Improving Health Mention Classification Through Emphasising Literal Meanings: A Study Towards Diversity and Generalisation for Public Health Surveillance

Olanrewaju Tahir Aduragba
Durham University
Durham, UK
Kwara State University
Malete, Nigeria
olanrewaju.m.aduragba@durham.ac.uk

Jialin Yu
Durham University
Durham, UK
University College London
London, UK
jialin.yu@durham.ac.uk,jialin.yu@ucl.ac.uk

Alexandra I. Cristea
Durham University
Durham, UK
alexandra.i.cristea@durham.ac.uk

Yang Long
Durham University
Durham, UK
yang.long@durham.ac.uk

ABSTRACT

People often use disease or symptom terms on social media and online forums in ways other than to describe their health. Thus the NLP health mention classification (HMC) task aims to identify posts where users are discussing health conditions literally, not figuratively. Existing computational research typically only studies health mentions within well-represented groups in developed nations. Developing countries with limited health surveillance abilities fail to benefit from such data to manage public health crises. To advance the HMC research and benefit more diverse populations, we present the *Nairaland health mention dataset (NHMD)*, a new dataset collected from a dedicated web forum for Nigerians. NHMD consists of 7,763 manually labelled posts extracted based on four prevalent diseases (HIV/AIDS, Malaria, Stroke and Tuberculosis) in Nigeria. With NHMD, we conduct extensive experiments using current state-of-the-art models for HMC and identify that, compared to existing public datasets, NHMD contains out-of-distribution examples. Hence, it is *well suited for domain adaptation studies*. The introduction of the NHMD dataset imposes better diversity coverage of vulnerable populations and generalisation for HMC tasks in a global public health surveillance setting. Additionally, we present a *novel multi-task learning approach for HMC tasks by combining literal word meaning prediction as an auxiliary task*. Experimental results demonstrate that the proposed approach outperforms state-of-the-art methods statistically significantly ($p < 0.01$, Wilcoxon test) in terms of F1 score over the state-of-the-art and shows that our new dataset poses a strong challenge to the existing HMC methods.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Language resources**; • **Information systems** → **Web mining**.

KEYWORDS

Datasets, Health Mention Classification, Public Health Surveillance, Multi-task Learning

ACM Reference Format:
Olanrewaju Tahir Aduragba, Jialin Yu, Alexandra I. Cristea, and Yang Long. 2023. Improving Health Mention Classification Through Emphasising Literal Meanings: A Study Towards Diversity and Generalisation for Public Health Surveillance. In *Proceedings of the ACM Web Conference 2023 (WWW '23), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3543507.3583877

1 INTRODUCTION

Public health emergencies have become a major concern globally because of the negative impact they have on economic growth and stability, as well as the way of life of a population. Particularly in low and middle-income countries, public health poses a significant threat to their economic development. For example, in Africa, half of the human mortality is attributed to infectious diseases [11]. These illness outcomes add to the already high unemployment rates in these countries, thereby affecting economic productivity.

To mitigate the consequences of these public health emergencies, gathering, analysing and interpreting health-related data through surveillance are critical [36]. The widespread use of social media has allowed greater access to real-time data at a low cost, making it suitable for digital public health surveillance [47]. Social media users frequently discuss personal experiences on various health-related topics, such as prescription drugs, treatment, symptoms and general experience with the disease. Such data can be aggregated and analysed to provide population-level insights. Leveraging these data sources can contribute to achieving the third goal of the Sustainable Development Goal (SDG) - ensuring good health and well-being [4].

Previous works on using social media data for public surveillance have focused on a range of applications, including detecting and

WWW '23, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04.
https://doi.org/10.1145/3543507.3583877

monitoring outbreaks [48], monitoring adherence to public health guidelines [26] and tracking health and well-being during a global pandemic [3]. Compared to traditional surveillance methods, such as analysing clinical records from hospitals, laboratory reports and surveys, social media data has the advantage of being a real-time data source. They can also act as early detection systems for disease outbreaks and frequently pick up on patterns that more traditional techniques of health surveillance might otherwise overlook [24]. An essential step to using social media data for public health surveillance is identifying content related to health reports. This task has been formulated as a Health Mention Classification (HMC) [7]. The goal of the HMC task is to determine whether or not a particular social media post reports a health condition.

Most research on public health surveillance on social media, including HMC, has focused on social media sites, e.g. Twitter and Reddit, mainly used in developed nations [2]. They have failed to explore other online data sources popular within underrepresented communities, particularly in low- and middle-income countries with low public health surveillance capabilities and rising social media usage [34]. Although internet penetration continues to grow in Africa, with a 23% increase between 2019 and 2021[1], no work has explored data for public health surveillance from online communities in these countries. This results in a data bias for HMC tasks and impinges on their original goals towards diversity and generalisation for public health surveillance purposes.

In this paper, to address this bias in existing datasets, we focus on creating **a dataset mainly used by people from underrepresented communities to address the significant gap in the availability and quality of health-related data between developed and developing countries**. We construct our data from Nairaland[2], a dedicated online forum for Nigerians. We focus on HIV/AIDS, malaria, stroke and tuberculosis. HIV/AIDS, malaria and tuberculosis account for 27% of the disease burden from communicable diseases in Nigeria [42]. We perform extensive analysis on our proposed dataset, NHMD and other popular publicly available HMC datasets, to demonstrate how NHMD addresses the issue of data bias. Additionally, we proposed a novel multi-task learning framework, converting a novel label augmentation method to explore the literal meaning of disease-related words given a context as an auxiliary task, to improve the public health classification task performance. Our contributions include:

- Introduce NHMD, which is, to the best of our knowledge, the first health mention dataset for underseved populations that is publicly available - thus addressing the missing distribution of existing publicly available HMC datasets and mitigating the data bias problem.
- Propose a novel literal emphasised multi-task learning framework for the HMC task and achieve state-of-the-art performance across various HMC datasets.
- Study the generalisation ability of HMC models across existing HMC datasets, and investigate the language variations across the datasets.

