UCE-FID: Using Large Unlabeled, Medium Crowdsourced-Labeled, and Small Expert-Labeled Tweets for Foodborne Illness Detection

Ruofan Hu
Data Science Program
Worcester Polytechnic Institute
Worcester, USA
rhu@wpi.edu

Dongyu Zhang
Data Science Program
Worcester Polytechnic Institute
Worcester, USA
dzhang5@wpi.edu

Dandan Tao
Vanke School of Public Health
Tsinghua University
Beijing, China
dandantao@mail.tsinghua.edu.cn

Huayi Zhang*
ByteDance
San Jose, USA
huayi.zhang@bytedance.com

Hao Feng
College of Ag&Environ Sciences
North Carolina A&T State University
Greensboro, USA
haofeng@illinois.edu

Elke Rundensteiner
Computer Science/Data Science Program
Worcester Polytechnic Institute
Worcester, USA
rundenst@wpi.edu

***Abstract*—Foodborne illnesses significantly impact public health. Deep learning surveillance applications using social media data aim to detect early warning signals. However, labeling foodborne illness-related tweets for model training requires extensive human resources, making it challenging to collect a sufficient number of high-quality labels for tweets within a limited budget. The severe class imbalance resulting from the scarcity of foodborne illness-related tweets among the vast volume of social media further exacerbates the problem. Classifiers trained on a class-imbalanced dataset are biased towards the majority class, making accurate detection difficult. To overcome these challenges, we propose EGAL, a deep learning framework for foodborne illness detection that uses small expert-labeled tweets augmented by crowdsourced-labeled and massive unlabeled data. Specifically, by leveraging tweets labeled by experts as a reward set, EGAL learns to assign a weight of zero to incorrectly labeled tweets to mitigate their negative influence. Other tweets receive proportionate weights to counter-balance the unbalanced class distribution. Extensive experiments on real-world *TWEET-FID* data show that EGAL outperforms strong baseline models across different settings, including varying expert-labeled set sizes and class imbalance ratios. A case study on a multistate outbreak of Salmonella Typhimurium infection linked to packaged salad greens demonstrates how the trained model captures relevant tweets offering valuable outbreak insights. EGAL, funded by the U.S. Department of Agriculture (USDA), has the potential to be deployed for real-time analysis of tweet streaming, contributing to foodborne illness outbreak surveillance efforts.**

***Index Terms*—foodborne illness, social media, semi-supervised learning, learning with noisy labels, text classification**

I. Introduction

Motivation. Foodborne illnesses pose a significant public health threat, affecting millions of Americans annually. These illnesses result in productivity loss, high medical expenses, and even fatalities [1], [2]. Early foodborne illness detection is crucial for risk reduction, outbreak control, and public health safeguarding. Consumer-generated data from social media to internet search, a valuable resource for surveillance, has led to the creation of surveillance tools based on conventional supervised machine learning. These tools have been tested by local health agencies – by using Twitter data in New York City [3], Chicago [4], and Las Vegas [5], Yelp reviews in San Francisco [6] and New York City [7], as well as Google search queries in Las Vegas and Chicago [8].

Classification models were commonly employed to detect foodborne illness incidents within social media data, including tweets [5], [9] in the aforementioned surveillance systems. Subsequently, inspectors carried out case inspections based on these cases flagged as potential incidents by the system. Sound machine learning models that enhance precision in detecting foodborne illness incidents can potentially reduce the human resource demands involved in the case inspection process. However, supervised models require high-quality labeled training data, which are extremely resource-intensive and often prohibitively so to collect. Crowdsourcing has been explored as a less resource-intensive approach to gather more labels [10]. However, ensuring label quality with anonymous labelers tends to be challenging [11]. Models trained on data with low-quality labels may overfit to label noises and struggle to generalize. Furthermore, budget constraints often hinder collecting an adequate number of labels even via crowdsourcing, leaving substantial unlabeled data unused when relying exclusively on supervised learning.

Problem Definition. In this study, our focus is thus to train an effective foodborne illness detection model using tweet data

*This work was performed while Huayi Zhang was at Worcester Polytechnic Institute.

979-8-3503-2445-7/23/$31.00 ©2023 IEEE

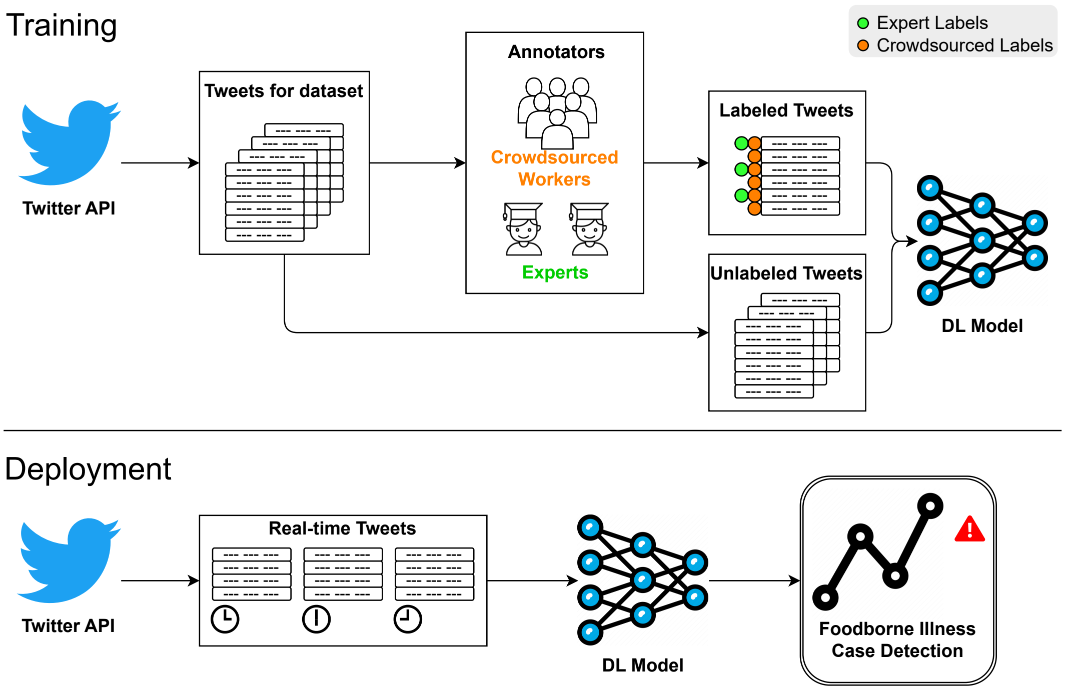

Fig. 1. Training and deployment procedure of EGAL. [1]

with low-quality labels curated under resource constraints such as a limited budget and limited support from experts. As shown in Figure 1, we collect a large volume of tweets using the Twitter API as the foundation of our dataset. However, due to limited resources, the majority of these tweets remain unlabeled, while only a small portion is labeled by crowdsourced workers, and an even smaller portion has been labeled by experts. Our objective is to utilize this tweet dataset to train a foodborne illness detection model. The trained model is designed for possible integration into a surveillance system capable of detecting foodborne illness cases in streaming tweets.

