Crowdsourcing for machine learning in public health surveillance: lessons learned from Amazon Mechanical Turk

Zahra Shakeri Hossein Abad[1,2]; Gregory P. Butler[3]; Wendy Thompson[3]; Joon Lee[1,2,4]

[1]Data Intelligence for Health Lab, Cumming School of Medicine, University of Calgary, Canada
[2]Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Canada
[3]Centre for Surveillance and Applied Research, Public Health Agency of Canada, Canada
[4]Department of Cardiac Sciences, Cumming School of Medicine, University of Calgary, Canada

Abstract

Background: Crowdsourcing services such as Amazon Mechanical Turk (AMT) allow researchers to use the collective intelligence of a wide range of online users for labour-intensive tasks. Since the manual verification of the quality of the collected results is difficult due to the large volume of data and the quick turnaround time of the process, many questions remain to be explored regarding the reliability of these resources for developing digital public health systems.

Objective: The main objective of this study is to explore and evaluate the application of crowdsourcing, in general, and AMT, in specific, for developing digital public health surveillance systems.

Methods: We collected 296,166 crowd-generated labels for 98,722 tweets, labelled by 610 AMT workers, to develop machine learning (ML) models for detecting behaviours related to physical activity, sedentary behaviour, and sleep quality (PASS) among Twitter users. To infer the ground truth labels and explore the quality of these labels, we studied four statistical consensus methods that are agnostic of task features and only focus on worker labelling behaviour. Moreover, to model the meta-information associated with each labelling task and leverage the potentials of context-sensitive data in the truth inference process, we developed seven ML models, including traditional classifiers (offline and active), a deep-learning-based classification model, and a hybrid convolutional neural network (CNN) model.

Results: While most of the crowdsourcing-based studies in public health have often equated majority vote with quality, the results of our study using a truth set of 9,000 manually labelled tweets show that consensus-based inference models mask underlying uncertainty in the data and overlook the importance of task meta-information. Our evaluations across three PASS datasets show that truth inference is a context-sensitive process, and none of the studied methods in this paper was consistently superior to others in predicting the truth label. We also found that the performance of the ML models trained on crowd-labelled data is sensitive to the quality of these labels, and poor-quality labels lead to incorrect assessment of these models. Finally, we provide a set of practical recommendations to improve the quality and reliability of crowdsourced data.

Conclusion: Findings indicate the importance of the quality of crowd-generated labels in developing machine learning models designed for decision-making purposes, such as public health surveillance decisions. A combination of inference models outlined and analyzed in this work could be used to quantitatively measure and improve the quality of crowd-generated labels for training ML models.

Introduction

In the past years, social media data have been used extensively in different areas of public health [1–3], such as detecting outbreaks and emerging diseases [4, 5], monitoring adverse drug reaction [6], and predicting or modelling health-related behaviours and outcomes [7–9]. Since 2011, Twitter has been the most popular form of social media used for public health communication [10, 11]. In 2020, Twitter alone reported 500 million tweets generated per day from 145 million daily active users. A recent scoping review of 755 articles on digital public health surveillance shows that Twitter is the most studied of all platforms and most utilized platform to study communicable diseases, behavioural risk factors, mental health, drug utilization, and vaccine [11]. Apart from the inherent limitations of social media data such as lack of demographics data and biased populations, when integrated with complex data-

driven models such as artificial neural networks (ANNs), these publicly accessible resources can be utilized for population-level surveillance to complement traditional public health surveillance (e.g., surveys) with faster and less costly longitudinal information.

While linguistic annotation is crucial for developing machine learning (ML) and natural language processing (NLP) models, manual labelling of a large volume of data is a notorious problem due to its high cost and labour-intensive nature. In recent years, this problem has been tackled using crowdsourcing technologies such as Amazon Mechanical Turk (AMT) [12], Crowdflower (CF) [13], and Prolific Academic (ProA) [13] to obtain relatively low-cost labelled data more quickly and easily. AMT is a software service operated by Amazon that allows users to crowdsource work–broken into micro-tasks called human intelligence tasks (HITs), to a large number of workers who are compensated for each HIT completed. With the wide potential applications of crowdsourcing in public health [14–16], the research community has seen steady growth in the use of AMT in the past ten years. The number of studies indexed in PubMed using the search term `'Amazon Mechanical Turk' AND 'Public Health'` has increased sharply from 42 studies in 2015 to 118 studies in 2019.

However, due to the uncertain quality of AMT workers with unknown expertise, their labels are sometimes unreliable, forcing researchers and practitioners to collect information redundantly, which poses new challenges in the field. Given that in large-scale crowdsourcing tasks the same workers cannot label all the examples, measuring inner-annotators agreement and managing the quality of workers differ from those of a team of in-house expert workers. Despite the growing popularity of AMT for developing ML models in public health research, the reliability and validity of this service have not yet been investigated. At least several public health studies have utilized AMT for training data-driven ML models without external gold standard comparisons [17–21]. Ayers et al. used AMT to create a gold standard dataset to develop predictive models to detect electronic nicotine delivery systems on social media [17]. Yin et al. developed a scalable classifier to detect personal health mentions on Twitter based on a gold standard dataset generated by AMT workers [18]. The reliability of the crowd-labelled dataset in this study was measured based on the agreement between workers.

Similarly, to characterize sleep quality using Twitter, McIver et al. utilized AMT for sentiment annotation of text data and used inter-annotator agreement to assess the reliability of workers [19]. Reece et al. used AMT to build a dataset and develop a prediction model to detect depression emergence and post-traumatic stress disorder in Twitter users [20]. To control the collected data quality, they required the workers to have completed at least 100 tasks, with a minimum 95% approval rating. While research has supported the efficacy of using reputation to evaluate the quality of crowdsourced data [22], the reliability of using this metric in developing ML-based digital public health systems has not yet been investigated. Thus, in this study, in addition to defining qualification requirements for AMT workers, we studied the reliability of crowd-generated training data for developing ML models in the context of public health surveillance. We used AMT to collect 296,166 labels for 98,722 unique tweets, labelled by 610 AMT workers to develop ML models that can detect physical activity, sedentary behaviour, and sleep quality (PASS) of Twitter users.

Our primary objective in this work was to evaluate the application of AMT for training data-driven ML models by analyzing the quality of crowd-generated labels. Our approach involved evaluating the performance of four consensus methods–which do not involve task features in their truth inference and to explore their feasibility in improving the quality of crowd-labelled data. As these methods are modelled purely as a function of worker behaviours in relation to labelling tasks, they cannot leverage the value of context-sensitive information (i.e., the task's meta-information) in their inference decisions. Thus, we collected additional features for our labelling dataset and developed seven ML models, including a deep learning (DL) model and a hybrid convolutional neural network (CNN) architecture to couple worker behaviours with the task's meta-information when inferring the truth label. To detect and correct noisy labels, we also developed five pool-based active learners to iteratively detect the most informative samples (i.e., samples with more uncertainty) and remove them from the validation set. Finally, we used Shapley additive explanations (SHAP) [23] to explore the contribution of different features, including worker behaviours and context-sensitive features, to the results of our supervised inference models.