2 RELATED WORK

2.1 Public Health Surveillance on Social media

The task of surveillance on social media focuses on extracting health-related posts, to estimate cases of diseases or monitor disease spread. Prior work showed that classifying social media posts as being related (or not) to a disease or health condition provided promising surveillance results [20]. Among all public social media sources, Twitter was the most widely used platform for public health surveillance [8]. This could be partly attributed to its relatively open data policy compared to other platforms, such as Facebook. The availability of geolocation information in tweets, which may be used to highlight geographical patterns of public health cases, was another advantage of Twitter data. Another platform that has been growing in use for public health applications is Reddit. Reddit also provides a publicly accessible API to collect its data.

Existing datasets on Health Mention Classification include one presented by Jiang et al. [15]. They manually annotated 12,331 tweets with either personal experience tweets (PET) or non-PET. The tweets were collected using the names of drugs. Karisani and Agichtein [16] introduced another dataset related to six health conditions: Alzheimer's disease, heart attack, Parkinson's disease, cancer, depression, and stroke. Biddle et al. [7] extended the dataset in [16] to cover four additional health conditions: cough, fever, headache and migraine. In contrast to these datasets that focused on Twitter, Nassem et al. [29] created an HMC dataset using Reddit. Their dataset consists of 10,015 Reddit posts collected using 15 disease or symptom terms. Several other datasets also exist that are an approximation of health mentions, including a Twitter dataset focused on distinguishing awareness of flu reports (*negative*) from actual infection reports (*positive)* [20], a dataset for detecting illness in tweets [17] and dataset for identifying medical self-disclosure in online health communities, such as *patient.info* [40].

Relatively few researchers have focused on using social media to improve public health surveillance in developing countries, where public health crises are most prevalent. This potentially leads to data bias and discourages generalisation in HMC tasks for public surveillance purposes. Some of the existing research aimed to address the issue of data bias, includes work by Odlum & Yoon [31] that explored using tweets to track the Ebola outbreak in Nigeria. They collected tweets using relevant keywords and performed a content analysis on them. Similarly, Shahid et al. [35] used tweets to understand the spread of the Dengue epidemic in Bangladesh and how such insight can be used to guide public health policy.

Our work contributes to this literature by exploring a social media platform specifically built and used by people from a developing country such as Nigeria. We present the Nairaland health mention dataset (NHMD) as our primary contribution to this paper; the introduction of the NHMD dataset imposes better diversity coverage of vulnerable populations and generalisation for HMC tasks in a global public health surveillance setting.

2.2 Health Mention Classification

In terms of approaches used for health mention classification, baseline results have utilised simple techniques, including WESPAD - Word Embedding Space Partitioning and Distortion [16] - that learns to distort the word embedding space, to more effectively

[1] https://www.itu.int/en/ITU-D/Statistics/Documents/facts/FactsFigures2021.pdf
[2] www.nairaland.com

distinguish cases of actual health reports from non-health reports. Deep learning approaches, including Long Short-Term Memory Networks (LSTM) [15] and bidirectional LSTM [7, 29] have also been used. Given the dominance of contextual word representations in NLP tasks, more recent work [29] have focused on using contextual word representations such as BERT [10] and ELMo [33] for health mention detection. For example, [7] compared the performance of contextual and non-contextual word representations for health-mention tweets. They found that contextual representations provided more helpful information for the HMC task and emphasised the importance of understanding the context in which disease or symptoms words are used. Thus, we have selected pretrained contextual language models as a strong baseline in this paper.

Some works have also shown that incorporating auxiliary information, such as sentiment and emotion information [7], user behaviour [29] or detecting the use of disease or symptom words figuratively [7, 14] can improve the performance on HMC tasks. Multi-task learning (MTL) is a popular approach to incorporating auxiliary information into the primary task. MTL is a transfer learning paradigm that involves jointly training multiple tasks to share knowledge between these tasks, in order to improve generalisation and the performance on a primary task [46]. Multi-task learning can be useful to compensate for low-resource settings, such as limited annotated data availability, by sharing knowledge from related tasks [27]. Multi-task learning has been successfully applied to several NLP tasks, such as explanation generation [44], pandemic information classification [45], identification of crisis location [19], and they have demonstrated improvements where tasks are related.

In this work, we explore detecting the literal use of a disease word, as an auxiliary task, in a multi-task setting to improve HMC primary task performance. Existing literature on HMC has demonstrated that combining linguistic phenomena such as figurative or literal usage of a disease word improves performance on HMC tasks [14, 29]. However, they have mainly focused on extracting them as features and using them in combination with task-specific features, or the tasks have been trained independently. In addition to the HMC task, we jointly learn the literal usage of a disease word. Following [28], the literal usage detection predicts whether a disease word in a given post is used literally or not.

3 NAIRALAND HEALTH MENTION DATASET

In this section, we present our health mention dataset from the largest Nigerian online community, Nairaland: NHMD. We detail the data collection and filtering, annotation procedures, and present an analysis of the dataset.

3.1 Data Collection and Filtering

There are no publicly accessible health mention datasets for underserved populations at the moment. Thus, developing such a dataset is crucial for the equality and diversity of the health mention research community, and we consider this as our primary contribution. We selected Nairaland, since it is the most popular online community used by Nigerians [5]. The forum is the most visited indigenous site in Nigeria and the ninth most visited site in the country[3]. With 33 million visits per month[4], the site provides a general discussion platform for Nigerians. As of October 2022, the forum has 2,946,061 members and 7,136,617 topics[5].

In contrast to popular social media sites such as Twitter or Reddit, which offer official APIs, Nairaland does not provide any API for collecting data from the site. Therefore, we crawled the Nairaland website using Beautiful Soup[6] (see section 6 for ethical considerations). We select forums where health-related topics are likely to be discussed, i.e. Health and Politics forums. We retrieved all the posts in these forums from March 2005, when Nairaland was created, until April 2022. In total, we collected 20,995,525 posts from both forums.

We filtered the posts to contain only keywords (i.e. HIV, AIDS, tuberculosis, malaria and stroke) for the diseases we are interested in. We apply length filtering to only include posts between 3 to 120 tokens long (length matching with existing HMC datasets). We further sample randomly from the remaining posts for annotation, while maintaining the distribution across diseases. This resulted in 7,763 posts - an acceptable number, slightly more than the datasets introduced by Karisani and Agichtein [16].

