventions, and Outcomes ("PIO") frames from social media (Reddit) text.[1] This shared task comprised two subtasks: (1) Causal claim identification; and (2) PIO frame extraction. In total, seven teams participated in the task. Of the seven, six provided system descriptions which we summarize here. For the first subtask, the best approach yielded a macro-averaged F-1 score of 78.40, and for the second subtask, the best approach achieved token-level F-1 scores of 40.55 for Populations, 49.71 for Interventions, and 30.08 for Outcome frames.

1 Introduction

Social media allows individuals to discuss, potentially rare, medical conditions with others and to find "long-tail" information about condition trajectories and treatment strategies (Wadhwa et al., 2023).

While users may feel empowered, the unvetted nature of such online discourse makes it vulnerable to *misinformation* (factually incorrect statements, though not necessarily motivated by the aim to mislead) and *disinformation* (falsehoods crafted explicitly to shape public opinion) (Swire-Thompson and Lazer, 2019). The problem of medical misinformation was particularly visible at the height of the COVID-19 pandemic, during which many unsubstantiated claims about treatments for the disease circulated. For example, Ivermectin might be an effective treatment for COVID-19.

[1] https://causalclaims.github.io/

Here, we propose a task which enlists NLP to detect health related conversations on social media. This is a crucial first step for countering health misinformation (among other potential applications). More specifically, we propose a multi-stage task. The first objective is to identify causal medical claims (Khetan et al., 2022) made within a given text snippet from a Reddit post, consisting of either a single or multiple sentences, i.e., to identify the text spans that contains claims. The second objective is to extract key clinical elements from identified causal claims, in particular: The *Population* (i.e., the condition), the *Intervention*, and the *Outcome* i.e., the *PIO* elements (Richardson et al., 1995; Nye et al., 2018). For example, if we identify a claim such as *"Drinking bleach can cure COVID-19!"*, both the Population and Outcome would be "COVID-19", while the Intervention would be "bleach".[2] In an identified claim, the extracted interventions and the outcomes have a causal relationship communicated between them, either explicitly or implicitly.

This task generated a lot of interest from researchers in the NLP community both because of the timely application for the pressing issue of medical misinformation (Zuo et al., 2021), and the challenging nature of the task, which entails identifying claims on social media (Risch et al., 2021; Ahne et al., 2022) and then extracting (Gi et al., 2021) key elements from these identified claims. Additionally, the current advancements and widespread availability of Large Language Model (LLM) based systems for text generation in medical and clinical applications (Singhal et al., 2022; Feng et al., 2022; Rajagopal et al., 2021) also underscore the need for research in this area.

[2] Note that the Outcome, in this case, is implicit, and is also vague as stated (it is unclear whether this refers to symptoms of COVID-19 or some measure of viral load in the nose, for instance); such imprecision is common on social media however.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 2266–2274

Reddit post	Extracted sentence classes	PIO elements from claims
Rheumatologist says cellcept failed to protect my kidneys and now I have developed lupus. Prednisone messed up my hips. Just wondering if anyone tried Xanax for pain?	**Claim:** Rheumatologist says cellcept failed to protect my kidneys **Personal Experience:** Prednisone messed up my hips **Question:** Just wondering if anyone tried Xanax for pain?	**P:** Rheumatoid Arthritis **I:** cellcept **O:** failed to protect kidneys
Specialist recommends chemo for my thyroid cancer even though we've told them we're trying to get pregnant. Spoke to my primary about it and he agrees that either of those two should be delayed but *specialist seems insistent that any pregnancy, even immediately after chemo, should not pose any problems whatsoever*. Am confused how to approach this, has anyone experienced this?	**Claim:** specialist seems insistent that any pregnancy, even immediately after chemo, should not pose any problems whatsoever **Question:** Am confused how to approach this, has anyone experienced this?	**P:** thyroid cancer **I:** chemo **O:** pregnant
Getting a lot of mixed signals information about what I can and can't eat. *One source tells me beans and plat proteins are fine, other says they're terrible. One article says cherry juice lowers uric acid, another says it does nothing.*	**Claim$_1$:** One source tells me beans and plat proteins are fine, other says they're terrible. **Claim$_2$:** One article says cherry juice lowers uric acid, another says it does nothing.	**I$_1$:** beans and plant proteins **I$_2$:** cherry juice **O$_2$:** uric acid

Table 1: Example of a dataset with a Reddit post, identified sentence class and extracted PIO elements; reproduced from Wadhwa et al. 2023. Sentence classes are identified in the first stage of annotations corresponding to *pure-claims*, personal experiences, experiences that are also claims, and questions. Populations, Interventions, and Outcomes corresponding to each individual claim are further identified in the second stage.

