A Linguistic Grounding-Infused Contrastive Learning Approach for Health Mention Classification on Social Media

Usman Naseem
usman.naseem@sydney.edu.au
School of Computer Science, University of Sydney
Sydney, Australia

Jinman Kim
jinman.kim@sydney.edu.au
School of Computer Science, University of Sydney
Sydney, Australia

Matloob Khushi
matloob.khushi@brunel.ac.uk
Department of Computer Science, Brunel University
London, UK

Adam G. Dunn
adam.dunn@sydney.edu.au
School of Medical Sciences, University of Sydney
Sydney, Australia

ABSTRACT

Social media users use disease and symptoms words in different ways, including describing their personal health experiences figuratively or in other general discussions. The health mention classification (HMC) task aims to separate how people use terms, which is important in public health applications. Existing HMC studies address this problem using pretrained language models (PLMs). However, the remaining gaps in the area include the need for linguistic grounding, the requirement for large volumes of labelled data, and that solutions are often only tested on Twitter or Reddit, which provides limited evidence of the transportability of models. To address these gaps, we propose a novel method that uses a transformer-based PLM to obtain a contextual representation of target (disease or symptom) terms coupled with a contrastive loss to establish a larger gap between target terms' literal and figurative uses using linguistic theories. We introduce the use of a simple and effective approach for harvesting candidate instances from the broad corpus and generalising the proposed method using self-training to address the label scarcity challenge. Our experiments on publicly available health-mention datasets from Twitter (HMC2019) and Reddit (RHMD) demonstrate that our method outperforms the state-of-the-art HMC methods on both datasets for the HMC task. We further analyse the transferability and generalisability of our method and conclude with a discussion on the empirical and ethical considerations of our study.

CCS CONCEPTS

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

KEYWORDS

Health Mention Classification, Public Health Surveillance, Contrastive Learning, Social Media

WSDM '24, March 4–8, 2024, Merida, Mexico

ACM ISBN 979-8-4007-0371-3/24/03
https://doi.org/10.1145/3616855.3635763

ACM Reference Format:
Usman Naseem, Jinman Kim, Matloob Khushi, and Adam G. Dunn. 2024. A Linguistic Grounding-Infused Contrastive Learning Approach for Health Mention Classification on Social Media. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM '24), March 4–8, 2024, Merida, Mexico.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3616855.3635763

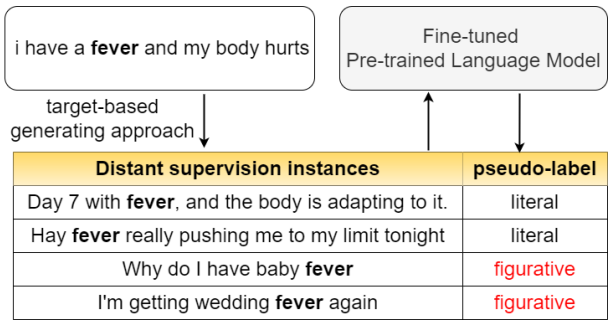
1 INTRODUCTION

Health mention classification (HMC) seeks to distinguish between user-generated content posted on social media platforms that discuss the personal experience with a symptom or condition and other, non-health related discussions that use the same symptom or condition keywords. A post on social media that includes a personal health mention is then assumed to mean that the author or the subject of the post has the health issue or symptom [21]. Data-driven public health surveillance might benefit from individuals sharing health experiences using disease or symptom terms [30].

A major challenge in HMC is distinguishing the literal usage of disease or symptom terms from the figurative usage of disease or symptom terms [2]. For example, in a tweet, "I have a **cold**, and need to see a doctor," *"cold"* is used literally, where a user is describing their health condition. Whereas, in "when America **sneezes**, the world catches a **cold**." *"sneezes"* and *"cold"* are used in figurative sense. Figurative use of symptoms and conditions can introduce errors and flawed conclusions in cases where social media-based natural language processing (NLP) epidemic intelligence tools depend on counts of keyword occurrences.

With the rapid development of contextualized representations, recently proposed HMC methods [2, 30, 31] use pretrained language models (PLMs) to extract context-dependent representations of disease or symptom terms and then fine-tune the PLMs for HMC tasks to achieve state-of-the-art (SOTA) results.

Using PLMs has shown promising performance for HMC [1, 19, 20, 29]. However, existing methods lack contrast between the literal and figurative use of target terms, that can be improved through analogical analysis in a given context depending on the linguistic context [10]. Second, fine-tuned LMs require relatively large volumes of labelled data to obtain SOTA accuracy on downstream tasks, including the HMC task [6, 17, 39]. Existing HMC datasets that have been made available are relatively small because annotation can be time-consuming to ensure robustness. Also, figurative phrases may require an expert for annotation and can be



Distant supervision instances	pseudo-label
Day 7 with **fever**, and the body is adapting to it.	literal
Hay **fever** really pushing me to my limit tonight	literal
Why do I have baby **fever**	figurative
I'm getting wedding **fever** again	figurative

Figure 1: Examples of selected candidate instances generated by target-based approach

biased [37], making distinguishing figurative usage of disease or symptom terms for HMC tasks a significant challenge.

Another challenge in the area relates to the generalisability and transferability of classifiers. For most previous experiments in HMC methods, the training and testing is done on the same social media platform (i.e., Twitter [2, 15, 18] or Reddit [30]). Understanding how different methods perform across domains and in multi-domain scenarios remains an important gap.

In this study, we aim to overcome the above limitations for improving the HMC by presenting a novel method that uses a contrastive loss to represent the contrast between the literal and figurative meanings of the target term, improving generalisation performance through self-training using unlabelled data generated by a simple approach. We used a transformer-based PLM to extract contextual information from a post containing a target term. If the target term is figurative, the semantic meaning is context-dependent and distinct from the literal usage. The literal usage of the target term can be explained using non-figurative examples. We use a contrastive loss to improve context-based representation between the target term's literal and figurative usage, making it more identifiable and allowing the classifier to make informed decisions.

