TWEET-FD: A Dataset for Multiple Foodborne Illness Incident Detection Tasks

Ruofan Hu*, Dongyu Zhang*, Dandan Tao+, Hao Feng+, and Elke Rundensteiner*
(*): Data Science Program, Worcester Polytechnic Institute, Worcester, MA
(+): Dept. of Food Science & Human Nutrition, Univ. of Illinois, Urbana-Champaign, IL
{rhu,dzhang5,rundenst}@wpi.edu, {dtao2,haofeng}@illinois.edu

Abstract

Foodborne illnesses are a serious but preventable public health problem – with delays in detecting these outbreaks resulting in productivity loss, expensive recalls, public safety hazards, and even loss of life. While social media are a promising source for identifying unreported foodborne illnesses early, there is a dearth of relevant labeled datasets available to the community for developing effective detection models. To accelerate the development of machine learning-based models for foodborne illness detection, we thus present TWEET-FD (TWEET-Foodborne illness Detection), the first publicly available dataset for multiple foodborne illness incident detection tasks. TWEET-FD collected from Twitter is annotated with three facets: (i) Tweet Class: tweet classification as foodborne illness incident, (ii) Entity Type: entities of interest for foodborne illness detection in the tweet, and (iii) Entity Relevance: relevancy of entities to the foodborne illness incident. We then introduce several domain tasks leveraging these three facets: text relevance classification (TRC), entity mention detection (EMD), and entity relevance classification (ERC). Additionally, we derive the relevant entity detection (RED) task by combining entity type and entity relevance facets. We describe the end-to-end methodology for dataset design, creation, and labeling for supporting model development for these tasks. We provide baseline results for these tasks leveraging state-of-the-art NLP-based deep learning methods. This dataset opens opportunities for promising future research in foodborne illness incident detection and resolution.

1 Introduction

Foodborne illnesses continue to threaten public health. Approximately 1 in 6 Americans (or 48 million people) are sickened by foodborne illness each year [1]. Foodborne illnesses lead to productivity loss, medical expenses, and even loss of life. The annual economic cost caused by foodborne illnesses in the United States is estimated to be between $14 to $60 billion [2][3].

Early detection of foodborne illness provides a means to reduce the risk and curtail the outbreak. Over the past decade, it has been recognized that user-generated public posts on social media or review platforms can play a significant role in tracking foodborne illness cases for surveillance. Surveillance applications were piloted by local health agencies, including mining data from Twitter in New York City [4], Las Vegas [5], and consumer review sites such as Yelp in San Francisco [6] and New York City [7].

However, these applications tend to focus only on a coarse-grained inspection, i.e., identifying food poisoning tweets or complaint reviews with machine learning models, followed by a labor-intensive manual-phase conducted by inspectors in person using a fine-grained protocol. This procedure, while potentially effective, is a slow and rather costly. Instead, a surveillance system should automate as
many tasks as possible. For instance, as shown in Figure 1 an example tweet: “I had a cole slaw at the KFC. After I came back to the hotel, I started vomit.” This tweet indicates a possible foodborne illness incident, where the entity cole slaw might be the source of contaminant, KFC the place where the person had the contaminated food, and vomit the symptom of food poisoning. While the hotel is also a location entity, it is not related to this incident. Information about these entities related to the incident can help staff (regulators, government officials to companies) trace the cause and source of foodborne illnesses. A successful method thus should not only determine if a given tweet indicates a possible foodborne illness incident but also automatically extract critical entities from the tweet so that they can be aggregated into trends and acted upon.

To accelerate this inspection process, we propose four tasks to be automated by such a system: (1) identify if the tweet indicates the existence of a foodborne illness incident; (2) find and extract entities in the tweet; (3) determine the relevance of entities to the foodborne illness incident in the tweet; and (4) isolate and only extract entities related directly to the foodborne illness incident. The four tasks correspond to text relevance classification (TRC), entity mention detection (EMD), entity relevance classification (ERC), and relevant entity mention detection (RED). While TRC helps the user to quickly identify possible foodborne illness incidents, EMD extracts key information from tweets needed for taking actions and/or aggregating findings into overall trends, whereas ERC/REM filters out non-relevant information.

In this work, we introduce a carefully curated dataset TWEET-FD (TWEET-Foodborne illness Detection) to support these tasks by covering multi-grained information about foodborne illness incidents. As depicted in Figure 1 our dataset provides tweets with multiple task labels that can be leveraged for training models for the EMD, ERC, RED, and TRC tasks proposed above. To select the keywords used to develop TWEET-FD, food safety experts in our team conducted preliminary studies and retrieved English posts from Twitter based on a batch of foodborne illness keyword candidates established in [7]. After analyzing the retrieved tweets, they carefully selected a set of keywords from those candidates to ensure we can collect a sufficient number of relevant tweets.

Using an iterative design methodology, we use crowdsourcing to generate annotations for each tweet that cover the above three facets. For this, we establish a protocol and carefully designed interface for crowd-sourced based labeling. Annotators are asked to complete three subtasks: rate the tweet using a Likert scale, tag all entity types in the tweet, and subsequently decide the entity relevance of each entity to the foodborne illness incident (if any). We set the restriction that at least one entity must be a relevant entity if and only if the tweet indicates a possible foodborne illness incident.