To preserve users' privacy, we replaced all usernames or references to names with the <USER> token. We also removed any website links, emails or phone numbers from all posts.

3.2 Data Annotation

We used two annotators to label the dataset. The annotators are Nigerian undergraduate students fluent in English and their local language, with one studying a health-related course. They also are proficient in Nigerian Pidgin, an English-based creole language spoken across Nigeria (see section 3.3.2). Additionally, these annotators are well-versed in Nigerian culture and humour, which is vital for understanding contexts.

We adopt the annotation guidelines in [7] and define 3 classes: *health mention*, *other mention* and *non health mention*. See Table 1 for examples in each class. Each post can only be annotated with a single label based on annotation agreements. We asked the authors to skip any instance they were unsure about. The dataset is annotated in two steps: the preparation step and the production step. In the preparation step, for quality control and training, the authors first annotated 100 posts as a guideline for the annotations to begin with. Then, the annotators are asked to annotate the same batch of 100 posts. Both annotators achieved at least 70% agreement with the authors' annotations in the preparation batch (100 posts). For the 30% dataset with non-agreement reached, we manually went through the examples with the annotations and discussed the mislabelled instances to ensure they understood the label categories fully. We emphasised the significance of basing assessments solely on the details expressly contained in a given post and avoiding any further assumptions. In the production step, we sent the whole dataset to the annotators with our first annotated examples.

We consider only posts that both annotators have labelled. For instances where both annotators disagreed, we first consider the level of disagreement between annotators. For example, suppose one of the annotators selects *non health mention*, and the other annotator selects *health mention*. In that case, we assume this instance

[3] https://techcabal.com/2021/10/11/the-next-wave-wrestling-us-cyber-dominance/
[4] https://www.similarweb.com/website/nairaland.com/#overview
[5] www.nairaland.com
[6] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Table 1: Example of annotations and corresponding label descriptions

Sample post	Disease	Label	Description
i am HIV + and to tell u there's no need to commit suicide or what have u.all u have to do is get committed to taking your drugs religiously ,eat well and stay healthy.	HIV/AIDS	Health mention	The post contains a health mention using a disease term. The author of the post or someone has a certain disease, or has corresponding symptoms
If you are that knowledgeable about Tuberculosis..You should know that being infected with the bacteria is not the same thing as being a Tuberculosis patient.	Tuberculosis	Other mention	The post contains the disease term but does not mention a specific person health. Discuss disease or symptom or discuss prevention of disease or symptoms in general.
OP's English fit give pesin Malaria sef.	Malaria	Non health mention	The post contains the disease terms used metaphorically, departing from the literal meaning, not aligning with commonsense, mock usage, or sarcastic expression

Table 2: Inter-Annotator Agreement across diseases

Disease	Kappa (κ)
HIV/AIDS	0.6033
Malaria	0.6517
Stroke	0.7806
Tuberculosis	0.6866

is difficult, and we discard the post. For cases with a smaller annotation difference, we forward these to a third annotator (a Nigerian with a Bachelor's degree qualification) to label and determine the final label, based on the majority vote. In the event there is no majority, we remove the post.

We measure the inter-annotator agreement using Cohen's kappa [9]. The average Cohen's kappa across the entire dataset is $\kappa = 0.67$. According to [22], the score indicates a strong agreement between the annotators. We also calculate the Kappa score (κ) per disease, to verify the agreement across the diseases (see Table 2). As can be seen, the agreement is consistent across diseases, with stroke-related posts having the highest agreement. This suggests that no disease-specific posts are more difficult to annotate than others.

3.3 Dataset Analysis

In this section, we conduct an extensive analysis of our proposed data on the following aspects: data statistics and language distribution.

3.3.1 Dataset Statistics. Table 3 shows the statistics of the dataset. We observe that the majority of the posts (64%) are labelled as *other mention.* This is the overall trend across diseases, except for stroke, where posts labelled as *non health mention* are the majority. Our label distribution is similar to a popular public HMC dataset created by Briddle et al. [7].

In terms of coverage of diseases, posts related to HIV/AIDS and Malaria are the majority, with 54% (4137/7763) and 37% (2860/7763) of the posts, respectively. Tuberculosis-related posts are the least represented, at only 1% (93/7763). Posts related to stroke account for 8% (673/7763) of the posts. The uneven distribution of posts across the diseases shows the focus of the discussion on Nairaland on HIV/AIDS over other diseases considered in this research.

Table 3: Dataset Statistics

Disease	Health mention	Other mention	Non health mention	Total
HIV/AIDS	221	2,855	1,061	4,137
Malaria	820	1,742	298	2,860
Stroke	90	288	295	673
Tuberculosis	17	54	17	93
Total	1,148	4,939	1,676	7,763

3.3.2 Language Distribution. Nigeria is a multilingual society, and English is the common language adopted as the official language to enhance communication. However, the contact of indigenous languages with the English language has led to the development of Nigerian Pidgin [38]. Nigerian Pidgin is spoken widely across Nigeria, and it has been suggested that it makes communication easier on the Nairaland forum [39]. To determine the proportion of our datasets that contain Nigerian Pidgin, we use Franc[7], a Language Identification Tool trained on 403 languages, including Nigerian Pidgin. Franc has shown superior performance on the Nigerian Pidgin dataset [1]. Of the 7,763 posts, 1,527 (20%) are in Nigerian Pidgin, while 6,233 posts (80%) are in English[8]. Although Franc can detect other major Nigerian languages, we observe that none of the posts was identified as any of the widely spoken languages (Hausa, Yoruba or Igbo) We suspect this phenomenon occurs because the forums are for the general public, so posters will use languages most people understand in Nigeria.

3.3.3 Dataset Split. We randomly split our dataset into training (80%), validation (10%) and test (10%) sets to promote reproducibility and facilitate comparisons between HMC models. The training set was used to train our models, while the validation set was used to choose hyperparameters, and the test set was used to evaluate the performance of the models. We make our dataset split publicly available[9] and the breakdown of the splits is provided in Table 4.

4 EXPERIMENTS

In this section, we detailed our experiments, including baseline models, proposed methods, evaluation metrics, hyperparameter search, results and discussions.