We suggest using a target-based generating approach [24] motivated by a distantly supervised concept [13, 27] for automatically generating training data to solve label scarcity challenge. Specifically, all posts in the collected data that include the detection target term are collected and considered as potential candidate instances. We leverage the transformer-based PLM to produce pseudo-labels and include them in training examples to increase the amount of training data (Figure 1). To enhance the generalisation, we iteratively update the pseudo-labels using self-training. Our main contributions are summarised as follows:

- We propose a novel HMC method that uses contrastive loss for capturing the semantic inconsistencies in the use of disease or symptom terms used figuratively based on linguistic theories.
- Our method integrates semi-supervised learning with self training to deal with the label scarcity issue for HMC.
- Experimental results on two public benchmark HMC datasets show that the proposed method outperforms state-of-the-art HMC methods and is robust, generalisable, and transferable for the HMC task.

2 RELATED WORK

2.1 Existing HMC Methods

HMC Methods For Twitter: Karisani and Agichtein [18] was the first to release a personal health mentions (PHMs) dataset in 2017. They collected 7,192 English tweets using disease or symptom terms for binary classification of personal health mentions. They demonstrated that lexical and syntactic associations are high-utility features, which they combined with word embedding-based, and context-based features. Their method, WESPAD (Word Embedding Space Partitioning and Distortion), tackles the data scarcity problem; they distort and partition the word representations based on labels. However, for figurative mentions in HMC, the WESPAD's division and distortion of word embedding can reduce performance if a text contains a 'noisy' region and a filtration is applied [2].

Jiang et al. [15] released 12,331 manually annotated English tweets. The task was a binary classification of tweets being personal experience tweets (PETs) or non-PETs. They used Long Short Term Memory Network (LSTM) [12] to predict PET/non-PET. In addition to generic tweet preprocessing, in their method (JiangLSTM), they introduced some feature engineering resulting in improving the results over decision tree (DT) [36], support vector machine (SVM) [3] and k-nearest neighbours algorithm (kNN) [11] models.

In the HMC task, figurative mentions of symptom terms or diseases are important, which is firstly addressed by Iyer et al. [14]. They presented FeatAug+, where they first computed the average similarity score between words learned using an external database, i.e., the Sentiment140 dataset [9]. They further concatenated similarity scores with language-based features like part-of-speech tags and abstractness to detect figurative mentions of disease or symptom terms to enhance the HMC task.

Biddle et al. [2] released Twitter HMC2019 and included an additional label of a figurative mention of disease of symptom terms in the PHM2017 [18] dataset. To improve performance, their developed method BiLSTM-Senti employs features from sentiment-based linguistics and context word embeddings. They showed that the tweets which are misclassified are the ones where health is mentioned figuratively. Therefore, they argue that context and sentiments must be accounted for in the HMC task. The method improved on prior HMC methods and could be further improved by considering contextual information for sub-domain tasks [28].

Naseem et al. [31] addressed the shortcomings of BiLSTM-Senti and presented a new method where they combined domain-specific PLM and part-of-speech to improve the representation. The improved tweet representation was forwarded to bidirectional-LSTM with attention to disease or symptom terms and word-level linguistic features for the HMC. Experiments on Twitter HMC2019 showed that their method improved the overall performance.

HMC Methods For Reddit: Naseem et al. [30] argued that the existing HMC studies mostly focused on Twitter data, with limited disease or symptom terms coverage, ignoring how people interact on social media and leaving their user behaviour signatures in their text. To address the problem, they presented a Reddit health mention dataset (RHMD) covering 15 disease or symptom terms. By including emotional and domain-specific features with the specification of disease and symptom terms, they proposed Health-Mention

Classification Network (HMCNET). The HMCNET has three linguistic features i) neighbouring words and their part-of-speech tags, ii) detection of the presence or absence of subordinate clauses, and iii) health-related words derived from corpus features. In this work, the authors note that differentiating between symptoms or disease terms is still challenging and prone to errors.

2.2 Existing Pretrained Language Models

Pretrained language models (PLMs) have increased the accuracy on many downstream NLP tasks with appropriate fine-tuning [4, 25]. In HMC, to model the semantics and contextual features of target terms, previous studies used BERT and its domain-specific variants like PHS-BERT [32], COVID Twitter BERT (CT-BERT) [28] that are trained on social media data to model the contextual representation of the user-generated textual content on social media [2, 30, 31, 33]. These BERT-based PLMs are widely used to encode user-generated text in previous HMC studies. For example, Khan et al. [20] evaluated the performance of various BERT-based PLMs for health mention classification and showed that BERT trained on general corpus performed better than other domain-specific PLMs such as BioBERT and BertTweet. Karisani et al. [19] trained BERT with tweets to evaluate the performance of their method on the HMC task and showed that BERT trained on Wikipedia performs better compared to the domain-specific variants of a BERT. A recent study by Aduragba et al. [1] also showed that fine-tuned versions of BERT pretrained on Wikipedia performed better compared to domain-specific variants of BERT,i.e., trained on social media data.

Despite the efficacy of PLMs, one major constraint for fine-tuning PLMs is the need for a large amount of labelled training data. When labelled training data is insufficient, the performance of fine-tuned PLMs is degraded, and the number of parameters might result in overfitting [5, 38]. However, manually annotating and labelling high-quality, large-scale training data is labour-intensive for HMC.

2.3 Contrastive learning

Contrastive learning seeks to learn an embedding space with positive pairs nearby and negative pairs far apart. Supervised contrastive loss-based methods have demonstrated considerable success in various tasks [7, 8, 24, 26]. To the best of our knowledge, no prior work has used supervised contrastive learning for HMC on social media. We fill this gap and introduce an optimised loss function that uses supervised contrastive loss to improve the HMC.

3 METHOD

Given a social media post $P = \{t_1, t_2, ..., t_n\}$, where n indicates the number of tokens in a post with a target (i.e., disease or symptom) term $t_t \in P$. HMC task aims to classify whether a target term t_t is used figuratively or otherwise, i.e., a personal health mention or a non-personal health mention. Figure 2 illustrates an overview of the proposed method.

3.1 Pretrained Language Model

Our model uses a BERT-based pretrained language model (PLM)[1], as a sentence encoder to effectively capture the overall semantics and

[1]We used both general and domain-specific BERT-based PLM (see results section).

