Journal of the American Medical Informatics Association, 00(0), 2018, 1–7
doi: 10.1093/jamia/ocx093
Research and Applications

Research and Applications

Discovering foodborne illness in online restaurant reviews

Thomas Effland,[1] Anna Lawson,[1] Sharon Balter,[2] Katelynn Devinney,[2] Vasudha Reddy,[2] HaeNa Waechter,[2] Luis Gravano,[1] and Daniel Hsu[1]

[1]Computer Science Department, Data Science Institute, Columbia University, New York, NY, USA and [2]Bureau of Communicable Disease, New York City Department of Health and Mental Hygiene, Queens, NY, USA

Corresponding Author: Thomas Effland, Mudd Building, 500 W 120th St., New York, NY 10027, USA. E-mail: teffland@cs.columbia.edu. Phone: +1-301-395-7542

Received 10 January 2017; Revised 7 July 2017; Editorial Decision 27 July 2017; Accepted 25 September 2017

ABSTRACT

Objective: We developed a system for the discovery of foodborne illness mentioned in online Yelp restaurant reviews using text classification. The system is used by the New York City Department of Health and Mental Hygiene (DOHMH) to monitor Yelp for foodborne illness complaints.
Materials and Methods: We built classifiers for 2 tasks: (1) determining if a review indicated a person experiencing foodborne illness and (2) determining if a review indicated multiple people experiencing foodborne illness. We first developed a prototype classifier in 2012 for both tasks using a small labeled dataset. Over years of system deployment, DOHMH epidemiologists labeled 13 526 reviews selected by this classifier. We used these biased data and a sample of complementary reviews in a principled bias-adjusted training scheme to develop significantly improved classifiers. Finally, we performed an error analysis of the best resulting classifiers.
Results: We found that logistic regression trained with bias-adjusted augmented data performed best for both classification tasks, with F1-scores of 87% and 66% for tasks 1 and 2, respectively.
Discussion: Our error analysis revealed that the inability of our models to account for long phrases caused the most errors. Our bias-adjusted training scheme illustrates how to improve a classification system iteratively by exploiting available biased labeled data.
Conclusions: Our system has been instrumental in the identification of 10 outbreaks and 8523 complaints of foodborne illness associated with New York City restaurants since July 2012. Our evaluation has identified strong classifiers for both tasks, whose deployment will allow DOHMH epidemiologists to more effectively monitor Yelp for foodborne illness investigations.

Key words: machine learning, social media, foodborne diseases, text mining, classification

BACKGROUND AND SIGNIFICANCE

Foodborne illness remains a major public health concern nationwide. The Centers for Disease Control and Prevention (CDC) estimates that there are 48 million illnesses and >3000 deaths caused by the consumption of contaminated food in the United States each year.[1] Of the approximately 1200 foodborne outbreaks reported and investigated nationally, 68% are restaurant-related.[2] Most restaurant-associated outbreaks are identified via health department complaint systems. However, there are potentially valuable data sources emerging that could be incorporated in outbreak detection. Specifically, the increasing use of social media has provided a public platform for users to disclose serious real-life incidents, such as food poisoning, that may not be reported through established complaint systems.

As a result of the increasing interest and potential value of social media data, research institutions are partnering with public health agencies to develop methods and applications to use data from social media to monitor outbreaks of infectious diseases. Textual data from Internet search engines and social media have been used to monitor outbreaks of various infectious diseases, such as influenza.[3]

An evaluation comparing the use of informal and unconventional outbreak detection methods against traditional methods found that the informal source was the first to report in 70% of outbreaks, supporting the usefulness of such systems.[4] The incorporation of social media data into public health surveillance systems is becoming more common. Multiple projects focus on identifying incidents of foodborne illness using data from Twitter. Harvard Medical School developed and maintains a machine learning platform, HealthMap Foodborne Dashboard, to identify complaints and occurrences of foodborne illness and send a survey link where Twitter users can provide more information; this platform is freely available for research.[5] The Chicago Department of Public Health partnered with the Smart Chicago Collaborative to develop Foodborne Chicago, which also uses machine learning to identify tweets indicating foodborne illness and also sends a survey link where Twitter users can provide more information.[6] The Southern Nevada Health District developed nEmesis, an application that associates a user's previous locations with subsequent tweets indicating foodborne illness.[7]