2 Data and Resources

Data Sources The data we release as part of this shared task stems from a larger corpus of over 22k richly annotated social media posts from Reddit spanning 24 health conditions; this is described at length in Wadhwa et al. (2023). For this task, we released 5,695 annotated Reddit posts of this data focused on 10 of those health conditions. These include 597 posts containing PIO annotations for claims.

For the full dataset in Wadhwa et al. (2023), we identified a set of 24 condition-focused health communities ("subreddits") on Reddit, ranging from common health conditions such as diabetes (generating very high online activity) to relatively rare chronic diseases like Multiple Sclerosis (MS). For each subreddit, we extracted the most recent posts (up to 1000) to include in the dataset. To annotate for gold labels, we relied on Amazon Mechanical Turk (MTurk). Annotations were collected in two stages. In the first stage (stage-1), we obtain sentence-level annotations by asking workers to identify sentences (text spans) that correspond to *claims, personal experiences, claims based on personal experiences, and questions*. For the second stage (stage-2), we considered only instances where we find pure-claims, i.e., broad claims that are *not* related to a personal experience ($\sim$7.7% of posts). Here, we collected annotations to identify the relevant PIO elements that correspond to those claims. The annotations from the first stage have an average span length of $\sim$23.0 tokens while the the second stage covers entities with an average span length of $\sim$2.0 tokens.

For the first stage, we use the following definitions to identify sentences belonging to each category:

- **Claims:** A span is classified as a claim if and only if there exists any explicit or implicit relationship (regardless of directionality) between an *intervention* and an *outcome* (e.g., my friend took X, and Y happened; I was having X symptoms and my doctor prescribed me Z for treatment). Operationally, we are interested in claims that could potentially change someone's perception about the efficacy of an intervention for a particular condition and/or outcome (i.e., the relationship between X, Y, and Z). An independent claim on average spans $\sim$19.6 tokens.
- **Question:** If a span of text contains a question (e.g.: Is this normal?; Should I increase/decrease my dosage?; etc). The average length of a question in this data is $\sim$10.5 tokens.
- **Personal Experience:** If a span describes a personal experience related to specific outcomes/symptoms or populations/interventions. The average length of a standalone personal experience is $\sim$27.0 tokens while those containing a claim are on average $\sim$30.6 tokens long.

Obtaining high quality annotations was one of our top priorities and challenges. To that end, we ran three pilot experiments: the first, was an internal experiment with a very small sample of about 100 Reddit posts which were annotated for stage-1 by two people with expertise in analysing biomedical data. We evaluated the quality of annotations through token-wise label agreement between our internal annotators. Then, we conducted two pilot experiments on mTurk with $\sim$6000 samples to identify and recruit workers. Recruited workers were paid periodic bonuses based on the quality of a random subset of their annotated samples. In total, it took us a little over six weeks to accumulate stage-1 annotations and approximately two weeks to accumulate stage-2 annotations. This includes the time it took for us to evaluate the annotated data in batches and provide feedback to the workers. We provide additional details on quality validation of our data in Wadhwa et al. (2023).

To account for user consent, we sent a short message to every Reddit user whose public post we scraped to inform them about the potential inclusion in this corpus, the intended purpose of the data and to provide them the option to opt-out by responding within a period of 30 days. Every user who responded within the 30 day period had their data completely removed from the broader dataset.

3 Task Description and Evaluation

3.1 Task Description

Participants of SemEval-2023 task 8 were invited to develop systems to automatically identify medical claims, questions, personal experiences, and associated PIO elements made within a text snippet (single or multi-sentence Reddit Post). The proposed task was divided into the two following subtasks.

Subtask 1: Causal claim identification Given a text snippet (single or multi-sentence Reddit Post), the first subtask aims to identify all the spans of text containing `Claims`, `Personal Experiences`, `Claims based on Personal Experiences`, `Questions`, and `Other`. This can be framed as a sentence-level multiclass classification task. Nevertheless, there are instances in which the desired target spans represent only a portion of the sentence.

Subtask 2: PIO frame extraction Given a text snippet (single or multi-sentence Reddit Post) and an identified claim in that snippet, the goal here is to extract related `Population`, `Intervention`, and `Outcome` frames. In rare cases, there may be more than one claim in a given text snippet. In any case, the task is to identify the PIO elements associated with a *particular claim*; this can be framed as a sequence tagging task.