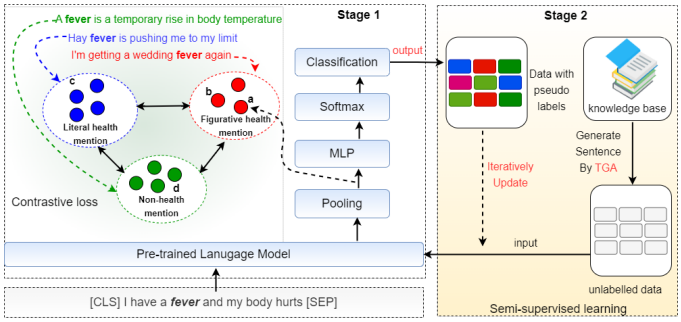


Figure 2: Overall architecture of proposed method.

contextual information of a given post P containing a target term t_t. Following [4], we placed "[CLS]" at the beginning and "[SEP]" at the ending of the input post P and fed the post P into BERT PLM (PLM_{BERT}) to extract the final hidden states $\mathbf{H}$ ((equation 1).

$$H = PLM_{BERT}([\texttt{CLS}], t_1, t_2, ..., t_n, [\texttt{SEP}]) \quad (1)$$

Since the target of our task is to determine if the semantic usage of a target term t_t in the post P is figurative or not. We extracted the context-specific representation of a target term t_t to correctly identify the use of a target term and used the average of a tokenized word to derive a fixed-size feature vector. Considering that the hidden states associated with the target term's subwords are from h_i to h_j, we average these hidden states (equation 2).

$$c = \frac{1}{j - i + 1} \sum_{k=i}^{j} h_k, \quad (2)$$

where c denotes the contextualised representation of a target term t_t. Then, to predict the figurative use of a term term t_t, we inject context enriched representation c to a multi-layer perceptron (MLP) with *tanh* as a activation function and a softmax layer (equation 3).

$$p = Softmax(W_2(tanh(W_1 c + b_1) + b_2) \quad (3)$$

where $W_1 \in \mathcal{R}^{dxd}$, $b_1 \in \mathcal{R}^d$, $W_2 \in \mathcal{R}^{2xd}$, and $b_2 \in \mathcal{R}^d$ and d represents BERT's hidden state size. The cross-entropy loss minimisation is used to tune the parameters (equation 4).

$$L_{cls} = \frac{1}{M} \sum_{m=1}^{M} y_m log(p_m) \quad (4)$$

where M denotes the dataset's size.

3.2 Contrastive Loss

We hypothesised that a figurative term could be recognised when the literal usage of a target term contrasts with the meaning that the term adopts in context. The difference between a word's contextual and literal sense is a key element for determining the figurative use of a target term.

The contrastive loss captures the contrastive relationship, making the classifier recognisable. The contrastive loss allows the figurative usage of the target term to have a closer semantic representation and separates literal usage. As illustrated in Figure 2, the target

word "*fever*" in cases "a" and "b" is used figuratively rather than its literal meaning to describe the health condition of a user, as shown in instance "c". Therefore, we expect the contextual representation of the target word "*fever*" in post *"a"* and *"b"* to be closer and farther away from the representation in post *"c"* and *"d"*.

Specifically, given a post P_a with a target term t_t as an anchor, P_p is a positive instance with target term t_t belonging to a similar category as P_a in batch $\mathcal{B}$, whereas P_n is a negative instance with target term t_t corresponding to another category in batch $\mathcal{B}$. We use equation 2 to compute their contextualised representations c_a, c_p and c_n (equation 5).

$$L_{co} = \sum_{(a,p,n)\in\mathcal{B}} d(c_a, c_p) + [\gamma - d(c_a, c_n)]_+, \tag{5}$$

where $[\cdot]_+$ denotes the functions $f(x) = max(0, x); d(\cdot,\cdot)$ denotes the L2-normalized euclidean distance, whereas γ is a controlling parameter that determines the margin.

The contrastive loss requires identifying similarities between instances of a similar class (label) and comparing them to instances from other classes (labels). When the data samples represent different classes, the contrastive loss widens the gap and maintains them apart by at least a margin γ. Capturing the separation in embedding representation between the target term's literal and figurative meanings is an essential feature of the HMC tasks.

3.3 Semi-supervised Learning

Motivated by [24], we use Target-based Generating Approach (TGA) to construct a large-scale training dataset without needing experts or sophisticated pre-defined rules using Wikipedia as a knowledge base (see section 4.2 – TGA implementation for details).

Target-based Generating Approach (TGA): The TGA is designed on a methodology in which if a target term acts as the identification target in a post, all other posts within a specific corpus having this target term serve as potential candidate examples. This approach efficiently generates a large-scale candidate set U using the target terms in the labelled data as heuristic seeds that can cover a wider range of topics without requiring any specific manual design. The fine-tuned PLM may then determine the labels of candidate examples, and high-confidence samples can be chosen to serve as the extended data. However, this depends on the PLM's performance, which may result in predictive error and noise.

Self training: To reduce the error and the noise in U, we employ self-training (ST) [22, 35] to construct pseudo-labels for candidate examples using the fine-tuned PLM and include these in the training data, where the pseudo-labels and PLM are iteratively updated. We create pseudo-labels $\hat{y_i} \in \mathcal{R}^K$ for every example $u_i \in U$ (equation 6).

$$\hat{y_{ij}} = \frac{p_{ij}^2 / f_j}{\sum_j p_{ij}^2 / f_j} \tag{6}$$

where p_{ij} is the j-th class prediction of u_i and $f_j = \sum_i p_{ij}$ is the summation of the soft frequencies of class j. Equation 6 generates $\hat{y_i}$ by enhancing high-confidence predictions and decreasing ones with less confidence by squaring and normalising the existing predictions and preserving more information. The ST objective is defined as a Kullback–Leibler (KL)-divergence loss between the current prediction P and the pseudo-label distributions (equation 7).

$$L_{st} = KL(\hat{Y}||P) = \sum_{i=1}^{|U|}\sum_{j=1}^{K} \hat{y}_{ij} log \frac{\hat{y}_{ij}}{p_{ij}} \tag{7}$$

3.4 Training Process

The overall loss function (L) incorporates contrastive loss L_{CO}, classification loss L_{CLS} for labelled data, and KL loss for unlabeled data U. The overall loss function is shown in equation 8.

$$L = L_{CLS} + \alpha L_{CO} + \beta L_{ST} \tag{8}$$

Where *alpha* and *beta* are hyperparameters used to optimise the contrastive and KL losses, respectively. Following [24], our method has a two-stage training process (Algorithm 1). First, we use labelled data to fine-tune the PLM with the first two terms of equation 8, which can effectively learn contrastive relations and improve the performance of the HMC tasks. The fine-tuned PLM is then used to determine soft pseudo-labels for all unlabelled data obtained by TGA. In the second stage, we use an ST approach to complement the data for training with pseudo-labelled data and iteratively optimise the PLM. We iteratively calculate soft pseudo-labels using current predictions during ST and adjust model parameters (equation 8).