[7] https://github.com/wooorm/franc

[8] The tool could not determine the language of 3 posts.

[9] Code and data are available at https://github.com/tahirlanre/nairaland_hmc

Table 4: Train, validation and test splits per class

Label	Train	Validation	Test
Health mention	923	112	113
Other mention	3,950	491	496
Non health mention	1,335	173	168

4.1 Baseline Models

Several machine learning models have been applied to the task of HMC [14, 16], and current state-of-the-art models for HMC tasks are based on pre-trained language models (PLMs) [18, 30]. We consider the following PLMs as our baseline models: BERT, ROBERTa and ALBERT.

- **BERT**: Bidirectional Encoder Representations from Transformers [10] is a language model pre-trained on unlabelled English texts Transformers [41]. The pre-training objective of BERT focuses on learning contextualised representations of words that can be useful for downstream applications. BERT has achieved exceptional performance across many natural language understanding tasks [3, 10].
- **RoBERTa**: Robustly optimized BERT approach (RoBERTa) [25] is a descendent of BERT introduced with modified pre-training objectives to create a more robust model. RoBERTa outperformed BERT on several NLP benchmark tasks [25].
- **ALBERT**: A Lite BERT (ALBERT) [21] was proposed to reduce the size of parameters and lower memory consumption. ALBERT has 12 million parameters compared to BERT, which has 110 million parameters. This is well-suited for low-resource settings where computing memory is limited.

4.2 Datasets

We use the three publicly available HMC datasets for our experiments:

- **PHM2017**: this dataset is a collection of English tweets related to Alzheimer's disease, heart attack, Parkinson's disease, cancer, depression, and stroke [16]. The dataset contains 4,987 instances labelled with either *personal health mention*, *awareness*, *other mention* and *non health mention*.
- **HMC2019**: This dataset is an extension of the PHM2017 dataset. The creators [7] of the dataset added tweets related to four additional health condition: cough, fever, headache and migraine. The dataset contains 14,051 posts labelled as *health mention*, *other mention* and *figurative mention*.
- **RHMD**: Unlike the other datasets, this dataset is a Reddit-based data that covers 15 disease or symptom terms [29]. Generally, the posts in this dataset are longer than the Twitter-based datsets. The dataset consists of 10,015 posts annotated as either *figurative mention*, *non-pesoronal health mention* or *health mention*.
- **NHMD**: Our proposed dataset in this paper with a detailed description in section 3.

4.3 Label Mapping

There are slight differences in the labels used for these datasets. Thus, we map the original labels to three classes: *health mention*, *other mention* and *non health mention* to create a uniform label distribution[10].

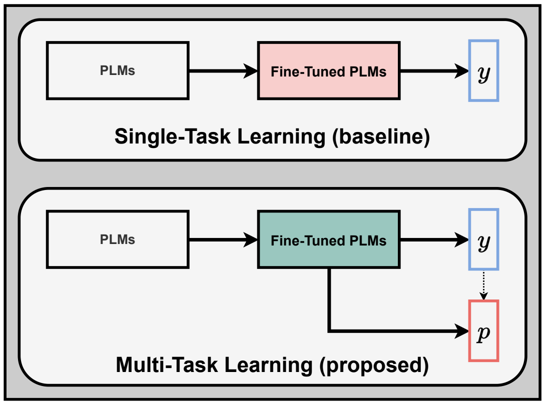
PLMs
Fine-Tuned PLMs
y
Single-Task Learning (baseline)
PLMs
Fine-Tuned PLMs
y
p
Multi-Task Learning (proposed)

Figure 1: Multitask learning framework to emphasise literal meanings as an auxiliary task (demonstrated as the red block) for personal health mention classification tasks.

4.4 Compared Method: Fine-Tuning PLMs

Before introducing our novel multi-task learning approach, we start with the standard finetuning approach with pretrained language models (PLMs), shown as the single-task learning method in Figure 1. In this section, we describe the fine-tuning method, using BERT as an example; all other models presented in section 4.1 can be interchangeably adopted. From this point, we represent data in its vectorised form and ignore the number suffix, for clarity.

For the fine-tuning approach, given data $(\boldsymbol{x}, \boldsymbol{y})$, we first pass $\boldsymbol{x}$ through the PLM, here, a BERT model, and retrieve its contextual representation $\boldsymbol{h}$:

$$\boldsymbol{h} = BERT(\boldsymbol{x}) \tag{1}$$

Then we directly map the contextual representation $\boldsymbol{h}$ to its label $\boldsymbol{y}$ through an affine transformation:

$$\hat{\boldsymbol{y}} = Softmax(\boldsymbol{W_1} * \boldsymbol{h} + \boldsymbol{b_1}) \tag{2}$$

Finally, we calculate the cross-entropy loss between the prediction label $\hat{\boldsymbol{y}}$ and the ground truth label $\boldsymbol{y}$. We denoted this loss from fine-tuning the model as the HMC loss:

$$L_{ft} = L_{HMC} \tag{3}$$

Although the method of fine-tuning PLMs is very intuitive and simple, it is the state-of-the-art method for health mention classification tasks and is considered as a very strong baseline method to compare with.

4.5 Proposed Method: Literal Emphasised Multi-task Learning

In this section, we present our proposed novel multi-task learning framework for health mention classification tasks, as shown in Figure 1. We propose to explicitly model the *literal* usage of a disease word in the text context as an auxiliary task.

[10] We follow the annotation descriptions provided in each dataset.

Table 5: Main Results Between PLMs Fine-Tuning Baselines and Our Proposed Framework.