In this study, we use data from consumer reviews obtained from the popular website Yelp. A comparison of food vehicles associated with outbreaks from the CDC Foodborne Outbreak Online Database and data extracted from Yelp reviews indicating foodborne illness and implicating a specific food item found that the distribution of food categories was very similar between the 2 sources, supporting the usefulness of these data in public health responses.[8] Furthermore, Yelp reviews can be directly linked with individual restaurant locations, allowing for targeted and timely response.

Since 2012, the Computer Science Department at Columbia University has been collaborating with the New York City (NYC) Department of Health and Mental Hygiene (DOHMH) to develop a system that applies data mining and uses text classification to identify restaurant reviews on Yelp indicating foodborne illness, which are later manually reviewed and classified by DOHMH epidemiologists. This system was used in a pilot study from July 1, 2012, to March 31, 2013, and found 468 Yelp reviews that described a foodborne illness occurrence.[9] Of these 468 reviews, only 3% of the illness incidents had been reported to the DOHMH by calling NYC's citywide complaint system, 311. Investigations as a result of these reviews led to the discovery of 3 previously unknown foodborne illness outbreaks, approximately 10% of the total number of restaurant-associated outbreaks identified during the pilot project's time period. This highlighted the need to mine Yelp reviews to improve the identification and investigation of foodborne illness outbreaks in NYC. Due to the success of the pilot study, DOHMH integrated Yelp reviews into its foodborne illness complaint surveillance system and continues to mine Yelp reviews and investigate those pertaining to foodborne illness; this process has been instrumental in the identification of 10 outbreaks and 8523 reports of foodborne illness associated with NYC restaurants since July 2012.

OBJECTIVE

In this study, we aimed to evaluate the performance of several classifiers on both tasks: the prototype classifiers used by the deployed DOHMH system and multiple well-known state-of-the-art classification models. Additionally, we sought to investigate the impact of training the classifiers using data collected by the prototype system over years of deployment. This process, however, must be treated with care, as the data collected from the prototype system suffer, unavoidably, from a selection bias. To resolve this issue, we derived principled bias-adjusted training and evaluation objectives and designed training regimes that incorporate data sampled from the complement of this biased set to produce improved classifiers. We investigated the impact of these biased vs bias-adjusted training regimes and identified strong final models for both tasks.

MATERIALS AND METHODS

We first describe the overall DOHMH system design. We then describe the classification models used in our evaluation. Finally, we describe the data used in the evaluation and discuss bias-adjusted training and evaluation objectives.

Yelp system design

The system runs a daily process to pull Yelp reviews of NYC restaurants from a privately available application programming interface (API) and applies text classification techniques to classify reviews according to 2 criteria. The first criterion, referred to as the "Sick" task, corresponds to whether the review mentions the occurrence of a person experiencing foodborne illness from the restaurant. The second criterion, the "Multiple" task, corresponds to whether there was a foodborne illness event experienced by more than one person; although they are quite rare, these cases constitute significant evidence of a foodborne illness outbreak and are of special interest to DOHMH epidemiologists. After automatically classifying all new reviews according to these criteria, all reviews classified as "Sick" (ie, having a "Sick" probability>0.5) are then presented to DOHMH epidemiologists in a user interface for manual review. Upon reviewing a document, the epidemiologists record the gold standard label for both criteria.

Yelp messages are sent to the authors of reviews that appear to report true incidents of foodborne illness, and an interview is attempted with each author to collect information regarding symptoms, other illnesses among the author's dining group, and a 3-day food history. All sources of restaurant-associated foodborne illness complaints are aggregated in a daily report; outbreak investigations are initiated if multiple complaints indicating foodborne illness are received within a short period of time for one establishment, or if a complaint indicates a large group of individuals experiencing illness after a single event.