3.2 Evaluation

Data for subtask 1 was annotated at the *sentence-level* while data for subtask 2 was annotated at the *token-level*. We therefore required participants to submit test files with *token-level* labels. For subtask 1, we evaluated macro-averaged F1 scores across five classes. These were evaluated at the *sentence-level* as opposed to exact span/token matches since differences in annotated spans often depend on differences in where annotators (and consequently, trained models) decide to mark span boundaries. However, sentences covering those spans can reasonably be assumed to belong to the same given class.

For subtask 2, we evaluated *token-level* F1 individually for each class (effectively treating it as an entity-tagging task). Note that most tokens in any given post will not belong to any PIO element.

4 Results and Discussion

During the evaluation phase of our task, a total of seven teams participated with 48 valid submissions for subtask 1 and six teams participated with 43 valid submissions for subtask 2. The teams were given the opportunity to make an unlimited number of submissions for each subtask. The results presented in Table 2 and Table 3 reflect the outcomes of the final submission for subtask 1 and subtask 2, respectively.

4.1 Summary of Participating Systems

We provide detailed information about the top three performing systems, and some insights gleaned from other systems.

Team MaChAmp (van der Goot, 2023) posed both the subtasks as a sequence tagging task at the token level for respective categories. They used a unified multi-task learning toolkit, MaChAmp (van der Goot et al., 2020), to model multiple SemEval 2023 tasks including our subtasks. MaChAmp consists of a shared transformer-based encoder and 8 different task-specific decoder heads named as SEQ, SEQ_BIO, STRING2STRING, MULTISEQ, etc. In this work, the authors have used SEQ decoder head for token-level sequence tagging using greedy decoding with a softmax output layer, and SEQ_BIO decoder head for token-level sequence prediction using CRF as decoder. They utilized intermediate multi-task training (Gururangan et al., 2020a; Muller et al., 2020; Phang et al., 2018), i.e. training on all the SemEval 2023 text-based tasks datasets, before finetuning the same shared encoder on our task data. They achieved the best F1 score using SEQ_BIO decoder for subtask-1 and SEQ decoder for subtask-2. This team ranked first for both subtasks on our leaderboard.

Team NCUEE-NLP (Lee et al., 2023) posed the first subtask as a sentence classification task and the second subtask as a sequence tagging task at the token level. They segmented each text snippet into sentences using Trankit (Nguyen et al., 2021), a transformer-based NLP toolkit. For both the subtasks, they performed five-fold cross-validation for hyperparameter search and finetuned various encoders including BERT (Devlin et al., 2019), DeBERTa (He et al., 2020, 2021), BioBERT (Lee et al., 2019), and RoBERTa (Liu et al., 2019). For subtask-1, they utilized CLS based classifier head while for subtask-2 they utilized the token-level classifier. This team ranked second for both subtasks on our leaderboard.

Team MasonNLP (Ramachandran et al., 2023) posed both subtasks as sequence tagging using token-level classifiers to predict the begin-inside-outside (BIO) tags for the respective categories. They finetuned general domain as well as mixed domain encoders, for both the subtasks, and performed a grid search to tune the hyperparameters. BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) were selected as the general domain encoders; BioMedRoBERTa (Gururangan et al., 2020b) and BioRedditBERT (Basaldella et al., 2020) were used as mixed domain encoders. Team MasonNLP also incorporated external knowledge by identifying disease and chemical entities in text snippets and annotating these with special tokens. To that end, they used Scispacy[3](Neumann et al., 2019) to augment the provided text snippet data. This approach yielded improvements in the F1 score for subtask-2. This team ranked third for both subtasks on our leaderboard.

Other participating teams built systems using a variety of methods including weak supervision, ensemble-based modeling, and data augmentation methods. **Team HEVS-TUW (Dhrangadhariya et al., 2023)** posed both the subtasks as a sequence tagging task at the token level for respective categories. They leveraged majority voting for an ensemble approach. **Team CAISA (Karimi and Flek, 2023)** posed the first subtask as a sentence classification task and the second subtask as a sequence tagging task at the token level. They experimented with various data augmentation methods including AEDA (Karimi et al., 2021), entity replacement (Zeng et al., 2020), using YouChat to produce diverse and counterfactual sentences to mitigate class imbalance in the provided dataset. Finally, **Team Togedemaru (Oica et al., 2023)** only participated in subtask-1 and posed it as a sentence classification task.