Labeling the tweet with three facets has several advantages. First, for each tweet, it provides labels for all our proposed tasks (TRC, EMD, ERC/REM). Therefore, both single- or multi-task models can be trained on our TWEET-FD dataset. In other words, TWEET-FD can serve as a valuable resource for training foodborne illness detection and information extraction model(s). Second, the three labeling tasks enable us to detect and prevent low-quality crowdsource workers as further discussed in our methods section. Establishing an effective quality control mechanism, we collect annotations with high inter-annotator agreement. Then we conduct aggregations to create gold standard labels for tweet class and entity type independently. Subsequently, we construct entity relevance label based on the created entity type and tweet class.
Thereafter, we set out to conduct experimental studies with deep learning models to realize each of our proposed tasks TRC, EMD, ERC, and RED. For this, we utilize the curated TWEET-FD data set to train and then test several state-of-the-art single- as well as multi-task deep learning models such as RoBERTa [8], BERTweet [9], MGADE [10], and IMGJM [11]. We observe that multi-task methods can match or exceed the performance of single-task methods. We hypothesize that multi-task methods can benefit from interrelationships among the tasks. Our experimental findings illustrate that indeed effective models for these critical tasks can be found, while also opening the opportunity for future research studies for which our work can then serve as baseline performance.

Our contributions are as follows:

- We developed TWEET-FD, the first publicly available social media dataset with multiple task labels representing a valuable resource for model development for food safety.
- TWEET-FD covers three facets: (i) tweet class of foodborne illness, (ii) entity types in the tweet, and (iii) relevance of entities to the incident. We create gold standard labels for these three facets based on crowd-sourced annotations with high inter-annotator agreement.
- We characterize four domain tasks, and then experiment with single-task and multi-task deep learning-based models for solving these tasks on the TWEET-FD dataset. Our findings provide a much needed benchmark to research in model development for food safety.

## 2 Dataset creation

It is challenging to collect multi-grained annotations for a specific domain. First, the tasks should be illustrated clearly to collect high-quality labels, especially for annotating entity relevance. Furthermore, it is even more difficult to merge the three interconnected labels from multiple annotators into an integrated gold standard. In this section, we describe the data collection process, provide the annotation strategies, and present the data aggregation procedure.

### 2.1 Data collection

**Data sampling.** We collect streaming data with the keyword search mechanism through the use of Twitter API with keyword filtering since January 2019. The search keywords are based on food safety literature best practices [7], and include common terms and hashtags that are intuitively indicative of foodborne illness. The search keywords used to create this dataset are 

- 
- 
- 
- 
- 
- 
- 

Based on our preliminary study, very ambiguous words (e.g., "sick", "fever") are not in the keyword list to exclude tweets about other diseases. Here, as suggested by experts in the food safety domain, we only keep tweets containing carefully selected keywords since these tweets are more likely to be related to foodborne illness incidents. Given that the fact that Twitter API only supports collecting information via keyword search, and a large number of tweets are generated every moment across numerous topics and a foodborne illness detection system is unlikely to process every tweet. It’s more efficient to narrow down the search scope by keywords and select tweets with keywords as candidate instances for the dataset at the first phase.

**Data preprocessing.** For data cleaning, we then filtered out the retweeted tweets and the ones with less than 3 tokens. As for preprocessing, we normalize and anonymize the tweets by converting user mentions and url links to @USER, HTTPURL, respectively. Since the emojis may carry important information, we keep the emojis and translate the icons into text strings by using the emoji module.

### 2.2 Annotation by crowdsourcing

We use the Amazon Mechanical Turk (MTurk) crowdsourcing platform to elicit multiple annotations per each tweet with a custom interface. Participants are asked to complete three tasks: 1) rate the tweet on how much they agree with the sentence that the tweet indicates a possible foodborne illness, 2) rate the tweet class of foodborne illness, and 3) rate the relevance of entities to the incident. For each task, we ask workers to complete a practice task before they are able to complete the data collection tasks.

We note that for space reasons the full interface design can be found in the Appendix A.2. We kept this part of the design during our data collection phase. It’s fine to either building a regression model or classification model after converting the rating to binary label.
illness incident on a Likert scale of 0-5 (0 means "not at all" and 5 means "very sure"), 2) Highlight all words/phrases belonging to specific labels (food, location, symptoms, and foodborne illness keywords) to indicate entity types, and 3) Decide for each highlighted word/phrase if or if not it is related to the foodborne illness incident.

2.2.1 Pilot study

In-house study for the design of annotation instructions and the interface. We first conduct an in-house study to assure the usability and clarity of our labeling interface and instructions for annotation elicitation. In our preliminary user interface, we provide verbal instructions that explain the project objective, display the screenshot of one completed annotation to help the annotators understand the task, and show the definition of an entity that is deemed to be "relevant" to foodborne illness incidents. We asked 12 graduate students to participate in this in-house study. Participants were shown the interface design. They were asked to complete the three annotation tasks mentioned in the second to last paragraph for several tweets. Then we asked them for constructive feedback and suggestions on the interface design. Many of them said that our tasks appear fairly complex to them. Moreover, that they were not familiar with the expected annotations nor the process of providing annotations for the first time. Thus, in addition, we embedded a two-minute video into the interface with a detailed explanation of each task as well as showcasing each of the manipulations available to the annotators.