Algorithm 1: Training Process

Input: labelled instances S; candidate instances U collected by TGA; PLM $f(\cdot;\theta)$.
Stage 1: *fine-tune PLM with labelled training data.*
Use the first 2 terms in the equation 8 to update θ on S.
Stage 2: *Optimise PLM using unlabelled data.*
for $t = 1, 2, ..., T$ **do**
  Generate pseudo-labels for U using equation 6.
  Use equation 8 to update θ on S.
end
Output: The final fine-tuned PLM $f(\cdot;\theta)$.

4 EXPERIMENTS

4.1 Datasets

We used two benchmark HMC datasets (i.e., HMC2019 and RHMD) that are widely used in previous HMC studies (Table 1).

Health mention classification 2019 (HMC2019): HMC2019 [2] consists of 15,393 English tweets and includes ten different disease

Table 1: Total is the number of posts in the HMC and RHMD datasets. HM, FHM and NHM represent the number of posts labelled as health mentions, figurative health mentions and non-health mentions, respectively. '# of disease' represents the total number of disease or symptom terms.

Dataset	Source	# of disease	FHM	NHM	HM	Total
HMC2019	Twitter	10	4073	7199	4121	15393
RHMD	Reddit	15	3225	3430	3360	10015

or symptom terms such as headache, parkinson, stroke, cough, depression, fever, cancer, migraine, heart attack and alzheimer. The proportion of disease or symptom terms that are used figuratively varies greatly in HMC2019, with a standard deviation of 20.48%; the percentage of figurative disease or symptom use varies from 3.11% for Parkinson's disease to 65.30% for a heart attack.

Reddit health mention dataset (RHMD): RHMD [30] consists of 10,015 Reddit posts that mention 15 common disease or symptom terms such as migraine, asthma, diabetes, OCD, cough, depression, fever, addiction, allergy, stroke, alzheimer, PTSD, headache, cancer, and heart attack. There is considerable variation in the proportion of disease or symptom terms among the labels in RHMD dataset. For example, the number of posts mentioning depression is 41 (6.46%), whereas posts mentioning heart attacks were 582 (66.90%). The standard deviation of disease or symptom terms used figuratively is 16.92% in RHMD. We used the public version of the dataset that combines figurative and hyperbolic health mention classes.

4.2 Experimental Settings

Preprocessing: Consistent with previous studies [31], to enhance the quality of the informal nature of posts, decrease the impact of out-of-vocabulary words and fix social media-specific consumer health vocabulary (CHV) terms [16] such as 'massiveheadache,' 'PSTD' 'OCD' and 'Migrane,' our preprocessing steps included a spelling correction, and emoticons/emojis were replaced.

Parameter settings and other training details: We used a grid search optimisation technique to derive the best parameters of our model. Specifically, we used AdamW [34] to optimise the parameters of our model. We trained our classifier for 30 epochs, stopping early after ten epochs. The contrastive objective's margin γ is fixed at 1.0. After empirical evaluation, we fixed α to 0.2 and β to 0.05. For a fair comparison, we kept our settings the same as the previous studies [2, 30, 31]. We used the base version of PLMs using the HuggingFace Python library. All models used the same experimental parameters and 10-fold cross-validation (CV) for consistency, and reported results are averaged across folds.

TGA implementation: We initially harvest target phrases set as triggers in all datasets and then leverage TGA to retrieve huge target-related candidate examples from the same corpora for semi-supervised learning. We leverage Wikipedia[2] as the knowledge base (KB) since it comprises a broad range of topics, making it a suitable database that is typically easy and inexpensive to access. We retrieve and select content from the English Wikipedia dump[3] to build huge candidate sets and then leverage the NLTK package to convert documents into sentences and deduplicate them.

4.3 Evaluation Metrics

The performance of our model is evaluated using F1-score, precision, recall and a custom metric TN_{FHM} used in previous similar works on HMC [2, 30, 31].

$$TN_{FHM} = \frac{tnfhm}{fhm}, \tag{9}$$

[2]We also used data from other sources to generate target-related candidate instances but empirically found that using Wikipedia performed better. Due to page limit restrictions, results obtained using Wikipedia data are reported here.

[3]https://dumps.wikimedia.org/enwiki/20210201/

Table 2: Proposed v/s the baselines. F1, Precision (Pre), and Recall (Rec) scores are averaged across ten folds. * shows that our method obtained a significant ($p < 0.05$) improvement over HMCNET under Mann–Whitney U test.

Model\Dataset	HMC2019			RHMD		
	F1	Pre	Rec	F1	Pre	Rec
BERT	0.76	0.75	0.77	0.65	0.68	0.63
BioBERT	0.73	0.71	0.75	0.63	0.65	0.62
CT-BERT	0.79	0.78	0.80	0.67	0.65	0.68
PHS-BERT	0.80	0.78	0.79	0.68	0.67	0.69
WESPAD	0.52	0.53	0.40	0.59	0.60	0.60
FeatAug+	0.57	0.52	0.53	0.51	0.51	0.51
JiangLSTM	0.70	0.70	0.66	0.63	0.63	0.63
BERT-MTL	0.77	0.76	0.78	0.67	0.69	0.65
BiLSTM-Senti	0.81	0.81	0.80	0.68	0.67	0.68
BiLSTM-Attn+Senti	0.85	0.84	0.85	0.71	0.70	0.71
HMCNET	0.89	0.89	0.89	0.75	0.75	0.75
Proposed (+BERT)	0.93*	0.93*	0.93*	0.79*	0.78*	0.79
Proposed (+PHS-BERT)	0.95*	0.95*	0.95*	0.81*	0.81*	0.81*

where TN_{FHM} is referred to as "the percentage of FHMs accurately identified as non-health mentions, i.e., true negatives."