	NHMD			PHM2017			HMC2019			RHMD		
Model	**Precision**	**Recall**	**Macro F_1**	**Precision**	**Recall**	**Macro F_1**	**Precision**	**Recall**	**Macro F_1**	**Precision**	**Recall**	**Macro F_1**
Single-task Learning Models (baseline)												
BERT	80.56	76.29	78.08	86.51	84.12	85.23	89.47	88.91	89.12	80.47	80.31	80.32
RoBERTa	82.94	80.55	81.25	83.00	86.02	84.33	90.00	89.35	89.58	81.27	80.84	80.87
ALBERT	81.7	78.44	79.86	84.41	84.24	84.29	88.26	87.49	87.84	78.32	78.38	78.23
Multi-task Learning Models (proposed)												
BERT-MTL	81.75 ↑	78.62 ↑	79.98 ↑	86.53 ↑	84.19 ↑	85.28 ↑	89.65 ↑	89.17 ↑	89.35 ↑	80.69 ↑	80.34 ↑	80.43 ↑
RoBERTa-MTL	83.08 ↑	81.08 ↑	81.91 ↑	85.14 ↑	86.51 ↑	85.77 ↑	90.46 ↑	89.95 ↑	90.16 ↑	81.60 ↑	81.10 ↑	81.18 ↑
ALBERT-MTL	82.04 ↑	79.65 ↑	80.74 ↑	85.67 ↑	84.83 ↑	85.16 ↑	88.65 ↑	87.77 ↑	88.18 ↑	78.44 ↑	78.50 ↑	78.39 ↑

4.5.1 Pseudo-Literal Label Generation. Since the literal label is unknown, we create a pseudo literal label $\boldsymbol{p}$ (as demonstrated with a dashed arrow in Figure 1) for a given text pair $(\boldsymbol{x}, \boldsymbol{y})$ from its existing label $\boldsymbol{y}$, based on the following rule: if a post is originally labelled as either *health-mention* or *other mention* in $\boldsymbol{y}$, then we assume the disease word is labelled as literally in $\boldsymbol{p}$. Otherwise, if the original label $\boldsymbol{y}$ is *non health*, we assume the use of the disease word is labelled as non-literal in $\boldsymbol{p}$. By assigning the pseudo-literal label, we are essentially using the same amount of data in its original form, and no additional human labelling process is introduced. With the pseudo-literal label induced, we convert the original dataset from $(\boldsymbol{x}, \boldsymbol{y})$ to the following form $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{p})$.

4.5.2 Literal Emphasised Multi-task Learning. In this section, we describe our multi-task learning method, same as in section 4.4, using BERT as an example; all other models presented in section 4.1 can be interchangeably adopted.

For our multi-task learning approach, given data $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{p})$, we first pass $\boldsymbol{x}$ through the PLM - here a BERT model - and retrieve its contextual representation $\boldsymbol{h}$:

$$\boldsymbol{h} = BERT(\boldsymbol{x}) \tag{4}$$

Then we directly map the contextual representation $\boldsymbol{h}$ to its label $\boldsymbol{y}$ through an affine transformation and to its pseudo label $\boldsymbol{p}$ through a complex non-linear transformation:

$$\begin{aligned} \hat{\boldsymbol{y}} &= Softmax(\boldsymbol{W}_1 * \boldsymbol{h} + \boldsymbol{b}_1) \\ \hat{\boldsymbol{p}} &= \sigma(\boldsymbol{W}_3(tanh(\boldsymbol{W}_2 * \boldsymbol{h} + \boldsymbol{b}_2) + \boldsymbol{b}_3) \end{aligned} \tag{5}$$

Finally, we calculate the cross-entropy loss between the prediction label $\hat{\boldsymbol{y}}$ and the ground truth label $\boldsymbol{y}$, we denoted this loss from multi-task learning model as the HMC loss, L_{HMC}; we calculate the cross-entropy loss between the prediction label $\hat{\boldsymbol{p}}$ and the ground truth label $\boldsymbol{p}$, we denoted this loss from multi-task learning model as the Literal loss, $L_{literal}$:

$$L_{mtl} = L_{HMC} + \lambda * L_{literal} \tag{6}$$

Where λ is a tunable weight hyperparameter that controls the importance placed on the auxiliary task.

4.6 Evaluation Metrics

We evaluate the performance of the models on our dataset using precision, recall and F1-score following previous work on HMC [7, 16], and with all performance reported on the test set. To account for variability, we perform five independent runs using different seeds for each model and report the average results over five runs.

4.7 Hyperparameter Selection

We select the best hyperparameters based on the average validation F1-score across 5 seeds. The range of hyperparameters is summarised as follows: batch size ε {16, 32}, learning rate ε {1e-5, 2e-5, 3e-5, 5e-5}, loss weight parameter λ ε [0.0, 1.0] for multi-task experiments. All models were trained for 5 epochs using the Adam optimiser. We use a dropout of 0.2 for all models.

4.8 Results and Discussion

Table 5 presents the results of the single-task and multi-task models. The table shows precision, recall and F1 scores for the test set. In terms of the single-task models, RoBERTa is superior to the other PLMs on all datasets except PHM2017 where BERT achieved the best performance. ALBERT (lite version of BERT), a smaller model, achieved significantly better results than BERT on NHMD. This is a promising result, particularly in low-income countries like Nigeria, with limited access to powerful computing resources.

In general, the multi-task models, where we jointly model the literal usage of disease words with the HMC task, consistently outperform their corresponding single-task models across all datasets in terms of precision, recall and F1 scores. The improvements are generally statistically significant based on the Wilcoxon test ($p < 0.01$) over five runs with random seeds. We speculate that this is because the model learns to identify the context in which the disease word is used to determine whether a text is a health mention. The RoBERTa-based multi-task model, RoBERTa-MTL achieves the best performance across all datasets.

For our proposed method, the NHMD dataset had the highest gains, with the improvement ranging from 0.7% - 1.9% on the F1 score. We suggest that the literal meaning of the disease words, e.g. HIV/AIDs and malaria, used to collect NHMD contains more information that is beneficial to the HMC task. We also note that these words are less likely to be used in figurative contexts when compared to other disease or symptoms words, such as headache, and depression, used in the other datasets. This phenomenon can be justified from an information theory perspective: if one event is less likely to happen, it generates more information when it happens. Hence, modelling the health mention dataset (NHMD) when the disease keywords are less likely to be used in figurative contexts results in most information gain and better improvements in performance. On the remaining datasets, the performance improvements of the multi-task models over the single-task models are also substantial. For instance, on the PHM2017 dataset, RoBERTa-MTL and Albert-MTL improve on their corresponding single-task models by at least 0.9% on the F1 score. Overall, the results demonstrate the feasibility and generalisation of our proposed approach model if a disease word is used literally or not.