First pilot study for preliminary investigation of the quality of the annotations. Next, we conduct a pilot study to investigate workers’ performance on MTurk. Although MTurk provides a mechanism to only allow workers with a high approval rate to take a task, this screening does not reliably filter out workers who may provide low quality annotations for our complicated task. Thus, we set out to develop an effective quality control mechanism while also calibrating an equitable payment strategy. For this, we published an initial batch of 200 tweets. We set the reward as $0.1 per HIT based on an estimation of a reasonable working time of roughly 50 seconds per overall task for one tweet. Each tweet was assigned to five unique annotators, resulting in a total of 1000 annotations collected from 110 annotators. Intuitively, the task of tagging entities requires more effort from annotators compared to rating tweet relevance. Furthermore, the entity relevance task depends upon it. Thus, we focused on analyzing the entity annotation task to filter out low-quality work that may negatively affect the quality of our to-be-derived gold standard labels. We measured the average macro-F1 score over all pairs of crowd annotations. We only keep the ones with the average F1 score exceeding a specified threshold (0.7). The inter-annotator agreement Krippendorff’s $\alpha$ was raised from 0.35 to 0.61 by dropping the low quality annotations.

Findings on working time and compensation. The average working time in the in-house study was 15s. Thus we deemed 50s on average as a reasonable time for the workers and set the corresponding compensation to $0.1 per HIT for the first pilot. The results of this pilot study showed an average working time of 47s for good-quality jobs - confirming our compensation strategy.

Findings on annotation quality and control mechanism. In the first pilot study, two of the 110 annotators took more than 80 HITs on MTurk, but over 50% of their work was identified as low quality. We thus need to address this risk that a few workers that perform low-quality work may do a great deal of damage to the quality of our generated labels. To assure the quality of the collected annotation set and lower the annotator rejection rate, our strategy thus is to split our dataset into small batches and control the HITs available for a worker in each batch - allowing us to confirm their agreement with other workers iteratively. The improvement of the inter-annotator agreement confirms the effectiveness of our quality control mechanism for identifying and preventing low-quality work.

---

4HIT refers to a single, self-contained, virtual task that a worker can work on, submit an answer, and collect a reward for completing. In our case, the three questions (see the screenshot of interface in Appendix A.1) shown in the same interface is a HIT.

5This corresponds to the US federal minimum wage ($7.25/hour).

6We note that the detailed algorithm is shown in the Appendix A.2

https://github.com/LightTag/simpleedorff
2.2.2 Main study

Based on the findings from the above pilot study, we then were ready to conduct the main study as following. Five different workers annotated each tweet. Then, when some were rejected due to low-quality work, these tasks were reassigned to other workers. Since the rejection may ultimately affect a worker’s reputation, we divided the whole dataset into six batches and workers were only allowed to take at most 10 tasks in each batch. In this way, we can both alleviate the negative impact of our rejection on the annotators and concurrently reduce the potential bias in the collected annotations. We publish a new batch only after each tweet in the previous batch has a large enough number of good annotations for the final gold standard creation. By filtering out the low-quality annotations, the entity-level inter-annotator agreement was raised from 0.55 to 0.73.

2.3 Gold standard creation

Since the entity relevance label is connected with the entity label and the Likert label of the tweet, we first create gold standard labels for the entity and the overall tweet in parallel. Subsequently, given the entity label and the sentence label, we can create the entity relevance label.

**Tweet class label.** The class label (relevant to foodborne illness or not) of a tweet is decided based on majority voting. Since we collected the sentence score on a scale of 0-5, we need to convert the rating score to binary label i.e., we transformed the rating score to the class label 1 if it is higher than 2, and class 0 otherwise. Then we applied majority voting on the binary labels (ties are broken as the relevant tweet).

**Entity label.** The entity label is determined by employing a Bayesian sequence combination (BSC) method along with a probabilistic model of annotator noise and bias, referred to as BSC-seq in [13]. BSC-seq has been shown to perform better on crowdsourced sequence labeling aggregation than majority voting (MV) and HMM-Crowd [14]. In this way, we create gold standard entity labels.

**Entity relevance label.** If the tweet is irrelevant to a foodborne illness incident, we ignore the relevance indication of all entities included in the tweet. Otherwise, we did majority voting to determine the entity relevance. First, we drop the annotations which have inconsistent boundary and entity labels with the gold standard entity label. Then, we only take into account the relevance labels in the left annotations.

**Assuring quality labels.** To inspect the quality of the derived gold standard labels, we randomly picked 500 instances and generate expert labels by food safety researchers in our team. We then measure the agreement between gold standard labels and expert labels for these tweets. For the tweet class label, they agree on 437 (85%) tweets. For entity labels, we measure the agreement with the \( \text{cu}_\alpha \) method introduced in [13]. \( \text{cu}_\alpha \) designed for continuum data fits our problem, e.g., an entity is a continuum with multiple words. Words not identified as entities are valued as \( \phi \). \( \text{cu}_\alpha \) (with \( k \) denoting the entity type) enables us to look into the reliability associated with each of the four entity types. We find that the agreement \( \text{cu}_\alpha \) of all entity types between the expert label and gold standard label is very high, namely, 0.915. While the agreements of food, location, symptom, and other entities are also high, namely, 0.89, 0.85, 0.91, and 0.94, respectively.