Challenges. Despite the availability of a large volume of tweets that can be collected using keyword search via the Twitter API, constructing a high-quality labeled training set to develop a reliable model remains a challenge. High-quality training data should possess two crucial properties: (1) a balanced class distribution and (2) a sufficient number of accurate labels. Due to the scarcity of relevant tweets, the training data, even when carefully curated with domain-driven semantic rules, still consists mainly of irrelevant tweets, resulting in an imbalanced training set. In such cases, classification models tend to label all data as belonging to the majority class, which is contrary to our core objective here to identify the items in the minority class of foodborne illness-relevant tweets. Additionally, limited resources restrict the size and quality of the training set. The training set consists of either a small expert-labeled dataset only [12] or a relatively larger one collected via crowdsourcing [13], with the latter potentially augmented with a small number of samples whose labels have been verified by experts [14]. With abundant tweet data available for access yet remaining unused in our context, we risk limiting the effectiveness of training machine learning models for food safety detection.

While some studies have been conducted to develop machine models for detecting food poisoning incidents, they have primarily relied on high-quality datasets labeled by experts or have employed crowd-workers to adaptively label unlabeled tweets through active learning [13], [15]. Some approaches have been developed to build models utilizing both unlabeled data and imbalanced labeled data, *i.e.*, imbalanced semi-supervised learning [16]. However, they typically assume a sufficient amount of accurate labels for the initial labeled data, an assumption difficult to guarantee in real-world applications. *The real-world scenario of a sufficient number of having a median number of unreliable labels and a small number of accurate labels in the context of imbalanced semi-supervised learning* presents an even more challenging problem, which is the focus of this work.

Proposed Method. To address the aforementioned challenges, we design a novel framework called EGAL (Expert Labels Guided Approach Learning with Crowdsourced and Unlabeled Tweets). EGAL harnesses a vast amount of unlabeled tweets by assigning pseudo-labels. Simultaneously, it employs a small number of tweets with expert labels as a reward set to rebalance the class distribution and filter out falsely labeled instances. Note that this reward set does not necessarily have a balanced distribution across classes. An instance reweighting strategy is thus employed to address label bias caused by incorrect labels and class imbalance. This reweighting process

[1] Twitter has been renamed as X. We note that the data collection and framework design were carried out when the Twitter API was accessible for academic research.

is guided by the performance of the reward set utilizing robust criteria for imbalanced data [17], [18]. Our main contributions are:

- We propose EGAL, a practical solution for training a classifier to detect foodborne illness. It uses crowdsourced and massive unlabeled data, guided by a small amount of expert-labeled tweets, even if they are not class-balanced. This approach effectively reduces the impact of noisy labels and rebalances the class distribution in the scenario of semi-supervised learning.
- We extensively evaluate EGAL and strong state-of-the-art methods on the real-world dataset *Tweet-FID* [14]. The results demonstrate EGAL's superior performance compared to strong baselines, even when varying both expert-labeled set sizes and imbalance ratios.
- We perform a case study on a multistate outbreak of *Salmonella* Typhimurium infection associated with packaged salad greens. Our method identifies informative tweets that offer insights into the outbreak trend, showcasing the effectiveness of our model in foodborne illness surveillance.

Our work is part of the USDA-funded ***FACT*** project[2], which aims to develop innovative big data analytics technologies for ensuring the safety of fresh produce. This research explores the use of social media analysis for early food safety warnings. By deploying EGAL in real-time tweet analysis, we contribute to the development of a comprehensive foodborne illness outbreak surveillance system, enabling early detection and timely response to outbreaks for enhanced public health.

II. Related Work

The problem is the intersection of class imbalance, semi-supervised learning, and learning with noisy labels. In this section, we briefly review the related literature on machine learning methods used for foodborne illness detection, semi-supervised (SSL), and learning with noisy label learning (LNL). And also include literature on class imbalance scenario in the context of SSL and LNL, respectively.

A. Machine Learning Methods in Foodborne Illness Detection

Previous studies primarily used supervised classifiers (e.g., SVM, Naive Bayes, Decision Trees) with text content features (unigrams, bigrams) to identify relevant posts. However, these methods require parameter optimization and can be sensitive to chosen parameters [10], [19]. Alternatively, supervised classification models based on pre-trained language models have proven effective in classifying foodborne illness tweets [9], [14]. Sadilek et al. [13] adopt a multi-step strategy to build a high-quality model from crowds. It first collects a small set of crowdsourced labels and trains an initial model. Subsequently, to address the class imbalance, it employs active learning, using crowd workers to adaptively label unlabeled tweets under the assumption of accurate crowdsourced labels.

TABLE I
Summary of notation

Notation	Description
x	Feature vector of instance
y	Label of instance
N^c	Number of instances in crowdsourced set
M	Number of instances in unlabeled set
N^e	Number of instances in expert-labeled set
$D^c = \{(x_i^c, y_i^c)\}_{i=1}^N$	Crowdsourced set
$D^u = \{x_i^u\}_{i=1}^M$	Unlabeled set
$D^e = \{(x_j^e, y_j^e)\}_{j=1}^S$	Expert-labeled set
γ	Imbalance ratio
θ	Model parameters
$l()$	Supervised training loss function
$l^u()$	Unsupervised training loss function
$L^e(',')$	Loss function on expert-labeled set
w_i	Weight for the i-th training instance

B. Semi-supervised Learning (SSL)

Semi-supervised learning [20]–[23] is a well-studied field that significantly reduces the requirements on laborious annotations by leveraging abundant unlabeled data. Among existing methods, pseudo-labeling [20], [21], [24], in particular, the methods utilizing self-supervised learning [22], [23], [25], [26] have achieved great advances. The main idea is to assign pseudo labels to unlabeled data with the model's predictions and add them to the labeled data. Despite the great success, these methods commonly assume that the labeled and/or unlabeled data are class-balanced and also the labels of the labeled data are accurate.

Typical SSL methods usually fail to generalize well on the minority classes under class imbalance. Imbalanced semi-supervised learning has drawn more attention, and those methods can be divided into two categories. One category aims to acquire high-quality pseudo labels and rebalance the class distribution [27]–[30]. The other one is to learn a balanced classifier. Lee et al. [31] proposed an auxiliary balanced classifier that addresses class imbalance by introducing an additional regularization term. CoSSL [32] adopts a co-learning framework to decouple the representation learning and balanced classifier learning and share the learned representation and generated pseudo labels.

C. Learning with Noisy Labels (LNL)

Existing methods for learning with noisy labels require either a large-scale dataset [33], [34] or a meticulously curated, class-balanced validation set for training guidance [35]–[37]. Limited resources make it impractical to collect numerous labels from crowds or to build a clean and balanced validation set using expert knowledge. CSWL [38] tackles label noise and class imbalance through reweighting based on the AUC criteria. However, it remains susceptible to noisy information, especially when the labeled data is scarce.