4.4 Baselines

We evaluated the performance of our method with existing SOTA methods (discussed in section 2). For PLMs we used BERT [4], BioBERT [23], CT-BERT [28] and PHS-BERT [32]. For HMC methods, we used *FeatAug+* [14], WESPAD [18], JiangLSTM [15], BERT-MTL [1], BiLSTM-Senti [2], BiLSTM-Attn+Senti [31] and HMCNET [30]. We fine-tuned the PLMs to encode the posts and utilised the grid-search CV technique to obtain the optimised parameters.

5 RESULTS

5.1 Comparison with Baselines

Overall comparison: Results in Table 2 demonstrate that our method outperforms all previous HMC methods on the HMC task, with an F1 score of 0.93 on HMC2019 and an F1 score of 0.79 on RHMD using BERT PLM (an absolute increase of 4% on both datasets than the results of the next best baseline, i.e., HMCNET) and an F1 score of 0.95 on HMC2019 and an F1 score of 0.81 on RHMD using PHS-BERT, a domain-specific PLM (an absolute increase of 6% on both datasets than HMCNET). HMCNET additionally uses linguistic features in their model, whereas our method does not utilise language-based features. We show that PLMs (e.g., BERT, BioBERT, CT-BERT and PHS-BERT) that are trained on general and domain-specific corpus perform better compared to methods (e.g., WESPAD, FeatAug+, JiangLSTM) that use non-contextual methods to encode text; however these PLMS are less desirable for HMC tasks due to their inability to contrast between the literal and figurative sense of target (disease or symptom) terms. We also note that the methods (e.g., FeatAug+, BERT-MTL, BiLSTM-Senti, BiLSMT-Attn+Senti) that are designed to capture figurative mentions were less able to understand the context in which a target term was used. Our

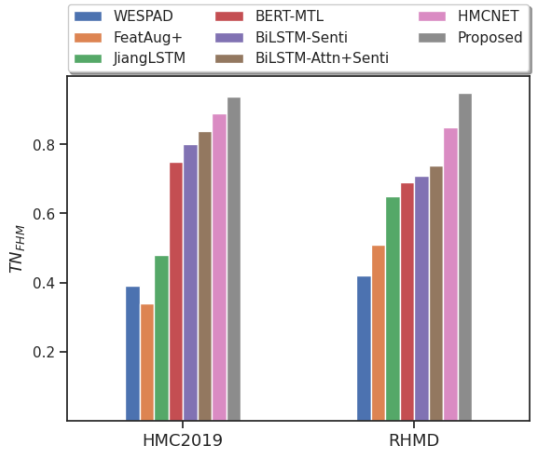


Figure 3: Correct prediction of figurative (TN_{FHM}) use of disease or symptom terms

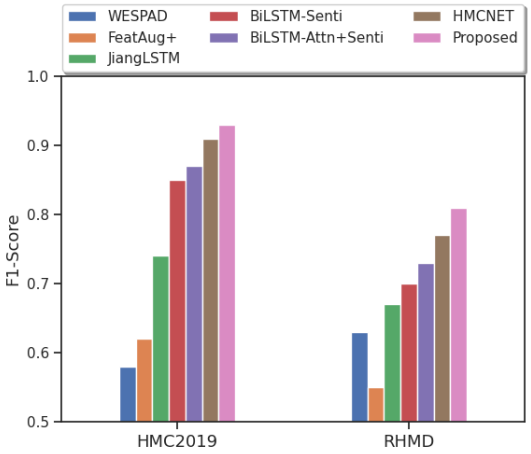


Figure 4: Comparison using additional training data (ATD) used in proposed method

method performs better because it can fully use unlabelled data harvested by the proposed TGA that enhances the model generalisation by ST and explicitly capture the contrast between the literal and figurative mentions of disease or symptom terms by a contrastive objective. Not surprisingly, the approaches based on transformer-based PLMs (e.g., proposed, HMCNET, BiLSTM-Senti, BERT-MTL and BiLSTM-Attn+Senti) are consistently better compared to the other HMC methods (e.g., WESPAD, FeatAug+, JiangLSTM) due to the strong expressive power of transformer-based PLMs to capture rich semantics and contextual information into the representations. Thus, below we will compare the performance of the proposed method with only HMC methods.

Identification of figurative health mentions: Figure 3 demonstrates the effectiveness of our method in capturing figurative health mentions (TN_{FHM}) in comparison to the HMC baselines. Our method consistently outperforms all other baselines in identifying a disease or symptom term used as figurative health mentions with an TN_{FHM} score of 0.94 on HMC2019 and an TN_{FHM} score of 0.95 on RHMD, which is an absolute increase of 5% on HMC2019 and 10% increase on RHMD than the HMCNET. We attribute this increase in performance to the effective target-based generating approach (TGA) that improves the robustness of our method using self-training and is designed to capture the contrast between the figurative and other mentions of disease or symptom terms by a contrastive loss.

Baselines with additional training data: We also investigated the effect of additional training data (ATD) generated using a target-based generating approach (TGA) and self-training in baselines (Figure 4). Although we observed an increase in the overall performance (F1-Score) of each tested baseline on both datasets; however, the proposed method still outperforms all the baselines. We postulate this to the use of contrastive loss in our method that helps our model to the contrast between literal and figurative use of target terms. Hence, we conclude that the proposed target-based generating approach (TGA) and self-training address the data scarcity issue. The contrastive loss further improves the performance due to its ability to capture a literal and non-literal sense of target words.

