LocalTweets to LocalHealth: A Mental Health Surveillance Framework Based on Twitter Data

Vijeta Deshpande[1], **Minhwa Lee**[2], **Zonghai Yao**[3], **Zihao Zhang**[3], **Jason Brian Gibbons**[4], **Hong Yu**[1,3,5]
[1]University of Massachusetts Lowell,
[2]University of Minnesota,
[3]University of Massachusetts Amherst,
[4]University of Colorado Anschutz Medical Campus,
[5]University of Massachusetts, Chan Medical School,
vijeta_deshpande@student.uml.edu, lee03533@umn.edu, {zonghaiyao, zihaozhang}@umass.edu,
jason.gibbons@cuanschutz.edu, hong_yu@uml.edu

Abstract

Prior research on Twitter (now X) data has provided positive evidence of its utility in developing supplementary health surveillance systems. In this study, we present a new framework to surveil public health, focusing on mental health (MH) outcomes. We hypothesize that locally posted tweets are indicative of local MH outcomes and collect tweets posted from 765 neighborhoods (census block groups) in the USA. We pair these tweets from each neighborhood with the corresponding MH outcome reported by the Center for Disease Control (CDC) to create a benchmark dataset, LocalTweets. With LocalTweets, we present the first population-level evaluation task for Twitter-based MH surveillance systems. We then develop an efficient and effective method, LocalHealth, for predicting MH outcomes based on LocalTweets. When used with GPT3.5, LocalHealth achieves the highest F1-score and accuracy of 0.7429 and 79.78%, respectively, a 59% improvement in F1-score over the GPT3.5 in zero-shot setting. We also utilize LocalHealth to extrapolate CDC's estimates to proxy unreported neighborhoods, achieving an F1-score of 0.7291. Our work suggests that Twitter data can be effectively leveraged to simulate neighborhood-level MH outcomes.
Keywords: Social Media Processing, Corpus (Creation, Annotation, etc.), Evaluation Methodologies

1. Introduction

For effective design of public health interventions, it is critical to have surveillance systems that are reliable and fast-acting. Traditional health surveillance systems often resort to survey-based reporting of health outcomes hence, are subject to response bias, and significant temporal lag (Bitsko et al., 2022). For the timely design and implementation of health intervention programs, real-time data monitoring, processing, and estimation systems are required (Simonsen et al., 2016). Electronic Health Records (EHR) based surveillance systems carry the potential to overcome the disadvantages of the traditional systems (Greco et al., 2023; Simonsen et al., 2016). While EHRs offer valuable insights, operational challenges, their expensive nature, and relatively delayed updates in information compared to social media platforms reduce their effectiveness for real-time public health surveillance (Kataria and Ravindran, 2020; Menachemi and Collum, 2011). Hence, the exploration of supplementary data sources for health surveillance is needed.

Social media platforms as a data source are proving to be important for various surveillance applications (Shakeri Hossein Abad et al., 2021), with Twitter (now X[1]) being one of the most explored platforms for population health surveillance applications (Greco et al., 2023; Mavragani, 2020; Jordan et al., 2018; Pilipiec et al., 2023). The previous decade evidenced a spectrum of research efforts to highlight the utility of Twitter data for health surveillance (Coppersmith et al., 2015; Naseem et al., 2022; Nguyen et al., 2017a; Athanasiou et al., 2023; Klein et al., 2022; Coppersmith et al., 2014; Shakeri Hossein Abad et al., 2021). Numerous studies conducted correlation analysis to emphasize that Twitter activities are highly correlated with the reported outcomes, at the national, state, and even county level (Coppersmith et al., 2015, 2014; Paul and Dredze, 2011). Several studies developed Twitter surveillance systems with advanced Natural Language Processing (NLP) methods to identify tweets indicating serious health concerns (Naseem et al., 2022; Barbieri et al., 2020; Rosenthal et al., 2019; Yadav et al., 2020). However, the research efforts to develop population-level health outcome prediction systems have been quite limited (Nguyen et al., 2017a, 2016b, 2017b, 2016a; Wang et al., 2020; Athanasiou et al., 2023). Recently conducted studies by Barbieri et al. (2020); Naseem et al. (2022) show that the population-level inference tasks are absent in the current collection of Twitter evaluation benchmarks.

Following responsible data practices, we will share data for requests that align with our privacy policy. Corresponding Author: vijeta_deshpande@student.uml.edu

[1]We refer to X by its older name 'Twitter' and refer to the messages posted on X as tweets.

LREC-COLING 2024, pages 10698–10715
20-25 May, 2024.

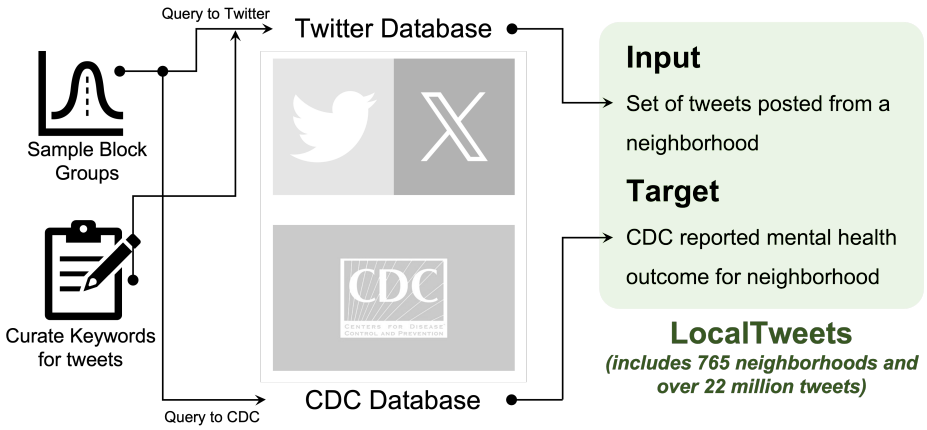

Figure 1: **Data Collection Process.** In this figure, we present a simple schematic of our data curation process. First, we sample 1K neighborhoods (i.e., block groups or BGs) and curate a list of keywords for three categories of tweets, to form queries. Secondly, we query the CDC and Twitter databases to collect the desired data. Lastly, for each BG, we join the set of tweets posted from the BG with the reported health outcome from the CDC database. The final cleaned version of LocalTweets includes 765 unique BGs, spans over five years, and includes over 22 million tweets.

Exacerbating the issue, several analytical limitations within the population-level studies constrain the transferability of findings. First, keyword-based tweet filtering is hypothesized to improve prediction systems, but this has not been tested. Second, studies have focused on larger geographical areas (census tract being the smallest area considered in Nguyen et al. (2016b)) hence, indirectly normalizing worsened health conditions in smaller, resource-deprived areas. Lastly, most population-level prediction systems presented in literature employ rule-based or count-based feature extraction to encode tweets, which lacks the benefits of advanced pre-trained language models.

To overcome the above-mentioned limitations in the literature, we first present a benchmark dataset; LocalTweets; that enables the **prediction of neighborhood-level mental health (MH) outcomes**[2] from locally posted tweets. Compared to previous studies we **focus on a much smaller geographical unit, Census Block Group (BG)** [3] in LocalTweets, refer to Figure 1. LocalTweets includes data for 765 unique BGs, spans over a period of five years (2015-2019), and includes more than 22 million tweets. Furthermore, we propose an efficient and effective analytical framework, LocalHealth, that **leverages language models to encode locally posted tweets** and predicts MH outcomes based on the encodings. We evaluate LocalHealth with extensive experiments and find that the **unfiltered tweets present better generalization properties** compared to filtered tweets (containing MH-related keywords). With LocalHealth we achieve an F1-score of 0.7429 in predicting future outcomes and 0.7291 in predicting outcomes for a proxy set of unreported BGs.

Our work thus lays the groundwork for the development of a neighborhood-level, real-time MH surveillance system and holds substantial benefits for public health decision-making. For example, the presented work in this study can directly be used to identify neighborhoods that can benefit from additional MH care resources and the establishment of community MH programs. In the following sections, we delineate details of our analysis.

