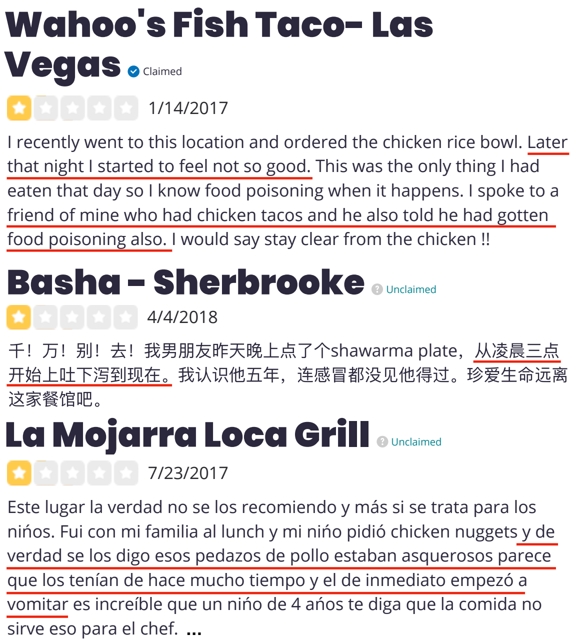
Detecting Foodborne Illness Complaints in Multiple Languages Using English Annotations Only

Ziyi Liu, Giannis Karamanolakis, Daniel Hsu, Luis Gravano
Columbia University, New York, NY 10027, USA
zl2888@columbia.edu, {gkaraman, djhsu, gravano}@cs.columbia.edu

Abstract

Health departments have been deploying text classification systems for the early detection of foodborne illness complaints in social media documents such as Yelp restaurant reviews. Current systems have been successfully applied for documents in English and, as a result, a promising direction is to increase coverage and recall by considering documents in additional languages, such as Spanish or Chinese. Training previous systems for more languages, however, would be expensive, as it would require the manual annotation of many documents for each new target language. To address this challenge, we consider cross-lingual learning and train multilingual classifiers using only the annotations for English-language reviews. Recent zero-shot approaches based on pre-trained multi-lingual BERT (mBERT) have been shown to effectively align languages for aspects such as sentiment. Interestingly, we show that those approaches are less effective for capturing the nuances of foodborne illness, our public health application of interest. To improve performance without extra annotations, we create artificial training documents in the target language through machine translation and train mBERT jointly for the source (English) and target language. Furthermore, we show that translating labeled documents to multiple languages leads to additional performance improvements for some target languages. We demonstrate the benefits of our approach through extensive experiments with Yelp restaurant reviews in seven languages. Our classifiers identify foodborne illness complaints in multilingual reviews from the Yelp Challenge dataset, which highlights the potential of our general approach for deployment in health departments.

1 Introduction



Wahoo's Fish Taco- Las Vegas

1/14/2017

I recently went to this location and ordered the chicken rice bowl. Later that night I started to feel not so good. This was the only thing I had eaten that day so I know food poisoning when it happens. I spoke to a friend of mine who had chicken tacos and he also told he had gotten food poisoning also. I would say stay clear from the chicken !!

Basha – Sherbrooke

4/4/2018

千！万！别！去！我男朋友昨天晚上点了个shawarma plate，从凌晨三点开始上吐下泻到现在。我认识他五年，连感冒都没见他得过。珍爱生命远离这家餐馆吧。

La Mojarra Loca Grill

7/23/2017

Este lugar la verdad no se los recomiendo y más si se trata para los nińos. Fui con mi familia al lunch y mi nińo pidió chicken nuggets y de verdad se los digo esos pedazos de pollo estaban asquerosos parece que los tenían de hace mucho tiempo y el de inmediato empezó a vomitar es increíble que un nińo de 4 ańos te diga que la comida no sirve eso para el chef. ...

Figure 1: Examples of Yelp restaurant reviews discussing food poisoning in different languages.

With the rise of social media, more and more users post online documents where they disclose serious incidents, such as getting food poisoning from a restaurant. As many of those incidents may not be reported through established complaint systems, health departments have deployed text classification systems for the identification of social media documents, such as Yelp reviews and tweets, that discuss foodborne illness episodes. Figure 1 shows examples of Yelp restaurant reviews discussing food poisoning in English, Chinese, and Spanish.