Classification methods

Prior to classification, the reviews, or documents, are converted into a representation that is usable by the classification algorithms, known as the featurization of documents. This is done using a bag-of-words (BOW) approach by converting each document into a vector with the counts for each word in the vocabulary.

The classifiers built for the operational system at DOHMH, further referred to as "prototype" classifiers, were J4.8[10] decision tree models, chosen for the interpretability of their decision functions. These models were trained using 500 reviews, labeled by DOHMH epidemiologists for both criteria. The 500 reviews were selected using a mix of an unbiased sample of reviews and reviews from keyword searches for terms that are intuitively indicative of foodborne illness, such as "sick," "vomit," "diarrhea," and "food poisoning."

To identify the most effective classifiers for our classification tasks, we experimentally evaluated several standard document classification techniques in addition to the prototype classifiers. First, we considered improvements to the document featurization over basic BOW by including *n*-grams (*n* consecutive words) for $n = 1$, 2, and 3, and term frequency-inverse document frequency (TF-IDF) weights for the terms.[11] For both classification tasks, "Sick" and

"Multiple," we evaluated 3 well-known supervised machine-learning classifiers: logistic regression,[12] random forest,[13] and support vector machine (SVM).[14] Logistic regression is a classical statistical regression model where the response variable is categorical. Random forest is an ensemble of weak decision tree classifiers that vote for the final classification of the input document. SVM is a nonprobabilistic classifier that classifies new documents according to their distance from previously seen training documents.

By definition, the positive examples for the "Multiple" task are a subset of the positive "Sick" examples, since at least one person must have foodborne illness for multiple people to have foodborne illness. Using this notion, we additionally designed a pipelined set of classifiers, further referred to as "Sick-Pipelined" classifiers, for the "Multiple" task, which first condition their predictions on the best "Sick" classifier. If the "Sick" classifier predicts "Yes," then the "Multiple" classifier is run. Intuitively, this allows the "Multiple" classifier to focus more on the number of people involved than on whether there was a singular foodborne illness event at all. We evaluated logistic regression for this model class.

Enhanced dataset and selection bias–corrected training

Since July 2012, DOHMH epidemiologists have labeled 13 526 reviews selected for manual inspection by the prototype "Sick" classifier. These reviews are balanced for the "Sick" task, with 51% "Yes" and 49% "No" documents, but are imbalanced for the "Multiple" task, with only 13% "Yes" and 87% "No" documents. For training and evaluation, we split the data chronologically at January 1, 2017, to mirror future performance when training on historical data. This resulted in 11 551 training reviews and 1975 evaluation reviews. The training and evaluation sets have equal class distributions: 51%/49% for "Sick" and 13%/87% for "Multiple."

While these reviews contain useful information, having been selected by the prototype "Sick" classifier before labeling heavily biases them, and so they are not representative of the full (original) Yelp feed. To understand and correct for the impact of such bias, we derived a bias-adjusted training objective and augmented the training and evaluation datasets with a sample of reviews from the complement of the biased datasets in the full Yelp feed.

Selection-bias correction

To account for the selection bias of the prototype "Sick" classifier in the labeled data, we augment the training data with reviews from the set of Yelp reviews that were labeled "No" by the prototype "Sick" classifier. Reviews from this set, further referred to as "complement-sampled" reviews, likely have nothing to do with foodborne illness, but instead serve as easy "No" examples that the classifiers should predict correctly.

Exactly how these 2 datasets are merged, however, requires principled consideration. For classifiers that learn to reduce classification error in training, we can formally model the joint likelihood of the classifier misclassifying some review and that review being selected by the prototype "Sick" classifier. Then, by marginalizing this joint distribution over the indicator that a review is selected by the prototype "Sick" classifier, we arrive at an unbiased estimate of the classification error. The end result is that we weigh classification mistakes for the biased and complement-sampled reviews by the inverses of their respective probabilities of being selected at random from the full Yelp dataset.