2. Related Works

Numerous studies evaluated Twitter data for surveillance applications (Greco et al., 2023; Mavragani, 2020; Jordan et al., 2018; Proserpio et al., 2016; Klein et al., 2022; Kim et al., 2023; De Choudhury et al., 2016; Abdellaoui et al., 2017; Simonsen et al., 2016; Shakeri Hossein Abad et al., 2021). Twitter-based surveillance studies can be divided into three categories: (1) **Correlation Studies:** studies that investigate the agreement between Twitter data and reported cases of health conditions (Paul and Dredze, 2011; Broniatowski et al., 2013; Coppersmith et al., 2015; Velardi et al., 2014; Paul et al., 2015; Schwartz et al., 2013; Culotta, 2014; Jashinsky et al., 2014); (2) **Tweeet/User-level Studies:** studies that develop tweet-level or user-level categorization systems to identify tweets or users relevant to a particular health condition (Braith-

[2] For a precise definition of the MH outcome refer to (for Disease Control, 2023; for Disease Control et al., 2022)

[3] United States Census Bureau has defined geographical units to collect data from. Census Block Groups are areas with a population ranging from 600 to 3000. More detailed definitions can be found at Bureau (2023a)

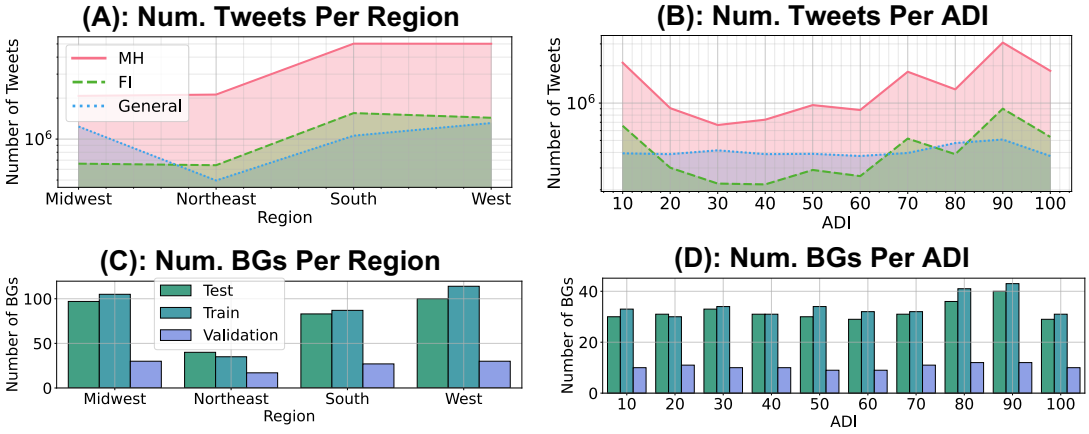

Figure 2: **Distributional Properties of LocalTweets.** (A): Region vs. Number of Tweets: MH tweets are the most numerous, while FI and general tweets have comparable volumes. Tweet volume is slightly skewed toward the South and West regions. (B) ADI vs. Number of Tweets: MH and FI tweets are slightly skewed toward ADIs ≥ 70 and ≤ 20. (C) Region vs. Number of BGs: The distribution of data splits over regions is approximately the same. The number of BGs from the Northeast region is less than other regions. Refer to Appendix E for more discussion. (D): ADI vs. Number of BGs: The number of BGs is fairly balanced over the ADI values and across the data splits.

waite et al., 2016; De Choudhury et al., 2017; Coppersmith et al., 2014, 2015; Barbieri et al., 2020; Naseem et al., 2022; Yadav et al., 2020); and (3) **Population-level Studies:** methods that process a set of tweets to make population health inferences. The setup of our study is closest to the third type.

Population-level Studies: Recently conducted studies by Barbieri et al. (2020); Naseem et al. (2022) show that the population-level inference tasks are not present in the current evaluation benchmarks for Twitter data-based Natural Language Processing (NLP) systems. However, there are a few notable studies. In studies conducted by Culotta (2014); Schwartz et al. (2013); Giorgi et al. (2018), the authors encode tweets using a count-based system and then use the encodings to make county-level inferences. In studies conducted by Nguyen et al. (2016b,a, 2017b), the authors develop machine learning systems that can extract essential indicators of health from tweets and show that extracted indicators are associated with state or census tract-level health outcomes. Athanasiou et al. (2023) conducted clustering analysis and encoded tweets to represent the presence of word clusters. The authors later used encoded tweets along with other data sources for country-level prediction of influenza-like illness outcomes. In a recent study conducted by Zhang et al. (2022), the authors first developed a tweet-level identification system to focus on COVID-related tweets and then used the filtered pool to make country-level COVID outcome prediction. In the above-mentioned studies, either in the data collection or in the encoding process, a focus is put on a specific set of keywords or features. In addition, the authors consider large geographic areas (the smallest area being the census tract in Nguyen et al. (2016b)) for making predictions.

3. Data

For the presented analysis, we collected tweets posted from 1,000 census block groups in the United States. We refer to a block group as a neighborhood and use both terms interchangeably. Then, we coupled the Twitter data with mental health outcome estimates reported by the Center for Disease Control (CDC). We refer to the final cleaned version of the collected data as the LocalTweets dataset. In the following subsections, we discuss our data collection process in detail.

3.1. Sampling of Block Groups

We started by sampling 1,000 block groups (BGs) from the contiguous United States. Specifically, we stratified the BGs by geographic region (northeast, south, midwest, west) (Bureau, 2023b; Wikipedia, 2023a), and Area Deprivation Index (ADI) (for Health Disparities Research, 2023) [4]. We created 40 strata (four regions and ten ADI bins) and sampled 25 BGs from each stratum.

[4]ADI values are calculated based on the data collected in the American Community Survey (ACS) 5-year estimates at Census Block Group level, representing the socio-economic profile of a respective block-group. ADI values are between one to a hundred, the highest value being the most undesirable.

3.2. Collection of Twitter data

Identification and collection of tweets that are relevant to the mental health (MH) status (i.e., outcome of interest) of a population is challenging. Population MH status is an expansive construct influenced by a multitude of demographic, socio-economic, infrastructural, and other factors. Thus, for Twitter data collection, it may not be possible to create a set of keywords that exhaustively cover all possible linguistic expressions of MH-related distress across demographic and cultural features of the population. Hence, we hypothesize that unfiltered tweets may present better datasets for population-level MH surveillance tasks. However, to test our hypothesis we collect keywords-based filtered tweets as well. Overall we collect three subsets of Twitter data, each subset corresponding to a unique category of tweets. The categories are defined by three mutually exclusive sets of keywords used for the collection of tweets. First is the MH category i.e., a subset of tweets that contain keywords directly related to the outcome of interest (MH outcome in our case). For the second category, we list keywords that map to a risk factor of the outcome of interest. In our analysis, we focus on food insecurity, a well-proven contributor to MH (Pourmotabbed et al., 2020; Elgar et al., 2021). Lastly, we collect "general" tweets i.e., tweets that contain only one keyword, a space character. We provide our manually curated keyword lists for MH, FI, and general categories in Appendix A. Using the Twitter Developer API and twarc (Summers et al., 2023) library, we collect all three subsets of tweets from BGs selected in stratified sampling. For general tweets, we upper bound our collection to 1,000 tweets per BG per year due to the sheer volume of general tweets. We repeat this process for each year from 2015 to 2019. We also collect counts of tweets for each category without any upper bound, using the "twarc count" command. Additional details of the twarc library parameters are provided in the Appendix A.

3.3. Coupling the data with ground truth labels

After collecting tweets for each sampled BG, we pair the Twitter data with health outcomes published by CDC (also referred to as ground truth) (for Disease Control et al., 2022). For the presented study, we primarily focus on one outcome, the percentage of the adult population with MH not good for more than 14 days of the last 30 days (for Disease Control, 2023). These prevalence estimates are reported annually at the BG level. Hence, for every unique pair of a year and a BG, we couple the collected set of tweets with the prevalence estimate value available from CDC data. In Figure 1, we provide a schematic of our data collection process.

3.4. Cleaning and splitting the data for model development

We filter the Twitter-CDC coupled data to remove BGs with no tweets (for any category and any year), and BGs not included in the CDC data. This filtering process removed 235 BGs, resulting in a final dataset of 765 BGs with corresponding Twitter and CDC data for five years. We refer to the cleaned version of the data as LocalTweets. Figure 2 shows the distributional properties of the LocalTweets. We provide a detailed description of data properties in Appendices B and C. In the cleaned version of LocalTweets, we observe significantly fewer BGs from the Northeast region. Hence, we analyzed the effect of fewer BGs from the Northeast region on the space generalizability of the data. We observe that fewer BGs from the Northeast region do not hamper the space generalizability of the data and provide details in Appendix E.

We split the data differently for our two experimental settings namely, the forecasting setting and spatial extrapolation setting. For forecasting, we use all 2019 data as the test dataset and consider 2015 to 2018 data for model development. For the spatial extrapolation, we first divide the final set of 765 BGs into three splits (test, train, and validation) such that the distribution over ADI values and geographic regions is approximately held the same across splits, refer to Figure 2, panels C and D. We remove the BGs in test split, to create a proxy set of "unreported" BGs. We use 2015-19 data from the train and validation split for model development and report the performance on the unreported BGs (test split) of the year 2019.

We denote LocalTweets as D, defined as follows:

$$D = \{(t_{b,y}^{(k,s)}, c_{b,y}^{(k,s)}, g_{b,y}^{(s)}, r_{b,y}^{(s)})\}$$

where the subscripts b and y represent the BG and year, the superscripts k and s denote the tweet category and split. The variables t, c, and g represent the tweets, tweet counts, and ground truth outcome values, respectively. We create a risk category variable r, such that $r_{b,y}^{(s)} = 1$ (high-risk BG) if $g_{b,y}^{(s)} \geq 75^{th}$ percentile, otherwise $r_{b,y}^{(s)} = 0$ (low-risk BG). In other words, r is nothing but a flag indicating whether a BG is high-risk or not. We leverage the variable r to evaluate regression models' performance with discrete metrics (e.g., accuracy and F1-score). Note that the superscript k is not present for g and r variables because we only consider the MH outcome as the target variable.