[3]model en_ner_bc5cdr_md

Team	P	R	F1
MaChAmp	78.14	78.65	78.40
NCUEE-NLP	72.97	67.36	70.05
MasonNLP	71.16	65.78	68.59
HEVS-TUW	68.73	62.90	65.70
CAISA	60.68	55.71	58.09
Togedemaru	34.93	31.14	32.93

Table 2: Subtask-1 evaluation based on final submission.

Team	F1 (POP)	F1 (INT)	F1 (OUT)
MaChAmp	40.55	49.71	30.08
NCUEE-NLP	37.78	43.58	30.67
MasonNLP	34.96	42.16	20.83
HEVS-TUW	17.44	26.39	22.78
CAISA	17.67	21.05	20.31

Table 3: Subtask-2 evaluation based on final submission.

5 Related Work

Causality expressed in health-related text data Understanding causality expressed in the text has been an area of interest for a long time (Talmy, 1987; Wolff, 2007). Researchers have proposed guidelines to represent (Mostafazadeh et al., 2016), built datasets to capture (Mirza and Tonelli, 2014; Dunietz et al., 2017), and methods to extract (Khetan et al., 2020) causality in text data from various domains (Bethard and Martin, 2008).

Gurulingappa et al. (2012) studied causality communicated in medical case reports by developing a dataset of Adverse Drug Effects. Whereas, Mihaila et al. (2012) annotated various causal events as arguments and the connectives between them as triggers from biomedical scientific articles to capture causality. More recently, Khetan et al. (2022) defined causal typology and built a dataset to understand types and directions of causal interaction communicated in clinical notes.

Health-related corpora from social media posts Social media posts can act as complementary sources to obtain data for research on various topics, including healthcare (Chen et al., 2018; Aragón et al., 2019; Yadav et al., 2020).

Various past works have built corpora from health-related Reddit and Twitter posts. Coppersmith et al. (2014) studied the quantification of mental health signals using Twitter posts. Jiang et al. (2020) introduced a dataset of Reddit posts to evaluate models for automatically detecting psychiatric disorders. Shen and Rudzicz (2017) studied anxiety disorders through Reddit posts, whereas, Ahne et al. (2022) built a dataset of cause-effect pairs from Twitter posts specifically for the diabetes distress study.

Crowd-sourcing annotation of scientific and medical texts

Crowdsourcing has been an acceptable approach for parsing and obtaining annotations for many NLP tasks in a variety of domains, including scientific and medical datasets (Dumitrache et al., 2013; Drutsa et al., 2021). Nye et al. (2018) annotated texts from PubMed via crowdsourcing. Similarly, Bogensperger et al. (2021) leveraged crowdworkers to build a dataset of drug mentions on the darknet. For our dataset, we also relied on crowdworkers to identify claims and annotate related PIO labels.

6 Conclusion and Future work

We presented SemEval-2023 shared task 8, a novel task to address the important and timely problem of identifying medical causal claims on social media posts. This was formulated as a multi-step process involving two subtasks: 1) causal claim identification, which consisted in classifying sentences as containing medical claims, personal experiences, claims based on personal experiences, or questions; and 2) PIO frame extraction, aiming to extract spans corresponding to Populations, Interventions and Outcomes associated to the identified claims. SemEval participants were asked to build and evaluate systems to either or both subtasks given a dataset of Reddit posts discussing 10 different health conditions.

In total, seven teams participated in our shared task and six submitted a paper describing their systems. While there was a clear variation across different submitted systems, all of them used transformer-based models. Overall, the final results show that both our subtasks are difficult and there is considerable room for improvement.

The current advancements and widespread availability of LLMs capable of generating text that is indistinguishable from human-written text exacerbated the risk of mass production and dissemination of medical mis- and disinformation. Therefore there is a growing imperative to conduct further research into methods that can help to detect and combat the spread of such misleading and potentially harmful content. A promising approach in this direction is to combine systems for the proposed subtasks and integrate them with a module

to retrieve trustworthy evidence from the scientific literature that can help to validate or refute medical claims on social media.

7 Limitations

We have presented a novel task of identifying key medical elements in social media (i.e. Reddit) text and highlighted some potential applications that these tasks might enable. However, there are certain important limitations of our work. First, the data released for this shared task is only a subset of the dataset collected in (Wadhwa et al., 2023) containing a small number of instances per medical population. Second, the entire data collection focused solely on well-formed social media posts in the English language. Third, high-quality reference samples are key to building effective machine learning models. However, our reference instances are obtained via manual annotations of free text by laypersons (Amazon Mechanical Turk workers). While we took steps to ensure annotation quality we do acknowledge that these references may contain some noise.