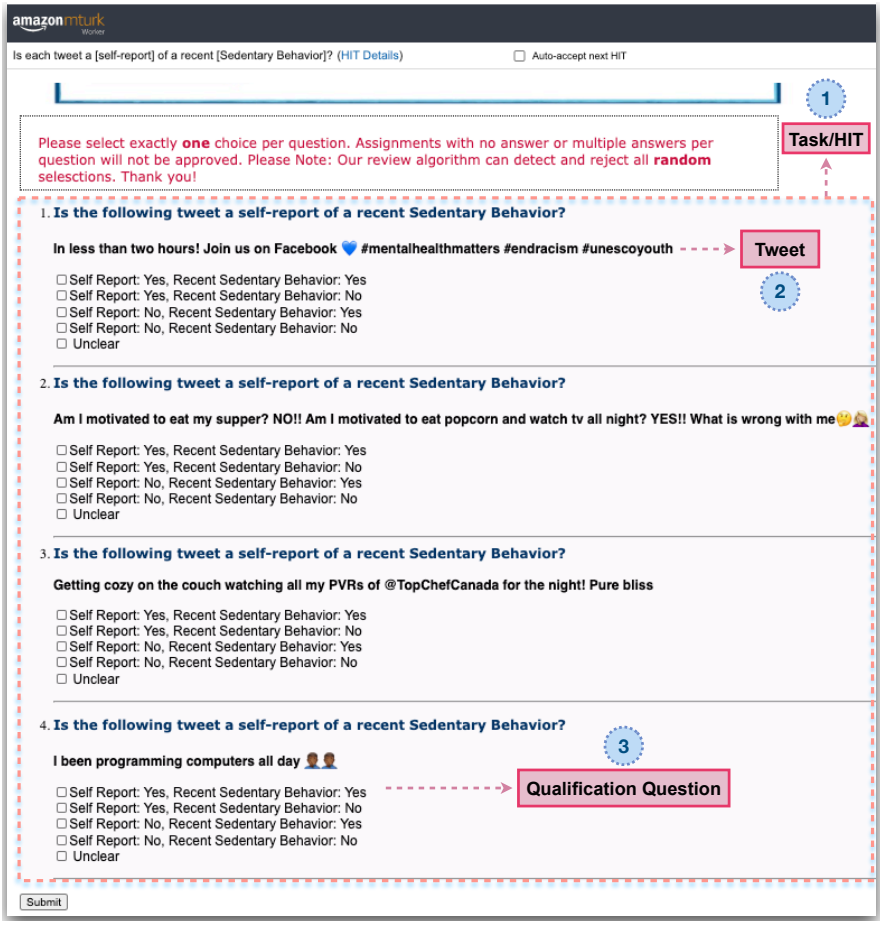

Figure 1. A sample labelling task (i.e., HIT) for the sedentary behaviour category, designed for this study. Each HIT contains four questions (section ①), and each asks if the presented tweet is a self-reported PASS-related behaviour (section ②). The fourth question is a pre-defined qualification question that was designed in addition to the qualification requirements defined by AMT (section ③). The answer to this question was always choice#1, and it was easy enough to detect spammers or irresponsible workers. Also, each HIT contains an illustrative example that explains each choice of the questions (see figure 4). Workers were asked to select exactly one choice, and HITs with zero or more than one label were rejected during the approval process.

Methods

The crowdsourcing tasks, referred to as HITs by AMT, were designed to collect five labels, based on two conditions ⟨`self-reported, recent PASS experience`⟩, to develop binary and multi-class classification models that can detect PASS-related behaviour in Twitter users. The labels of the multi-class prediction models were defined as: `11`, `10`, `01`, `00`, based on the value of each condition, respectively (supplementary figure 1). We also let workers choose a fifth option, called `Unclear`, to ensure they do not give random labels to tasks that they are not confident of performing successfully (figure 1). We excluded this label for both inference and classification tasks. We define the binary labels as `1` if both conditions are met; `0`, otherwise. The binary labels did not directly come from the AMT workers and were generated by dichotomizing the collected labels.

Crowdsourcing Workflow

We implemented a pipeline to create the HITs, post them on AMT, collect the labels through a quality check process, approve/reject the HITs, and store the results. To minimize noisy and low-quality data, we added a qualification requirement to our tasks and granted the labelling access to workers who had demonstrated a high degree of success

in performing a wide range of HITs across AMT (i.e., master qualification). In addition to this, we added a simple qualification question to each HIT to detect spammers or irresponsible workers. Each HIT contained four questions, including the qualification question and was assigned to three workers (figure 1 and supplementary figures 2 and 3). Through different iterations of data labelling, workers were paid from $0.03 USD to $0.05 USD after completing each HIT. We collected the labels for the 98,722 tweets used in this study through different iterations, from April 2019 to June 2020. We regularly checked the quality of submitted tasks to detect low-quality workers during each iteration and revoke their access to our tasks. Before the formal initiation of the process, we pilot tested the design, response time, and complexity of the HITs through two different iterations and revised the workflow accordingly. We did not collect any personally identifiable information from the workers (participants) during the data labelling task. The experiments were carried out in accordance with relevant guidelines and the University of Calgary Conjoint Faculties Research Ethics Board's (CFREB) regulations. We implemented the entire workflow in `python` and used `Boto3 python SDK` to connect to and work with AMT.

Table 1. The characteristics of the dataset used to develop and evaluate the un(supervised) inference models. [PA]: Physical Activity, [SQ]: Sleep Quality, [SB]: Sedentary Behaviour.

Variable	PA	SB	SQ
N	4,000	3,000	2,000
Labels			
Binary (%)			
Yes	1,629 (41%)	726 (36%)	1,063 (35%)
No	2,371 (59%)	1,274 (64%)	1,937 (65%)
Multi-class			
YY	1,629 (41%)	726 (36%)	1,063 (35%)
YN	550 (14%)	395 (20%)	862 (29%)
NY	179 (4%)	19 (1%)	52 (2%)
NN	1,642 (41%)	860 (43%)	1,023 (34%)
Gender			
Female (%)	1,240 (36%)	777 (39%)	1,097 (50%)
Male (%)	2,760 (64%)	1,223 (61%)	1,109 (50%)
Age range (%)			
$\leq$ 18	262 (7%)	229 (11%)	446 (15%)
(19-29)	955 (24%)	641 (32%)	1085 (36%)
(30-39)	1,153 (29%)	493 (25%)	717 (24%)
($\geq$ 40)	1,629 (41%)	637 (32%)	752 (25%)
Weekday (n)			
Sunday	683	344	426
Monday	626	312	456
Tuesday	496	241	459
Wednesday	489	264	403
Thursday	499	275	426
Friday	509	268	415
Saturday	683	297	416
Time, (24) hour [Q1,Q3]	[10,19]	[10-19]	[5-18]
Month [min,max]	[Jul, Feb]	[Sep, Apr]	[Aug, Jan]
Source			
Organization	563 (14%)	179 (9%)	97 (3%)
Users	3,437 (86%)	1,821 (91%)	2,903 (97%)

Data Collection

We collected the data of this study from Twitter using `Twitter livestream application programming interface (API)` for the period between 28th November 2018 to 30th June 2020. The dataset was filtered to include only Canadian tweets relevant to PASS: physical activity, sedentary behaviour, or sleep quality. 103,911 tweets were selected from 22,729,110 collected Canadian tweets using keywords and regular expressions related to

PASS categories. Each of these 103,911 tweets was labelled by three AMT workers, from which 98,722 tweets received three valid labels, with almost half of them related to physical activity.