2.4 Dataset description and dataset split

**Dataset description.** There are 3173 tweets in TWEET-FD, including 2152 (67.8%) foodborne illness relevant tweets and 1021 (32.2%) irrelevant tweets. Since we retrieve tweets with words relevant to foodborne illness, more relevant tweets than irrelevant ones are included. Only 189 tweets doesn’t any entity inside and 96 tweets include all these four types of entities. The number of other entity and symptom are larger than the amount of food or location entities - which also can be attributed to our choice in retrieval keywords. Detailed statistics of each entity category are given in Table1. Each relevant tweet contains 2.3 relevant entities and 0.17 irrelevant entities on average. For irrelevant tweets, the average number of relevant and irrelevant entities is 0.0 and 1.53, respectively.
Given the goal of foodborne illness characterization, we care more about the relevant entities inside the relevant tweets. With an entity frequency analysis, we find the top 3 foodborne illness relevant food categories to be chicken, sushi, and sandwich – all three popular food categories. One exciting finding is that peach, which was recalled in August 2020 in many states in the US, appears in our dataset. The tweet is “I’ve stayed in to avoid getting Corona for months only to get food poisoning from a peach?!” This match of a recalled food captured in our data set indicates the potential of leveraging social media data for foodborne illness detection. The most frequent relevant locations are McDonald’s, Burger King, and Taco Bell. Although there are some place names like Illinois or NYC, the extraction of more fine-grained locations by leveraging geolocating tools will be needed. The top 3 entities of symptom are diarrhea, throwing up, and vomiting. There are a total of 23 appearances of relevant emojis e.g., face-vomiting, nauseated face, etc.

**Dataset split.** We do a train-validation-test split for TWEET-FD. The training set consists of 1721 relevant tweets and 817 irrelevant tweets. The validation set includes 215 relevant tweets and 102 irrelevant tweets. The test set consists of 216 relevant tweets and 102 irrelevant tweets. The split ratio is close to 8:1:1. This split is stratified by the tweet-level relevance class. The ratios of relevant to irrelevant tweets in these three splits are the same.

### 3 Experiments

#### 3.1 Evaluation tasks

Based on the labels collected, we design four tasks and evaluate state-of-the-art deep learning methods on these four tasks on our data set. The resulting models can be applied for detecting tweets relevant to foodborne illness incidents and extracting relevant information from these tweets.

**Text relevance classification (TRC).** This task is to predict if a given tweet does indicate a possible foodborne illness incident. As described in Section 2.3 in our dataset, each tweet comes with a binary class label, which denotes if the tweet is relevant to foodborne illness or not. These binary class labels are the gold standard labels for TRC tasks. Models trained on this task can be applied to detect possible foodborne illness incidents mentioned in online posts.

**Entity mention detection (EMD).** In our data set, we highlight all words or phrases belonging to specific entity classes (food, location, symptom, and foodborne illness keywords). The EMD task aims to identify all the mentions of the four types of entities listed above. EMD is different from the classical Named Entity Recognition (NER) task, in that the latter only focuses on the named entities. Here, EMD requires extracting both named and unnamed entities. Actually, in many tweets, people may mention many entities without providing their actual names. Therefore, the EMD task may help us find more information related to foodborne illness incidents compared to the NER task.

**Entity relevance classification (ERC).** This task is to determine the relevance of a given entity to the foodborne illness incident. Let us consider an example: *I had a cole slaw at the KFC. After I*

### Table 1: Statistics of entity distribution over relevant tweets and irrelevant tweets. Relevant and Irrelevant stand for "relevant entity" and "irrelevant entity" respectively

<table>
<thead>
<tr>
<th>Entity category</th>
<th>Food</th>
<th>Location</th>
<th>Symptom</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Train</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relevant tweets</td>
<td>94</td>
<td>120</td>
<td>50</td>
<td>12</td>
<td>276</td>
</tr>
<tr>
<td>Irrelevant tweets</td>
<td>0</td>
<td>128</td>
<td>324</td>
<td>0</td>
<td>1272</td>
</tr>
<tr>
<td><strong>Val</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relevant tweets</td>
<td>90</td>
<td>76</td>
<td>4</td>
<td>3</td>
<td>206</td>
</tr>
<tr>
<td>Irrelevant tweets</td>
<td>0</td>
<td>19</td>
<td>38</td>
<td>64</td>
<td>152</td>
</tr>
<tr>
<td><strong>Test</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relevant tweets</td>
<td>85</td>
<td>71</td>
<td>85</td>
<td>221</td>
<td>462</td>
</tr>
<tr>
<td>Irrelevant tweets</td>
<td>10</td>
<td>24</td>
<td>18</td>
<td>1</td>
<td>53</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>1267</td>
<td>1109</td>
<td>1481</td>
<td>2956</td>
<td>6813</td>
</tr>
</tbody>
</table>

---

*Here, we cannot describe the type of foodborne illness since citizens would not have any knowledge which particular germ has actually caused them to be poisoned.*

came back to the **hotel**, I started **vomit**. Here, even **hotel** is a location entity, but it is not likely to relate to the foodborne illness incident. Hence, **hotel** is labeled as an irrelevant entity. All other four entities are labeled as relevant entities. In many cases, we care in particular about entities specifically related to the foodborne illness incidents, as they might contain valuable information about the source or cause of the foodborne illness in question. Then this ERC task can help us filter out less important information. Hence, the ERC and EMD tasks can be solved jointly.

**Relevant entity mention detection (RED).** This task is to detect entities that are related to the foodborne illness incident mentioned in a given tweet. As described in the last paragraph, entities that are related to the foodborne illness incident contain more valuable information than unrelated entities. RED can be seen as a one-step alternation of the combination of EMD and ERC that directly outputs relevant entities. Compared to EMD + ERC, RED is more efficient and can avoid the risk of mismatch between entities boundary predicted on EMD and ERC task.