[2] Project link: https://www.nal.usda.gov/research-tools/food-safety-research-projects/fact-innovative-big-data-analytics-technology-microbiological-risk-mitigation-assuring-fresh-produce

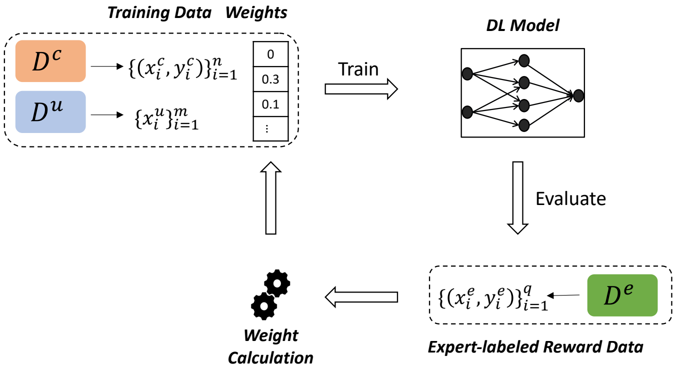

Fig. 2. Overview of EGAL.

III. PROPOSED FRAMEWORK: EGAL

A. Problem Definition

Let D be a Tweets Dataset collected via the Twitter API using foodborne illness-related keywords. $D^c = \{(x_i^c, y_i^c)\}_{i=1}^{N^c}$ represent the crowdsourced set, where x_i^c is a tweet selected from D based on semantic rules, $y_i^c \in \{0, 1\}$ is its aggregated crowdsourced label (1 indicates a foodborne illness relevant tweet, while 0 indicates an irrelevant tweet), and $N^c =\| D^c \|$. $D^u = \{x_i^u\}_{i=1}^M$ denotes the unlabeled set, where $M =\| D^u \|$. And $D^e = \{(x_j^e, y_j^e)\}_{j=1}^{N^e}$ is the small expert-labeled set randomly chosen from $\{x_i^c\}$, with accurate labels $\{y_i^e\}$ provided by domain experts, where $N^e =\| D^e \|$. We denote the relevant and irrelevant tweets with the subscripts $_+$ and $_-$, respectively.

In real-world scenarios, the number of irrelevant tweets is usually much higher than the number of relevant tweets, resulting in imbalance ratios: $\gamma_c := \frac{N_-^c}{N_+^c} \gg 1$, $\gamma_u := \frac{M_-}{M_+} \gg 1$, and $\gamma_e := \frac{N_-^e}{N_+^e} \gg 1$. Here, $\gamma_u \succ \gamma_c \simeq \gamma_e$ with γ_u and γ_c unknown because the goal is to select as many relevant tweets as possible during crowdsourcing, while the huge unlabeled data set tends to be less curated due to its size.

Our objective, given $\{D^c, D^u, D^e\}$, is to learn a model $\Phi(\theta)$ capable of accurately classifying tweets as being foodborne illness relevant or irrelevant.

B. Overview

EGAL leverages the small expert-labeled set as a reward set to guide training. The key idea is to assign weights to losses of the crowdsourced tweets and unlabeled tweets according to their influence on the model's performance on the expert-labeled reward set. The underlying hypothesis is that a model trained with accurate labels and balanced class distribution will reduce the loss of the reward set. We describe the process of EGAL depicted in Figure 2. Initially, the model is trained with a regular semi-supervised learning procedure, and each labeled tweet and unlabeled tweet is weighted equally, respectively. Then it learns the weight for the loss of each training sample by solving the meta-optimization problem that minimizes the loss of the expert-labeled reward set. Based on the learned weights, EGAL first filters out the samples deemed to be false-labeled. Finally, the model is updated with the weighted losses of the correctly labeled samples. To fully make use of these false-labeled tweets, we now also add them to the unlabeled set with the aim of having new pseudo-labels generated. EGAL iterates through the weight learning and model update steps, respectively, until the performance of the reward set no longer improves or reaches the iteration limit.

C. Objective Functions

1) Training Loss: In semi-supervised learning, the training loss defined as Eq. (1) is composed of supervised loss l_i and unsupervised loss l_j^u, where l_i denotes the per-sample supervised loss, *e.g.*, cross-entropy loss, while l_j^u denotes the per-sample unsupervised loss, $w_i = \frac{1}{n}$ and $w_j = \frac{1}{m}$, $\beta \in R_{\succ 0}$ denotes the trade-off. Given the weights $\boldsymbol{w} = [w_1, ..w_n, w_{n+1}, ...w_{n+m}]$ as hyperparameters, the objective function is defined in Eq. (2). We note that the optimal weights can be learned based on the performance of the reward set.

$$L(\theta; \boldsymbol{w}) = \sum_{i=1}^{n} w_i l_i(\theta) + \beta \sum_{j=1}^{m} w_j l_j^u(\theta). \quad (1)$$

$$\theta^*(\boldsymbol{w}) = \arg\min_{\theta} \sum_{i=1}^{n} w_i l_i(\theta) + \beta \sum_{j=1}^{m} w_j l_j^u(\theta). \quad (2)$$

2) Loss of Reward Set: Usually, accuracy is used as the evaluation metric for classification tasks. However, in a class imbalance scenario, accuracy could be misleading. AUC score is a more informative measure for an imbalanced dataset. However, it has been shown that the algorithm maximizes the accuracy of a model but does not necessarily maximize the AUC score [18]. Here, we adopt both accuracy and AUC to evaluate the performance of the reward set. Usually, maximizing the accuracy is to minimize the cross-entropy loss. The non-parametric estimator of AUC is non-convex and discontinuous, as defined in Eq. (3), where n_+ and n_- are the numbers of relevant tweets and irrelevant tweets in a mini-batch.

$$AUC(\Phi; D) = \frac{1}{n_+ n_-} \sum_{x_i \in D_+} \sum_{x_j \in D_-} \mathbb{I}(\Phi(x_i) \succ \Phi(x_j)). \quad (3)$$

To maximize the AUC score with stochastic gradient descent, a convex and differentiable surrogate loss function $f(\Phi(x^-) - \Phi(x^+))$ replaces the indicator function in Eq. (3). Here, f is the pairwise squared loss $f(s) = (1 + s)^2$. EGAL defines the loss of the reward set as:

$$\begin{aligned} L^e(\theta) &= \lambda \sum_{i=1}^{q} l_i(\theta, x_{i=1}^e, y_i^e) \\ &+ (1 - \lambda) \frac{1}{q_- q_+} \sum_{i=1}^{q_-} \sum_{j=1}^{q_+} f(\Phi(x_i^-, \theta) - \Phi(x_j^+, \theta)). \end{aligned} \quad (4)$$

EGAL aims to find the optimal weights that minimize the loss of the reward set.

$$\boldsymbol{w}^* = \arg \min_{\boldsymbol{w}, w_i \in [0,1]} L^e(\theta^*, \boldsymbol{w}). \quad (5)$$

D. Reweighting and Parameters Updating

1) Bi-level Optimization: The EGAL employs a bi-level optimization strategy, wherein one optimization objective is encapsulated within another objective. In this particular scenario, the outer objective is to minimize $L^e(\theta)$, which represents the loss of the reward set. The insight is that the performance of the reward set can serve as an indicator of the quality of the trained model. The inner objective is to minimize $L(\theta)$, the loss of the training set. The bi-level optimization problem can be formulated as follows:

$$\begin{aligned} &\min_{w, w\in[0,1]} L^e(\theta^*, \boldsymbol{w}) \\ &\text{s.t.} \quad \theta^* = \arg\min_{\theta} L(\theta, \boldsymbol{w}). \end{aligned} \tag{6}$$