Transferability Test: We also performed the transportability test where we first performed cross-domain (i.e., train on HMC2019 and test on RHMD and vice-versa) and multi-domain (i.e., combined HMC2019 and RHMD as HMC+RHMD) evaluation of our method to validate the effectiveness and robustness of our method (Figure 5). It is clear from Figure 5 that the performance of our method drops less compared to other methods. Our method outperforms previous methods when trained on HMC2019, which contains 10 target terms and tested on RHMD, which contains 15 disease terms, i.e., 5 new target terms that are not seen during the training phase and consistently outperforms other baselines on both settings (i.e., cross-domain and multi-domain). In the first setting (i.e., cross-domain), when we trained on HMC2019 and tested on RHMD, our method achieved the highest performance compared to the baselines. The highest performance achieved by our method is

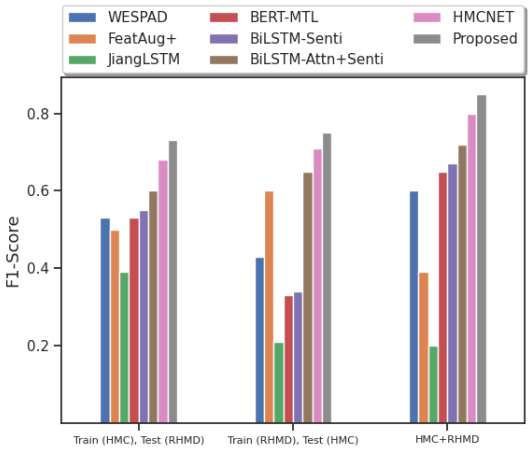


Figure 5: Transferability test.

Table 3: Binary scheme: Comparing performance of our method with the SOTA baselines. F1, Precision (Pre), and Recall (Rec) scores are averaged over 10 folds. * represents that our method obtained a significant ($p < 0.05$) improvement over HMCNET under Mann–Whitney U test.

Model\Dataset	HMC2019			RHMD		
	F1	Pre	Rec	F1	Pre	Rec
WESPAD	0.67	0.67	0.67	0.73	0.74	0.73
FeatAug+	0.75	0.76	0.73	0.71	0.72	0.71
JiangLSTM	0.72	0.73	0.72	0.73	0.74	0.73
BERT-MTL	0.80	0.79	0.81	0.74	0.74	0.74
BiLSTM-Senti	0.85	0.85	0.85	0.76	0.79	0.77
BiLSTM-Attn+Senti	0.88	0.88	0.88	0.78	0.78	0.78
HMCNET	0.92	0.92	0.91	0.81	0.81	0.80
Proposed	0.96*	0.96*	0.95*	0.85*	0.87*	0.85*

73% which is highest than the baselines. We observed a similar improvement in performance when using RHMD for training and HMC2019 for testing. The highest performance achieved is 75%, the highest among all tested results. In our second setting (multi-domain), where we combined HMC2019 and RHMD and tested our method on combined HMC+RHMD data. We observed that our method outperforms all the baselines in multi-domain settings and achieved an F1-Score of 85%. This transportability test on both cross-domain and multi-domain settings validates the robustness and generalisability of our method.

Labels segregation (Binary label scheme): Following previous HMC studies [2, 30], we observe results for an alternative scheme (i.e., binary label) where we used only the figurative health mention (FHM) and health mention (HM) classes (Table 3). The result shows that the predictive models perform better when we adopt this binary scheme than when we use the original three-class scheme. We attribute this increase in performance to the potential of reducing the level of freedom of the output variables (binary classes instead of three). These findings also indicate that all methods perform poorly when evaluated on terms used in three classes of fine-grained use of disease or symptom terms. We also notice that our method outperforms all previous HMC methods when tested on a binary label scheme with an F1-Score of 0.96 on HMC2019 and 0.85 on RHMD, which is an absolute increase of 4% on both datasets than the HMCNET (Table 3).

Disease or symptom wise comparison: Our method obtained the highest F1-Score on both datasets for each of the disease or symptom terms (increases in F1-Score range from 0.68% to 11.16%) compared to the HMC baselines (Table 4). The highest increase in performance on HMC2019 is observed for 'Cancer' (3.18%) and 'Cough' (3.31%), whereas for RHMD, the highest increase in performance is noted for 'OCD' (11.16%) and 'Allergic' (9.83%). We postulate this increase to the target-based generating approach (TGA), which constructs diverse training data from Wikipedia containing various topics and avoids the model's tendency to be biased toward a specific domain (disease/symptom). The lowest increase was observed for the 'fever' (0.92%) term on HMC2019 and for 'Alzheimer' (0.68%) on RHMD. This poor performance is due to the infrequent use of these disease or symptom terms, as figurative health mentions. We conclude that our method is robust for determining the use of disease or symptom terms in HMC because it outperforms previous HMC methods in determining an individual disease or symptom term.

5.2 Analysis

Ablation analysis: We conduct an ablation analysis to evaluate the effectiveness of each component of our method (Table 5). We compare our method variants without the contrastive loss (w/o CO) and the self-training (w/o ST). The result shows that each component is essential for our method, as removing any of them would significantly decrease performance. When the self-training is excluded, the F1-score drops by 4% on HMC2019 and 3% on RHMD, demonstrating the importance of coupling semi-supervised learning to enhance the generalisation. The contrastive loss captures the contrast between the literal and figurative semantics of the target terms and is useful for our method, and removing CO results in a drop of 3% on HMC2019 and 2% on RHMD. Hence, we infer that our method's strengths lie in using both self-training and contrastive loss, which contributes to increased performance.

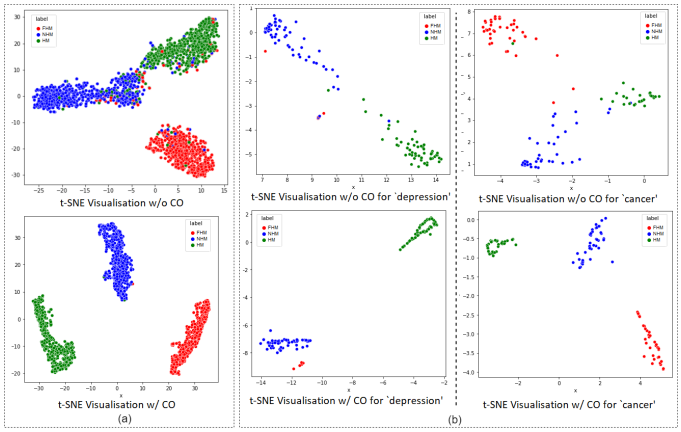


Figure 6: Embedding Visualisation of (a) labels w/o CO (top) and w/ CO (bottom) and (b) target words 'cancer' and 'depression' w/o CO (top) and w/ CO (bottom). Red represents the figurative health mention (FHM), blue represents the non-health mention (NHM) and green represents the health mention (HM) use of disease or symptom terms.