5 FURTHER ANALYSIS

5.1 Domain Shift and Generalisation

The domain of existing HMC datasets varies in terms of where they are extracted from (e.g. Twitter and Reddit), the disease or health condition they focus on (e.g. cancer, heart attack, HIV/AIDS) and their target population (e.g. mainly based on text from developed countries). The distinction in the data domain imposes high selection bias and potentially leads to domain shift. The domain shift in data harms the generalisation of the models when tested on an out-of-distribution dataset in a text classification setting [43]. For public health classification tasks, it is critical for the systems to react to unseen diseases from other domains [23]. To address this challenge, previous research has proposed to use domain adaptation to leverage datasets from related domains [13]. Domain adaptation is particularly useful in public health research, where the availability of labelled data is limited, as a result of the cost or expert annotators and sudden-onset of a public health emergency, such as the global COVID-19 pandemic. Nevertheless, the generalisation performance of models is expected to drop under domain shift due to underlying distributional shifts [37]. Recent work by Harrigian et al. [12] showed that mental health models generalise poorly across multiple social media platforms. To this end, we evaluate the domain generalisation for current HMC datasets and discuss whether our proposed dataset, NHMD alleviates this issue.

5.2 Analysis Setting

For the domain adaptation, we explore the following settings:

5.2.1 Single-Source (In-Domain) -> Single-Target (In-Domain).

In this setting, we perform the standard fine-tuning, as described in section 4.4, using single-source data and report results based on its corresponding in-domain single target dataset (i.e. we report results on the test split of PHM2017 when it is trained on the training split of PHM2017). The results for this experiment are denoted as 'S(I) -> S(I)' in Table 7.

5.2.2 Multiple-Source (In-Domain) - Single-Target (In-Domain):

In this setting, we again perform the fine-tuning, as described in section 4.4, using multiple-source data and report results for each individual target dataset (i.e. we report results on the test split of PHM2017 when it is trained on a combination of training split of PHM2017, HMC2019, RHMD and NHMD). The results for this experiment are denoted as 'MI) -> S(I)' in Table 7.

5.2.3 Single-Source (In-Domain) - Single-Target (Out-Domain):

In this setting, we perform out-of-domain experiments by training a model on a single HMC dataset (source), e.g. PHM2017 and test on another HMC dataset (target), e.g. NHMD. For this experiment, we aim to understand and quantify the effect of out-domain generalisation on HMC tasks with unseen examples. The results for this experiment are denoted as 'S(In) ->S(O)' in Table 7.

5.2.4 Multiple Source (In Domain) - Single Target (Out Domain):

In this setting, we perform another set of out-of-domain experiments by training a model on a single HMC dataset (source), e.g. PHM2017 and test on another HMC dataset (target), e.g. NHMD. For this experiment, we aim to understand and quantify the effect of out-domain generalisation on HMC tasks with unseen examples. The results for this experiment are denoted as 'S(In) ->S(O)' in Table 7.

5.3 Analysis Results and Discussion

Table 7 shows the results of the domain adaptation experiments using the RoBERTa model, which is the overall best performing architecture based on the results presented in Table 5. We report the average F1 score from five independent runs using different seeds.

In the first part of the table, we examine the in-domain generalisation of the datasets: from *Single-Source (In-Domain) -> Single-Target (In-Domain)* and *Multiple-Source (In-Domain) -> Single-Target (In-Domain)* experiment, in most cases, the in-domain generalisation performance drops statistically significantly ($p < 0.01$ based on the Wilcoxon test), with the exception of the RHMD dataset.

In the second part of the table, we examine the out-domain generalisation of the datasets: for *Single-Source (In-Domain) -> Single-Target (Out-Domain)*; we present results with source dataset used on the left column with respect to its out-domain test datasets. Additionally, we report the mean average of the out-domain performance as an indication on the average generalisation performance. For *Multiple-Source (In-Domain) -> Single-Target (Out-Domain)*, we combine all the datasets as training and test their out-domain performance on each single target dataset. The results suggest that, in most cases, the out-domain generalisation performance improves statistically significantly ($p < 0.01$ based on the Wilcoxon test) with the exception of the RHMD dataset. We also observe that models trained with Twitter-based datasets (PHM2017 & HMC2019) transfer to the Reddit-based dataset (RHMD) better than models trained on our dataset (NHMD) in the *Single-Source (In-Domain) -> Single-Target (Out-Domain)* setting. The negative transfer to NHMD from other datasets is notably higher, with a 14 - 23% decrease in the F1 score. The results are similar in the reverse direction, except for PHM2017, where the transfer from NHMD performs better than RHMD. These results demonstrate the importance of our dataset, which aims to mitigate the data selection bias in HMC tasks.

In summary, we can confidently claim that our proposed dataset, NHMD, imposes better diversity coverage of vulnerable populations and generalisation for HMC tasks in a global public health surveillance setting.

5.4 Linguistic Analysis

Understanding the underlying language variations can highlight the differences between the datasets. Thus, we conducted a further analysis, by comparing the topic distribution of the collected posts in our datasets, NHMD, with three other popular public health mention datasets (PHM2017, HMC2019 and RHMD), based on the LIWC package [32][11]. We report the Pearson correlation of the top 10

[11] https://www.liwc.app/

Table 6: LIWC feature correlations across classes for all datasets, sorted by Pearson correlation (r).