4. Methodology

Our methods of analysis are mainly divided into two parts. First, we conduct a correlation analysis to investigate the agreement between Twitter data-based statistics and the reported CDC outcome values. In the second part, we present a regression analysis. Specifically, we develop a model that processes the Twitter data (set of tweets) to predict the MH outcome value for respective neighborhoods, refer to Appendix G for a simple schematic. In the following subsections, we discuss each part in detail.

4.1. Correlation analysis

We use the Pearson correlation coefficient (Wikipedia, 2023b; Virtanen et al., 2020; Scipy, 2023) to measure the correlation between the ground truth values $g_{b,y}$ (the reported MH outcomes) and the Twitter activity i.e., tweet counts $c_{b,y}^{(MH)}$, $c_{b,y}^{(FI)}$ and $c_{b,y}^{(General)}$. In addition, we also measure the correlation between the ground truth and the ADI values. The correlation is measured separately for each year and over the 765 BGs in the LocalTweets.

4.2. Regression analysis

In the regression analysis, our goal is to develop a parametric function that can predict the continuous and scalar-valued MH outcome for a BG ($g_{b,y}$), based on the set of tweets posted from the same BG. We develop two types of models, one utilizing the tweet count values and the other utilizing the set of tweets. For the count-based model, we utilize the normalized count values (normalized by the count of all tweets posted from the BG) for the MH and FI categories of tweets. We adopt a simple linear regression setting for the count-based model, where normalized count values act as input variables. For the case when we consider both, MH and FI counts, we treat each normalized count value as a separate variable in the linear regression model.

For the text-based model, we follow Algorithm 1, with four main steps, sampling, encoding, aggregation, and prediction. In the sampling step, if the total number of tweets exceeds the 4K mark, we uniformly sample 4K tweets. The sampled tweets are encoded with a language model and then aggregated across sequence length and number of tweets. Finally, we employ a convolutional neural network ($f_{conv}(\cdot)$) followed by a fully connected neural network ($f_{fcn}(\cdot)$) to predict the MH outcome value based on aggregated encodings ($\bar{v}_b^{(k)}$), therefore,

$$\hat{g}_b = f_{fcn}(f_{conv}(\bar{v}_b^{(k)}; \theta_{conv}); \theta_{fcn}) \quad (1)$$

Input: D, $f_{\text{LM}}(\cdot)$, $f_{\text{conv}}(\cdot)$, $f_{\text{fcn}}(\cdot)$
Require : $B = \{\text{BGs in LocalTweets}\}$,
$Y = \{2015, \ldots, 2019\}$

Step : Sampling
for *each* $b \in B$, $y \in Y$ **do**
  Sample $t_{b_s,y}^{(k)} \sim \text{Uniform}(t_{b,y}^{(k)})$
end
Such that, $|t_{b_s,y}^{(k)}| = \min(4000, |t_{b,y}^{(k)}|)\ \forall\ b, y, k$
Step : Encoding
for *each* $b \in B$, $y \in Y$ **do**
  $v_{b_s,y}^{(k)} = f_{\text{LM}}(t_{b_s,y}^{(k)}; \theta_{\text{LM}})$;
end
Such that, $v_{b_s,y}^{(i)} = [v_{b_s,y,1}^{(k)}, v_{b_s,y,2}^{(k)}, \ldots, v_{b_s,y,n}^{(k)}]$,
where $v_{b_s,y,j}^{(k)}$ is the representation vector of the j^{th} tweet in the $t_{b_s,y}^{(k)}$ set and $n = |t_{b_s,y}^{(k)}|$
Step : Aggregation
for *each* $b \in B$, $y \in Y$ **do**
  $\bar{v}_{b_s,y}^{(k)} = \frac{1}{|t_{b_s,y}^{(k)}|} \cdot \sum_{j=1}^{|t_{b_s,y}^{(k)}|} v_{b_s,y,j}^{(k)}$
end

Step : Prediction
$\hat{g}_{b,y} = f_{\text{fcn}}(f_{\text{conv}}(\bar{v}_{b_s,y}^{(k)}; \theta_{\text{conv}}); \theta_{\text{fcn}})$

Algorithm 1: LocalHealth Approach. In this table, we present the LocalHealth algorithm to predict mental health outcome values based on a set of tweets. There are four main steps namely, sampling, encoding, aggregation, and prediction of outcome value. Superscript k denotes the tweet category.

Where, θ_{conv} and θ_{fcn} represent parameters of the $f_{conv}(\cdot)$ and $f_{fcn}(\cdot)$, respectively. We predict the reported MH outcome based on various categories of tweets by simply changing the $\bar{v}_b^{(k)}$. When we consider both, MH and FI tweets to make predictions ($k = \{MH \text{ and } FI\}$), we add the vectors for MH and FI to compute the final aggregated vector i.e., $\bar{v}_b^{(k)} = \bar{v}_b^{(MH)} + \bar{v}_b^{(FI)}$. We refer to our approach presented in Algorithm 1 as LocalHealth.

5. Experimental Setup

5.1. Sets of experiments

We divide our experiments into four sets.

Set-1: Effect of input information type. In this set of experiments, we compare the effects of different information priors (ADI values, tweet counts, tweet texts, and tweet categories) on forecasting MH outcomes. Hence, we use data from 2015 to 2018 for developing the model 2019 data for testing (forecasting data splits discussed in Section 3.4).

Year	MH	FI	General	ADI
2015	0.1640	0.1460	0.1299	0.6767
2016	0.1366	0.1332	0.1215	0.7074
2017	0.1123	0.1132	0.0969	0.7257
2018	0.0928	0.0937	0.0863	0.7162
2019	0.0922	0.0954	0.0832	0.7318

Table 1: **Correlation Results.** In this table, the columns MH, FI, General, and ADI represent the Pearson Correlation Coefficient between the CDC-reported MH outcome i.e., $g_{b_s,y}$ and count of mental health, food insecurity, general tweets, and the ADI values, respectively. All correlation coefficients are statistically significant with $p < 0.05$.

We also augment the count-based and text-based models with ADI information to measure the impact of combined information. To augment ADI, we concatenate the normalized ADI value ($ADI/100$) with the scalar output of LocalHealth ($f_{fcn}(f_{conv}(\bar{v}))$) and pass the vector to a linear layer to predict the target outcome value.

Set-2: Effect of text encoder. In this experiment, we replace the language model ($f_{LM}(\cdot)$) in Algorithm 1 with various pre-trained language models, including Twitter-RoBERTa (Barbieri et al., 2020), PHS-BERT(Naseem et al., 2022), and GPT-3.5 (OpenAI, 2021), and assess the changes on forecasting performance. We work with "general" category of tweets and consider 2015 to 2018 data for developing the model and, 2019 data for testing. We also measure the zero-shot performance of GPT-3.5, refer to Appendix F for further details.

Set-3: Effect of data availability. In the first two experiments, we utilize data from 2015 to 2018 for training and validating the model. Here in the third set, we gradually reduce the data availability from the prior four years (2015-18) to the prior year (2018) and examine the changes in the forecasting performance subject to data availability. Specifically, we create four train and validation sets based on the data from 2015 to 2018, 2016 to 2018, 2017 to 2018, and only 2018, respectively. We keep the test set (2019 data) constant for all data availability scenarios and compare two language models in this setting, RoBERTa-base, and GPT3.5.

Set-4: Spatial extrapolation capabilities. In the fourth set of experiments, we evaluate the LocalHealth approach based on the capabilities to extrapolate CDC outcomes to the unreported BGs (refer to Section 3.4). We consider five data availability scenarios, 2015 to 2019, 2016 to 2019, and likewise till 2019-only. Based on data availability we vary the training and validation data while keeping the test data fixed to 2019 data for proxy unreported BGs.

5.2. Language models and regression head

In Set-1 experiments, we use the RoBERTa (base configuration) model (Liu et al., 2019) to encode the tweets. In Set-2, we provide results for multiple language models e.g., Twitter-RoBERTa (Barbieri et al., 2020), PHS-BERT (Naseem et al., 2022), GPT-3.5 (OpenAI, 2021), etc. We do not update the parameters of the pre-trained language model. Only the parameters corresponding to the regression head of our framework i.e., θ_{conv} and θ_{fcn}, are updated. For all our experiments, we fixed the structure of the convolutional head to have a single channel, a kernel size of 16, and a stride of four.

5.3. Baseline models

We provide four baseline models. First is the majority baseline i.e., predicting all BGs in the test set as non-high-risk BGs ($r_{(b,2019)} = 0,\ \forall\ b$). Second, we report the performance of the linear regression model making predictions based only on normalized ADI values. For the third and fourth baselines, we make use of Logistic Regression (LoR) and Support Vector Machine (SVM) models, respectively, along with aggregated general tweet encodings ($\bar{v}$) (from RoBERTa-base model) to directly predict the risk category (r) of the BGs.

5.4. Hyperparameters and evaluation

We train all models for 1,600 epochs with a batch size of 512. We use a linear learning rate schedule with a 20% warmup and peak learning rate of 1×10^{-3}. We minimize mean squared error (MSE) using AdamW (Loshchilov and Hutter, 2017; PyTorch, 2023) with a weight decay of 0.1. In order to leverage standard classification metrics for model evaluation, we employ a thresholding technique to convert the continuous-valued model outputs ($\hat{g}_{b,y}$) into binary risk category predictions. This allows us to directly compare predicted risk categories with the ground truth labels (variable r, Section 3.4) using established metrics like accuracy and F1-score (macro-averaged). Model selection is conducted based on the macro-F1 score achieved on the validation set. Following training, we evaluate the best model on the test split and report averaged macro-F1 and accuracy across 10 random seeds. For both, SVM and LoR, we change the loss function to binary cross-entropy and we use a classification threshold of 0.15 to identify high-risk BGs.