Current classification systems have been applied for documents written in English and deployed in several health departments, including those in Chicago (Harris et al., 2014), Nevada (Sadilek et al., 2016), New York City (Effland et al., 2018), and St. Louis (Harris et al., 2018). Online documents flagged by the classifiers are typically analyzed by epidemiologists, who further investigate the incidents (e.g., by inspecting the corresponding restau-

Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, pages 138–146
November 20, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

rants). This process contributes to the early detection of previously unknown foodborne outbreaks. Given the success of current systems, a promising new direction is to extend these systems to use non-English languages, thus increasing their coverage and capacity to identify foodborne outbreaks.

Directly applying existing techniques for foodborne illness detection to other languages would be expensive and time-consuming. Current (supervised) classifiers have been trained on thousands of documents that were manually labeled with binary ("Sick" vs. "Not Sick") labels provided by epidemiologists, and it would be expensive to replicate this effort for new target languages. Furthermore, it is hard to collect documents for annotation for our task because most online documents do not discuss foodborne illness. Alternative approaches beyond supervised learning are thus required to efficiently scale to multiple languages.

To address the costly requirement of supervised learning approaches, we train multilingual classifiers through a less expensive *cross-lingual* text classification approach. For a given non-English target language, our approach does not require manually annotated in-language documents but instead trains classifiers using the already available English annotations. We follow recent techniques for cross-lingual text classification and employ pre-trained multi-lingual BERT (mBERT) representations (Wu and Dredze, 2019; Pires et al., 2019). However, while pre-trained mBERT representations have been shown to be effective for tasks such as cross-lingual sentiment classification (Wu and Dredze, 2019), we show that such representations are less effective for capturing the nuances of foodborne illness, which is required by our application of focus. To improve performance, we translate labeled English reviews to the target language and fine-tune mBERT *jointly* for both languages, which turns out to be more effective than fine-tuning on either language separately. Furthermore, we show that fine-tuning mBERT for multiple languages in parallel leads to additional improvements for some target languages such as German and Italian.

Our work makes the following contributions:

1. We present a cross-lingual learning approach for foodborne illness detection in non-English social media documents. Our approach is efficient and requires only English labeled data.

2. We show how to improve the performance of pre-trained mBERT for our rare classification task. Our preliminary results show that generating additional artificial training data in multiple languages through machine translation leads to promising improvements over zero-shot mBERT.

3. We evaluate our approach on Yelp reviews in English, Spanish, Chinese, French, German, Japanese, and Italian. Our approach substantially outperforms previous techniques and baselines for this task. Our multilingual classifiers successfully identify foodborne illness across languages in reviews from the Yelp Challenge dataset, which highlights the potential of our approach for successful, real-world deployment in health departments.

The rest of this paper is organized as follows. In Section 2, we provide the necessary background for our work. In Section 3, we describe our approach for cross-lingual foodborne detection. In Section 4, we present the experimental setup and results. In Section 5, we conclude and suggest future work.

2 Background

In this section, we provide background on foodborne illness detection (Section 2.1) and cross-lingual text classification (Section 2.2).

2.1 Foodborne Illness Detection in English Documents

Foodborne illness detection in online documents has been addressed as a binary text classification task: the goal is to train a classifier that, given the text of a document, predicts a binary ("Sick" vs. "Not Sick") label, corresponding to whether the document is mentioning foodborne illness or not. Sadilek et al. (2016) trained support vector machine classifiers (based on unigram, bigram and trigram features) using 8,000 tweets that were independently labeled by five human annotators. Effland et al. (2018) trained classifiers using more than 10,000 Yelp reviews that were manually annotated by epidemiologists. The paper compares several methods and found that logistic regression had the best performance. Karamanolakis et al. (2019) trained a weakly-supervised neural network that predicts a label for each individual sentence of a review and improves the recall of foodborne illness complaints compared to the best performing classifier in Effland et al. (2018).

2.2 Cross-Lingual Text Classification

Cross-lingual text classification trains a classifier on a target language T by leveraging labeled documents in a source language S. We focus on the challenging cross-lingual classification setting where only unlabeled documents are available in T.

Some effective approaches address cross-lingual classification by relying on cross-lingual word embeddings (Gouws and Søgaard, 2015; Ruder et al., 2019), which represent words from different languages in the same vector space, where words across languages with similar meanings are represented as similar vectors. Cross-lingual word embeddings facilitate cross-lingual model transfer as a classifier trained on labeled documents in S could be directly applied for test documents in T.