The demographics variable of age and gender and the information about the source of each tweet (e.g., organization vs. real users) were not available within the dataset collected from Twitter. We estimated these variables for each tweet using the `M3inference` package in `python` [24], which uses a multimodal deep neural architecture for joint classification of age, gender, and information-source of social media data. The text (tweet) field and each of the daytime, weekday, and month variables were extracted from the metadata provided by the Twitter API.

Data Processing

Tweets have a bounding box of coordinates, which enabled spatial mapping to their respective city locations. As Twitter API returns `datetieme` values in UTC time, we used `timezonefinder` in `python` and adjusted the time of each tweet based on its spatial data. Given that daytime, month, and weekday can be influential factors in twitting about each of the PASS categories, and to better utilize the datetime data (`%a %b %d %H:%M:%S %Y`), we extracted `a:` weekday, `b:` month, and `H:` hour fields and stored them as separate features.

We cleaned the `text` column by eliminating all special characters (e.g., #, &, @), punctuations, weblinks, and numbers. We also replaced common contractions with their un-contracted form. For example, `I'll` is resolved to `I will`. While developing and evaluating our NLP models, we noticed that the impact of removing stop words, stemming, and converting the text to lower case on the performance of our predictive models is not noticeable. This could relate to the ability of transfer-learning techniques (i.e., GloVe embeddings) to generalize on unseen data. Thus, we applied neither stop-word removing nor lexical cleaning on the textual features of our dataset. Moreover, given that hashtags and emojis can be used as independent words and facilitate emotional expressions, we did not remove them during the cleaning process.

To develop the ML models, all categorical data were encoded into dummy variables using one-hot encoding, and as we only approved HITs with complete answers, this dataset did not contain any missing data.

Label Consistency

To measure the consistency of answers given by workers, we calculated label consistency ($\mathcal{LC}$) as the average entropy of the collected labels for each PASS category [25]. For each tweet $t_i \in \mathcal{T}_s$, where $\mathcal{T}_s$ denotes the set of all tweets related to surveillance category $s \in \{$`physical activity, sleep quality, sedentary behaviour`$\}$, n_{ij} defines the number of answers given to the j^{th} choice ($j \in \{1,2,3,4,5\}$, as we have five choices for each tweet). We calculate $\mathcal{LC}_s$ as:

$$\mathcal{LC}_s = 1 - \left(-\frac{1}{|s|} \times \sum_{i=1}^{i=|s|} \sum_{j=1}^{j=5} \frac{n_{ij}}{3} \times \log_3 \frac{n_{ij}}{3} \right) \qquad (1)$$

$|s|$ denotes the size of the surveillance category s and as we collect three labels for each tweet, the denominators in the entropy formula receive the constant value of three. $\mathcal{LC}$ ranges from 0 to 1, and the values close to one show more consistency between the workers' input.

Inference Models

The majority voting (MV) approach estimates the actual ground truth based on the majority of labels submitted by different workers. For example, defining the estimated label as $\hat{l}_i$, and the submitted label by worker w as l_w, the MV approach, for a binary labeling task assigns 1 to $\hat{l}_i$ if $\frac{1}{3} \times \sum_{i=1}^{3} l_w^i > 0.5$, and assigns 0, otherwise. While individual workers' reliability coming from different backgrounds with different quality levels varies, the MV approach

assumes equal expertise among the workers, and it does not model worker behaviours [26]. Given that this approach is completely task-independent, it does not involve task properties in the inference process, and thus, it is fast.

David and Skene (DS) [27] approach uses expectation-maximization (EM) to simultaneously estimate the error rate of annotators (workers) and latent label classes, when, similar to MV, the ground truth is unknown, and workers are assumed to operate independently. Unlike MV that is agnostic of worker behaviour, DS models worker *k*'s behaviour as a function of each task's true label by creating a confusion matrix π^k with size $L \times L$, where L is a fixed number and represents the number of possible labels for a single-labelled classification task. DS defines worker k's error-rate $\pi_{jl}^{(k)}$ $(j = 1, \ldots, L;\ l = 1, \ldots, L)$ as:

$$\pi_{jl}^{(k)} = \frac{\#\text{ of times worker } k \text{ records } l \text{ when } j \text{ is correct}}{\#\text{ of tweets seen by worker } k \text{ where } j \text{ is correct}} \qquad (2)$$

Given that not all workers need to label all the tasks, and a worker may label the same task more than once, sparsity can be a problem in large-scale labelling tasks when using the DS approach [26]. DS iteratively estimates the true label of each task based on worker's quality and estimates worker's error rate (quality) based on the inferred labels until it converges. While the worker-specific confusion matrix generates the quality score of each worker, it may not be sufficient to measure the actual contribution of each worker [28]. The inherent complexity of a task, especially in NLP, or worker's bias may result in wrong labels while the worker is quantitatively accurate.

Generative model of Labels, Abilities, and Difficulties (GLAD) [29] models the quality of workers as a function of the input task using parameter α. The quality parameter ranges from $-\infty$ to $+\infty$ implying that the worker always labels the tasks incorrectly or correctly, respectively. When $\alpha = 0$, the worker cannot distinguish between the labels and their input does not contribute to the task's correct label. To estimate the ground truth, in addition to the workers' quality, GLAD models the difficulty of task t_i as $d_i = 1/\beta_i$, where $\beta_i > 0$. The difficulty index ranges from 0 to ∞, where $d_i = \infty$ classifies t_i as the most difficult task, and $d_i = 0$ means the task always receives a correct label, even from the workers with $\alpha \leq 0$. GLAD uses the EM approach to obtain the maximum likelihood estimation of α and β, and models the probability that worker *k* correctly labels t_i using $p\left(\hat{l}_i^k = l_i | \beta_i, \alpha_k\right) = 1/(1 + e^{-\alpha_k \beta_i})$.

Like DS, Raykar's algorithm (RY) [30] forms a confusion matrix to model a worker's quality. In addition, in the case of binary classification, it models worker's bias toward the positive class (i.e., sensitivity) and towards the negative class (i.e., specificity) using beta prior [26]. Worker bias in this context usually occurs when a worker underestimates or overestimates the truth of a task [25]. As with DS and GLAD, RY uses an unsupervised EM approach to estimate each of the model parameters and truth labels. Depending on the availability of task-specific features, RY can either use automatic supervised classifiers or fall back to unsupervised EM models to estimate the truth label.

Predictive Models

As the meta-information associated with each task may reveal its underlying complexity and, thus, help to model worker behaviours, we developed a set of machine learning models to involve this metadata in the inference process. Models are trained based on quintuple $\mathcal{F}$: $(\mathcal{W}, \mathcal{I}, \mathcal{M}, t, l)$, where $\mathcal{W} = \{w_1, \ldots, w_k\}$ represents labels collected from AMT workers, $\mathcal{I} = \{\text{MV}, \text{DS}, \text{RY}, \text{GLAD}\}$ denotes the results of inference models, and $\mathcal{M}$ denotes metadata associated with each tweet, including time (i.e., weekday, month, daytime), gender, age group, and the source of the tweet (i.e., organization vs. real people). The text of each tweet is presented by *t* and *l* denoted the truth label (table 1).