6. Results

We started the assessment of the surveillance utility of Twitter data with correlation tests, between the reported MH outcomes and tweet counts. We

Input information	F1-score	Acc. (%)
Majority baseline	0.4336	76.56
Text-based (LoR)	0.5224	52.75
Text-based (SVM)	0.5510	**76.73**
ADI-only (LR)	0.6406	72.81
Count-based (LR)		
MH only	0.5052	61.95
FI only	0.4465	67.88
MH and FI	0.5133	63.84
General only	–	–
Text-based (LocalHealth)		
MH only	0.5668	60.05
FI only	0.5602	64.43
MH and FI	0.5853	64.16
General only	0.5984	66.76
Count-based (LR) with ADI		
MH-only	0.5647	69.36
FI only	0.5545	71.57
MH and FI	0.6138	69.31
General only	–	–
Text-based (LocalHealth) with ADI		
MH only	0.7089	74.52
FI only	0.7117	75.36
MH and FI	0.7085	74.33
General only	**0.7236**	76.48

Table 2: **Effect of Input Information Type.** Here, we present the F1-score and accuracy (Acc.) for identifying the risk category of BGs. Within text-based models, general tweets present better performance than other tweet categories. LoR and LR stand for logistic and linear regression, and SVM for support vector machine.

find that tweet counts moderately correlate with the MH outcome, refer to Table 1. The correlation strength for general tweet counts is consistently lower than the MH or FI tweet counts. In other words, a higher volume of general tweets moderately correlates with worse MH outcomes but, the higher count of MH or FI tweets correlates with worse outcomes marginally better. For validation, we also conducted a correlation test between MH outcome and ADI value and confirmed much higher correlation strengths compared to all categories of tweets. To further scrutinize the utility of Twitter data, in the latter part of our analysis, we conducted four sets of experiments to evaluate the LocalHealth approach. We will discuss the results of each set of experiments one by one.

Set-1: Effect of input information type. In the first set of experiments, we compared various information priors available in LocalTweets: tweet count, tweet texts, tweet categories, and ADI values. The count-based regression models failed to exceed any of the baselines except the majority baseline, refer to Table 2. The text-based LocalHealth models performed better than the count-based models but fell short of the ADI baseline. The F1-score improved significantly, on average by 18% for count-based models and by 24% for text-based models, after augmenting with ADI values. Notably, the text-based model augmented with ADI outperformed the individual counterparts, with an F1-score of 0.7236 and an accuracy of 76.48%, refer to Table 2. This result highlighted the complementary nature of the information contained in tweets compared to the ADI values.

For text-based models, a comparison within the tweet categories revealed interesting insights. The model with general tweets performed better than other categories of tweets. This finding supports our hypothesis (refer to Section 3.2) and highlights the better generalization capabilities of the general tweets for population-level MH outcome prediction, compared to the keyword-derived tweets. Hence, our finding motivates the usage of general tweets for the prediction of population-level MH outcomes. We provide additional comparison of the statistical properties of tweet categories in Appendix D.

Language Model	Train Par.	F1-score	Acc. (%)
Majority baseline	–	0.4336	76.56
GPT3.5 (0-shot)	0	0.4675	76.21
ADI only	2	0.6406	72.81
RoBERTa-base	210	0.7236	76.48
RoBERTa-large	274	0.7228	76.04
Twitter-RoBERTa-base	210	0.7245	76.44
PHS-BERT	274	0.7301	76.97
GPT3.5	402	**0.7429**	**79.78**

Table 3: **Effect of Text Encoder.** Here, we present the F1-score and accuracy (Acc.) for identifying the risk category of BGs. RoBERTa presents competitive results compared to the best-performing GPT3.5. Train Par.: Trainable parameters in LocalHealth setting.

Set-2: Effect of text encoder. By focusing our attention on the text-based model (general tweets) augmented with ADI information, we measure the effect of changes in the language model ($f_{LM}(\cdot)$) used for encoding tweets.

We experiment with five language models, RoBERTa-base, RoBERTa-large, Twitter-RoBERTa-base, PHS-BERT, and GPT3.5, and present our results in Table 3. The effect of the size of the language model was mixed. We observed a minor reduction of 0.0008 in the F1-score for the RoBERTa-large compared to the RoBERTa-base. For the domain-adapted models,

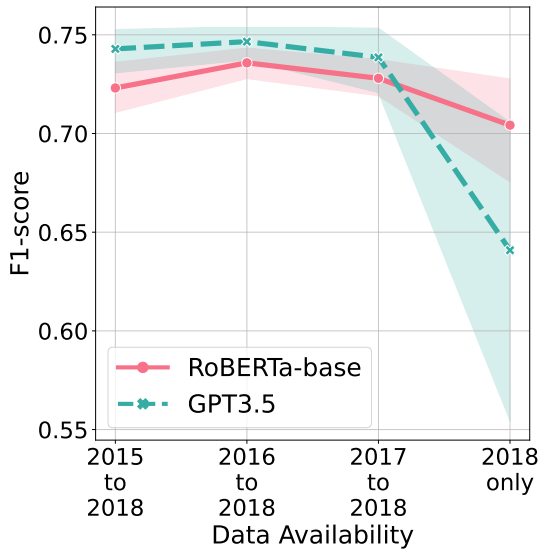

Figure 3: **Effect of Data Availability on Prediction of Future Outcomes.** In the figure, we present the effect of data availability (x-axis) on the prediction of future i.e., 2019 (all 765 BGs in LocalTweets), MH outcomes. We evaluate models on the correct identification of the BG risk category and plot the F1-score on the y-axis. The lines and shaded regions represent the average value and range of F1-scores, calculated over 10 seeds.

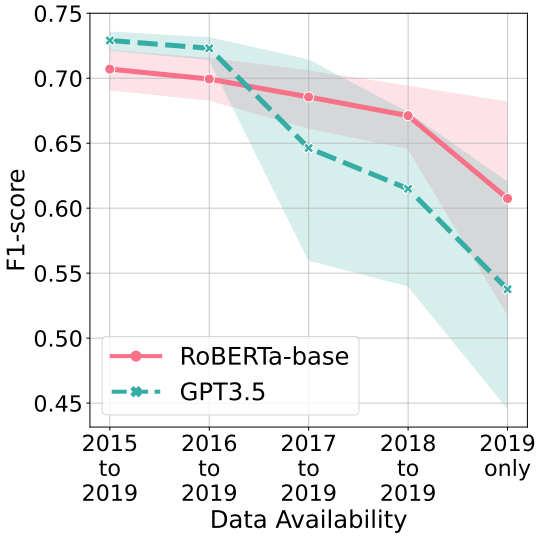

Figure 4: **Effect of Data Availability on Prediction of Outcomes for Unreported Neighborhoods.** In the figure, we present the effect of data availability (x-axis) on the prediction of MH outcomes for a set of proxy unreported BGs (test split in 2019, 320 BGs). We evaluate models on the correct identification of the BG risk category and plot the F1-score on the y-axis. The lines and shaded regions represent the average value and range of F1-scores, calculated over 10 seeds.

we observed an increment of 0.0056 in F1-score for the PHS-BERT (250 mil. parameters) compared to Twitter-RoBERTs (120 mil. parameters). The effect of domain adaptation was consistent for various sizes of the language models. Twitter-RoBERTa improved F1-score and accuracy by 1.2% and 1.0% compared to RoBERTa-base. The same improvements were 1% and 0.6% for PHS-BERT compared to RoBERTa-large. Interestingly, we observed a striking 59% improvement in the F1-score for GPT-3.5 compared to GPT3.5 in a zero-shot setting. This highlights the difficulty of the MH outcome prediction task, especially for the zero-shot setting. Of all language models evaluated, we observed the best F1-score and accuracy of 0.7429 and 79.78% for the GPT-3.5 model when used in the LocalHealth approach.

Set-3: Effect of data availability. In this set of experiments, we investigate the impact of varying data availability on the performance of the models. Performance trends for both models, GPT-3.5 and RoBERTa-base, reveal valuable insights.

Contrary to our expectations, we observed a slight declination in the F1-score when we augmented the 2015 data with the data from 2016 to 2018 for developing the model, refer to Figure 3. The declination is 0.0036 for GPT3.5 and 0.0127 for RoBERTa-base. This unexpected dip may be attributable to the possibility that Twitter posts in 2015 may not accurately represent the population MH status in 2019. Interestingly, we found that RoBERTa (with an F1-score of 0.7042) outperforms GPT-3.5 (F1-score: 0.6406) when only 2018 data is available to develop the model. In other cases, when more data is available, GPT-3.5 presents as a better choice of tweet encoder. To this end, we compared the statistical properties of RoBERTa-base and GPT-3.5. We observe a higher variance in the encodings taken from the RoBERTa model and speculate that the more spread out data potentially helps LocalHealth model to detect underlying patterns for MH prediction (refer to Appendix C).