More recent approaches addressed cross-lingual transfer using Multilingual BERT (Wu and Dredze, 2019; Pires et al., 2019; Karthikeyan et al., 2019; Rogers et al., 2020). Multilingual BERT, or mBERT, is a version of BERT (Devlin et al., 2019) that was trained on 104 languages in parallel. Training mBERT on English documents was shown to achieve impressively high performance on different target languages for several document classification tasks such as sentiment classification or topic detection (Rogers et al., 2020). The successful application of mBERT for various cross-lingual tasks inspired us to employ mBERT for our public-health application, as we describe next.

3 Foodborne Illness Detection in Multiple Languages

We now define our problem of focus (Section 3.1) and describe our cross-lingual learning approach (Sections 3.2 and 3.3).

3.1 Problem Definition

Our goal is to address foodborne illness detection in non-English languages where labeled documents are not available. As the collection of manual annotations for each new language is an expensive and time-consuming proposition, we focus on training multilingual classifiers using only already available English documents. More formally, we assume access to a source language S (English) with a labeled dataset $D_S = \{(x_i^S, y_i^S)\}$, where x_i^S is a source language document and y_i^S is the corresponding binary ("Sick" vs. "Not Sick") label. For a target language T we assume access to a dataset D_T of unlabeled target documents x^T. Our goal is to train a classifier for the target language T that, given an unseen test document x^T in T, predicts a binary ("Sick" vs. "Not Sick") label.

3.2 Fine-Tuning mBERT on S and T

To address the task mentioned in Section 3.1, we use pre-trained mBERT representations, which effectively align representations of different languages (Section 2.2).

It has been shown that mBERT achieves impressive zero-shot performance for tasks such as sentiment classification and topic detection (Wu and Dredze, 2019; Pires et al., 2019): fine-tuning mBERT on the labeled dataset D_S in S leads to accurate classification of unlabeled documents x^T in T, possibly because representations across languages are well aligned with respect to the target sentiment or topic. However, in contrast to previous tasks, we show that zero-shot mBERT is not effective for foodborne detection. We hypothesize that this discrepancy is observed because pre-trained mBERT representations are not effectively aligned across languages with respect to the aspect of foodborne illness, which may be rarely mentioned in documents used for pre-training mBERT.

To address this issue and improve classification performance for our task, we do not consider zero-shot training but fine-tune mBERT in both S and T. Our main idea is that fine-tuning mBERT in documents from both S and T will encourage a stronger alignment of the cross-lingual representations with respect to the aspect of foodborne illness. The main challenge associated with our approach is that labeled documents are not available in the target language T.

To generate training documents in T, we translate labeled documents x_i^S from S (English) to T using machine translation. In particular, we assume that machine translation is sufficiently accurate to the extent that the translated document $x_i^{S\to T}$ has the same label as the original document x_i^S. Under this assumption, we generate a weakly annotated dataset $D'_T = \{(x_i^{S\to T}, y_i^S)\}$ by translating all documents x_i^S annotated as "Sick" and an equal number of documents randomly sampled from "Not Sick" documents in D_S. Then, we increase the size of D'_T by sampling unlabeled documents x_i^T from D_T uniformly at random. Each sampled document is assigned the "Not Sick" label as the chance of randomly choosing a document mentioning foodborne illness is very low. The number of sampled