8 Ethics Statement

Our proposed task aims to motivate research towards understanding how social media users perceive and discuss various health conditions online. To that end, we created a dataset consisting of personal experiences, claims, and questions of Reddit users along with key clinical elements related to the claims. Given the nature of the dataset and related privacy concerns, we made sure that any individual user could choose not to be included in our corpus. First, we notified every Reddit user whose post we scraped, informed them about the corpus and the intended purpose, and provided them an option to opt-out within a period of 30 days. Second, instead of releasing the dataset directly, we only provide Reddit post identifiers, related annotations, and a script to download and combine them. The script ensures that if a user deletes their posts they can no longer be retrieved.

References

Adrian Ahne, Vivek Khetan, Xavier Tannier, Md Imbesat Hassan Rizvi, Thomas Czernichow, Francisco Orchard, Charline Bour, Andy E. Fano, and Guy Fagherazzi. 2022. Extraction of explicit and implicit cause-effect relationships in patient-reported diabetes-related tweets from 2017 to 2021: Deep learning approach. *JMIR Medical Informatics*, 10.

Mario Ezra Aragón, Adrian Pastor Lopez-Monroy, Luis Carlos González-Gurrola, and Manuel Montes y Gómez. 2019. Detecting depression in social media using fine-grained emotions. In *North American Chapter of the Association for Computational Linguistics*.

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.

Steven Bethard and James H. Martin. 2008. Learning semantic links from a corpus of parallel temporal and causal relations. In *Annual Meeting of the Association for Computational Linguistics*.

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.

Xuetong Chen, Martin D. Sykora, Thomas W. Jackson, and Suzanne Elayan. 2018. What about mood swings: Identifying depression on twitter with temporal measures of emotions. *Companion Proceedings of the The Web Conference 2018*.

Glen A. Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *CLPsych@ACL*.

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.

Anjani Dhrangadhariya, Wojciech Kusa, Henning Müller, and Allan Hanbury. 2023. HEVS-TUW at SemEval-2023 Task 8: Ensemble of Language Models and Rule-based Classifiers for Claims Identification and PICO Extraction. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

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.

Anca Dumitrache, Lora Aroyo, Chris Welty, Robert-Jan Sips, and Anthony Levas. 2013. "dr. detective": combining gamification techniques and crowdsourcing to create a gold standard in medical text. In *CrowdSem*.

Jesse Dunietz, Lori S. Levin, and Jaime G. Carbonell. 2017. The because corpus 2.0: Annotating causality and overlapping relations. In *LAW@ACL*.

Steven Y. Feng, Vivek Khetan, Bogdan Sacaleanu, Anatole Gershman, and Eduard H. Hovy. 2022. Chard: Clinical health-aware reasoning across dimensions for text generation models. *ArXiv*, abs/2210.04191.

In-Zu Gi, Ting-Yu Fang, and Richard Tzong-Han Tsai. 2021. Verdict inference with claim and retrieved elements using RoBERTa. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 60–65, Dominican Republic. Association for Computational Linguistics.

Harsha Gurulingappa, A. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius, and L. Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of biomedical informatics*, 45 5:885–92.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020a. Don't stop pretraining: Adapt language models to domains and tasks. *ArXiv*, abs/2004.10964.

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

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *ArXiv*, abs/2111.09543.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *ArXiv*, abs/2006.03654.

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.

Akbar Karimi and Lucie Flek. 2023. CAISA at SemEval-2023 Task 8: Counterfactual Data Augmentation for Mitigating Class Imbalance in Causal Claim Identification. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2021. AEDA: An easier data augmentation technique for text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2748–2754, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vivek Khetan, Roshni Ramnani, Mayuresh Vivekanand Anand, Shubhashis Sengupta, and Andrew E.Fano. 2020. Causal-bert : Language models for causality detection between events expressed in text. In *Sai*.

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.

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*, 36:1234 – 1240.

Lung-Hao Lee, Yuan-Hao Cheng, Jen-Hao Yang, and Kao-Yuan Tien. 2023. NCUEE-NLP at SemEval-2023 Task 8: Identifying Medical Causal Claims and Extracting PIO Frames Using the Transformer Models. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. 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.