2) Parameters and Weights Updating: In EGAL, it adopts the widely-used online updating strategy from the meta-learning literature [35], [36], [39], [40] to update w and θ. To employ SGD to optimize Eq.(1), in each training iteration, a batch of labeled samples $\{(x_i^c, y_i^c)\}_{i=1}^n$ and unlabeled samples $\{x_i^u\}_{i=1}^m$ are sampled. Then consider approximating θ^* with one gradient descent step updated value via a first-order Taylor expansion of the loss function. The updating equation of θ is formulated as:

$$\hat{\theta}^t(\boldsymbol{w}) = \theta^t - \alpha(\sum_{i=1}^{n} w_i \nabla_\theta l_i(\theta) \mid_{\theta^t} + \beta \sum_{j=1}^{m} w_j \nabla_\theta l_j^u(\theta) \mid_{\theta^t}). \tag{7}$$

where α is the descent step size. Subsequently, the formulated model parameters $\hat{\theta}^t$ are utilized to get the optimal selection of weights w at step t with the objective function Eq. (5). Here, similarly, we exploit a first-order Taylor approximation of Eq. (5) at $\boldsymbol{w} = \boldsymbol{0}$:

$$\hat{\boldsymbol{w}}^{t+1} = -\gamma \nabla_{\boldsymbol{w}} L^e(\hat{\theta}^t(\boldsymbol{w})) \mid_{w^t=0} . \tag{8}$$

It has been proven in [39] that

$$\nabla_{w_i} L^e(\hat{\theta}^t(\boldsymbol{w})) \propto -\nabla_\theta L^e(\theta^t)^T \cdot \nabla_\theta L_i(\theta^t)$$

, where the latter term is the inner product of the gradient of the loss of the reward set and the training loss of training sample x_i. Thus $\hat{\boldsymbol{w}}^{t+1} \propto \nabla_\theta L^e(\theta^t)^T \cdot \nabla_\theta L_i(\theta^t)$, A positive inner product means the labeled training sample (x_i, y_i) can also optimize the loss of reward set, and it should have a positive and large weight. Otherwise, it would degrade the performance of the reward set, and the model should not learn from it. Based on this, we rectify $\hat{w}^{t+1}$ as non-negative weights to make the model ignore the tweets with incorrect labels. Then we can normalize the weights to realize the objective of maximizing the performance of the reward set by taking into account both accuracy and AUC.

$$\tilde{w}_i^{t+1} = \max(\hat{w}_i^{t+1}, 0). \tag{9}$$

$$w_i^{t+1} = \frac{\tilde{w}_i^{t+1}}{\sum_{i=1}^{n+m} \tilde{w}_i^{t+1} + \sigma}. \tag{10}$$

where σ=1 if $\sum_{i=1}^{n+m} \tilde{w}_i^{t+1} = 0$, otherwise equals to zero. Then, the updated weights $\boldsymbol{w}^{t+1}$ are utilized to optimize the model parameters θ with Eq. (11). The overall bi-level optimization procedure is summarized in Algorithm 1.

$$\theta^{t+1} = \theta^t - \alpha(\sum_{i=1}^{n} w_i^{t+1} \nabla_\theta l_i(\theta) \mid_{\theta^t} + \beta \sum_{j=1}^{m} w_j^{t+1} \nabla_\theta l_j^u(\theta) \mid_{\theta^t}). \tag{11}$$

Algorithm 1: Bi-level Optimization Procedure of EGAL

```
Data: $D^c, D^u, D^e, n, m, q, T, \beta$
Result: Model parameters $\theta^T$
Initialize model parameters $\theta^0$;
for $t \leftarrow 0$ to $T-1$ do
    $\{(x_i^c, y_i^c)\}_{i=1}^n \leftarrow$ BatchSampler($D^c, n$);
    $\{x_i^u\}_{i=1}^m \leftarrow$ BatchSampler($D^u, m$);
    $\{(x_i^e, y_i^e)\}_{i=1}^q \leftarrow$ BatchSampler($D^e, q$);
    $\{\hat{y}^u\} \leftarrow \Phi(x^u, \theta^t)$;
    $L(\theta; \boldsymbol{w}) \leftarrow \sum_{i=1}^n w_i l_i(\theta) + \beta \sum_{j=1}^m w_j l_j^u(\theta)$;
    $\boldsymbol{w} \leftarrow \boldsymbol{0}$; Compute $\hat{\theta}^t(w)$ by Eq. (7);
    Update $\boldsymbol{w}^{t+1}$ by Eq. (8)-(10) ;
    Update $\theta^{t+1}$ by Eq. (11) ;
end
```

IV. Experiments

A. Experiment Settings

1) Text Relevance Classification (TRC) Task: This task, introduced by [14], aims to identify tweets that refer to a foodborne illness incident. Each tweet in Tweet-FID has a binary label indicating its relevance to foodborne illness. This task can help detect potential outbreaks of foodborne diseases from social media posts.

2) Dataset and Metrics: We utilized the publicly available Tweet-FID dataset from [14], comprising 1,362 (33%) relevant and 2,760 (67%) irrelevant tweets for the TRC task. Each tweet received both an expert label and an aggregated crowdsourced label. The dataset was divided into a train-validation-test set based on the expert labels. The training set consists of 1,088 relevant tweets and 2,210 irrelevant tweets. The validation set includes 137 relevant tweets and 275 irrelevant tweets. The test set consists of 137 relevant tweets and 275 irrelevant tweets. The imbalance ratios of the training set, validation set, and test set are $\gamma_c = \gamma_v = \gamma_t \approx 2$. Aggregated crowdsourced labels identified 1,625 tweets as relevant and 1,673 tweets as irrelevant. The noise ratio, defined as the difference between aggregated crowdsourced labels and expert labels divided by the total number of labels in the training dataset, is 20.29%.

In January 2023, we collected tweets from 2016 to the end of 2022, using the same domain-specific keywords as in [14], resulting in a total of approximately 600,000 tweets. We filtered out tweets with fewer than five tokens and those already present in the Tweet-FID dataset. Then, we sampled 50,000 tweets as the unlabeled set used for our experiment. Without associated labels, the exact imbalance ratio γ_u of the unlabeled set is unknown. But, unlike the carefully curated training set, which is relatively balanced, the unlabeled set is naturally imbalanced, and $\gamma_u \gg 1$.

Evaluation Metrics. We use standard accuracy (Accuracy), F1, and *balanced accuracy* (bACC) [41] to measure each method's performance. bACC works well with imbalanced datasets when the standard accuracy leads to misleading results. These three metrics can illustrate a method's performance from various perspectives, given an imbalanced class distribution.