To mitigate the risk of biased results caused by a specific learning algorithm and also to overcome the over-fitting problem, we developed and evaluated five standard machine learning classifiers with different architectures, including generalized linear (Logistic Regression (LR)), kernel-based (Support Vector Machines (SVM)), decision-

tree based (Random Forest (RF), and XGBoost), and sample-based (K-Nearest Neighbours (KNN)) classifiers. Moreover, to incorporate textual features into our analysis, we developed a hybrid deep learning architecture in which a CNN based on long short-term memory (LSTM) learns textual data t and a multilayer perceptron deep neural network (DNN) learns metadata $\langle \mathcal{W}, \mathcal{I}, \mathcal{M} \rangle$. The cleaned text, represented as an integer encoded vector, is converted into pre-trained tweet word embeddings using GloVe [31] (containing two billion tweets, 27 billion tokens, and 1.2 million vocabularies) in the embedding layer. The output of this layer is passed through an LSTM layer for sequence modelling, followed by one dropout layer to avoid over-fitting and two dense ReLU layers. At the same time, the metadata of each tweet is passed through three fully connected layers with ReLU activation. The outputs of these networks are concatenated into a dense layer, followed by two fully connected dense layers, terminating at an output layer with `softmax` activation, and cross-entropy loss, and `adam` optimizer. The high-level presentation of this architecture is shown in Figure 2.

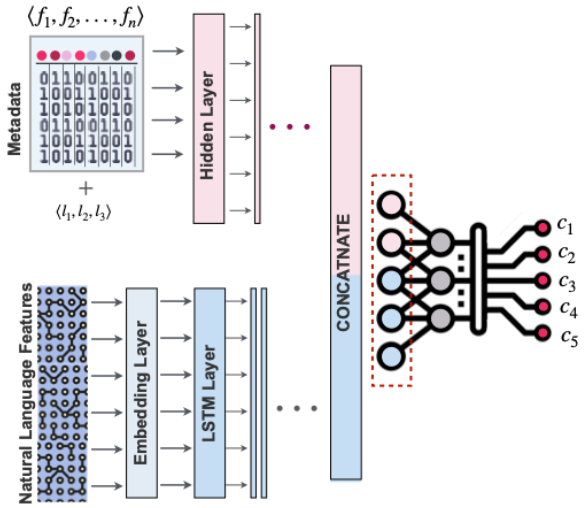

Figure 2. The pipeline of the deep learning model used to predict labels using both textual and meta-information

To counter the bias caused by class imbalance, for both multi-class and binary classification tasks, we used the class-weight approach to incorporate the weight of each class into the cost function by assigning higher weights to minority classes and lower weights to the majority classes. We also used the Synthetic Minority Over-sampling TEchnique-Nominal Continuous (SMOTE-NC) [32] approach to oversample the minority classes by creating synthetic samples based on their feature-space. However, we did not notice much difference between using SMOTE or not. Thus, the model presented in Table 3 are trained using the `class_weight` approach. Hyperparameters for each method were determined using a nested 10-fold cross-validation Bayesian Optimization [33].

As the main goal of both supervised and unsupervised label inference models was minimizing the number of false negative (FN) and False positive (FP) inferences, to evaluate the models developed in this study, we used precision, recall, F1, and precision-recall area under the curve (AUC_{PR}) metrics.

All the computations and predictive models were implemented using Python 3.7 with TensorFlow 2.0 [34], Keras [35], and Scikit-learn [36] libraries. To facilitate the replication of our study, the code repository of this study is publicly available on GitHub: https://github.com/data-intelligence-for-health-lab/AMT-for-Intelligent-Public-Health.

Results

In total, 610 unique workers participated in our data labelling tasks and completed 103,911 HITs, from which 5,189 HITs were removed as they did not receive three valid answers. We approved 98,722 tasks for further analysis. The majority of workers (87%, 530) completed less than 100 HITs, among which 164 completed only one HIT. Among

the workers who completed more than 5,000 HITs, one worker completed 21,801 HITs, and three workers completed between 5,000 and 10,000 HITs (figure 3). The calculated label consistency ($\mathcal{LC}$) for each PASS category, for the multi-class labelling, was 0.54, 0.58, and 0.55, respectively, and was 0.75, 0.77, and 0.74 for the binary labelling, respectively. This implies a high level of label inconsistency, prompting a need for further label quality analysis for the development of ML models (table 2).

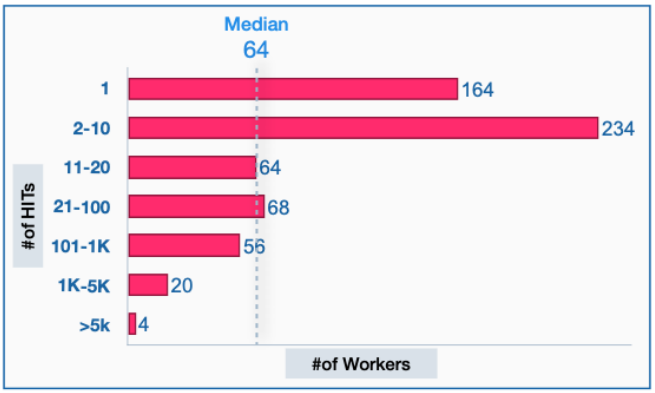

Figure 3. The number of workers completed different ranges of HIT numbers. This data follows the long-tail distribution: the majority of workers labeled a few tasks and only a few workers labeled the majority of tasks.

Table 2. Details of the collected labels and the label consistency score for each of the physical activity, sleep quality, and sedentary behaviour categories. $\mathcal{LC}$ ranges from 0 to 1, and the values close to one show more consistency between the workers' input.

Type	#Tweets	$\mathcal{LC}_{\text{multi}}$	$\mathcal{LC}_{\text{binary}}$	#Workers
Physical Activity	48,576	0.54	0.75	232
Sedentary Behaviour	17,367	0.55	0.74	157
Sleep Quality	32,779	0.58	0.77	221
Total	**98,722**	**0.56**	**0.75**	**610**

Truth Inference

To investigate the viability of unsupervised inference models in predicting truth labels from crowd-labelled data and to compare it with that of supervised predictive models, we used a random sample of our dataset (i.e., 9000 tweets: 4000 physical activity, 3000 sleep quality, and 2000 sedentary behaviour). Six data scientists manually labelled this sample, and the entire labelled dataset was manually reviewed and re-labelled by an experienced in-house domain expert in both ML and public health surveillance. The disagreements between this dataset and the crowd-labelled dataset were manually checked to exclude any labelling bias that could impact the results of this study. Table 3 lists the inference results obtained from four unsupervised models and seven supervised predictive models, including two deep learning models. Each model was evaluated over both binary and multi-class versions of the dataset for each PASS category. Among the unsupervised models for physical activity and sleep quality, DS and RY performed better than MV and GLAD for all performance metrics, while MV outperforms the other models on the sleep quality dataset. Interestingly, for binary inference across all PASS categories, MV outperformed or performed just as well as the other methods, indicating the impact of task complexity on the performance of inference methods.

Table 3 presents the results of supervised models across all datasets. DL_{meta} is seen to outperform other methods with the minimum number of false positives (precision: 78%) for the multi-class classification task, but other methods performed better with respect to recall, F1, and AUC_{PR} metrics. Performance on each PASS dataset, for binary classification, did not highlight any individual method constantly performing best. For example, while SVM showed the best performance for physical activity, KNN and LR outperformed other models for sleep quality and sedentary behaviour, respectively. LR achieves superior performance across all datasets for the multi-class inference task. To analyze this further, we modified the hyper-parameters of the LR algorithm presented in Table 3 to stochastic average gradient (SAG) solver and *l2* regularization, and the optimizer of the hybrid neural network to stochastic gradient

descent (SGD) and repeated the comparisons. LR still outperforms the neural network model by over 2% in all metrics. The poor performance of the neural networks in this study could be attributed to the imbalanced ratio of data (per class) to the model parameters (i.e., high variance).