Set-4: Spatial extrapolation capabilities. In this set of experiments, distinct from the forecasting task in previous iterations, our focus shifts to predicting MH outcomes for unreported neighborhoods (BGs). Opposed to the findings of Set-3 experiments (forecasting task), we observe that more data is always beneficial for both, RoBERTa-base and GPT3.5, refer to Figure 4. Because the proxy set of unreported BGs is never seen by the model, training on more data likely helps the model to find generalizing patterns. Similar to the findings of Set-3, we observe that for limited data availability, RoBERTa stands out as a superior choice of text encoder. In addition, we find RoBERTa to be more robust to changes in the data availability compared to GPT-3.5. The F1-score values

for various data availabilities span over a narrow range of 0.0997 ($[0.6074, 0.7071]$) for RoBERTa, while the same range is almost double; 0.1915 ($[0.5376, 0.7291]$); for GPT-3.5 (Figure 4).

7. Conclusion and Future Work

In this study, we introduce LocalTweets, a novel dataset for mental health (MH) surveillance at the neighborhood level, based on locally posted tweets. We present a simple and efficient approach, LocalHealth, to predict health outcomes based on tweets. Our findings suggest that general category tweets generalize better than the tweets filtered with MH-related keywords. Our results also emphasize RoBERTa-base's effectiveness in data-limited settings.

Our work thus lays the groundwork for a more nuanced and responsive approach to population MH surveillance, fostering advancements in natural language processing methodologies. Alongside surveillance, our work can guide public health resource allocation decisions. For example, presented data and methods can be utilized directly to identify neighborhoods that can benefit from the establishment of community health programs.

Extending our analysis, in the future we hope to investigate resource allocation decisions for specific MH and other health conditions. Furthermore, we also plan to broaden our dataset to include a balanced representation of features that impact the care continuum. We believe improvements in this direction can help us understand the care needs of various communities in a better way.

8. Limitations

Our study has several limitations that should be taken into account when interpreting the results. First, for the stratified sampling of BGs, we do not consider features such as the availability of healthcare facilities in the neighborhood, insurance-holding population, urban-rural status, educational level, etc. As a result, our data may not capture a balanced view of the population along these features that potentially impact health outcomes. Second, the tweets collected under the "general" category are not randomly sampled due to the chronological ordering of the Twitter data. This may skew the distribution of the data over time of the year and may limit the applicability of our work for seasonal health conditions. Third, our framework can not make inferences for the population unable to access the internet or Twitter. However, based on the estimates of the size of the population not using the internet (Meeker and Wu, 2018), we speculate that this limitation only minimally affects our contributions. Lastly, the cost of our presented framework may increase based on Twitter's data pricing policy. However, our findings can help users focus on general tweet data and reduce the volume, time, and cost of data collection.

9. Ethics Statement

While this study demonstrates the potential utility of Twitter data for supplementary mental health surveillance, we acknowledge important ethical considerations. In this section, we describe the procedure we adopted to ensure rightful data access, privacy preservation, and gated sharing of the data.

Data Access: To access Twitter data we followed the Twitter Developer Account application procedure[5]. Our application for accessing Twitter data was reviewed, scrutinized, and approved by Twitter, based on an academic research proposal focused on leveraging Twitter data for public health applications.

Data Privacy Preservation: The use of social media data raises privacy concerns. We took rigorous steps throughout our analysis to protect the privacy of the data. Firstly, we selected tweets that are publicly available and did not collect data from any profiles or tweets that are marked private. For model development, we focused on tweet texts only and did not make use of any additional features of the tweet or the user such as, location or demographic features. We used twarc library to fetch tweets posted from a specific location (BGs) but we did not access or utilize the location feature of the tweets for model development. Furthermore, in the presented study we ensured privacy preservation through encoding and aggregation of the tweets. In the first step textual information gets encoded into high-dimensional vectors and thousands of such vectors are aggregated together for each block-group. With these two steps, we ensure that human-readable text cannot be excavated from the aggregated representation.

Data Bias: The final version of LocalTweets likely is biased along demographic characteristics. However, to maintain the privacy of users and ethical usage of the collected data we did not explore bias along demographic features. We assume that the bias exists and we improve the performance of LocalHealth within the constraints of the demographic bias. Nonetheless, we maintain a fairly balanced distribution across ADI values. We clearly show the existing regional bias and conduct detailed analysis to show that it does not affect results majorly. Furthermore, we show that the bias derived from the commonly used keyword-based data collection methods does not generalize well.

[5]Updated procedure and terms and conditions can be found HERE.

Hence, we make the best effort to address bias-related issues while maintaining privacy of the data.

Usage of LocalHealth: We study the usage of LocalTweets and the application of LocalHealth solely as a supplementary system for traditional health surveillance systems. While supplementary surveillance has benefits, it cannot capture lived experiences. Thus our findings should be considered preliminary and complemented by qualitative, participatory research methods. In addition, mental health is a sensitive topic, and care must be taken not to further stigmatize mental illness. While our work aims for the betterment of mental health public policies, we acknowledge that the findings of our study could be used to develop algorithms that can target distressed areas or populations at risk with discriminatory or harmful content. Hence, we will provide gated access to LocalHealth. We will release the model and the data based on individual requests that adhere to, (1) focus usage of the data for research on public health research questions (2) follow Twitter's data privacy policy.

Reproducability: Lastly, on the technical front, we made our best efforts to reduce the technical barrier to research by considering economic language models, and training lean (≤ 402 parameters) systems. Our goal was to encourage community participation for the benefit of the community. However, we recognize that the barrier is also contingent upon Twitter's privacy policies.

10. Bibliographical References

Redhouane Abdellaoui, Stéphane Schück, Nathalie Texier, Anita Burgun, et al. 2017. Filtering entities to optimize identification of adverse drug reaction from social media: how can the number of words between entities in the messages help? *JMIR public health and surveillance*, 3(2):e6577.

Maria Athanasiou, Georgios Fragkozidis, Konstantia Zarkogianni, and Konstantina S Nikita. 2023. Long short-term memory–based prediction of the spread of influenza-like illness leveraging surveillance, weather, and twitter data: Model development and validation. *Journal of Medical Internet Research*, 25:e42519.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.

Rebecca H Bitsko, Angelika H Claussen, Jesse Lichstein, Lindsey I Black, Sherry Everett Jones, Melissa L Danielson, Jennifer M Hoenig, Shane P Davis Jack, Debra J Brody, Shiromani Gyawali, et al. 2022. Mental health surveillance among children—united states, 2013–2019. *MMWR supplements*, 71(2):1.

Scott R Braithwaite, Christophe Giraud-Carrier, Josh West, Michael D Barnes, and Carl Lee Hanson. 2016. Validating machine learning algorithms for twitter data against established measures of suicidality. *JMIR mental health*, 3(2):e4822.

David A Broniatowski, Michael J Paul, and Mark Dredze. 2013. National and local influenza surveillance through twitter: an analysis of the 2012-2013 influenza epidemic. *PloS one*, 8(12):e83672.

United States Census Bureau. 2023a. Census blocks and block groups.

United States Census Bureau. 2023b. Statistical groupings of states and counties.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 51–60.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 1–10.

Aron Culotta. 2014. Estimating county health statistics with twitter. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1335–1344.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2098–2110.

Munmun De Choudhury, Sanket S Sharma, Tomaz Logar, Wouter Eekhout, and René Clausen Nielsen. 2017. Gender and cross-cultural differences in social media disclosures of mental

illness. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 353–369.

Vijeta Deshpande, Dan Pechi, Shree Thatte, Vladislav Lialin, and Anna Rumshisky. 2023. Honey, i shrunk the language: Language model behavior at reduced scale. *arXiv preprint arXiv:2305.17266*.

Frank J Elgar, William Pickett, Timo-Kolja Pförtner, Geneviève Gariépy, David Gordon, Kathy Georgiades, Colleen Davison, Nour Hammami, Allison H MacNeil, Marine Azevedo Da Silva, et al. 2021. Relative food insecurity, mental health and wellbeing in 160 countries. *Social science & medicine*, 268:113556.

Center for Disease Control. 2023. Health status metrics.

Centers for Disease Control, Prevention, Centers for Disease Control, Prevention, et al. 2022. Places: local data for better health.

Center for Health Disparities Research. 2023. Neighborhood atlas.

Salvatore Giorgi, Daniel Preoţiuc-Pietro, Anneke Buffone, Daniel Rieman, Lyle Ungar, and H. Andrew Schwartz. 2018. The remarkable benefit of user-level aggregation for lexical-based population-level predictions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1167–1172, Brussels, Belgium. Association for Computational Linguistics.

Candida M Greco, Andrea Simeri, Andrea Tagarelli, and Ester Zumpano. 2023. Transformer-based language models for mental health issues: A survey. *Pattern Recognition Letters*, 167:204–211.

Jared Jashinsky, Scott H Burton, Carl L Hanson, Josh West, Christophe Giraud-Carrier, Michael D Barnes, and Trenton Argyle. 2014. Tracking suicide risk factors through twitter in the us. *Crisis*.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2021. Mentalbert: Publicly available pretrained language models for mental healthcare. *arXiv preprint arXiv:2110.15621*.