3) Baseline Methods: For a fair comparison, we adopt a pre-trained RoBERTa [42] as the backbone model for all compared methods. RoBERTa is an improved version of BERT [43] that removes the next-sentence prediction task and uses larger learning rates and mini-batches for pre-training. RoBERTa has achieved state-of-the-art performance on multiple tasks [42], including the TRC task [14]. We compare EGAL with following methods:

Fully supervised learning. (Sup. Learning) Sup. Learning trains a text classification model on the given labeled data without any special treatment for addressing label noise. This method also does not use any unlabeled data.

SoftMatch. Softmatch, proposed by [30], is a state-of-the-art semi-supervised learning method for balanced and imbalanced classification, assuming the labeled data and unlabeled data have the same class distribution. It balances the quality and quantity of pseudo-labels by using a truncated Gaussian function based on sample confidence. Softmatch also encourages diverse pseudo-labels using a uniform alignment approach. It has achieved significant improvements, particularly in tasks with imbalanced class distributions.

CWSL. This is a weakly-supervised learning method proposed by [38]. It is designed to cope with severe label noise by assigning small weights to noisy instances. The instance reweight process is under the guidance of performance on a clean reward dataset. CWSL adopts robust AUC criteria as performance measurement on the reward set to conquer the issue that the label distributions in the training and testing data are different.

COSINE. This self-training approach [44] fine-tunes a pre-trained language model using both weakly-labeled and unlabeled data. COSINE first fine-tunes the pre-trained model with the weakly-labeled data and then generates pseudo-labels for the unlabeled data. COSINE applies contrastive regularization and confidence-based sample reweighting to enhance model performance and mitigate error propagation during the self-training procedure.

4) Methodology: For our tweet-fid dataset, we treat expert labels as ground-truth labels and consider crowdsourced labels as noisy labels. The expert-labeled reward set D^e is randomly derived from the training set D^c with the expert labels ratio λ, *i.e.*, $\| D^e \| = \lambda \| D^c \|$ and $\gamma_e \simeq \gamma_c$. In our experiments, since the test set in Tweet-FID is small, we combine the validation and test sets to create a new validation set D^v, which is used exclusively for evaluation purposes. For those experimental settings involving expert labels, we create ten different D^e, which means selecting different tweets to assign expert labels.

In our experimental study, we utilize the Adam optimizer [45] with a learning rate of 1×10^{-5}, β_1 of 0.9, β_2 of 0.999, weight decay of 1×10^{-4}, and layer decay of 0.75 to train each neural network model. For EGAL, we set the β value in Eq. (1) to 1. Each model is trained for a total of 10,000 steps. Throughout the training process, we utilize the separate expert-labeled validation dataset D^v to evaluate the model's performance every 512 steps. We report the average of the model's best performance of five random seeds under each scenario. All experiments are conducted on a server with an A100-80G GPU. All code is developed with Python 3.9 on PyTorch 1.12.0.

B. Effect of Expert-labeled Data

Many weakly-supervised learning only leverage weak supervision from heuristic rules or crowdsourcing to train a model. Here, we want to investigate how much improvement can be made by utilizing a small expert-labeled set in the training process.

1) Setup: We compare two training set choices for each method using either (1) using all crowdsourced labels in D^c as the training labels or (2) merging λ expert labels into D^c. The specific data settings for each method are shown in Table II. Here, we run experiments on the two training label choices with the fixed value of expert labels ratio λ as 10% and measure their test performance.

2) Results: Table III provides a comparison of different methods' performance when the training data includes both expert labels and crowdsourced labels. It is evident that SoftMatch, a semi-supervised method that incorporates unlabeled data, outperforms the supervised learning method that relies solely on labeled data. However, COSINE, another method designed for learning with weakly-labeled and unlabeled data, does not demonstrate any advantage over the supervised method, likely due to its limitations in handling imbalanced classification problems and lack of guidance from expert-labeled data. Methods utilizing a small reward set with expert

TABLE II
LABELED TRAINING SET (D^c), UNLABELED TRAINING SET (D^u), AND EXPERT-LABELED REWARD SET (D^e) SETTINGS FOR EACH METHOD. ✓INDICATES USAGE, ✗INDICATES NON-USAGE.

Method	Expert labels ratio in D^c	D^u	D^e
Sup. Learning	0%	✗	✗
	λ	✗	✗
SoftMatch	0%	✓	✗
	λ	✓	✗
COSINE	0%	✓	✗
	λ	✓	✗
CWSL	0%	✗	✓
	λ	✗	✓
EGAL	0%	✓	✓
	λ	✓	✓

TABLE III

Performance comparison of all methods when expert labels are not in the labeled training dataset. *: the method utilizes the unlabeled training set. $^{\#}$: the method utilizes expert-labeled reward set.

Method	F1	Accuracy	bACC
Sup. Learning	0.796 ± 0.006	0.838 ± 0.008	0.866 ± 0.003
Softmatch*	0.798 ± 0.008	0.841 ± 0.006	0.866 ± 0.007
COSINE*	0.778 ± 0.009	0.826 ± 0.009	0.849 ± 0.007
CWSL$^{\#}$	0.804 ± 0.005	0.847 ± 0.003	0.873 ± 0.005
EGAL*$^{\#}$	**0.817 ± 0.006**	**0.863 ± 0.004**	**0.879 ± 0.005**

labels show superior performance compared to the supervised learning method. Notably, our method EGAL achieves the most promising results by effectively leveraging both unlabeled data and the reward set.

TABLE IV

Performance comparison of all methods when expert labels are not in the labeled training dataset.

Method	F1	Accuracy	bACC
Sup. Learning	0.789 ± 0.007	0.828 ± 0.009	0.862 ± 0.005
Softmatch*	0.797 ± 0.004	0.837 ± 0.003	0.862 ± 0.006
COSINE*	0.767 ± 0.011	0.811 ± 0.014	0.843 ± 0.008
CWSL$^{\#}$	0.798 ± 0.005	0.839 ± 0.007	0.870 ± 0.003
EGAL*$^{\#}$	**0.813 ± 0.009**	**0.856 ± 0.009**	**0.879 ± 0.006**

In Table IV, we present the results obtained when the training data lacks expert labels. It is important to note that all methods perform worse compared to their counterparts in Table III, highlighting the negative influence of label noise. However, even in this scenario, SoftMatch outperforms the supervised method. Additionally, methods that leverage the small expert-labeled reward set demonstrate superior performance compared to the supervised learning method. Once again, our method EGAL maintains its position as the top-performing approach among all methods. These results demonstrate the effectiveness of utilizing an expert-labeled set in the training.

C. Effect of Expert-labeled Data Ratio

In this experiment, we investigate the how the number of expert labels affects performance. The ideal method should be label-efficient, improving the performance a lot with a small number of expert labels.

1) Setup: Specifically, we vary the expert labels ratio λ from 10% to 50% and create five different expert-labeled sets with each imbalance ratio value. We repeat the experiments of 5 random seeds and report the average performance on each dataset.

2) Results: As illustrated in Figure 3, the performance of all methods improves with an increase in the number of expert labels. EGAL consistently outperforms other methods, regardless of the expert label ratio λ. Notably, even with only 10% expert labels, EGAL outperforms other methods using 50% expert labels, highlighting its label-efficiency. The weakly supervised learning method COSINE achieves comparable accuracy with 50% expert labels. However, its F1 score and balanced accuracy (bACC) are consistently lower than other methods, indicating limitations in handling scenarios with varying class distributions between weakly-labeled samples and unlabeled data.