Across all datasets, supervised models consistently performed better than unsupervised methods. This highlights the value of context-sensitive information that was used as meta-information when training supervised models. However, on sleep quality, a dataset with the same features and same level of complexity as physical activity and sedentary behaviour datasets, MV appears sufficient for the binary inference task, with supervised models providing little or no improvement.

The hybrid CNN architecture did not provide any gain over either the unsupervised inference models or the supervised predictive models (i.e., LR, KNN, SVM, RF, XGBoost, and DL_{meta}), and in some ways, underperformed them. It is possible that the LSTM stream could not capture the underlying dynamics of the features due to the inconsistencies between the poorly labelled tasks and the textual features.

Table 3. Performance of the truth interference methods using a ground truth dataset of 9,000 labeled tweets: 4,000 physical activity, 2,000 sedentary behaviour, and 3,000 sleep quality tweets. [M]: multi-class, [B]: binary. The top four rows of each PASS category represent the results of the applied unsupervised truth inference models.

PHYSICAL ACTIVITY								
Method	**Precision**		**Recall**		**F1**		**AUC $_{PR}$**	
	M	**B**	**M**	**B**	**M**	**B**	**M**	**B**
MV	72%	85%	70%	85%	71%	84%	56%	85%
DS	74%	85%	68%	85%	70%	84%	54%	85%
GLAD	73%	84%	70%	84%	71%	83%	57%	84%
RY	74%	85%	68%	85%	70%	84%	54%	84%
LR	74%	85%	**75%**	**85%**	**74%**	**85%**	**61%**	87%
KNN	74%	85%	74%	85%	73%	84%	60%	**88%**
SVM	72%	**86%**	73%	**85%**	73%	**85%**	61%	**88%**
RF	73%	85%	74%	84%	73%	85%	60%	87%
XGBoost	72%	81%	72%	81%	71%	81%	58%	83%
DL_{meta}	**79%**	84%	68%	84%	73%	84%	60%	78%
$DL_{text/meta}$	78%	84%	70%	84%	73%	84%	60%	78%
SEDENTARY BEHAVIOUR								
Method	**Precision**		**Recall**		**F1**		**AUC $_{PR}$**	
	M	**B**	**M**	**B**	**M**	**B**	**M**	**B**
MV	71%	82%	68%	82%	68%	82%	54%	80%
DS	70%	81%	62%	81%	65%	81%	48%	79%
GLAD	71%	79%	68%	79%	68%	79%	54%	77%
RY	70%	81%	62%	81%	65%	81%	48%	79%
LR	72%	**83%**	**72%**	**83%**	70%	**83%**	**58%**	**81%**
KNN	71%	82%	71%	82%	67%	82%	56%	80%
SVM	73%	83%	72%	83%	70%	83%	58%	81%
RF	72%	83%	72%	82%	69%	83%	57%	81%
XGBoost	68%	82%	69%	82%	67%	82%	54%	80%
DL_{meta}	**78%**	80%	65%	80%	**71%**	80%	56%	73%
$DL_{text/meta}$	78%	80%	65%	80%	71%	80%	56%	75%
SLEEP QUALIT75Y								
Method	**Precision**		**Recall**		**F1**		**AUC $_{PR}$**	
	M	**B**	**M**	**B**	**M**	**B**	**M**	**B**
MV	78%	89%	74%	89%	75%	89%	61%	87%
DS	80%	89%	74%	89%	77%	89%	62%	87%
GLAD	79%	85%	75%	85%	76%	85%	62%	82%
RY	80%	89%	74%	89%	76%	89%	62%	87%
LR	76%	88%	**77%**	87%	**77%**	88%	**64%**	88%
KNN	76%	**89%**	77%	**89%**	77%	**89%**	63%	**89%**
SVM	76%	88%	**77%**	88%	**77%**	88%	**64%**	88%
RF	75%	89%	76%	89%	76%	89%	63%	89%
XGBoost	72%	87%	72%	89%	72%	87%	58%	87%
DL_{meta}	**82%**	86%	72%	86%	76%	86%	63%	81%
$DL_{text/meta}$	80%	87%	72%	87%	76%	87%	65%	82%

Active Learning

To further explore the feasibility of correcting mislabelled samples, we used pool-based active learning (AL) [37] with uncertainty sampling. Pool-based AL assumes that only a small set of data is labelled, and a large pool of data still needs to be labelled through an iterative learning process. All samples in the pool are queried based on an informativeness measure, which improves the learner's discrimination ability [38]. In this study, our learners were modelled to query the most ambivalent and uncertain samples. For example, for the binary label inference task, samples for which $p(\hat{l} = l \mid f) \approx 0.5$ are the most informative samples that may help detect mislabelled samples of the dataset through different iterations. We used five different base learners with different architectures (i.e., RF, LR, KNN, SVM, and XGBoost) with a batch size of five and queried the unlabelled pool through 100 iterations.

Our results show that, during the learning process, the accuracy of the classifiers generally increased, slightly degraded at some iterations, and stabilized around iteration 60 for KNN and iteration 20 for other classifiers (figure 4). While the active learners in this study could improve their predictive ability through a self-learning process, they failed to correct mislabelled samples and stabilized at performance scores lower than those of the offline learners discussed earlier (table 3).

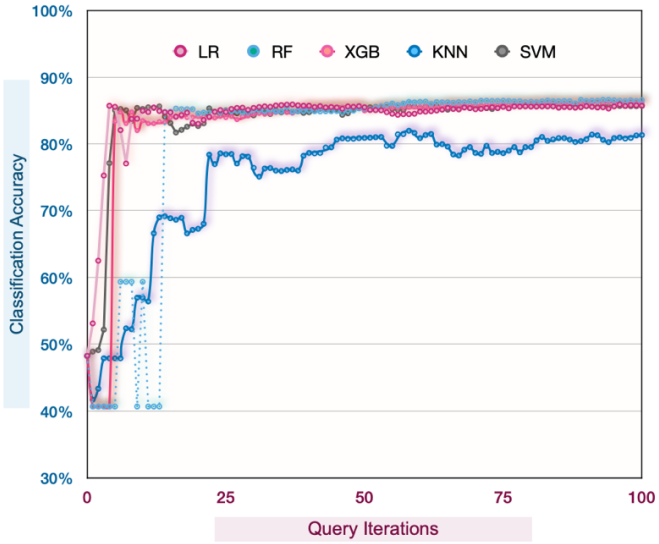

Figure 4. Incremental classification accuracy using pool-based active learning.

Discussion

Practical Recommendations

We start this section with some practical recommendations and guidelines on using AMT in specific and crowdsourcing in general for developing ML-based public health surveillance systems. These guidelines are supported by the results described earlier and the findings and further analysis discussed in the rest of this section.