Sophie E Jordan, Sierra E Hovet, Isaac Chun-Hai Fung, Hai Liang, King-Wa Fu, and Zion Tsz Ho Tse. 2018. Using twitter for public health surveillance from monitoring and prediction to public response. *Data*, 4(1):6.

Suchitra Kataria and Vinod Ravindran. 2020. Electronic health records: a critical appraisal of strengths and limitations. *Journal of the Royal College of Physicians of Edinburgh*, 50(3):262–268.

Nathan J Kim, Jessica Lin, Craig Hiller, Chantal Hildebrand, and Colette Auerswald. 2023. Analyzing us tweets for stigma against people experiencing homelessness. *Stigma and Health*, 8(2):187.

Ari Z Klein, Steven Meanley, Karen O'Connor, José A Bauermeister, and Graciela Gonzalez-Hernandez. 2022. Toward using twitter for prep-related interventions: An automated natural language processing pipeline for identifying gay or bisexual men in the united states. *JMIR Public Health and Surveillance*, 8(4):e32405.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Justin Littman, Daniel Chudnov, Daniel Kerchner, Christie Peterson, Yecheng Tan, Rachel Trent, Rajat Vij, and Laura Wrubel. 2018. Api-based social media collecting as a form of web archiving. *International Journal on Digital Libraries*, 19(1):21–38.

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*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Steven M Manson. 2020. Ipums national historical geographic information system: version 15.0.

Amaryllis Mavragani. 2020. Infodemiology and infoveillance: scoping review. *Journal of medical internet research*, 22(4):e16206.

Mary Meeker and Liang Wu. 2018. Internet trends 2018.

Nir Menachemi and Taleah H Collum. 2011. Benefits and drawbacks of electronic health record systems. *Risk management and healthcare policy*, pages 47–55.

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*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*.

Quynh C Nguyen, Suraj Kath, Hsien-Wen Meng, Dapeng Li, Ken R Smith, James A VanDerslice, Ming Wen, and Feifei Li. 2016a. Leveraging geotagged twitter data to examine neighborhood happiness, diet, and physical activity. *Applied Geography*, 73:77–88.

Quynh C Nguyen, Dapeng Li, Hsien-Wen Meng, Suraj Kath, Elaine Nsoesie, Feifei Li, and Ming Wen. 2016b. Building a national neighborhood dataset from geotagged twitter data for indicators of happiness, diet, and physical activity. *JMIR public health and surveillance*, 2(2):e5869.

Quynh C Nguyen, Matt McCullough, Hsien-wen Meng, Debjyoti Paul, Dapeng Li, Suraj Kath, Geoffrey Loomis, Elaine O Nsoesie, Ming Wen, Ken R Smith, et al. 2017a. Geotagged us tweets as predictors of county-level health outcomes, 2015–2016. *American journal of public health*, 107(11):1776–1782.

Quynh C Nguyen, H Meng, Dapeng Li, S Kath, M McCullough, D Paul, P Kanokvimankul, TX Nguyen, and F Li. 2017b. Social media indicators of the food environment and state health outcomes. *Public health*, 148:120–128.

OpenAI. 2021. OpenAI GPT-3.5 Documentation.

Michael Paul and Mark Dredze. 2011. You are what you tweet: Analyzing twitter for public health. In *Proceedings of the international AAAI conference on web and social media*, volume 5-1, pages 265–272.

Michael J Paul, Mark Dredze, David A Broniatowski, and Nicholas Generous. 2015. Worldwide influenza surveillance through twitter. In *AAAI workshop: WWW and public health intelligence*.

Patrick Pilipiec, Isak Samsten, and András Bota. 2023. Surveillance of communicable diseases using social media: A systematic review. *PLoS One*, 18(2):e0282101.

Ali Pourmotabbed, Sajjad Moradi, Atefeh Babaei, Abed Ghavami, Hamed Mohammadi, Cyrus Jalili, Michael E Symonds, and Maryam Miraghajani. 2020. Food insecurity and mental health: a systematic review and meta-analysis. *Public health nutrition*, 23(10):1778–1790.

Davide Proserpio, Scott Counts, and Apurv Jain. 2016. The psychology of job loss: using social media data to characterize and predict unemployment. In *Proceedings of the 8th ACM Conference on Web Science*, pages 223–232.

PyTorch. 2023. Adamw.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2019. Semeval-2017 task 4: Sentiment analysis in twitter. *arXiv preprint arXiv:1912.00741*.

Hansen Schwartz, Johannes Eichstaedt, Margaret Kern, Lukasz Dziurzynski, Richard Lucas, Megha Agrawal, Gregory Park, Shrinidhi Lakshmikanth, Sneha Jha, Martin Seligman, et al. 2013. Characterizing geographic variation in well-being using tweets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7-1, pages 583–591.

Scipy. 2023. Pearson r.

Zahra Shakeri Hossein Abad, Adrienne Kline, Madeena Sultana, Mohammad Noaeen, Elvira Nurmambetova, Filipe Lucini, Majed Al-Jefri, and Joon Lee. 2021. Digital public health surveillance: a systematic scoping review. *NPJ digital medicine*, 4(1):41.

Lone Simonsen, Julia R Gog, Don Olson, and Cécile Viboud. 2016. Infectious disease surveillance in the big data era: towards faster and locally relevant systems. *The Journal of infectious diseases*, 214(suppl_4):S380–S385.

Ed Summers, Igor Brigadir, Sam Hames, Hugo van Kemenade, Peter Binkley, tinafigueroa, Nick Ruest, Walmir, Dan Chudnov, David Thiel, Betsy, Ryan Chartier, celeste, Hause Lin, Alice, Andy Chosak, Mirko Lenz, R. Miles McCain, Ian Milligan, Andreas Segerberg, Daniyal Shahrokhian, Melanie Walsh, Leonard Lausen, Nicholas Woodward, eggplants, Ashwin Ramaswami, Boyd Nguyen, Darío Hereñú, Dmitrijs Milajevs, and Frederik Elwert. 2023. Docnow/twarc: v2.14.0.

Paola Velardi, Giovanni Stilo, Alberto E Tozzi, and Francesco Gesualdo. 2014. Twitter mining for fine-grained syndromic surveillance. *Artificial intelligence in medicine*, 61(3):153–163.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R.

Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Yufang Wang, Kuai Xu, Yun Kang, Haiyan Wang, Feng Wang, and Adrian Avram. 2020. Regional influenza prediction with sampling twitter data and pde model. *International journal of environmental research and public health*, 17(3):678.

Wikipedia. 2023a. List of regions of the united states.

Wikipedia. 2023b. Pearson correlation coefficient.

Wikipedia contributors. 2023. County statistics of the United States.

Shweta Yadav, Jainish Chauhan, Joy Prakash Sain, Krishnaprasad Thirunarayan, Amit Sheth, and Jeremiah Schumm. 2020. Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework. *arXiv preprint arXiv:2011.06149*.

Yiming Zhang, Ke Chen, Ying Weng, Zhuo Chen, Juntao Zhang, and Richard Hubbard. 2022. An intelligent early warning system of analyzing twitter data using machine learning on covid-19 surveillance in the us. *Expert Systems with Applications*, 198:116882.

A. Collection of Tweets

Twitter data collected in this study was retrieved using the Twitter Developer API. Specifically, we leveraged the twarc library (Summers et al., 2023) for querying data from Twitter. Multiple query features were used to retrieve the required data. Out of all the features, the list of keywords and location information were vital for our analysis. The lists of keywords were curated separately for mental health and food insecurity-related tweets, refer to Table 4. For the general category of tweets, we use only one space character to retrieve tweets. The location information feature was used to retrieve Tweets from a specific geographic area. We considered the census block groups as the geographical unit for the collection of Tweets. The centroid of a block group along with a radius value was used to define the block group area in the twarc data retrieval query. We varied the radius value based on the population density of the county that a block group belongs to. We primarily employ the variability in the radius value to focus on an area that is appropriately scaled according to the demographics. For example, in the locations with low population density a fixed radius may focus on a very small area and hence, we would not be able to collect any tweets. Likewise for a densely populated area if the fixed radius is set to too high a value, then we may collect an eccentrically high volume of tweets. Here, we assume that the population density (ρ) is uniform for any county, therefore,

$$\rho_{county} = \rho_{BG \in county} = \frac{population_{BG}}{\pi \cdot r^2}$$

We find the variable radius value as follows,

$$r = \sqrt{\frac{population_{BG}}{\pi \cdot \rho_{county}}}$$

We collect the BG population values from Manson (2020) and county density values from Wikipedia contributors (2023). Lastly, to avoid very large or very small values of the radius we keep an upper and a lower bound of 10 and 2 miles, respectively.