D. Effect of Imbalance Ratio

The imbalance ratio γ of the Tweet-FID dataset is 2, which means the dataset is rather balanced. However, in a more realistic scenario, the class labels could be extremely imbalanced, *i.e.*, the number of food poisoning relevant tweets is much smaller than the number of irrelevant ones. In this experiment, we investigate the robustness of all methods under different imbalanced ratios.

1) Setup: Except for the experiments with the original dataset ($\gamma = 2$), datasets with other imbalance ratios were created by reducing the number of relevant tweets according to the function $N^+ = N^-/\gamma$. It's important to note that the imbalanced datasets are created based on the true class labels. We conducted experiments across different imbalance ratios ($\gamma \in \{2, 5, 10, 50\}$) while maintaining a fixed expert labels ratio of 10%.

TABLE V

Performance under different imbalance ratios.

Imb.ratio	Method	F1	Accuracy	bACC
2	Sup. Learning	0.796 ± 0.006	0.838 ± 0.008	0.866 ± 0.003
	Softmatch	0.798 ± 0.008	0.841 ± 0.006	0.866 ± 0.007
	COSINE	0.778 ± 0.009	0.826 ± 0.009	0.849 ± 0.007
	CWSL	0.804 ± 0.005	0.847 ± 0.003	0.873 ± 0.005
	EGAL	**0.817 ± 0.006**	**0.863 ± 0.004**	**0.879 ± 0.005**
5	Sup. Learning	0.779 ± 0.003	0.840 ± 0.005	0.854 ± 0.002
	Softmatch	0.791 ± 0.004	0.837 ± 0.004	0.861 ± 0.004
	COSINE	0.765 ± 0.002	0.826 ± 0.007	0.836 ± 0.004
	CWSL	0.795 ± 0.011	0.839 ± 0.012	0.865 ± 0.007
	EGAL	**0.808 ± 0.008**	**0.861 ± 0.006**	**0.867 ± 0.007**
10	Sup. Learning	0.788 ± 0.012	0.838 ± 0.008	0.856 ± 0.009
	Softmatch	0.778 ± 0.012	0.825 ± 0.007	0.849 ± 0.013
	COSINE	0.763 ± 0.010	0.823 ± 0.006	0.834 ± 0.009
	CWSL	0.790 ± 0.018	0.838 ± 0.017	0.859 ± 0.015
	EGAL	**0.803 ± 0.006**	**0.852 ± 0.007**	**0.866 ± 0.005**
50	Sup. Learning	0.769 ± 0.006	0.821 ± 0.007	0.841 ± 0.003
	Softmatch	0.767 ± 0.012	0.819 ± 0.010	0.835 ± 0.006
	COSINE	0.746 ± 0.010	0.814 ± 0.003	0.823 ± 0.008
	CWSL	0.783 ± 0.015	0.834 ± 0.013	0.852 ± 0.011
	EGAL	**0.792 ± 0.006**	**0.844 ± 0.006**	**0.859 ± 0.006**

2) Results: Table V provides a comparative view of how different methods perform across various imbalance ratios. Sup.Learning, Softmatch, and COSINE exhibit progressively lower performance as the imbalance ratio increases, particularly in terms of balanced accuracy and F1, indicating their limitations in handling imbalanced datasets. CSWL is primarily designed to mitigate the effects of class imbalance; however, this advantage is not evident in relatively balanced datasets. EGAL's performance remains stable and competitive across different levels of class imbalance. This consistency highlights

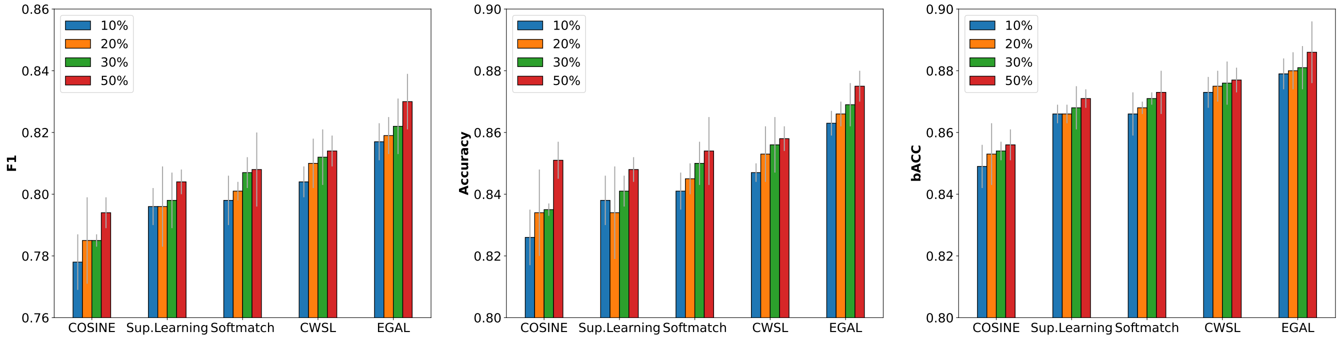

Fig. 3. Performance of methods across different expert label ratios.

its robustness and makes it a reliable choice for handling datasets with unreliable labels and unlabeled data in real-world scenarios.

V. Case Study

A. Predictions Analysis

We conducted an inspection of both correct and incorrect predictions made by the model trained on the dataset with an imbalance ratio of $\gamma = 5$ to gain a better understanding of its behavior. Table VI presents examples of correct and incorrect predictions. The first eight rows showcase tweets with correct predictions. While these tweets contain keywords like 'food poisoning' and 'stomach,' the model successfully captures the complicated semantic relationships within the tweets, resulting in accurate predictions. However, it occasionally struggles with more challenging tweets, leading to some incorrect predictions. Most of these incorrect predictions fall into the category of false positives. We found the model has difficulty in understanding some figurative expressions. For instance, the tweet in row 10 uses the term "economic food poisoning", which is a metaphorical use of "food poisoning". This terminology misleads the model to incorrectly classify the tweet as a foodborne illness incident. For the tweets in rows 10 and 11, the users express uncertainty about their situations, making it harder for the model to make the decisions. The last tweet is indeed about food poisoning but does not describe a personal experience. The model struggles to distinguish between personal experiences and other relevant content effectively.

B. Preliminary Comparison

The final goal is to identify foodborne illness cases and try to detect the early signal of the outbreak. Here, we take one outbreak as an example and conduct a preliminary analysis. In 2021, the CDC and FDA investigated a Salmonella Typhimurium outbreak [46], linking it to BrightFarms brand packaged salad greens in four states. Officially, 31 cases were reported, but the actual number is likely higher due to underreporting and mild cases not seeking medical care [46]. Using our trained model with EGAL, we identified tweets potentially related to this outbreak. Following the approach in [14], we collected geotagged tweets from 2021, filtering out non-U.S. locations. Our trained model predicted relevant foodborne illness tweets mentioning keywords like "salad" and "greens".