First, although the demographics of AMT workers are not available, we can still implement the crowdsourcing process in a way that accommodates a greater diversity of workers. A longitudinal labelling process, rather than one-time labelling, allows researchers to monitor the quality of the collected data over time and mitigate the impact of spammers, irresponsible workers and workers who are biased or mistake prone. Second, the overall quality of AMT workers can be context-sensitive and vary based on the type of labelling tasks. For example, the familiarity of the workers with the context of the tasks in the sleep quality dataset, contrasting the broad context of physical activity and sedentary behaviour concepts, resulted in higher data quality. Researchers should also be aware of the exclusion rate (e.g., 5,189, 5% in our study) and need to consider this when planning for their study's budget and design. Third,

our study results show that consensus-based inference models that do not take the task's features into account may not always be efficient for integrating crowdsourced labels and thus, negatively impact ML models' performance. Fourth, in addition to qualification requirements to filter crowdsourcing participants, sound and illustrative instruction is a less direct way to increase data quality. During the course of this project, we received nearly 70 emails from AMT workers, with the majority of them asked about scenarios that were mentioned in the instruction. This implies that the instruction has changed their default understanding of the tasks, improving the quality of the labels. Fifth, when controlling the quality of workers using a qualification question, we recommend not informing the worker that this technique is being used, as they might guess the questions based on their simplicity.

Key Findings

Information loss about label uncertainty– Despite all the alternative models developed in this study to improve the inference accuracy, there were still considerable discrepancies between workers and the truth labels. These disagreements may be attributable to the underlying uncertainty in the data. While reducing uncertainty through collecting more labels from more workers might simplify the process of label inference, it limits the learning ability of ML models in modelling the inherent uncertainty of data and prevents them from any possibility of recovering later from mistakes made early during the inference process [39].

Robustness of inference models– We observed from our inference results that, regardless of the type of the classification task, none of the eleven methods outperformed other methods across all datasets (table 3). This indicates that inference methods are sensitive to dataset characteristics. For example, the performance of all of the methods on the sleep quality dataset is better than that of physical activity and sedentary behaviour datasets, indicating the low robustness of these models against the task context.

The importance of task features– Compared to the supervised models that need a large volume of labelled data to integrate crowd-generated labels, using unsupervised inference models is simple and straightforward. However, this simplicity is gained through the cost of throwing away the contextual characteristics of tasks, which may sacrifice quality in context-sensitive scenarios. For example, the time that a tweet is posted during a day can contribute to the decision about its relevance to physical activity or sleep quality contexts. The importance of these characteristics was far more pronounced in the multi-class inference tasks than binary tasks (table 3), suggesting the need for more complicated models when inferring the truth label of tasks with a high level of uncertainty.

The effectiveness of qualification requirements– In this study, we used two levels of quality control: (1) through the task assignment process by accepting only workers with a master qualification, and (2) through the design and implementation of the tasks by adding a qualification question to our HITs and iteratively observing workers' performance based on their answer to this question. Our results show that even though defining these requirements improved the quality of crowd-generated labels to a great extent, 12% (498) of physical activity tweets were still mislabelled by all of the workers, regardless of their context or complexity level, indicating the need for further quality assessment of crowdsourced data. This number for sleep quality and sedentary behaviour tasks was 8% (231) and 13% (266), respectively. These mislabelled samples were not misclassified due to sample uncertainty or difficulty, and our further analysis shows that they were not informative enough (i.e., prediction scores) to improve the performance of predictive models through the iterative process of active learning (supplementary figure 5). Considering the sparsity of the `(workers, tasks)` matrix in large-scale crowdsourcing tasks, distinguishing irresponsible workers and removing their impact is a challenging task that should be carefully considered when training machine learning models based on crowd labelled data. A sample list of low-quality labels for all the PASS categories is provided in supplementary figure 6.

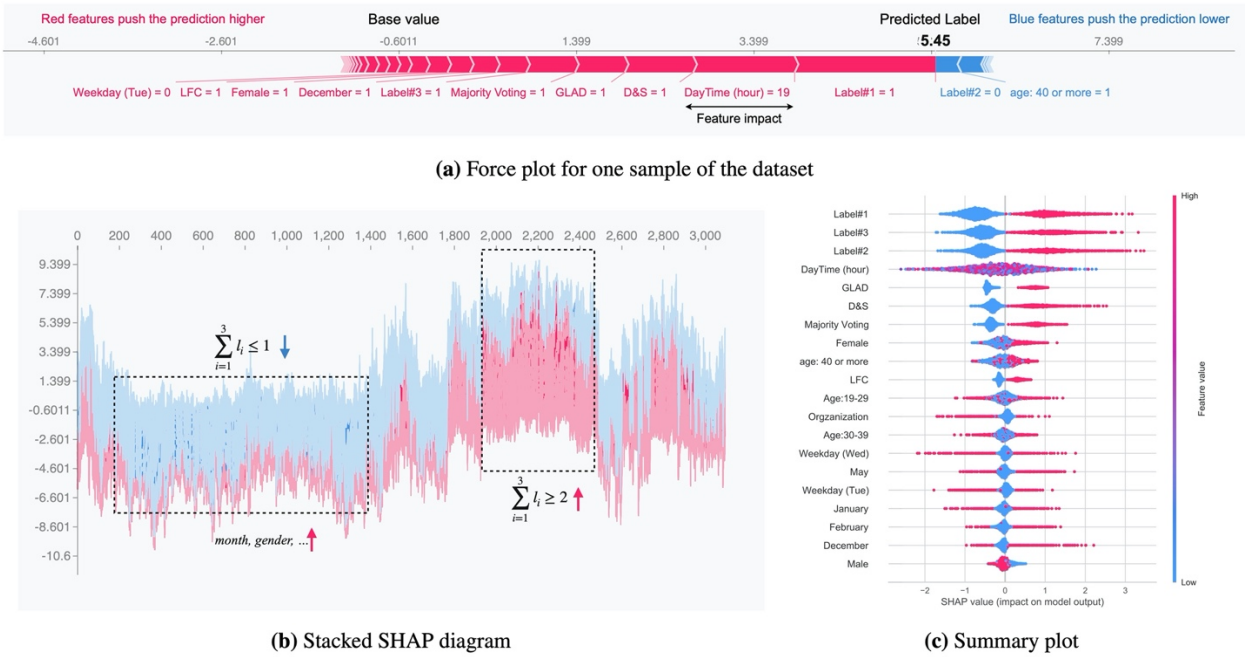

Figure 5. The estimated impact of each meta-information on XGBoost when predicting the truth label. The sample in **(a)** has a predicted value of 5.45, roughly ten times the base value. The base value in this plot shows the average model output over the physical activity training dataset. The major contributors in this sample are the label from worker#1, daytime=7pm, and the results of statistical inference models, where all of them inferred the truth label as 1. The predicted values in this force diagram are not in the range of [0,1], while they show the results of a binary XGBoost classifier. This is because the SHAP values in the XGBoost SHAP algorithm are log-odds values and are computed based on the margin, not the transformed probability. The stacked plot in **(b)** contains the sample-specific forced plots that are rotated vertically and integrated based on their similarity. This plot shows the similar behaviour of the features that make the most impact on the predicted value of the samples in our dataset. Samples with $\sum_{i=1}^{3} l \geq 2$ are receiving lower predictions and samples with $\sum_{i=1}^{3} l \geq 2$ (i.e., samples that received at least two 1's from the workers) are receiving higher predicted values. **(c)** presents the global distribution of each feature's impacts on the model output. The thickness of the line presents the frequency of samples mapped to each SHAP value. This diagram highlights the significant impact of crowd-generated labels on the results of the XGBoost classifier using the physical activity dataset.