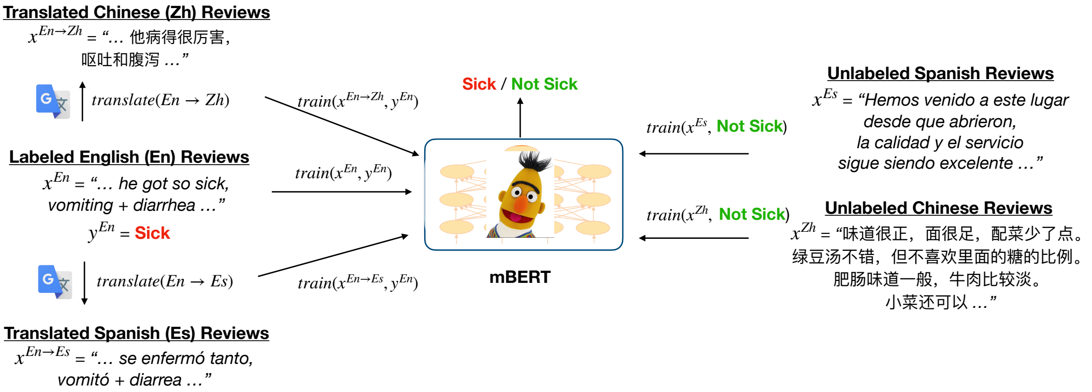

Translated Chinese (Zh) Reviews
$x^{En\rightarrow Zh}$ = "… 他病得很厉害，呕吐和腹泻 …"
translate(En → Zh)
Labeled English (En) Reviews
x^{En} = "… he got so sick, vomiting + diarrhea …"
y^{En} = Sick
translate(En → Es)
Translated Spanish (Es) Reviews
$x^{En\rightarrow Es}$ = "… se enfermó tanto, vomitó + diarrea …"
$train(x^{En\rightarrow Zh}, y^{En})$
$train(x^{En}, y^{En})$
$train(x^{En\rightarrow Es}, y^{En})$
Sick / Not Sick
mBERT
$train(x^{Es}$, Not Sick)
$train(x^{Zh}$, Not Sick)
Unlabeled Spanish Reviews
x^{Es} = "Hemos venido a este lugar desde que abrieron, la calidad y el servicio sigue siendo excelente …"
Unlabeled Chinese Reviews
x^{Zh} = "味道很正，面很足，配菜少了点。绿豆汤不错，但不喜欢里面的糖的比例。肥肠味道一般，牛肉比较淡。小菜还可以 …"

Figure 2: Our training procedure. We translate *labeled* English reviews to the target languages and use the translated reviews with the original labels as extra training samples. We also use a sample of *unlabeled* multilingual reviews as negative ("Not Sick") training examples.

documents is chosen so that the total number of "Not Sick" documents in D'_T is equal to that in D_S.

After creating the weakly labeled D'_T set we fine-tune our mBERT-based classifier jointly on D_S and D_T by concatenating and shuffling the two datasets. As we will show, this training procedure is more effective than fine-tuning mBERT on D_S or D_T separately.

3.3 Considering Multiple Source Languages

Classification performance in T may potentially improve using *multiple* source languages $\{S_1, \dots, S_K\}$ other than S (English) for which unlabeled documents and machine translation systems are available. The main idea behind this approach is that training signals from multiple source languages could prevent overfitting to a single source language and as a result encourage mBERT to learn better cross-lingual representations for our task. Therefore, we adapt the procedure described in Section 3.2 to consider more source languages in addition to S and T, as we describe next.

To train mBERT using multiple source languages $S, S_1, \dots, S_K$, we create a big training set that considers all source-language documents. In particular, first we create a weakly-labeled dataset D_{S_k} for each source language using machine translation, as we described in Section 3.2 for creating D_S. Then, we concatenate all source datasets $D_S, D'_{S_1}, \dots, D'_{S_K}$ and fine-tune mBERT across all languages $(S, S_1, \dots, S_K, T)$. Note that, in our preliminary experiments, we have treated all languages as equal but in the future it would be interesting to consider alternative approaches, such as using different weights for examples from different languages. Figure 2 shows our overall training procedure using English, Spanish and Chinese for training mBERT.

An important advantage of this approach is that the same mBERT classifier can be applied on any target language T supported in mBERT. As a result, deployment in health departments would be easier since it involves a single model for all languages and does not require extra pre-processing steps such as running a language detector[1] for each test document and applying language-specific models. Also, as we will show next, considering multiple source languages during training encourages better generalization to a new *unseen* test language.

4 Experiments

We evaluate our approach on foodborne detection in English (En), Spanish (Es), Chinese (Zh), French (Fr), German (De), Japanese (Ja), and Italian (It).

4.1 Experimental Settings

Datasets. We use the same corpus of labeled English reviews from Effland et al. (2018). This dataset contains English reviews with ground truth annotations provided by epidemiologists. Table 1 reports the number of reviews on the train and test set. For details, see Effland et al. (2018).

We collect unlabeled multilingual reviews from Yelp restaurants in New York City (NYC), Los Angeles (LA), as well as other metropolitan areas in the Yelp Challenge dataset.[2] As the language of the

[1] In our experiments, language detectors sometimes predicted the wrong language for the text of a test restaurant review, for example because of multiple mentions of Italian dishes in a non-Italian review.