**Combination Tasks.** In our setting, a tweet contains one or more relevant entities if and only if the tweet is labeled as a relevant tweet. Those relevant entities are always a subset of all entities. In this study, in addition to learning each task individually, we also design several combinations of these tasks that can be learned jointly to observe multi-task models’ performance on these tasks. These combinations are: TRC + EMD, TRC + ERC, TRC + RED, and EMD + ERC. We pair TRC with the other 3 tasks since we want to see if the sentence level and entity level can benefit from each other. Especially the ERC and RED tasks, given the fact that in our dataset, at least one entity must be a relevant entity if and only if the tweet indicates a possible foodborne illness incident. The model that can utilize such relationships may achieve better performance. We also included EMD + ERC as one combination since these two tasks have the same entity label boundary. A model that can leverage the shared boundary information may reach better performance. Here, we do not solve all 4 tasks at the same time. Since ERC + EMD will return both the entity type label and the entity relevancy label for each word, there is no need to do the RED task concurrently. Thus, it should not be necessary to invoke all three tasks jointly.

### 3.2 Compared methods

To establish a starting point for future study, we evaluate several baseline models on TWEET-FD. Configurations of each method and implementation details are available in the Appendix.

**RoBERTa.** This pretrained deep learning model, proposed by [8], builds on BERT [16], and refines the pre-training procedure by removing the next-sentence prediction tasks and training with larger learning rates and mini-batch. [8] implements RoBERTa on multiple tasks and shows that RoBERTa can achieve state-of-the-art results on several benchmarks.

**BERTweet.** This model, proposed by [9], is the first public large-scale pretrained language model for English Tweets. BERTweet model has the same architecture as BERT-base [16] and is pretrained using the RoBERTa pretraining procedure [8]. BERTweet has shown great performance improvements over previous state-of-the-art models on benchmark Tweet datasets.

**RoBERTa and BERTweet** are implemented for all tasks. For each task, we pair RoBERTa and BERTweet with a classification head as the top layer (linear layer on top of hidden state output). For EMD, ERC, and RED tasks, we also create a network with a CRF layer on top of the linear classification layer. For multi-task combinations (TRC + EMD, TRC + ERC, TRC + RED, EMD + ERC), we assign two classification heads on top of the model’s hidden state output, one head for one task.

**IMGJM.** Interactive Multi-Grained Joint Model (IMGJM), proposed as an interactive multi-grained joint model for targeted sentiment analysis [17], targets extraction and target sentiment classification jointly. The interaction mechanism in IMGJM can share information between the target and sentiment tagging results. IMGJM is utilized for our EMD + ERC dual-task problem due to its good fit for our goal of detecting entities and predicting the relevance of detected entities. In our setting, entity tagging and entity relevance tagging share the same boundary — similar to the original target sentiment analysis setting of IMGJM.

**MGADE.** MGADE, a multi-grained joint deep network model proposed by [10], was designed to concurrently solve both ADE entity recognition and ADE assertive sentence classification. The model is equipped with a dual-attention mechanism to construct multiple distinct sentence representations
to capture both task-specific and semantic information in a sentence. We applied this method to our TRC + EMD, TRC + ERC, TRC + RED tasks.

**BiLSTM.** Bidirectional LSTM (BiLSTM), is a widely used sequence architecture. In this study, we input the sequence of word embeddings of each tweet into the BiLSTM, the BiLSTM’s hidden states are fed into a classifier for final prediction. Same as RoBERTa and BERTweet, we also create a network with a CRF layer on top of the linear classification layer for EMD, ERC, and RED tasks. For multi-task combinations (TRC + EMD, TRC + ERC, TRC + RED, EMD + ERC), we assign two classification heads on top of the BiLSTM’s hidden states, one head for one task. The TRC task classification layer takes the concatenation of the last hidden states from both directions as input. EMD, ERC, RED tasks’ classifiers take each word’s hidden state from both directions as input.

**SVM.** Support Vector Machine (SVM), is a standard machine learning technique. In this study, we first transform a given tweet into feature vectors via Bags of words and term frequency-inverse document frequency (tf-idf). Then, feature vectors are fed into SVM for classification. We applied this method to our TRC task.