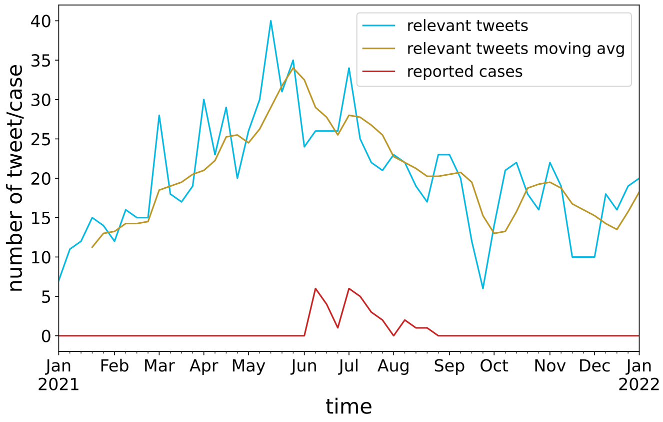

Fig. 4. Trend plot of weekly reported cases of *Salmonella* outbreak from prepackaged salads (red curve), tweets mentioning foodborne illness with the words 'salad' or 'greens' (light blue curve), and the 4-week moving average of tweets (yellow curve).

Figure 4 presents the weekly count of relevant tweets throughout 2021 (represented by the light blue curve) and its corresponding moving average (displayed as the yellow curve, with a sampling width of 4). The figure also includes the weekly number of reported cases recorded by the CDC [2] (indicated by the red curve). Notably, during the period from May to August 2021, there was a noticeable increase in the number of tweets referencing foodborne illness and containing the specified keywords. The reported cases of illness commenced between June 10, 2021, and August 18, 2021. However, it is important to acknowledge that the actual outbreak might have started prior to the first reported case and could have persisted beyond the detection of the last reported case. The tweets captured by EGAL offer insights into the trend of this

[2] Data source: https://www.cdc.gov/salmonella/typhimurium-07-21/epi.html

TABLE VI
EXAMPLES OF MODEL PREDICTIONS. ✓: CORRECT PREDICTION. ✗: INCORRECT PREDICTION

	Tweet	Prediction
1	@USER awwww thank you for caring but I know for a fact that it's not food poisoning or the flu :) I know how both those feel.	0 ✓
2	This is a stressful enough weekend as it is and then Sunday comes and it's #LIVMUN. No game knots my stomach like it	0 ✓
3	@USER @USER What a decade to be alive! Great designing decisions lead this game to the top! Like for a example make food poisoning worse and more common because of the filth on the floor. How players should overcome this without doormats or cleaning robots you ask? Don't build any floors!	0 ✓
4	i just drank so much water and now my stomach hurts coz the only way i know to drink water is to chug it all	0 ✓
5	I got food poisoning off an Italian dessert. I've a good mind to tiramisu that company.	1 ✓
6	I've spent the last few days with probably the worst case of food poisoning I've had in my life. I think that's the last time I eat food my housemates cook.	1 ✓
7	@USER I like cheese too but I have food poisoning today, I do not like food poisoning [EMOJI_nauseated_face]	1 ✓
8	I feel absolutely terrible. First I get food poisoning, and I guess I was catching a cold?? I've been sneezing and congested all day. Someone send cold medicine and tea	1 ✓
9	Robinhood gave economic food poisoning to its user base today, people generally don't come back after that, reviews and ratings aside. Next functioning market day for them will likely see redemption Seppuku. #robinhoodapp	1 ✗
10	That sandwich I made that I just ate is going to give me food poisoning I think :	1 ✗
11	I think I just had a bad experience with Great Steak[EMOJI_beaming_face_with_smiling_eyes].Or maybe I just ate too fast, I'll let y'all know in 23 hrs if I got food poison [EMOJI_skull]	1 ✗
12	@USER A little disappointed that you cropped out the riveting news about Panda Express. #foodpoisoning	1 ✗

outbreak. This highlights the potential of our method, EGAL, in detecting early signals of possible foodborne illness outbreaks.

VI. Discussion

In our work with social media data, we believe there are no glaring ethical consequences related to applying AI-based techniques for food safety surveillance. Even though tweets we accessed were are public data, we opted to obfuscate any mentions of users and URL links to `@USER` and `HTTPURL`, respectively, to reduce reference to specific tweeter users.

We note that, unfortunately, X (formerly known as Twitter) suspended academic research access to its API in March 2023. As a result, we are no longer able to maintain our developed surveillance system, at least as related to Twitter as a data source [47]. Nevertheless, we set out in this research to design a new model to uncover patterns in foodborne illness outbreaks by analyzing the historical social media data previously collected. This makes the assumption that as other social media platforms increasingly replace X, we will be able to redeploy our tool to these alternate sources with people's social interactions online continuing to be available for surveillance.

Additionally, we can leverage EGAL to develop new models using alternative sources of information, benefiting the public by providing early warnings about foodborne illness outbreaks—potentially saving lives and livelihoods. In general, this work contributes to the rapidly accelerating field of detecting disease spread through social media data.

VII. Conclusion

In this study, we introduce EGAL, a practical solution for detecting foodborne illnesses by leveraging a combination of crowdsourced-labeled, large labeled, and small expert-labeled tweet data sets. EGAL incorporates a reward set of expert-labeled tweets to assign weights to the training set, aiming to achieve a more balanced class distribution. Incorrectly labeled tweets are assigned zero weights to mitigate their negative influence, while correctly labeled tweets receive appropriate weights. This approach effectively improves the performance of the detection process. Through extensive experiments, we demonstrate the superior performance of EGAL compared to strong state-of-the-art models across various scenarios, including different sizes of the expert-labeled set and class imbalance ratios.

We also conduct a case study focusing on a multistate outbreak of *Salmonella* Typhimurium infection associated with packaged salad greens. Our method successfully captures relevant tweets that provide valuable insights into the outbreak trend.

Acknowledgment

This work was supported by the Agriculture and Food Research Initiative (AFRI) award no. 2020-67021-32459 from the U.S. Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA) and the Illinois Agricultural Experiment Station.

References

[1] S. Hoffmann and E. Scallan Walter, "Acute complications and sequelae from foodborne infections: informing priorities for cost of foodborne illness estimates," *Foodborne pathogens and disease*, vol. 17, no. 3, pp. 172–177, 2020.

[2] R. L. Scharff, "The economic burden of foodborne illness in the united states," in *Food safety economics*. Springer, 2018, pp. 123–142.

[3] C. Harrison, M. Jorder, H. Stern, F. Stavinsky, V. Reddy, H. Hanson, H. Waechter, L. Lowe, L. Gravano, and S. Balter, "Using online reviews by restaurant patrons to identify unreported cases of foodborne illness—new york city, 2012–2013," *MMWR. Morbidity and mortality weekly report*, vol. 63, no. 20, p. 441, 2014.

[4] J. K. Harris, R. Mansour, B. Choucair, J. Olson, C. Nissen, J. Bhatt, C. for Disease Control, and Prevention, "Health department use of social media to identify foodborne illness - chicago, illinois, 2013-2014." *MMWR. Morbidity and mortality weekly report*, vol. 63, 2014.