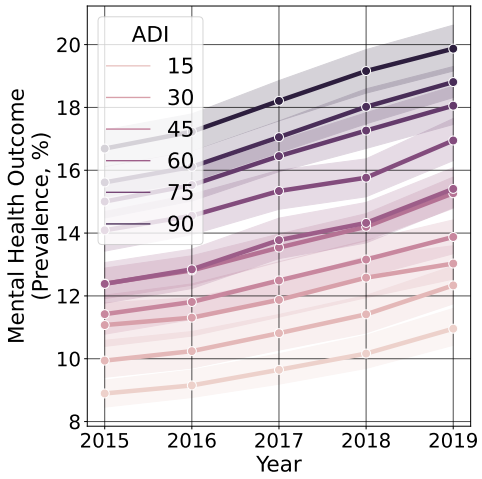

Figure 5: **Longitudinal Trend of Target Variable.** In this figure, we present the average values of MH outcomes for the years 2015 to 2019. The average value is calculated separately for each ADI value considered in the analysis. The increasing trend in the MH outcome values is observed across all BGs irrespective of their socio-economic status.

B. Data Properties

We calculate two main statistics for setting a few parameters in our analysis, the number of words per tweet and the number of tweets per block group (BG). For calculating the number of words per tweet, we simply count whitespace-separated words and use the distribution of this statistic to guide our decision of setting sequence length for encoding tweets. First, we calculate percentiles of the number of words per tweet separately for each year and tweet category, refer to Table 5 for the values.

Category	Keywords
Mental health	'bored', 'disgusting', 'sick of', 'tired of it', 'dont want to', 'so fucking miserable', 'tired of being', 'depressed', 'alone', 'isolate', 'given up', 'no friend', 'cant deal', 'want to talk', 'in my room', 'awake', 'sleepless', 'nightmares', 'insomnia', 'cant sleep', 'wish sleep', 'up all night', 'body is begging', 'exhausted', 'tired', 'my energy', 'dont have energy', 'tired to look', 'feel myself falling', 'binge', 'fasting', 'eating disorder', 'eat again', 'always eating', 'forced to eat', 'am eating ?', 'failure', 'ugly', 'worthless', 'hate myself', 'fat piece', 'self hatred', 'piece of shit', 'feel like trash', 'thoughts', 'confused', 'overthinking', 'am losing', 'losing mind', 'my mind off', 'quiet', 'attention', 'nervous', 'social anxiety', 'dead quiet', 'dont wanna move', 'cut', 'hang', 'blade', 'die', 'suicidal', 'rip skin', 'suicide attempt', 'car hit', 'kill myself', 'of the road'
Food insecurity	"food stamps", "SNAP", "food charities", "food pantry", "food voucher", "deficiency", "hunger", "hungry", "food insecurity", "poor diet", "junk food", "food desert", "poor nutrition", "starvation", "without food", "no food", "no groceries", "lack of food", "not enough food"
General	" "

Table 4: **Keywords for Collecting Tweets**. We use manually curated keywords for specific categories 'mental health', 'food insecurity', and 'general' category of tweets. For the 'general' category, we use a space character as the only keyword.

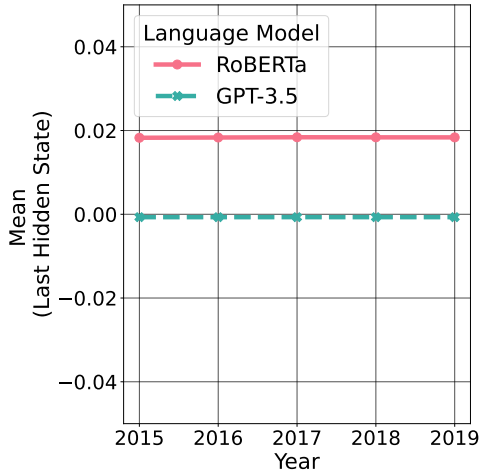

Figure 6: **Average of Activations.** In this figure we present the average value of the last hidden state for RoBERTa and GPT-3.5 model. The average values are calculated separately for each year, over all dimensions of the hidden vector and 765 block-groups. We observe fairly stable average values, primarily due to the normalization operations included in the language models.

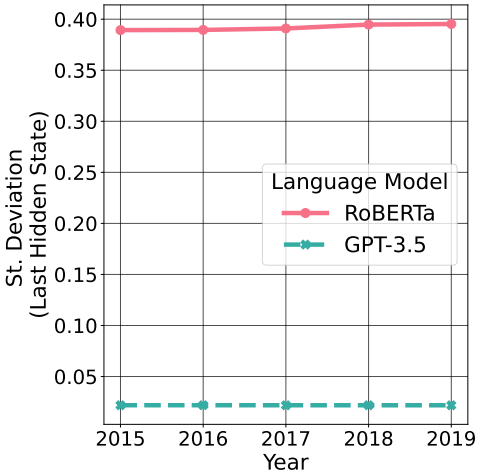

Figure 7: **Standard Deviation of Activations.** In this figure we present the standard deviation (STD) of the last hidden state for RoBERTa and GPT-3.5 model. The STD values are calculated separately for each year, the overall dimensions of the hidden vector, and 765 block-groups. We observe a significant difference between the STD values for RoBERTa and GPT-3.5. In addition, we also note a slight increase in STD for RoBERTa in the year 2018 and 2019.

We consider the maximum value of 75^{th} percentile i.e., 29 (for MH tweets in 2019), and estimate the number of ByteBPE (Liu et al., 2019) tokens per tweets, as $29 \times 1.32 = 38.28$ based on the findings presented by Deshpande et al. (2023). Finally, we set the sequence length parameter to the next power of 2 i.e. 64, for encoding tweets (encoding step in Algorithm 1). Similarly, we calculate percentile values for the number of tweets per BG and set the tweet sample size upper bound for each BG equal to 4,000, making sure we cover all 75^{th} percentile values. Lastly, we present the distribution of the MH outcome values reported by CDC in the Table 5. We defined risk categories for BGs

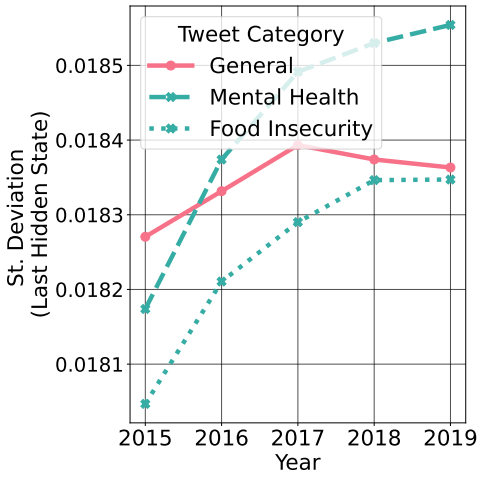

Figure 8: **Average Across Tweet Categories.** In this figure we present the average of the RoBERTa last hidden state for all tweet categories. The average values are calculated separately for each year, the overall dimensions of the hidden vector, and 765 block-groups. We observe negligible differences between tweet categories.

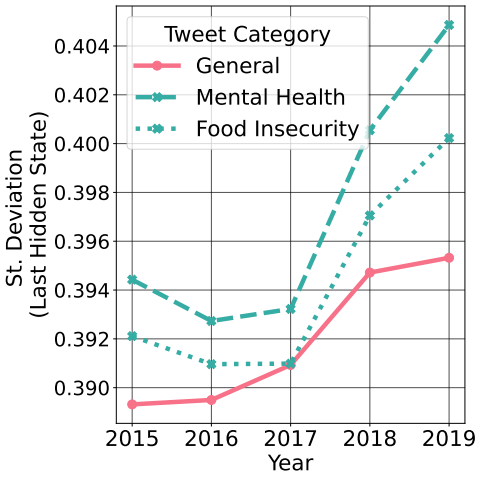

Figure 9: **Standard Deviation Across Tweet Categories.** In this figure we present the standard deviation (STD) of the RoBERTa last hidden state for all tweet categories. The STD values are calculated separately for each year, the overall dimensions of the hidden vector, and 765 block-groups. We observe a minor difference between tweet categories. Notably, the variance in the general category tweets is consistently observed as the lowest among all categories, for all years.

based on the distribution of the reported MH outcome values. Using the 75^{th} percentile value for each year we set the status of the BGs as high-risk if the reported outcome value is more than the 75^{th} percentile value.

C. Longitudinal Properties of Input and Target Variables

In this section, we present a few longitudinal properties of the input and target variables. To reiterate, the input variable is the encoding of a set of tweets from a specific language model (refer to Section 4.2). The target variable is the MH outcomes collected from the CDC database.

In Figure 5, we present the variation of the MH outcomes value in time. We calculate average values of MH outcomes across all BGs (765), separately for each year. We observe that MH outcomes have a consistent increasing trend over the years. Interestingly, this trend holds irrespective of the socio-economic status (ADI) of BGs. Hence, to effectively solve the problem of predicting MH outcomes, the model needs to recognize patterns in the tweets that eventually lead to an increasing pattern in the MH outcomes.

The input variables i.e., the encoding from language model and hence, high dimensional. Hence, to briefly understand the longitudinal patterns in the input variables, we calculate mean and standard deviation values, over all encoding dimensions and BGs, but separately for each year, refer to Figures 6 and 7. For both language models, the mean values are stable for all years. The mean values for GPT-3.5 are much closer to zero as compared to the RoBERTa model. The reason for such robust mean values is the normalization operations conducted in the language models. While mean values for both models are close to each other, there are noticeable differences in the standard deviation values. The standard deviation of the RoBERTa model is approximately 16 times higher than GPT-3.5. In addition, there also exists a slight increase in the standard deviation value in the RoBERTa model's activations for the year 2018 and 2019. Such increment in the standard deviation values is not observed in the GPT-3.5 model. We believe these statistical properties are the primary reason behind the performance differences between RoBERTa and GPT-3.5 models presented in Figures 3 and 4. The high variance of the RoBERTa encodings possibly helps the model to identify underlying hidden patterns effectively.