<table>
<thead>
<tr>
<th>Methods</th>
<th>TRC</th>
<th>EMD</th>
<th>ERC</th>
<th>RED</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Single-task Methods</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiLSTM - EMD</td>
<td>.54 ± 0.006</td>
<td>645 ± 0.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiLSTM (with CRF) - EMD</td>
<td>.67 ± 0.006</td>
<td>758 ± 0.009</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RoBERTa - EMD</td>
<td>.507 ± 0.004</td>
<td>609 ± 0.008</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERTweet - EMD</td>
<td>.448 ± 0.007</td>
<td>547 ± 0.019</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RoBERTa (with CRF) - EMD</td>
<td>.687 ± 0.003</td>
<td>786 ± 0.003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERTweet (with CRF) - EMD</td>
<td>.637 ± 0.11</td>
<td>748 ± 0.006</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BiLSTM - ERC</td>
<td></td>
<td></td>
<td>462 ± 0.007</td>
<td>597 ± 0.008</td>
</tr>
<tr>
<td>BiLSTM (with CRF) - ERC</td>
<td></td>
<td></td>
<td>573 ± 0.008</td>
<td>644 ± 0.01</td>
</tr>
<tr>
<td>RoBERTa - ERC</td>
<td></td>
<td></td>
<td>42 ± 0.005</td>
<td>517 ± 0.006</td>
</tr>
<tr>
<td>BERTweet - ERC</td>
<td></td>
<td></td>
<td>378 ± 0.13</td>
<td>48 ± 0.025</td>
</tr>
<tr>
<td>RoBERTa (with CRF) - ERC</td>
<td></td>
<td></td>
<td>593 ± 0.001</td>
<td>678 ± 0.006</td>
</tr>
<tr>
<td>BERTweet (with CRF) - ERC</td>
<td></td>
<td></td>
<td>592 ± 0.003</td>
<td>69 ± 0.008</td>
</tr>
<tr>
<td>BiLSTM - RED</td>
<td></td>
<td></td>
<td></td>
<td>445 ± 0.018</td>
</tr>
<tr>
<td>BiLSTM (with CRF) - RED</td>
<td></td>
<td></td>
<td></td>
<td>58 ± 0.01</td>
</tr>
<tr>
<td>RoBERTa - RED</td>
<td></td>
<td></td>
<td></td>
<td>409 ± 0.008</td>
</tr>
<tr>
<td>BERTweet - RED</td>
<td></td>
<td></td>
<td></td>
<td>389 ± 0.006</td>
</tr>
<tr>
<td>RoBERTa (with CRF) - RED</td>
<td></td>
<td></td>
<td></td>
<td>609 ± 0.002</td>
</tr>
<tr>
<td>BERTweet (with CRF) - RED</td>
<td></td>
<td></td>
<td></td>
<td>523 ± 0.013</td>
</tr>
<tr>
<td>BiLSTM - TRC</td>
<td>.786 ± 0.016</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RoBERTa - TRC</td>
<td>.858 ± 0.004</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERTweet - TRC</td>
<td>.844 ± 0.004</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM - TRC</td>
<td>.847 ± 0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| **Multi-task Methods** |              |              |              |              |
| BiLSTM - TRC+EMD      | .798 ± 0.034 | .553 ± 0.011 | 652 ± 0.005  |              |
| BiLSTM (with CRF) - TRC+EMD | .803 ± 0.012 | .665 ± 0.005 | 755 ± 0.007  |              |
| RoBERTa - TRC+EMD     | .85 ± 0.011  | 527 ± 0.003  | 626 ± 0.004  |              |
| BERTweet - TRC+EMD    | .833 ± 0.027 | 471 ± 0.019  | 573 ± 0.016  |              |
| RoBERTa (with CRF) - TRC+EMD | .862 ± 0.011 | 496 ± 0.004  | .788 ± 0.005 |              |
| BERTweet (with CRF) - TRC+EMD | .812 ± 0.018 | .638 ± 0.02  | .759 ± 0.015 |              |
| BiLSTM - TRC+ERC      | .806 ± 0.011 |              |              |              |
| BERTweet - TRC+ERC    | .809 ± 0.012 |              |              |              |
| RoBERTa - TRC+ERC     | .861 ± 0.007 |              |              |              |
| BERTweet - TRC+ERC    | .855 ± 0.013 |              |              |              |
| RoBERTa (with CRF) - TRC+ERC | .86 ± 0.01   |              |              |              |
| BERTweet (with CRF) - TRC+ERC | .838 ± 0.015 |              |              |              |
| BiLSTM (with CRF) - TRC+RED | .803 ± 0.012 |              |              |              |
| RoBERTa - TRC+RED     | .839 ± 0.018 |              |              |              |
| BERTweet - TRC+RED    | .831 ± 0.013 |              |              |              |
| BiLSTM (with CRF) - TRC+RED | .847 ± 0.026 |              |              |              |
| RoBERTa (with CRF) - TRC+RED | .847 ± 0.006 |              |              |              |
| BERTweet (with CRF) - TRC+RED | .847 ± 0.006 |              |              |              |
| BiLSTM - EMD+ERC      | N/A          | .399 ± 0.02  | .496 ± 0.013 | .372 ± 0.02  | .473 ± 0.014 |
| BiLSTM (with CRF) - EMD+ERC | N/A          | .577 ± 0.005 | .633 ± 0.12  | .532 ± 0.003 | .587 ± 0.014 |
| RoBERTa - EMD+ERC     | N/A          | .531 ± 0.017 | .628 ± 0.16  | .455 ± 0.01  | .542 ± 0.014 |
| BERTweet - EMD+ERC    | N/A          | .476 ± 0.013 | .589 ± 0.014 | .412 ± 0.012 | .516 ± 0.018 |
| RoBERTa (with CRF) - EMD+ERC | N/A          | .689 ± 0.015 | .786 ± 0.011 | .598 ± 0.007 | .687 ± 0.004 |
| BERTweet (with CRF) - EMD+ERC | N/A          | .648 ± 0.017 | .757 ± 0.017 | .597 ± 0.005 | .691 ± 0.011 |
| MGADE - TRC+EMD       | .855 ± 0.009 | .676 ± 0.017 | .77 ± 0.014  |              |
| MGADE - TRC+ERC       | .846 ± 0.004 |              |              |              |
| MGADE - TRC+RED       | .846 ± 0.011 |              |              |              |
| IMGJM - EMD+ERC       | N/A          | .411 ± 0.027 | .504 ± 0.007 | .425 ± 0.007 | .493 ± 0.009 |