The impact of crowd-generated labels on the performance of predictive models– To further investigate the reliability of using crowdsourcing for developing machine learning models, we used bidirectional encoder representations from transformers (BERT) [40] (i.e., `bert-base-uncased`), a transformer-based model with 12-layer, 768 hidden units, 12 heads, and 110M parameters as a contextual input to our deep learning model, to classify 4000 physical activity tweets, using our binary truth labels and crowd-generated labels. We used labels inferred by SVM for the crowd-generated labels, as it outperformed other models on the physical activity dataset (table 3). Interestingly, the model that was trained on our truth dataset outperformed the crowd-labelled dataset on all performance metrics by at least 8% (e.g., crowd-labelled: AUC_{PR} of 72%, expert-labelled: AUC_{PR} of 82%). This indicates the importance of the quality of crowd-generated labels in developing machine learning models designed for decision-making purposes, such as public health surveillance decisions.

Label Prediction Explanation

To interpret the results of our predictive models in terms of the individual contribution of each feature to the prediction results, we used SHAP [23, 41]. SHAP calculates the local, instead of global, feature importance for each sample of the dataset, which mitigates the risks associated with inconsistency problems in other feature importance techniques. Figure 5a illustrates the interpretation of the prediction using XGBoost on one randomly selected sample of the physical activity dataset using SHAP. The red arrows show the features that contribute to the increase, and the blue arrows present features that contribute to the decrease in the prediction. The width of each arrow shows the height of its impact. From this example, we can see that $l_1 = 1$ and `daytime= 7pm` have the most positive impact on the predicted label, while $l_2 = 0$ and `age≥40` have the most negative impact.

We further used Shapely values to cluster our dataset based on the explanation similarity of samples, using hierarchical agglomerative clustering (figure 5b). From this figure, we can see that the crowdsourced labels are the most influential features in grouping the samples in our dataset. The highlighted areas in this diagram show the samples that have similar force plots, implying the dominant and similar contribution of these features across the physical activity dataset.

Using the additive nature of Shapley values, we integrated all the local feature values for each data point and calculated the global contribution (I) of each feature. Considering $\varphi_j^i \in \mathbb{R}$ as the shapely value of feature j for sample i, we can calculate the global importance of this feature as $I_j = \sum_{i=1}^{n} \left|\varphi_j^i\right|$. Figure 5c shows the combination of feature importance (y-axis) and feature effects (coloured points) for the most influential features, ordered based on their importance. This plot shows that crowdsourced labels (l_1, l_2, l_3), followed by *daytime*, the results of the *inference models*, and *gender* are making the most impact on the decision making of XGBoost. From these results, which are extendable to the other predictive models developed in this study, it can be inferred that regardless of the complexity and the architecture of the predictive models, the crowd generated labels are the factors that most influence predictive models' prediction. While meta-information such as *daytime* and *gender* are amongst the most contributing features (figure 5c), they still cannot compete with the crowd generated labels in the majority of the samples. This can explain the vulnerability of our ML and DL models to the noisy labels of the dataset.

To triangulate the dominant impact of the crowdsourced labels, we excluded all the samples for which ($\sum_{i=1}^{3} \hat{l}_i = 1 \wedge l = 0$) or ($\sum_{i=1}^{3} \hat{l}_i = 0 \wedge l = 1$) from our dataset for both supervised and unsupervised techniques and achieved an F_1 score of ≈ 99%. This implies that inferring the truth label of crowdsourced data highly depends on the quality of the collected data from the crowd, and even advanced and complex predictive models might not be able to compensate for the poor quality of this data.

Limitations

Several limitations should be noted. First, the compensation amount paid to the workers could impact the quality of collected labels, and consequently, the evaluation results of this study. Workers may show higher quality in exchange for a higher payment. To investigate this, during the course of the project, we increased HITs' reward from $0.03 USD to $0.05 USD and did not notice any significant changes in quality. However, this is still debatable and needs further investigation.

Second, to develop the supervised models, we assumed that all the tasks share the same level of complexity, while in reality some examples are more difficult that other. For example, labelling `'I can't sleep'` to a self-reported sleep problem is more straightforward than labelling `'[...] I'm kind of envious of anyone who is able to fall asleep before 2am'`. We tried to address this by incorporating inherent task difficulties in the prediction models through developing a hybrid CNN model. However, crowd-generated labels dominated other features of our dataset, making the most impact on their inference decision. Building crowdsourcing models sensitive to the complexity of tasks to allocate more resources (workers) to more difficult tasks is a worthwhile direction for future research.

Third, the way we designed and presented the HITs on AMT could impact the performance of workers in various ways. Considering the central role of people in maximizing the benefits of crowdsourcing services, human factors should be taken into account when designing crowdsourcing tasks [39]. To address this, we added a succinct, precise, and demonstrative instruction to each task and explained each label with an illustrative example (e.g., supplementary figure 4). Also, through different iterations of data collection, we tweaked the design, presentation, and instructions to ensure we meet the basic usability requirements of tasks' design and presentation.

Fourth, we defined workers' qualifications based only on their historical performance in completing HITs across AMT (i.e., master qualification). Although this provided some degree of quality control on the collected labels,

alternative qualification requirements such as workers' education, work background, and language could have also impacted our study results. To further study the role of qualification filtering, we pilot tested the labelling process without any qualification requirements for 4,500 physical activity tasks. These tasks were completed in less than 12 hours with a consistency score ($\mathcal{LC}$) of less than 0.5, implying he importance of workers' quality in developing crowd-labelled intelligent systems.

Fifth, various physical activities, based on their energy requirements in metabolic equivalents (METs), can be categorized into different movement behaviours, such as light (1.6-2.9 METs), moderate (3-5.9 METs), and vigorous ($\geq$ 6METs) [42]. However, as the datils provided by social media data may not be enough to calculate METs value, in this study, we only used general terms related to physical activity (e.g., physical fitness, exercises, household, sports, or occupational activities) to filter and form the physical activity subset. To ensure the lists of contextual terms for filtering all the PASS categories are comprehensive enough, in addition to domain-specific ontologies and WordNet [43], we used natural language processing (NLP) techniques (e.g., topic modelling, language modelling, and lexical analysis) to detect latent word patterns that can be used to identify PASS-related contexts in unstructured text. However, with no impact on the methodology and results of this study, both data collection bias and population bias (inherent in social media data) should be considered when discussing the dataset used for this study.

Despite these limitations, our study is one of the first to rigorously investigate the challenges of using crowdsourcing for developing ML-based public health surveillance systems. Our findings support the argument that crowdsourcing, despite its low cost and short turnaround time, yields noisier data than in-house labelling. On the flip side, crowdsourcing can reduce annotation bias by involving a more diverse set of annotators [39]. This diversity, supported by the diversity of AMT workers [44], is highly beneficial to subjective labelling tasks such as detecting a sedentary behaviour based on a short text, which highly depends on the worker's understanding of sedentary lifestyles.

Acknowledgement

This work was supported by a postdoctoral scholarship from the Libin Cardiovascular Institute and the Cumming School of Medicine, University of Calgary. Also, this work was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (RGPIN-2014-04743). The Public Health Agency of Canada funded the Amazon Mechanical Turk costs. The funders of the study had no role in study design, data collection and analysis, or interpretation of results and preparation of the manuscript.

Conflict of Interest

The authors have no conflict of interest to declare.

Abbreviations

AMT: Amazon Mechanical Turk

CNN: Convolutional Neural Network

DS: David and Skene

GLAD: Generative model of Labels, Abilities, and Difficulties

HIT: Human Intelligence Task

METs: Metabolic equivalents

MV: Majority Voting

PASS: Physical Activity, Sedentary behaviour, and Sleep quality

RY: Raykar's algorithm

Author Contributions

Z.SH. was responsible for data collection and curation, model development, data analysis and visualization and wrote the paper. G.B. and W.T. reviewed the paper and provided comments. J.L. contributed to the conception and design of the study and revised the manuscript.