PHM2017						**HMC2019**					
Health mention		*Other mention*		*Non health mention*		*Health mention*		*Other mention*		*Non health mention*	
LIWC category	r	LIWC category	r	LIWC category	r	LIWC category	r	LIWC category	r	LIWC category	r
health	0.092	pronoun	0.426	health	0.247	Clout	0.218	pronoun	0.389	prep	0.180
Lifestyle	0.072	det	0.299	emo_neg	0.153	Culture	0.092	verb	0.333	adj	0.101
Clout	0.051	focuspast	0.292	Clout	0.143	Lifestyle	0.092	auxverb	0.282	Clout	0.100
sexual	0.050	verb	0.261	prep	0.135	curiosity	0.063	Authentic	0.278	Drives	0.084
curiosity	0.042	Authentic	0.213	tone_neg	0.134	attention	0.058	det	0.263	tone_pos	0.077
Culture	0.037	adverb	0.193	curiosity	0.113	sexual	0.051	focuspast	0.205	curiosity	0.074
attention	0.034	auxverb	0.192	Drives	0.098	socrefs	0.038	adverb	0.188	death	0.061
prep	0.017	socrefs	0.138	cogproc	0.067	socbehav	0.03	focuspresent	0.176	Lifestyle	0.051
tone_pos	0.016	swear	0.117	socbehav	0.056	health	0.026	acquire	0.124	time	0.049
want	0.010	conj	0.116	death	0.054	tone_pos	0.015	negate	0.120	tone_neg	0.047
adj	0.009	focuspresent	0.110	attention	0.034	cogproc	0.008	swear	0.108	feeling	0.046
RHMD						**NHMD**					
Health mention		*Other mention*		*Non health mention*		*Health mention*		*Other mention*		*Non health mention*	
LIWC category	r	LIWC category	r	LIWC category	r	LIWC category	r	LIWC category	r	LIWC category	r
Clout	0.287	pronoun	0.306	Authentic	0.144	sexual	0.190	pronoun	0.151	auxverb	0.162
health	0.192	verb	0.205	negate	0.120	Clout	0.148	focuspast	0.119	cogproc	0.116
Culture	0.079	auxverb	0.159	conj	0.108	Culture	0.081	time	0.091	focuspast	0.110
Lifestyle	0.064	Authentic	0.149	adj	0.099	cogproc	0.049	feeling	0.082	focuspresent	0.110
socrefs	0.055	focuspresent	0.134	prep	0.097	det	0.046	socbehav	0.081	verb	0.101
curiosity	0.043	focuspast	0.131	tone_pos	0.091	prep	0.044	swear	0.071	prep	0.098
sexual	0.042	det	0.13	cogproc	0.085	health	0.043	socrefs	0.067	health	0.096
attention	0.028	adverb	0.119	curiosity	0.066	swear	0.039	tone_pos	0.066	conj	0.071
prep	0.019	conj	0.085	feeling	0.065	negate	0.038	acquire	0.063	quantity	0.070
death	0.004	feeling	0.065	Drives	0.063	socrefs	0.038	Drives	0.063	adj	0.064
socbehav	0.004	socrefs	0.063	lack	0.059	death	0.034	Authentic	0.062	Authentic	0.060

Table 7: Macro F1 score for the domain adaptation experiments.

	PHM2017	**HMC2019**	**RHMD**	**NHMD**
In Domain Generalisation				
S(I) -> S(I)	85.28	90.69	80.87	81.25
M(I) -> S(I)	84.79 ↓	89.66 ↓	84.13 ↑	79.05 ↓
Out Domain Generalisation				
S(I) -> S(O)				
PHM2017	-	76.42	80.31	67.51
HMC2019	80.77	-	68.66	58.93
RHMD	74.04	80.59	-	58.34
NHMD	76.32	72.00	67.33	-
Average	77.04	76.34	72.10	61.59
M(I) -> S(O)	79.8 ↑	77.98 ↑	69.92 ↓	63.2 ↑

topics for each label in Table 6. A LIWC feature value measures the proportion of words used across posts in a specific label matching a given LIWC dimension. The version of LIWC (LIWC-22) we used covers over 100 language dimensions.

We note some similarities in the topics prevalent across all the HMC datasets. For example, *Health, e.g. illness* related topics are present in *health mention* posts for all datasets. This is unsurprising, as we expect the latter to cover health discussions. Word use related to other physical and health dimensions, *e.g. sexual*, are also prevalent, but they associate more with *health mention* posts in NHMD. However, we also note some differences between our dataset, NHMD, and the remaining datasets. For instance, topics related to *Lifestyle (e.g. home, work, money)*, *Perception (e.g. attention)* and *Motives (e.g. curiosity)* are present in PHM2017, HMC2019 and RHMD. However, in NHMD, we note more use of words related to *Negations, e.g. not, nothing*, *Determiners (i.e. det), e.g. articles and numbers* and association with *Quantities (e.g. all, one, more).*

6 ETHICAL CONSIDERATIONS

The dataset we use is from Nairaland, a web forum that is publicly available. Given the sensitive nature of data containing the health status of people, extra precautions were followed during all data collection and analysis in accordance with advice from Benton et al. [6]. Hence, ethical approval is not required for this research.

7 CONCLUSION

We propose here a multi-task learning approach combining health mention classification (HMC) with literal keyword use identification. Our experimental results demonstrate that our approach outperforms the state-of-the-art baseline approaches. A further key contribution is the construction and release of NHMD: a new benchmark dataset from underrepresented communities, extracted based on 4 types of disease from a public forum. Extensive analysis on its transferability and generalisation capacity suggests that our dataset contributes to the domain generalisation of the HMC task. Implications include the potential to improve HMC with *literal* identification as an auxiliary task; and also highlight the importance of introducing and using a dataset from the wider community, especially underrepresented groups, to ensure fairness, robustness and generalisation for public health surveillance.

REFERENCES

[1] Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2022. AfroLID: A Neural Language Identification Tool for African Languages. *arXiv preprint arXiv:2210.11744* (2022).

[2] Olanrewaju T Aduragba and Alexandra I Cristea. 2019. Research on Prediction of Infectious Diseases, their spread via Social Media and their link to Education. In *Proceedings of the 2019 4th International Conference on Information and Education Innovations*. 38–42.

[3] Olanrewaju Tahir Aduragba, Jialin Yu, Alexandra I Cristea, and Lei Shi. 2021. Detecting Fine-Grained Emotions on Social Media during Major Disease Outbreaks: Health and Well-being before and during the COVID-19 Pandemic. In *AMIA Annual Symposium Proceedings*, Vol. 2021. American Medical Informatics Association, 187.

[4] Yara M Asi and Cynthia Williams. 2018. The role of digital health in making progress toward Sustainable Development Goal (SDG) 3 in conflict-affected populations. *International journal of medical informatics* 114 (2018), 114–120.

[5] Idowu R Badmus, Simon A Okaiyeto, and Lambe K Mustapha. 2019. Agora for the diaspora: Exploring the use of Nairaland online forum for political deliberations among Nigerian emigrants. *The Nigerian Journal of Communication* 16, 1 (2019), 191–210.