[2] https://www.kaggle.com/yelp-dataset/yelp-dataset

	All Reviews	Sick	Not Sick
Train	21,551	5894	15,657
Validation	1500	1090	410
Test	2975	949	2026

Table 1: Number of Yelp reviews in the English dataset with ground-truth (Sick vs. Not Sick) annotations.

	NYC Area	LA Area	Yelp Challenge	Total
Spanish	6267	11,458	2658	20,383
Chinese	1624	1488	603	3715
French	3882	741	24,807	29,430
German	2912	657	1394	4963
Japanese	2161	1469	563	4193
Italian	1259	322	173	1754

Table 2: Number of unlabeled Yelp reviews from the New York City area, Los Angeles area, as well as other metropolitan areas in the Yelp Challenge dataset.

reviews is not mentioned in the metadata, we used Python's langdetect[3] library to automatically detect the language. For evaluation on non-English languages, we translate the 2975 English test reviews to the target languages using the Google Translate API.[4]

Model Comparison. We compare the following models for our task:

- **Monolingual LogReg**: the logistic regression classifier that achieved the best results in (Effland et al., 2018). We train LogReg for a non-English target language T by translating English reviews to T using Google Translate (see Section 3.2).
- **Monolingual BERT**: a monolingual BERT classifier. Similarly to LogReg, we train BERT for a non-English target language T by translating English reviews to T using Google Translate.
- **mBERT**: a multilingual BERT classifier. We train mBERT on several combinations of languages using our approach described in Section 3.

Model Configuration. For LogReg, we tokenize text using Spacy[5] and convert the text documents to TF-IDF vectors.

For monolingual BERT, we consider pre-trained BERT representations from huggingface[6]:

- English: bert-base-uncased
- Spanish: dccuchile/bert-base-spanish-wwm-cased
- Chinese: bert-base-chinese
- French: camembert-base
- German: bert-base-german-cased
- Japanese: cl-tohoku/bert-base-japanese
- Italian: dbmdz/bert-base-italian-xxl-cased

For mBERT, we consider pre-trained mBERT representations from huffingface: bert-base-multilingual-cased. We fine-tuned BERT and mBERT using the Python simpletransformers[7] library. We did a hyperparameter search with BERT on English data using the validation set. The best hyperparameters are a learning rate of 1e-05, a batch size of 512, and a maximum sequence length of 512. We fine-tune BERT/mBERT for up to 5 epochs with early stopping based on the validation loss.

Evaluation Procedure. For each model, we choose the best set of hyperparameters according to the F1 score on the validation set. We report the following classification metrics on the test set: accuracy (Acc), precision (Prec), recall (Rec), and macro-average F1 score (F1).

4.2 Experimental Results

Table 3 shows F1 scores on all languages for various methods.

Monolingual BERT outperforms previous systems. Monolingual BERT outperforms LogReg: leveraging pre-trained contextual representations captures foodborne illness effectively.

Monolingual BERT outperforms mBERT. Interestingly, monolingual BERT performs better than mBERT. We hypothesize that, by focusing on a single language, pre-trained monolingual BERT representations capture foodborne-related aspects more effectively than mBERT representations that were pre-trained for all languages in parallel.

[3] `https://pypi.org/project/langdetect/`

[4] The Google Translate API was used in February 2020.

[5] `https://spacy.io/api/tokenizer`

[6] `https://huggingface.co`

[7] `https://github.com/ThilinaRajapakse/simpletransformers`

	Train	Test Language							AVG
Model	**Language**	**En**	**Es**	**Zh**	**Fr**	**De**	**Ja**	**It**	**F1**
Monolingual LogReg	T	83.7	83.6	83.3	84.9	80.4	81.7	83.6	83.0
Monolingual BERT	T	**91.6**	**91.3**	**87.3**	**92.4**	**88.9**	**87.0**	**90.7**	**89.9**
mBERT	En	89.0	82.0	78.8	80.6	59.0	65.5	67.5	74.6
mBERT	T	89.0	87.1	87.0	88.6	87.3	88.8	**89.4**	88.2
mBERT	En + T	89.0	**89.8**	**89.7**	90.5	88.0	**89.4**	88.2	89.2
mBERT	ALL	**91.3**	89.6	88.0	**90.7**	**89.5**	86.8	89.2	**89.3**

Table 3: F1 scores for various approaches evaluated on different test languages. Monolingual LogReg and BERT are trained on the translated documents in the target language T. mBERT is trained with various language configurations. Training mBERT in English and T is more effective than training on either language separately. Training mBERT across all 7 languages ("ALL") leads to further improvements for En, Fr, and De. Results in red correspond to the best performance across all models.