Data availability

The dataset required to replicate this study contains the full text of tweets, which in accordance with Twitter policies of data sharing, will not be made publicly available. However, `Tweet_IDs` to retrieve the full Tweet objects from Twitter and crowd-generated labels for each tweet will be available on GitHub.

References

1. Mavragani A. Infodemiology and Infoveillance: Scoping Review. Journal of medical internet research. 2020;22(4):e16206.
2. Aiello AE, Renson A, Zivich PN. Social Media–and Internet-Based Disease Surveillance for Public Health. Annual Review of Public Health. 2020;41:101–118.
3. Sinnenberg L, Buttenheim AM, Padrez K, Mancheno C, Ungar L, Merchant RM. Twitter as a tool for health research: a systematic review. American journal of public health. 2017;107(1):e1–e8.
4. Bernardo TM, Rajic A, Young I, Robiadek K, Pham MT, Funk JA. Scoping review on search queries and social media for disease surveillance: a chronology of innovation. Journal of medical Internet research. 2013;15(7):e147.
5. Hossain L, Kam D, Kong F, Wigand R, Bossomaier T. Social media in Ebola outbreak. Epidemiology & Infection. 2016;144(10):2136–2143.
6. Lardon J, Abdellaoui R, Bellet F, Asfari H, Souvignet J, Texier N, et al. Adverse drug reaction identification and extraction in social media: a scoping review. Journal of medical Internet research. 2015;17(7):e171.
7. Hu H, Phan N, Chun SA, Geller J, Vo H, Ye X, et al. An insight analysis and detection of drug-abuse risk behavior on Twitter with self-taught deep learning. Computational Social Networks. 2019;6(1):10.
8. Cavallo DN, Tate DF, Ries AV, Brown JD, DeVellis RF, Ammerman AS. A social media–based physical activity intervention: a randomized controlled trial. American journal of preventive medicine. 2012;43(5):527–532.
9. Dunn AG, Mandl KD, Coiera E. Social media interventions for precision public health: promises and risks. NPJ digital medicine. 2018;1(1):1–4.
10. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health information science and systems. 2014;2(1):1–10.
11. Abad Z.S.H, Kline A, Sultana M, Noaeen M, Nurmambetova E, Lucini F, et al. Digital public health surveillance: a systematic scoping review. npj Digital Medicine. 2021;4(1):1–13.
12. Paolacci G, Chandler J, Ipeirotis PG. Running experiments on amazon mechanical turk. Judgment and Decision making. 2010;5(5):411–419.
13. Peer E, Brandimarte L, Samat S, Acquisti A. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. Journal of Experimental Social Psychology. 2017;70:153–163.
14. Brabham DC, Ribisl KM, Kirchner TR, Bernhardt JM. Crowdsourcing applications for public health. American journal of preventive medicine. 2014;46(2):179–187.
15. Kim SJ, Marsch LA, Hancock JT, Das AK. Scaling up research on drug abuse and addiction through social media big data. Journal of medical Internet research. 2017;19(10):e353.
16. Lu W, Guttentag A, Elbel B, Kiszko K, Abrams C, Kirchner TR. Crowdsourcing for Food Purchase Receipt Annotation via Amazon Mechanical Turk: A Feasibility Study. Journal of medical Internet research. 2019;21(4):e12047.
17. Ayers JW, Leas EC, Allem JP, Benton A, Dredze M, Althouse BM, et al. Why do people use electronic nicotine delivery systems (electronic cigarettes)? A content analysis of Twitter, 2012-2015. PloS one. 2017;12(3):e0170702.
18. Yin Z, Fabbri D, Rosenbloom ST, Malin B. A scalable framework to detect personal health mentions on Twitter. Journal of medical Internet research. 2015;17(6):e138.

19. McIver DJ, Hawkins JB, Chunara R, Chatterjee AK, Bhandari A, Fitzgerald TP, et al. Characterizing sleep issues using Twitter. Journal of medical Internet research. 2015;17(6):e140.

20. Reece AG, Reagan AJ, Lix KL, Dodds PS, Danforth CM, Langer EJ. Forecasting the onset and course of mental illness with Twitter data. Scientific reports. 2017;7(1):1–11.

21. Adrover C, Bodnar T, Huang Z, Telenti A, Salathé M. Identifying adverse effects of HIV drug treatment and associated sentiments using Twitter. JMIR public health and surveillance. 2015;1(2):e7.

22. Peer E, Vosgerau J, Acquisti A. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. Behavior research methods. 2014;46(4):1023–1031.

23. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in neural information processing systems; 2017;p. 4765–4774.

24. Wang Z, Hale S, Adelani DI, Grabowicz P, Hartman T, Flöck F, et al. Demographic inference and representative population estimates from multilingual social media data. In: The World Wide Web Conference; 2019;p. 2056–2067.

25. Zheng Y, Li G, Li Y, Shan C, Cheng R. Truth inference in crowdsourcing: Is the problem solved? Proceedings of the VLDB Endowment. 2017;10(5):541–552.

26. Sheshadri A, Lease M. Square: A benchmark for research on computing crowd consensus. In: First AAAI conference on human computation and crowdsourcing; 2013.

27. Dawid AP, Skene AM. Maximum likelihood estimation of observer error-rates using the EM algorithm. Journal of the Royal Statistical Society: Series C (Applied Statistics). 1979;28(1):20–28.

28. Ipeirotis PG, Provost F, Wang J. Quality management on amazon mechanical turk. In: Proceedings of the ACM SIGKDD workshop on human computation; 2010;p. 64–67.

29. Whitehill J, Wu Tf, Bergsma J, Movellan JR, Ruvolo PL. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In: Advances in neural information processing systems; 2009;p. 2035–2043.

30. Raykar VC, Yu S, Zhao LH, Valadez GH, Florin C, Bogoni L, et al. Learning from crowds. Journal of Machine Learning Research. 2010;11(4).

31. Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP); 2014;p. 1532–1543.

32. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002;16:321–357.

33. Snoek J, Larochelle H, Adams RP. Practical bayesian optimization of machine learning algorithms. In: Advances in neural information processing systems; 2012;p. 2951–2959.

34. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. Tensorflow: A system for large-scale machine learning. In: 12th Symposium on Operating Systems Design and Implementation; 2016. p. 265–283.

35. Chollet F, et al. Keras: The python deep learning library. Astrophysics Source Code Library. 2018.

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

37. Lewis DD, Gale WA. A sequential algorithm for training text classifiers. In: SIGIR'94. Springer; 1994;p. 3–12.

38. Laws F, Scheible C, Schütze H. Active learning with amazon mechanical turk. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing; 2011;p. 1546–1556.

39. Lease M. On quality control and machine learning in crowdsourcing. Human Computation. 2011;11(11).

40. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018.

41. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nature machine intelligence. 2020;2(1):2522–5839.

42. Caspersen CJ, Powell KE, Christenson GM. Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. Public health reports. 1985;100(2):126.

43. Miller GA. WordNet: a lexical database for English. Communications of the ACM. 1995;38(11):39–41.

44. Difallah D, Filatova E, Ipeirotis P. Demographics and dynamics of mechanical Turk workers. In: Proceedings of the eleventh ACM international conference on web search and data mining; 2018;p. 135–143.