[5] A. Sadilek, H. Kautz, L. DiPrete, B. Labus, E. Portman, J. Teitel, and V. Silenzio, "Deploying nemesis: Preventing foodborne illness by data mining social media," in *Twenty-Eighth IAAI Conference*, 2016.

[6] J. P. Schomberg, O. L. Haimson, G. R. Hayes, and H. Anton-Culver, "Supplementing public health inspection via social media," *PloS one*, vol. 11, no. 3, p. e0152117, 2016.

[7] T. Effland, A. Lawson, S. Balter *et al.*, "Discovering foodborne illness in online restaurant reviews," *Journal of the American Medical Informatics Association*, vol. 25, no. 12, pp. 1586–1592, 2018.

[8] A. Sadilek, S. Caty, L. DiPrete *et al.*, "Machine-learned epidemiology: real-time detection of foodborne illness at scale," *NPJ digital medicine*, vol. 1, no. 1, p. 36, 2018.

[9] D. Tao, D. Zhang, R. Hu, E. Rundensteiner, and H. Feng, "Crowdsourcing and machine learning approaches for extracting entities indicating potential foodborne outbreaks from social media," *Scientific reports*, vol. 11, no. 1, p. 21678, 2021.

[10] R. A. Oldroyd, M. A. Morris, and M. Birkin, "Identifying methods for monitoring foodborne illness: Review of existing public health surveillance techniques," *JMIR Public Health and Surveillance*, vol. 4, 6 2018.

[11] J. Zhang, V. S. Sheng, T. Li, and X. Wu, "Improving crowdsourced label quality using noise correction," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 5, pp. 1675–1688, 2017.

[12] M. A. H. Khan, M. Iwai, and K. Sezaki, "A robust and scalable framework for detecting self-reported illness from twitter," in *2012 IEEE 14th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE, 2012, pp. 303–308.

[13] A. Sadilek, S. Brennan, H. Kautz, and V. Silenzio, "nemesis: Which restaurants should you avoid today?" in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 1, 2013, pp. 138–146.

[14] R. Hu, D. Zhang, D. Tao, T. Hartvigsen, H. Feng, and E. Rundensteiner, "Tweet-fid: An annotated dataset for multiple foodborne illness detection tasks," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 6212–6222.

[15] X. Deng, S. Cao, and A. L. Horn, "Emerging applications of machine learning in food safety," *Annual Review of Food Science and Technology*, vol. 12, pp. 513–538, 2021.

[16] Q. Gui, H. Zhou, N. Guo, and B. Niu, "A survey of class-imbalanced semi-supervised learning," *Machine Learning*, pp. 1–30, 2023.

[17] Z. Yuan, Y. Yan, M. Sonka, and T. Yang, "Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3040–3049.

[18] T. Yang and Y. Ying, "Auc maximization in the era of big data and ai: A survey," *ACM Computing Surveys*, vol. 55, no. 8, pp. 1–37, 2022.

[19] D. Tao, D. Zhang, R. Hu, E. Rundensteiner, and H. Feng, "Epidemiological data mining for assisting with foodborne outbreak investigation," *Foods*, vol. 12, no. 20, p. 3825, 2023.

[20] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2, 2013, p. 896.

[21] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in neural information processing systems*, vol. 33, pp. 596–608, 2020.

[22] B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura, and T. Shinozaki, "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 408–18 419, 2021.

[23] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," *Advances in neural information processing systems*, vol. 33, pp. 6256–6268, 2020.

[24] H. Pham, Z. Dai, Q. Xie, and Q. V. Le, "Meta pseudo labels," in *Proceedings of the IEEE/CVF CVPR*, 2021, pp. 11 557–11 568.

[25] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, "S4l: Self-supervised semi-supervised learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1476–1485.

[26] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[27] J. Kim, Y. Hur, S. Park, E. Yang, S. J. Hwang, and J. Shin, "Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 14 567–14 579, 2020.

[28] C. Wei, K. Sohn, C. Mellina, A. Yuille, and F. Yang, "Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning," in *Proceedings of the IEEE/CVF CVPR*, 2021, pp. 10 857–10 866.

[29] Y. Oh, D.-J. Kim, and I. S. Kweon, "Daso: Distribution-aware semantics-oriented pseudo-label for imbalanced semi-supervised learning," in *Proceedings of the IEEE/CVF CVPR*, 2022, pp. 9786–9796.

[30] H. Chen, R. Tao, Y. Fan, Y. Wang, J. Wang, B. Schiele, X. Xie, B. Raj, and M. Savvides, "Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning," *ICLR*, 2023.

[31] H. Lee, S. Shin, and H. Kim, "Abc: Auxiliary balanced classifier for class-imbalanced semi-supervised learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 7082–7094, 2021.

[32] Y. Fan, D. Dai, A. Kukleva, and B. Schiele, "Cossl: Co-learning of representation and classifier for imbalanced semi-supervised learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 14 574–14 584.

[33] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[34] J. Li, R. Socher, and S. C. Hoi, "Dividemix: Learning with noisy labels as semi-supervised learning," *arXiv preprint arXiv:2002.07394*, 2020.

[35] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," in *International conference on machine learning*. PMLR, 2018, pp. 4334–4343.

[36] J. Shu, Q. Xie, L. Yi, Q. Zhao, S. Zhou, Z. Xu, and D. Meng, "Meta-weight-net: Learning an explicit mapping for sample weighting," *Advances in neural information processing systems*, vol. 32, 2019.

[37] H. Zhang, L. Cao, P. VanNostrand, S. Madden, and E. A. Rundensteiner, "Elite: Robust deep anomaly detection with meta gradient," in *Proceedings of the 27th ACM SIGKDD*, 2021, pp. 2174–2182.

[38] L.-Z. Guo, F. Kuang, Z.-X. Liu, Y.-F. Li, N. Ma, and X.-H. Qie, "Iwe-net: Instance weight network for locating negative comments and its application to improve traffic user experience," *Proceedings of the AAAI*, vol. 34, no. 04, pp. 4052–4059, Apr. 2020.

[39] S. Jenni and P. Favaro, "Deep bilevel learning," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 618–633.

[40] H. Zhang, L. Cao, S. Madden, and E. Rundensteiner, "Lancet: labeling complex data at scale," *Proceedings of the VLDB Endowment*, vol. 14, no. 11, 2021.

[41] R. Wang, X. Jia, Q. Wang, and D. Meng, "Learning to adapt classifier for imbalanced semi-supervised learning," *arXiv preprint arXiv:2207.13856*, 2022.

[42] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[44] Y. Yu, S. Zuo, H. Jiang, W. Ren, T. Zhao, and C. Zhang, "Fine-tuning pre-trained language model with weak supervision: A contrastive-regularized self-training approach," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 1063–1077.

[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[46] C. for Disease Control and Prevention, "Investigation details," Oct 2021. [Online]. Available: https://www.cdc.gov/salmonella/typhimurium-07-21/details.html

[47] D. Tao, R. Hu, D. Zhang, J. Laber, A. Lapsley, T. Kwan, L. Rathke, E. Rundensteiner, and H. Feng, "A novel foodborne illness detection and web application tool based on social media," *Foods*, vol. 12, no. 14, p. 2769, 2023.