Model	**Train**	**Es**	**Zh**	**AVG**
mBERT	En	82.0	78.8	80.4
mBERT	ALL-T	**84.7**	**84.0**	**84.4**

Table 4: Zero-shot performance under two different settings: training on English-only data (En) vs. training on all languages except the target language (ALL-T). The latter approach performs substantially better than the former.

Model	**Train**	**Acc**	**Prec**	**Rec**	**F1**
LogReg	En	88.1	74.1	96.2	83.7
BERT	En	**94.4**	88.1	**95.4**	**91.6**
mBERT	En	92.5	83.8	95.0	89.0
mBERT	ALL	94.3	**89.2**	93.6	91.3

Table 5: Evaluation on English Yelp reviews.

Zero-shot mBERT is not effective. Training zero-shot mBERT using only English training data (En) is not effective and performs substantially worse than monolingual LogReg. This result validates our argument that pre-trained mBERT representations do not effectively capture the aspect of food poisoning, which is rarely mentioned in documents used for pre-training mBERT.

Artificial training reviews in T improve mBERT's performance. Translating English reviews to T and using translated reviews to train mBERT on T is substantially better than zero-shot mBERT trained on English directly. This result highlights the importance of in-language training documents, even if those documents are artificially created. Furthermore, training mBERT jointly on English and the target language T leads to better performance compared to training on each language separately.

Training on all languages leads to the best performance for mBERT. On average across languages, mBERT trained on all languages jointly performs better than other mBERT configurations with a single source language, but comparably to mBERT trained on En and T. Interestingly, for Chinese (Zh) and Japanese (Ja) performance is worse if more languages are added to the training set, possibly because these languages are more distant from Romance languages such as Spanish or French, and as a result considering those languages in the training set is not helpful.

Using multiple source languages leads to higher zero-shot performance. Table 4 shows results for the setting where we assume that documents from the target language are not available for training. Crucially, training mBERT on all languages except this target language performs substantially better than training mBERT only on English data, validating the importance of training mBERT on multiple languages jointly. Also, F1 scores when ignoring those languages during training (ALL-T) are lower by about 5 absolute points compared to considering them during training (ALL): we could potentially apply our approach to any unseen language out of the 104 languages that are supported by mBERT.

Detailed English results. Table 5 shows results in English. BERT (monolingual) has the best F1 score. Training mBERT on all languages (En, Es, Zh, Fr, De, Ja, It) is more effective than training mBERT on English-only labeled data. This validates our hypothesis that, by considering all languages, mBERT generalizes better to test reviews.

Model	Train	Acc	Prec	Rec	F1
LogReg	Es	87.9	73.7	**96.4**	83.6
LogReg*	En	88.2	75.5	93.4	83.5
BERT	Es	**94.2**	87.1	96.0	**91.3**
BERT*	En	93.6	87.1	93.9	90.4
mBERT	En	89.7	**92.3**	73.8	82.0
mBERT	Es	90.9	79.5	**96.4**	87.1
mBERT	En+Es	93.2	85.9	94.1	89.8
mBERT	ALL	93.3	89.0	90.2	89.6

(a) Results on Spanish.

Model	Train	Acc	Prec	Rec	F1
LogReg	Zh	88.3	76.8	90.9	83.3
LogReg*	En	87.2	76.9	85.6	81.0
BERT	Zh	91.3	81.2	94.5	87.3
BERT*	En	92.4	88.6	87.6	88.1
mBERT	En	88.2	**91.9**	69.0	78.8
mBERT	Zh	90.9	80.2	95.0	87.0
mBERT	En+Zh	**93.2**	86.2	93.6	**89.7**
mBERT	ALL	91.7	81.7	**95.5**	88.0

(b) Results on Chinese.

Model	Train	Acc	Prec	Rec	F1
LogReg	Fr	89.4	77.8	93.6	84.9
BERT	Fr	**95.0**	89.6	**95.4**	**92.4**
mBERT	En	88.9	**91.4**	72.1	80.6
mBERT	Fr	92.1	82.6	**95.4**	88.6
mBERT	En+Fr	93.6	86.4	94.9	90.5
mBERT	ALL	94.0	89.8	91.6	90.7

(c) Results on French.