Qualitative analysis (Embedding Visualisation): We visualise the contextual embeddings (equation 2), coloured by their labels (Figure 6a) and their disease or symptom terms (Figure 6b). As shown in Figure 6, the boundary of different label dots in our method (bottom) is more pronounced than that in ablated models without a contrastive loss (top), which reveals that when the contrastive loss is removed, the literal and figurative and other mentions of disease or symptom terms are less distinguishable. Based on our hypothesis, a figurative health mention is detected if the literal usage of the target term contrasts with its contextual meaning. As expected, the proposed contrastive loss explicitly widens the

Table 4: Disease or symptom-wise performance. F1-Score is averaged across ten folds, and the following (i.e., second-best) result is underlined. Percentage Improvement (%) shows the increase in performance over the second-best result. *represents that our method obtained a significant ($p < 0.05$) improvement over the second best method (underlined) under Mann–Whitney U test.

Disease\Model		WESPAD	FeatAug+	JiangLSTM	BERT-MTL	BiLSTM-Senti	BiLSTM-Attn+Senti	HMCNET	Proposed	Performance Increase (%)
HMC2019	Alzheimer	0.49	0.61	0.59	0.65	0.69	0.78	0.75	**0.80***	2.17%
	Cancer	0.43	0.56	0.42	0.60	0.66	0.73	0.79	**0.82***	3.18%
	Cough	0.46	0.39	0.56	0.66	0.81	0.66	0.89	**0.93***	3.31%
	Depression	0.55	0.42	0.54	0.58	0.70	0.61	0.77	**0.79***	1.61%
	Fever	0.56	0.39	0.59	0.65	0.81	0.72	0.88	**0.89***	0.92%
	Headache	0.55	0.40	0.54	0.69	0.84	0.77	0.89	**0.91***	2.00%
	Heart attack	0.48	0.43	0.52	0.70	0.80	0.79	0.86	**0.88***	1.43%
	Migraine	0.65	0.41	0.69	0.72	0.83	0.80	0.90	**0.91***	1.01%
	Parkinson	0.48	0.68	0.44	0.61	0.67	0.80	0.78	**0.83***	2.91%
	Stroke	0.49	0.50	0.49	0.69	0.78	0.70	0.85	**0.87***	1.48%
RHMD	Addiction	0.57	0.65	0.43	0.56	0.65	0.57	0.73	**0.81**	7.60%
	Allergic	0.53	0.72	0.39	0.48	0.50	0.65	0.69	**0.79***	9.83%
	Alzheimer	0.58	0.80	0.23	0.39	0.41	0.81	0.70	**0.82***	0.68%
	Asthma	0.60	0.68	0.32	0.35	0.37	0.67	0.76	**0.77***	1.43%
	Cancer	0.61	0.65	0.32	0.45	0.49	0.70	0.79	**0.85***	6.44%
	Cough	0.51	0.59	0.28	0.31	0.35	0.50	0.72	**0.74***	2.46%
	Depression	0.58	0.67	0.31	0.39	0.41	0.74	0.68	**0.76***	2.86%
	Diabetes	0.61	0.83	0.33	0.45	0.49	0.70	0.76	**0.80***	3.50%
	Fever	0.54	0.62	0.37	0.43	0.46	0.67	0.75	**0.82***	7.71%
	Headache	0.55	0.65	0.38	0.50	0.37	0.73	0.65	**0.75***	2.98%
	Heart attack	0.65	0.67	0.46	0.51	0.56	0.76	0.80	**0.85***	5.26%
	Migraine	0.61	0.51	0.41	0.44	0.45	0.69	0.75	**0.82***	7.48%
	OCD	0.42	0.55	0.32	0.35	0.36	0.56	0.49	**0.68***	11.16%
	PSTD	0.58	0.73	0.42	0.49	0.56	0.76	0.66	**0.81***	5.04%
	Stroke	0.54	0.61	0.26	0.42	0.45	0.67	0.60	**0.72***	5.55%

gap between the target word's literal and figurative usage of target terms in embedding space and models more concise representations for data from the similar class (bottom of Figure 6a and Figure 6b).

5.3 Error Analysis

Below we present some of the errors made by our method and describe its limitations. For example, for a post: *'**coca cola is actual **cough** syrup URL**'* and "*call my girlfriend**asthma**breath away**" and "*group of kids**called**migraine**," our method was not able to correctly classify a figurative usage of '*cough*', '*asthma*' and '*migraine*'. We postulate these incorrect predictions due to a lower count of terms in our datasets, posts with a lack of information, and the use of an external link (i.e., *URL*).

Table 5: Ablation analysis: Proposed represents a complete model, Proposed w/o CO is a complete model without contrastive loss, and Proposed w/o ST represents a complete model without self-training. F1, Precision (Pre), and Recall (Rec) scores are averaged across 10-folds. *represents that our method obtained a significant ($p < 0.05$) improvement than other variants under Mann–Whitney U test.

Model\Dataset	HMC2019			RHMD		
	F1	Pre	Rec	F1	Pre	Rec
Proposed	0.93*	0.93*	0.93*	0.79*	0.78*	0.79*
Proposed w/o CO	0.90	0.91	0.91	0.77	0.77	0.75
Proposed w/o ST	0.89	0.90	0.90	0.76	0.76	0.76

6 CONCLUSION

We proposed a novel method that uses self-training to build a simple but effective approach to determine figurative uses of disease or symptom terms using the pre-trained backbone to extract contextualised information. Specifically, we leverage a simple approach that automatically builds large data for self training and integrates a contrastive loss to capture semantic inconsistencies in figurative mentions. Our results indicated that our method is robust, transferable across different social media platforms and outperform SOTA HMC benchmarks for the HMC task of distinguishing personal health mentions across a set of disease and symptom terms.