Table 2: Performance of each method on four tasks of text relevance classification (TRC), entity mention detection (EMD), entity relevance classification (ERC) and relevant entity mention detection (RED). Metrics: $F_1$ for TRC, Strict $F_1$ and Entity Type $F_1$ for other tasks.
3.3 Evaluation metrics

The four proposed tasks fall into the category of classical NLP tasks, namely, sentence classification and sequence labeling. TRC is a sentence classification problem and we use $F_1$ to measure its model performance. While EMD, ERC, and RED are sequence labeling problems. We thus evaluate these tasks with both Strict $F_1$ and Type $F_1$, both of which are evaluation metrics introduced in semEval-2013. A named entity usually has a clear boundary, however, people may segment the entity mentions at different positions. For example, "my stomach hurts really bad." describes the symptom and some people may only take "stomach hurts" but others may select "stomach hurts really bad" as the symptom. All of them have identified the entity successfully, but the boundaries do not exactly match. Under such a scenario, the Type $F_1$ is more appropriate as a metric.

3.4 Experimental results

Table 2 presents evaluation results of each method. When comparing single-task and multi-task methods, we observe that multi-task methods can match or exceed the performance of single-task counterparts. EMD, ERC, and RED reach the highest gain in $F_1$ scores when they are jointly learned with the TRC task. Also, the best performance of the TRC task is reached when it is jointly learned with the EMD task. It illustrates that the sentence-level and the token-level task can mutually benefit each other. Also, joint learning of EMD and ERC tasks also leads to higher $F_1$ scores on both tasks. Since EMD and ERC share entities boundaries, experiment results show that the model does benefit from this information.

By comparing methods’ performance across EMD, ERC, and RED tasks, we notice that methods reach the highest $F_1$ scores on the EMD task. They have lower scores on the ERC and RED tasks. This is not surprising. Because in ERC task, a word, for example, sandwich, is labeled as a relevant entity only when related to a foodborne illness incident, otherwise labeled as "out of relevant entity". But in the EMD task, in most of the cases, sandwich is labeled as a food entity, no matter if the word is related to a foodborne illness incident or not. It makes the ERC and RED tasks harder than the EMD task since the relevant information has to be incorporated when making predictions on ERC or RED tasks.

We observe that RoBERTa-based methods, equipped with the CRF layer, outperform other methods on many tasks. The BERTweet-based model also achieves good performance comparable to RoBERTa does. This is likely because both RoBERTa and BERTweet are pretrained on a large corpus, whereas MGADE and IMGJM are being trained from scratch on our dataset. An interesting finding is that, when methods are not equipped with CRF layer, BiLSTM sometimes can perform better on EMD, ERC, and RED tasks, compared with counterparts without CRF layer. But BiLSTM cannot keep this advantage when a CRF layer is equipped. On the TRC task, we notice that SVM can also make a good performance. It shows that even pre-trained transformer-based networks have become state-of-the-arts methods in the NLP domain, traditional and standard machine learning and deep learning framework still can be used in some settings.

Besides this main experiment, we conduct a case study to quantify the impact of the training dataset size on the models’ performance. Results of this experiment are referred to Appendix A.6.

4 Related work

Social media data has been identified as a great source of information for public health due to its timeliness and scalability. Many researchers have built effective surveillance systems in the food safety domain even with vanilla machine learning methods. Most previous work only collected the label of the posts (i.e., whether it’s relevant to foodborne illness events) and focused on identifying sick Yelp reviews or tweets with simple machine learning models (like logistic regression, SVM, and Random Forest) in a certain area [4,18,19,6,5,7]. Although the previous applications classify the relevant posts successfully, more detailed information are valuable to retrieve manually during the inspection process, as explained in Section 1. [11] constructed a Chinese corpus of food safety incident entities using Bi-LSTM-CRF from food safety incident related news. In our work, we not
only collected the sentence label, but also the entities mentioned in English tweets and their relevance to foodborne illnesses.

Since social media data is noisy and fragmented, social media related NLP techniques and datasets have been proposed. Some studies built systems to discover event categories and extract entities and event phrases [20]. The NLP community has shared tasks and datasets for enabling more research on social media data. [21, 22] showed promising results of the shared NER task. [23] presented the results of Paraphrase Identification (PI) and Semantic Textual Similarity (SS). Many datasets are emerging for new interesting tasks. [24] construct a large-scale question answering (QA) dataset over social media data. In particular, social media datasets related to Covid-19 have been published in the past year to reveal insights into societal perceptions of the pandemic and techniques for its prevention [25, 26]. In contrast, we constructed TWEET-FD, which can be used for multiple tasks: sentence classification, entity detection, entity relevance classification, and related entity detection.

5 Discussion

Data and Software Availability. Upon publication, other researchers can apply for access to our dataset under the restriction of Twitter redistribution policy. Our data aggregation code is available on Github.