Model	Train	Acc	Prec	Rec	F1
LogReg	De	85.2	69.5	95.1	80.4
BERT	De	92.4	83.3	**95.5**	88.9
mBERT	En	81.2	**97.1**	42.4	59.0
mBERT	De	91.1	80.4	95.4	87.3
mBERT	En+De	91.9	82.9	93.9	88.0
mBERT	ALL	**93.0**	86.0	93.4	**89.5**

(d) Results on German.

Model	Train	Acc	Prec	Rec	F1
LogReg	Ja	86.5	71.9	94.7	81.7
BERT	Ja	91.2	82.3	92.3	87.0
mBERT	En	83.0	**93.2**	50.5	65.5
mBERT	Ja	92.4	83.6	94.7	88.8
mBERT	En+Ja	**92.8**	84.2	95.3	**89.4**
mBERT	ALL	90.7	79.4	**95.7**	86.8

(e) Results on Japanese.

Model	Train	Acc	Prec	Recall	F1
LogReg	It	88.5	76.8	91.7	83.6
BERT	It	**93.7**	85.5	96.5	**90.7**
mBERT	En	83.5	**91.1**	53.6	67.5
mBERT	It	92.8	84.6	94.8	89.4
mBERT	En+It	91.7	81.1	**96.6**	88.2
mBERT	ALL	92.7	84.3	94.6	89.2

(f) Results on Italian.

Table 6: Results on different target languages. LogReg and BERT are trained on the translated target-language documents. LogReg* and BERT* are trained on English and applied on test reviews by translating the corresponding text from the target language to English. mBERT is trained with various configurations.

Detailed non-English results. Table 6 shows detailed results on non-English datasets. For Spanish and Chinese we evaluated an additional baseline where test reviews are translated to English and considered by LogReg ("Logreg*" baseline) or BERT ("BERT*" baseline) that were trained on English reviews only. This approach is less effective, as well as more expensive than the other approaches: to deploy in health departments, it would require each new test review to be translated to English. While BERT has the highest F1 score on average over all approaches, mBERT has higher recall than BERT on most non-English target languages.

We detect reviews mentioning foodborne illness. To demonstrate the potential of our approach for detecting foodborne illness, we ran mBERT on unlabeled restaurant reviews from the NYC Area, LA Area, and the Yelp Challenge dataset. Table 7 shows two examples that were classified as "Sick" by our classifier. Translating those two reviews to English and applying LogReg (trained in English) led to a (wrong) "Not Sick" prediction, possibly because the translated reviews are not matching the training distribution for LogReg.

5 Discussion and Future Work

We presented our cross-lingual learning method for scaling foodborne illness detection to languages beyond English without extra annotations for non-English languages. As most reviews do not discuss foodborne illness, it is challenging to create proper evaluation datasets for all languages.

In our preliminary experiments, we evaluated our approach on non-English languages by translating labeled test reviews from English to other languages. A caveat of this evaluation approach is that complaints of foodborne illness in native-language reviews may be expressed differently than