ETHICAL CONSIDERATIONS

This study aims to enhance social media applications and other health surveillance tools that automatically detect health mentions on social media. The annotated datasets we used are publicly available [2, 30] and include de-identified publicly available posts where users understand public access and there is no expectation of privacy. Hence, no ethical approval is required for this research.

REFERENCES

[1] Olanrewaju Tahir Aduragba, Jialin Yu, Alexandra Cristea, and Yang Long. 2023. Improving Health Mention Classification Through Emphasising Literal Meanings: A Study Towards Diversity and Generalisation for Public Health Surveillance. In *ACM Web Conference 2023-Proceedings of the World Wide Web Conference, WWW 2023*. ACM, 3928–3936.

[2] Rhys Biddle, Aditya Joshi, Shaowu Liu, Cecile Paris, and Guandong Xu. 2020. Leveraging Sentiment Distributions to Distinguish Figurative From Literal Health Reports on Twitter. In *Proceedings of The Web Conference 2020*. 1217–1227.

[3] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[4] 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[5] Jingfei Du, Edouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Ves Stoyanov, and Alexis Conneau. 2020. Self-training improves pre-training for natural language understanding. *arXiv preprint arXiv:2010.02194* (2020).

[6] Jingfei Du, Édouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Veselin Stoyanov, and Alexis Conneau. 2021. Self-training Improves Pre-training for Natural Language Understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 5408–5418.

[7] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. 2021. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9588–9597.

[8] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 6894–6910.

[9] Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford* 1, 12 (2009), 2009.

[10] Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and symbol* 22, 1 (2007), 1–39.

[11] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. 2003. KNN model-based approach in classification. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, 986–996.

[12] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735 arXiv:https://doi.org/10.1162/neco.1997.9.8.1735

[13] Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*. 541–550.

[14] Adith Iyer, Aditya Joshi, Sarvnaz Karimi, Ross Sparks, and Cecile Paris. 2019. Figurative usage detection of symptom words to improve personal health mention detection. *arXiv preprint arXiv:1906.05466* (2019).

[15] Keyuan Jiang, Shichao Feng, Qunhao Song, Ricardo A Calix, Matrika Gupta, and Gordon R Bernard. 2018. Identifying tweets of personal health experience through word embedding and LSTM neural network. *BMC bioinformatics* 19, 8 (2018), 210.

[16] Keyuan Jiang, CHEN Tingyu, Ricardo A Calix, and Gordon R Bernard. 2018. Identifying consumer health terms of side effects in Twitter posts. *Studies in health technology and informatics* 251 (2018), 273.

[17] Giannis Karamanolakis, Subhabrata Mukherjee, Guoqing Zheng, and Ahmed Hassan. 2021. Self-Training with Weak Supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 845–863.

[18] Payam Karisani and Eugene Agichtein. 2018. Did you really just have a heart attack? Towards robust detection of personal health mentions in social media. In *Proceedings of the 2018 World Wide Web Conference*. 137–146.

[19] Payam Karisani, Joyce C Ho, and Eugene Agichtein. 2020. Domain-guided task decomposition with self-training for detecting personal events in social media. In *Proceedings of The Web Conference 2020*. 2411–2420.

[20] Pervaiz Iqbal Khan, Imran Razzak, Andreas Dengel, and Sheraz Ahmed. 2022. Performance comparison of transformer-based models on twitter health mention classification. *IEEE Transactions on Computational Social Systems* (2022).

[21] Alex Lamb, Michael Paul, and Mark Dredze. 2013. Separating fact from fear: Tracking flu infections on twitter. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 789–795.

[22] Dong-Hyun Lee et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, Vol. 3. 896.

[23] 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. arXiv:1901.08746 [cs.CL]

[24] Zhenxi Lin, Qianli Ma, Jiangyue Yan, and Jieyu Chen. 2021. CATE: A contrastive pre-trained model for metaphor detection with semi-supervised learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 3888–3898.

[25] 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:1907.11692 [cs.CL]

[26] Orchid Majumder, Avinash Ravichandran, Subhransu Maji, Alessandro Achille, Marzia Polito, and Stefano Soatto. 2021. Supervised Momentum Contrastive Learning for Few-Shot Classification. *arXiv preprint arXiv:2101.11058* (2021).

[27] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. 1003–1011.

[28] Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter. *arXiv preprint arXiv:2005.07503* (2020).

[29] Usman Naseem, Matloob Khushi, Jinman Kim, and Adam G Dunn. 2022. RHMD: a real-world dataset for health mention classification on Reddit. *IEEE Transactions on Computational Social Systems* (2022).

[30] Usman Naseem, Jinman Kim, Matloob Khushi, and Adam G Dunn. 2022. Identification of disease or symptom terms in reddit to improve health mention classification. In *Proceedings of the ACM Web Conference 2022*. 2573–2581.

[31] Usman Naseem, Jinman Kim, Matloob Khushi, and Adam G Dunn. 2022. Robust Identification of Figurative Language in Personal Health Mentions on Twitter. *IEEE Transactions on Artificial Intelligence* (2022).

[32] Usman Naseem, Byoung Chan Lee, Matloob Khushi, Jinman Kim, and Adam G Dunn. 2022. Benchmarking for Public Health Surveillance tasks on Social Media with a Domain-Specific Pretrained Language Model. *NLP-Power 2022* (2022), 22.

[33] Usman Naseem, Surendrabikram Thapa, Qi Zhang, Liang Hu, Junaid Rashid, and Mehwish Nasim. 2023. Incorporating historical information by disentangling hidden representations for mental health surveillance on social media. *Social Network Analysis and Mining* 14, 1 (2023), 9.

[34] Matthew E Peters, Sebastian Ruder, and Noah A Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. *arXiv preprint arXiv:1903.05987* (2019).

[35] Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. 2005. Semi-supervised self-training of object detection models. (2005).

[36] S Rasoul Safavian and David Landgrebe. 1991. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics* 21, 3 (1991), 660–674.

[37] Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 248–258.

[38] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems* 33 (2020), 6256–6268.

[39] Yue Yu, Simiao Zuo, Haoming Jiang, Wendi Ren, Tuo Zhao, and Chao Zhang. 2021. Fine-Tuning Pre-trained Language Model with Weak Supervision: A Contrastive-Regularized Self-Training Approach. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1063–1077.