Limitations. Our dataset contains the content of tweet texts and corresponding labels. When we collected raw tweet data, we found that most tweet data does not come with location information (where the tweet was sent out from). Additionally, the location information mentioned in the text may not correspond to precise geospatial information (e.g., someone just mentioned the name of a restaurant, without any city or state). In the future, the creation of additional datasets with geospatial information to focus on spatial analysis of foodborne illness outbreaks would be of value. Due to the limited available information in the user profiles, we can not clearly state the bias in our dataset in terms of gender/age/subgroup. Moreover, as mentioned in [24], since we use the keywords highly related to foodborne illness to collect data, our dataset has a higher precision. For now, we focus more on the general terms relevant to foodborne illness. However, some tweets that are related to foodborne illness but do not contain any of our keywords were not included in our dataset. For example, some tweets that contain misspelled keywords, slang, and regional English varieties might be missed in data collection procedure. We have a detailed discussion on this issue in Appendix A.7.

Broader societal impact. In this work, we create this dataset to advance NLP-based model development to tackle the important problem of foodborne illness incident detection. Future works built based on this study can benefit the public by providing early warning about foodborne illness outbreaks - thus potentially saving lives and livelihoods.

Ethical considerations. The authors do not foresee the negative ethical consequences of this work. Since many tweets may be deleted by the tweeters, we have converted user mention and URL links to @USER, HTTPURL, respectively. Therefore, the privacy of tweeters is to some degree protected. To obey the restrictions described in Developer Terms of Twitter, we will give access to the researchers who send requests to us and agree to the Twitter Terms of Service, Privacy Policy, Developer Agreement, and Developer Policy before receiving Twitter content. Lastly, our IRB board informed us that no IRB was needed for collecting foodborne illness related annotations from workers on the content of public tweets. More details are referred to Appendix A.8.

6 Conclusion

In this work, we construct TWEET-FD, the first publicly available foodborne illness related tweet dataset for multiple foodborne illness incident detection tasks. TWEET-FD serves as training data for the following tasks: (1) identify if a given tweet indicates a foodborne illness incident; (2) find and extract entities in the tweet; (3) determine the relevance of entities to the foodborne illness incident mentioned in the tweet, or (4) only extract entities related to the foodborne illness incident. We conduct a crowdsourcing study to collect annotations for these four tasks. Thereafter, we create gold-standard labels from annotations with high inter-annotator agreement using a sound methodology. We then conduct an experimental study to evaluate several state-of-the-art single-task & multi-task models for realizing the four tasks on TWEET-FD. These results can now serve as baseline performances for future research work in this important area of food safety.
Acknowledgement

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. We thank Mengrui Luo for her assistance in implementing the crowdsourcing interface for this study. We also thank all reviewers for their valuable suggestions and feedback in helping us improve our work.

References


An T Nguyen, Byron C Wallace, Junyi Jessy Li, Ani Nenkova, and Matthew Lease. Aggregating

Klaus Krippendorff, Yann Mathet, Stéphane Bouvry, and Antoine Widlöcher. On the reliability
of unitizing textual continua: Further developments. Quality & Quantity, 50(6):2347–2364,
2016.

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

Da Yin, Xiao Liu, and Xiaojun Wan. Interactive multi-grained joint model for targeted sentiment
analysis. In Proceedings of the 28th ACM International Conference on Information and
Knowledge Management, CIKM ’19, pages 1031–1040, New York, NY, USA, 2019. ACM.

Jenine K Harris, Raed Mansour, Bechara Choucair, Joe Olson, Cory Nissen, and Jay Bhatt.
Health department use of social media to identify foodborne illness—chicago, illinois, 2013–

Jenine K Harris, Jared B Hawkins, Leila Nguyen, Elaine O Nsoesie, Gaurav Tuli, Raed Mansour,
and John S Brownstein. Research brief report: using twitter to identify and respond to food
poisoning: The food safety stl project. Journal of Public Health Management and Practice,

Alan Ritter, Sam Clark, Oren Etzioni, et al. Named entity recognition in tweets: an experimental
processing, pages 1524–1534, 2011.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. Results of the
wnut2017 shared task on novel and emerging entity recognition. In Proceedings of the 3rd

Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine De Marneffe, and Wei Xu.
Results of the wnut16 named entity recognition shared task. In Proceedings of the 2nd Workshop

Wei Xu, Chris Callison-Burch, and Bill Dolan. Semeval-2015 task 1: Paraphrase and semantic
similarity in twitter (pit). In Proceedings of the 9th international workshop on semantic
evaluation (SemEval 2015), pages 1–11, 2015.

Wenhan Xiong, Jiawei Wu, Hong Wang, Vivek Kulkarni, Mo Yu, Shiyu Chang, Xiaoxiao Guo,

Abraham C Sanders, Rachael C White, Lauren S Severson, Rufeng Ma, Richard McQueen,
Haniel C Alcântara Paulo, Yucheng Zhang, John S Erickson, and Kristin P Bennett. Unmasking
the conversation on masks: Natural language processing for topical sentiment analysis of

Juan M Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, and
Gerardo Chowell. A twitter dataset of 150+ million tweets related to covid-19 for open research.
Type: dataset, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s
       contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] see Section
(c) Did you discuss any potential negative societal impacts of your work? [Yes] see Section 5
(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] URL to code and data are in supplemental materials
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] see Appendix in supplemental materials
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] see Table 2
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] see Appendix in supplemental materials

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] see Appendix in supplemental materials
   (b) Did you mention the license of the assets? [Yes] see Appendix in supplemental materials
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] see supplemental material
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [Yes] see Appendix in supplemental materials
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] see Section 5 and Section 2.1

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] see Appendix in Supplemental materials
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] see explanation in Section 5 and Appendix in supplemental material
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] see Section 2.2.1