Spanish	**Original (Es) text:** Definitivamente mi peor experiencia, me intoxique con un ostra mala, llevo 4 días en muy malas condiciones, por favor tengan cuidado, los ostiones y mariscos no se pueden comer en cualquier lugar, yo aprendi por las malas, espero que mi experiencia le sirva a alguien **mBERT (train: ALL) prediction: "Sick"** ✓
	Translated (En) text: Definitely my worst experience, I got intoxicated with a bad oyster, I have been in very bad conditions for 4 days, please be careful, the oysters and shellfish cannot be eaten anywhere, I learned through the bad ones, I hope my experience will serve you someone **LogReg (train: En) prediction: "Not Sick"** ✗
Chinese	**Original (Zh) text:** 装修和服务都还不错，但味道极差：底料没有味道，我们自己加了几次盐和料才勉强能吃，菜品也非常不新鲜。一顿饭吃得我们四个人都很生气，然后回家三个人都拉肚子。绝对不会再去吃。Avoid!! **mBERT (train: ALL) prediction: "Sick"** ✓
	Translated (En) text: The decoration and service are good, but the taste is very bad: the base material has no taste, we added salt several times to make it barely edible, and the dishes are very fresh. Four of us were angry at a meal, and then all three got diarrhea. Will never eat again. Avoid !! **LogReg (train: En) prediction: "Not Sick"** ✗
German	**Original (De) text:** Wir haben hier 2 bowls mit Steak und einen Burger gegessen. Für unverschämte 70,03$ gab es recht kleine und nicht wirklich gute Portionen (besonders die bowls). Nachdem mein Sohn von der Bowl gegessen hat, musste er brechen. Auch meiner Tochter und mir war schlecht. Der Service wirkte lieblos und desinteressiert. Die bowls kamen gerade mal lauwarm an unseren Tisch und die Chips vom Burger schmeckten nach nichts. Nicht zu empfehlen!!! **mBERT (train: ALL) prediction: "Sick"** ✓
	Translated (En) text: We ate 2 bowls of steak and a burger here. For outrageous $70.03 there were quite small and not really good portions (especially the bowls). After my son ate from the bowl, he had to break. My daughter and I were also bad. The service seemed careless and uninterested. The bowls just came to our table lukewarm and the chips from the burger didn't taste like anything. Not recommendable!!! **LogReg (train: En) prediction: "Not Sick"** ✗

Table 7: Examples of Spanish, Chinese and German restaurant reviews in our dataset classified as "Sick" and their translations to English.

in automatically translated reviews and thus, performance numbers may not be fully indicative of performance in native reviews. Therefore, an important next step is to create better evaluation datasets.

Our exploratory results show that training mBERT in multiple languages jointly is more effective than training mBERT on English (zero-shot approach) or the target-language only. On average across languages mBERT is outperformed by monolingual BERT trained on (translated) target-language documents. On the other hand, deploying mBERT in health departments for daily inspections would be easier as it would not require extra pre-processing steps such as language detection that may introduce errors. Also, we showed that mBERT could potentially be applied for languages that were not seen in the training set, without extra translation efforts.

As another interesting direction for future work, we plan to evaluate the cross-lingual transfer approach of Karamanolakis et al. (2020), which applies even for low-resource languages that are not supported by mBERT or for which machine translation systems are not available. We also plan to extend our system for predicting which languages to use as source languages to achieve good performance on a target language (Lin et al., 2019).

Acknowledgments

We thank the anonymous reviewers for their constructive feedback. This material is based upon work supported by the National Science Foundation under Grant No. IIS-15-63785.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Thomas Effland, Anna Lawson, Sharon Balter, Katelynn Devinney, Vasudha Reddy, HaeNa Waechter, Luis Gravano, and Daniel Hsu. 2018. Discovering foodborne illness in online restaurant reviews. *Journal of the American Medical Informatics Association*.

Stephan Gouws and Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Jenine K Harris, Leslie Hinyard, Kate Beatty, Jared B Hawkins, Elaine O Nsoesie, Raed Mansour, and John S Brownstein. 2018. Evaluating the implementation of a Twitter-based foodborne illness reporting tool in the City of St. Louis Department of Health. *International Journal of Environmental Research and Public Health*, 15(5).

Jenine K Harris, Raed Mansour, Bechara Choucair, Joe Olson, Cory Nissen, and Jay Bhatt. 2014. Health department use of social media to identify foodborne illness-Chicago, Illinois, 2013-2014. *Morbidity and Mortality Weekly Report*, 63(32):681–685.

Giannis Karamanolakis, Daniel Hsu, and Luis Gravano. 2019. Weakly supervised attention networks for fine-grained opinion mining and public health. In *Proceedings of the 5th Workshop on Noisy User-Generated Text*.

Giannis Karamanolakis, Daniel Hsu, and Luis Gravano. 2020. Cross-lingual text classification with minimal resources by transferring a sparse teacher. In *Proceedings of the 2020 Findings of Empirical Methods in Natural Language Processing*.

Kaliyaperumal Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2019. Cross-lingual ability of multilingual BERT: An empirical study. In *International Conference on Learning Representations*.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, et al. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *arXiv preprint arXiv:2002.12327*.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Adam Sadilek, Henry A Kautz, Lauren DiPrete, Brian Labus, Eric Portman, Jack Teitel, and Vincent Silenzio. 2016. Deploying nEmesis: Preventing foodborne illness by data mining social media. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Shijie Wu and Mark Dredze. 2019. Beto, Bentz, Becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.