C. Mihaila, Tomoko Ohta, Sampo Pyysalo, and S. Ananiadou. 2012. Biocause: Annotating and analysing causality in the biomedical domain. *BMC Bioinformatics*, 14:2 – 2.

Paramita Mirza and Sara Tonelli. 2014. An analysis of causality between events and its relation to temporal information. In *International Conference on Computational Linguistics*.

N. Mostafazadeh, Alyson Grealish, Nathanael Chambers, James F. Allen, and Lucy Vanderwende. 2016. Caters: Causal and temporal relation scheme for semantic annotation of event structures. In *EVENTS@HLT-NAACL*.

Benjamin Muller, Antonis Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2020. When being unseen from mbert is just the beginning: Handling new

languages with multilingual language models. In *North American Chapter of the Association for Computational Linguistics*.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing. *ArXiv*, abs/1902.07669.

Minh Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A lightweight transformer-based toolkit for multilingual natural language processing. In *Conference of the European Chapter of the Association for Computational Linguistics*.

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.

Andra Oica, Daniela Gifu, and Diana Trandabat. 2023. Togedemaru at semeval-2023 task 8: Causal medical claim identification and extraction from social media posts. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 913–921, Toronto, Canada. Association for Computational Linguistics.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *ArXiv*, abs/1811.01088.

Dheeraj Rajagopal, Vivek Khetan, Bogdan Sacaleanu, Anatole Gershman, Andrew Fano, and Eduard Hovy. 2021. Template filling for controllable commonsense reasoning.

Giridhar Kaushik Ramachandran, Haritha Gangavarapu, Kevin Lybarger, and Ozlem Uzuner. 2023. Masonnlp+ at semeval-2023 task 8: Extracting medical questions, experiences and claims from social media using knowledge-augmented pre-trained language models. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 2143–2152, Toronto, Canada. Association for Computational Linguistics.

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.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12, Duesseldorf, Germany. Association for Computational Linguistics.

Judy Hanwen Shen and Frank Rudzicz. 2017. Detecting anxiety through reddit. In *CLPsych@ACL*.

K. Singhal, Shekoofeh Azizi, Tao Tu, Said Mahdavi, Jason Lee Kai Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather J. Cole-Lewis, Stephen J. Pfohl, P A Payne, Martin G. Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, P. D. Mansfield, Blaise Agüera y Arcas, Dale R. Webster, Greg S. Corrado, Y. Matias, Katherine Hui-Ling Chou, Juraj Gottweis, Nenad Tomaev, Yun Liu, Alvin Rajkomar, Joëlle K. Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large language models encode clinical knowledge. *ArXiv*, abs/2212.13138.

Briony Swire-Thompson and David Lazer. 2019. Public health and online misinformation: Challenges and recommendations. *Annual review of public health*.

Leonard Talmy. 1987. Force dynamics in language and cognition. *Cogn. Sci.*, 12:49–100.

Rob van der Goot. 2023. Machamp at semeval-2023 tasks 2, 3, 4, 5, 7, 8, 9, 10, 11, and 12: On the effectiveness of intermediate training on an uncurated collection of datasets. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pages 230–245, Toronto, Canada. Association for Computational Linguistics.

Rob van der Goot, A. Ustun, Alan Ramponi, and Barbara Plank. 2020. Massive choice, ample tasks (machamp): A toolkit for multi-task learning in nlp. *ArXiv*, abs/2005.14672.

Somin Wadhwa, Vivek Khetan, Silvio Amir, and Byron Wallace. 2023. Redhot: A corpus of annotated medical questions, experiences, and claims on social media. In *European Association of Computational Linguistics (EACL)*.

P. Wolff. 2007. Representing causation. *Journal of experimental psychology. General*, 136 1:82–111.

Shweta Yadav, Jainish Chauhan, Joy Prakash Sain, Krishnaprasad Thirunarayan, A. Sheth, and Jeremiah A. Schumm. 2020. Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework. In *International Conference on Computational Linguistics*.

Xiangji Zeng, Yunliang Li, Yuchen Zhai, and Yin Zhang. 2020. Counterfactual generator: A weakly-supervised method for named entity recognition. In *Conference on Empirical Methods in Natural Language Processing*.

Chaoyuan Zuo, Qi Zhang, and Ritwik Banerjee. 2021. An empirical assessment of the qualitative aspects of misinformation in health news. In *Proceedings*

of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, pages 76–81, Online. Association for Computational Linguistics.