D. Properties of Tweet Categories

Our experiment results presented in Table 2 highlight the advantage of using general category tweets over the other mental health (MH) associated categories namely, food insecurity (FI) and MH. Here, we present the mean and standard deviation of RoBERTa encodings for all three categories. In the LocalTweets dataset, we notice only minor dif-

Tweet Category	Year	Percentiles				
		0	25	50	75	100
Number of words per tweet						
FI	2015	1	6	11	16	93
FI	2016	1	7	11	17	89
FI	2017	1	8	13	18	176
FI	2018	1	8	14	23	179
FI	2019	1	9	15	25	196
MH	2015	1	6	11	18	122
MH	2016	1	7	12	18	104
MH	2017	1	8	14	20	163
MH	2018	1	9	16	28	202
MH	2019	1	9	17	29	191
General	2015	1	7	11	16	106
General	2016	1	8	12	16	91
General	2017	1	8	12	17	206
General	2018	1	8	13	22	172
General	2019	1	8	13	23	216
Number of tweets per BG per year						
FI	2015	20	207	496	1,229	25,841
FI	2016	1	64	274	1,135	41,595
FI	2017	1	46	197	771	29,628
FI	2018	1	35	160	621	26,053
FI	2019	1	35	143	539	25,403
MH	2015	28	584	1,462	3,549	88,717
MH	2016	1	78	591	2,878	130,157
MH	2017	2	58	436	2,172	109,464
MH	2018	2	58	440	2,229	121,132
MH	2019	1	52	370	1,854	115,160
General	2015	1000	1,079	1,092	1,097	1099
General	2016	443	1,079	1,092	1,097	1099
General	2017	405	1,082	1,092	1,097	1099
General	2018	285	1,072	1,089	1,096	1099
General	2019	196	1,061	1,088	1,095	1099
MH outcome reported by CDC						
–	2015	0.0580	0.1010	0.1260	0.1540	0.2300
–	2016	0.0590	0.1040	0.1310	0.1570	0.2320
–	2017	0.0590	0.1120	0.1380	0.1660	0.2610
–	2018	0.0690	0.1160	0.1450	0.1750	0.2820
–	2019	0.0780	0.1240	0.1540	0.1820	0.2910

Table 5: **Distributional Properties of LocalTweets.** In this table, we present the distribution of tweet length, tweet volume, and MH outcome for the data included in LocalTweets. For the properties related to the Twitter data, we present the statistics separately for each category of tweets.

ferences between the mean and standard deviation for various tweet categories, refer to Figure 8, 9. Nonetheless, we observe the standard deviation value for general category tweets to be the lowest, consistently over all years. We speculate that the lower variance in the general category tweets might help focus the model on a specific semantic latent space that is potentially beneficial for the prediction of MH outcomes.

E. Effect of Skewed Regional Distribution

In the data cleaning process, we removed the BGs with no tweets or are not reported in the CDC dataset (for Disease Control et al., 2022). In this data cleaning process, the number of BGs included under the Northeast region was reduced from 250 to 94. As a result, the number of BGs from the northeast region is considerably lower compared to other geographical regions in our analysis. Hence, to measure the impact of the skewed distribution we conducted an experiment. We split the cleaned data such that we do not use BGs from the Northeast region to train or validate the model. However, we tested our model on the 2019 data for the northeast region BGs. This setup is similar to our spatial extrapolation setup, refer to Section 3.2. We create training data from approximately 75% of the remaining BGs while using the rest for validation. Otherwise, we keep our experimental setup the same as that of our main experiments, discussed in Section 5. We find that the tweets posted from regions other than the Northeast region carry significant generalization capability. With a model trained on general tweets from regions other than the Northeast region, the high-risk BGs in the Northeast region can be identified with an F1-score and accuracy of 0.7450 and 75.09%. Hence, in conclusion, a skewed distribution over geographical regions will not significantly affect the geographic generalizability of LocalTweets.

F. GPT3.5 zero-shot experiment details

We conducted a zero-shot performance assessment using GPT3.5-16k (OpenAI, 2021). Our prompt consisted of the tweets posted from a specific BG and the task was to predict the risk category of the BG. We define the risk category as high-risk for the BGs with mental health outcome values over the 75th percentile. Similar to our experiments with the LocalHealth method, we considered 4,000 tweets for predicting the category of BG. Due to a stringent constraint on the input sequence size (16K tokens), we sampled 100 tweets 40 times instead of feeding 4,000 tweets to GPT3.5-16k. In other words, for each BG, we randomly sampled 100 tweets, with replacement, 40 times and utilized zero-shot prompting, separately for each sampled set. We selected a sample size of 100 based on the median length of the tweets (13, refer to Table 5) such that, 100 tweets and prompts will fit under the limit of 16,000 tokens. Out of the 40 responses if more than 20 responses categorize the BG as high-risk then we consider the GPT3.5-16k prediction to be high-risk for the respective BG. We adopted the following structure for the prompt:

```
Tweets: {["tweet1", "tweet2", ..., "tweet100"]}
Instruction: Above tweets are posted from a block-group in the United States in the year 2019. ADI values are representative of the socio-economic profile of a respective block-group. ADI values are between one to a hundred, the highest value being the most undesirable. The ADI index for this block group is {"adi"}
Question: Based on the above tweets and ADI value, what would be the prevalence of adults (>= 18 years) with mental health not good for more than 14 days in a period of 30 days? The range of reported values is from 5% to 30%. The 25th, 50th and 75th percentile values are, 11.1%, 13.9%, and 16.9%, respectively. Select your answer from following options.

Options:
A. High-risk (prevalence greater than the 75th percentile)
B. Low-risk (prevalence less than the 75th percentile)

You must output letter A or letter B

Output:
```

Lastly, we spent 215.62 USD in total for the zero-shot experiments with GPT3.5.

G. Simplified Schematic for the LocalHealth Method

Our proposed method primarily adopts four steps, sampling, encoding, aggregation, and prediction, as mentioned in the Algorithm 1. In this section we provide a simplified schematic for the Algorithm 1, refer to Figure 10.

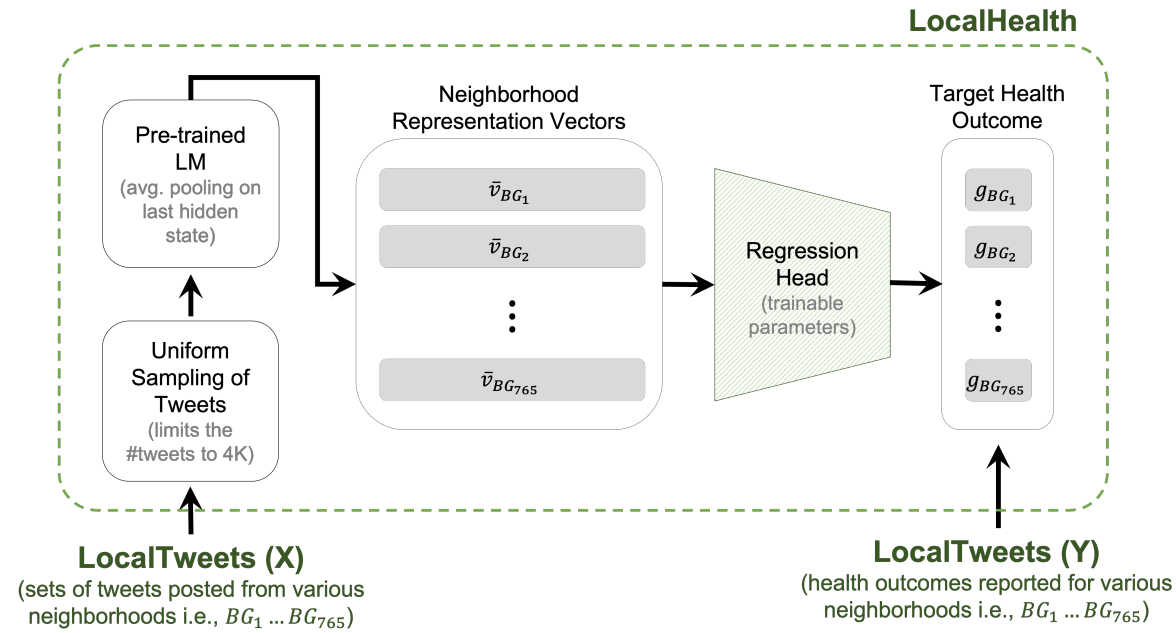

Figure 10: **LocalHealth Approach.** In this figure, we present a simplified schematic representation of the LocalHealth approach (Algorithm 1). The input to the LocalHealth method is the set of tweets contained in LocalTweets data. LocalHealth has four main steps: sampling, encoding, aggregation, and prediction. The vector $\bar{v}$ is an aggregated vector representation for individual BGs based on which the MH outcomes are predicted for respective BGs.