[6] Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the first ACL workshop on ethics in natural language processing*. 94–102.

[7] 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.

[8] Junhan Chen, Yuan Wang, et al. 2021. Social media use for health purposes: systematic review. *Journal of medical Internet research* 23, 5 (2021), e17917.

[9] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[11] Florence Fenollar and Oleg Mediannikov. 2018. Emerging infectious diseases in Africa in the 21st century. *New Microbes and New Infections* 26 (2018), S10–S18.

[12] Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020. Do models of mental health based on social media data generalize?. In *Findings of the association for computational linguistics: EMNLP 2020*. 3774–3788.

[13] Muhammad Imran, Prasenjit Mitra, and Jaideep Srivastava. 2016. Cross-language domain adaptation for classifying crisis-related short messages. *arXiv preprint arXiv:1602.05388* (2016).

[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), 67–74.

[16] 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.

[17] Payam Karisani, Negin Karisani, and Li Xiong. 2021. Contextual Multi-View Query Learning for Short Text Classification in User-Generated Data. *arXiv preprint arXiv:2112.02611* (2021).

[18] 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).

[19] Sarthak Khanal and Doina Caragea. 2021. Multi-task Learning to Enable Location Mention Identification in the Early Hours of a Crisis Event. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 4051–4056.

[20] 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.

[21] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).

[22] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[23] Egoitz Laparra, Steven Bethard, and Timothy A Miller. 2020. Rethinking domain adaptation for machine learning over clinical language. *JAMIA open* 3, 2 (2020), 146–150.

[24] Kathy Lee, Ankit Agrawal, and Alok Choudhary. 2017. Forecasting influenza levels using real-time social media streams. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 409–414.

[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 preprint arXiv:1907.11692* (2019).

[26] Yang Liu, Christopher Whitfield, Tianyang Zhang, Amanda Hauser, Taeyonn Reynolds, and Mohd Anwar. 2021. Monitoring COVID-19 pandemic through the lens of social media using natural language processing and machine learning. *Health Information Science and Systems* 9, 1 (2021), 1–16.

[27] Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. 2021. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. *arXiv preprint arXiv:2106.04489* (2021).

[28] Rui Mao, Chenghua Lin, and Frank Guerin. 2019. End-to-end sequential metaphor identification inspired by linguistic theories. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 3888–3898.

[29] 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.

[30] 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. *arXiv preprint arXiv:2204.04521* (2022).

[31] Michelle Odlum and Sunmoo Yoon. 2015. What can we learn about the Ebola outbreak from tweets? *American journal of infection control* 43, 6 (2015), 563–571.

[32] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.

[33] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 2227–2237. https://doi.org/10.18653/v1/N18-1202

[34] Jacob Poushter, Caldwell Bishop, and Hanyu Chwe. 2018. Social media use continues to rise in developing countries but plateaus across developed ones. *Pew research center* 22 (2018), 2–19.

[35] Farhana Shahid, Shahinul Hoque Ony, Takrim Rahman Albi, Sriram Chellappan, Aditya Vashistha, and ABM Alim Al Islam. 2020. Learning from tweets: Opportunities and challenges to inform policy making during Dengue epidemic. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–27.

[36] Gillian E Smith, Alex J Elliot, Iain Lake, Obaghe Edeghere, Roger Morbey, Mike Catchpole, David L Heymann, Jeremy Hawker, Sue Ibbotson, Brian McCloskey, et al. 2019. Syndromic surveillance: two decades experience of sustainable systems–its people not just data! *Epidemiology & Infection* 147 (2019).

[37] Adarsh Subbaswamy and Suchi Saria. 2020. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics* 21, 2 (2020), 345–352.

[38] Abiodun Balogun Temitope. 2013. In defense of Nigerian pidgin. *Journal of Languages and Culture* 4, 5 (2013), 90–98.

[39] Temple Uwalaka. 2015. Nairaland and the Reconstruction of the Public Sphere in Nigeria. In *Refereed Proceedings of the Australian and New Zealand Communication Association Conference: Rethinking Communication, Space and Identity, Queenstown, NZ, http://www. anzca. net/conferences/past-conferences/, ANZCA*.

[40] Mina Valizadeh, Pardis Ranjbar-Noiey, Cornelia Caragea, and Natalie Parde. 2021. Identifying medical self-disclosure in online communities. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4398–4408.

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

[42] Theo Vos, Stephen S Lim, Cristiana Abbafati, Kaja M Abbas, Mohammad Abbasi, Mitra Abbasifard, Mohsen Abbasi-Kangevari, Hedayat Abbastabar, Foad Abd-Allah, Ahmed Abdelalim, et al. 2020. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet* 396, 10258 (2020), 1204–1222.

[43] Danqing Wang, Pengfei Liu, Ming Zhong, Jie Fu, Xipeng Qiu, and Xuanjing Huang. 2019. Exploring domain shift in extractive text summarization. *arXiv preprint arXiv:1908.11664* (2019).

[44] Jialin Yu, Alexandra I Cristea, Anoushka Harit, Zhongtian Sun, Olanrewaju Tahir Aduragba, Lei Shi, and Noura Al Moubayed. 2022. INTERACTION: A Generative XAI Framework for Natural Language Inference Explanations. In *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[45] Xinchen Yu, Zhuoli Xie, Afra Mashhadi, and Lingzi Hong. 2022. Multi-task Models for Multi-faceted Classification of Pandemic Information on Social Media. In *14th ACM Web Science Conference 2022*. 327–335.

[46] Yu Zhang and Qiang Yang. 2021. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering* (2021).

[47] Bin Zou, Vasileios Lampos, Russell Gorton, and Ingemar J Cox. 2016. On infectious intestinal disease surveillance using social media content. In *Proceedings of the 6th International Conference on Digital Health Conference*. 157–161.

[48] Ovidiu Şerban, Nicholas Thapen, Brendan Maginnis, Chris Hankin, and Virginia Foot. 2019. Real-time processing of social media with SENTINEL: A syndromic surveillance system incorporating deep learning for health classification. *Information Processing & Management* 56, 3 (2019), 1166–1184.