Training regimes

Using the above sample weights, we incorporate both the biased label data and the complement-sampled data to train our classifiers under 3 different regimes. The first, "Biased," used only the data from the 11 551 reviews selected by the prototype "Sick" classifier. The second, "Gold," used the "Biased" data plus 1000 reviews sampled from the complement-sampled Yelp feed and labeled by DOHMH epidemiologists. In this sample of 1000 reviews, only 4 were labeled "Yes" for the "Sick" task and 1 was labeled "Yes" for the "Multiple" task. In the third regime, "Silver," we randomly sampled 10 000 reviews from the complement-sampled Yelp feed before January 1, 2017, and assumed all were negative examples of both tasks. Intuitively, this regime can be helpful if it regularizes out statistical quirks of the "Biased" data more than the noise it may introduce through false negatives.

Evaluation

The performance of each classifier was evaluated on the 1975 biased reviews from after January 1, 2017, along with another sample of 1000 reviews from the complement-sampled Yelp feed after January 1, 2017. These 1000 reviews were again labeled by DOHMH epidemiologists for both tasks. However, there were no positive examples of either task among the 1000 reviews.

We evaluated the models for both tasks using 4 performance metrics common to class-imbalanced binary classification problems: precision, recall, F1-score, and area under the precision-recall curve (AUPR). Precision (often called "positive predictive value") is the proportion of true positives out of the total number of positive predictions. Recall (often called "sensitivity") is the true positive rate. F1-score is the harmonic mean of precision and recall. Precision, recall, and F1-score were calculated at a classification threshold of 0.5, meaning that we classified reviews with "Yes" probabilities ≥ 0.5 as "Yes." The AUPR was measured by first graphing precision versus recall by varying the classification threshold from 0 to 1, then calculating the area under the curve. For all 4 metrics, 0 is the worst possible score and 1 is a perfect score.

Since our evaluation data are biased, the evaluation metrics as described would not reflect unbiased estimates of model performance on the full Yelp feed. We can again derive bias-corrected precision and recall quantities, as we did with the training objective, by weighing test examples from the biased and complement-sampled sets by the inverses of their respective probabilities of being selected from the full Yelp dataset.

For each model class, task, and training regime (21 variations total), we performed hyperparameter tuning experiments using 500 trials of random search from reasonable sampling distributions using 5-fold cross-validation on the training data, stratified by class label and biased/complement-sampled label. The details of the various featurization techniques and hyperparameter optimization experiments can be found in the Supplementary Appendix. After selecting the best hyperparameter settings for each model variation using best average bias-adjusted F1-score across the development folds, we retrained the models on their full training datasets.

We compared the resulting model variations to each other and the prototype classifiers on the 4 evaluation metrics. We calculated 95% confidence intervals for F1-score and AUPR using the percentile bootstrap method[15] with 1000 sampled test datasets. We then selected the best variation for both tasks based on test bias-adjusted F1-score as our final classifiers. We report the confusion matrices, perform a detailed error analysis, and identify insightful top features for the final classifiers on both tasks.

Table 1. Model performance on "Sick" task

Model	Training Regime	Precision	Recall	F1-Score (95% CI)	AUPR (95% CI)
J4.8	Prototype	0.48	0.99	0.65 (0.63-0.67)	0.83 (0.81-0.85)
Logistic regression	Biased	0.05	0.94	0.10 (0.09-0.11)	0.63 (0.55-0.76)
Logistic regression	Gold	0.83	0.88	0.85 (0.83-0.87)	0.90 (0.88-0.92)
Logistic regression	Silver	0.85	0.88	0.87 (0.85-0.88)	0.91 (0.90-0.93)
Random forest	Biased	0.04	0.91	0.07 (0.06-0.09)	0.59 0.54-0.70
Random forest	Gold	0.36	0.89	0.51 (0.38-0.68)	0.81 (0.78-0.84)
Random forest	Silver	0.70	0.88	0.78 (0.66-0.85)	0.87 (0.85-0.89)
SVM	Biased	0.09	0.95	0.16 (0.13-0.20)	0.82 (0.79-0.87)
SVM	Gold	0.33	0.93	0.49 (0.37-0.67)	0.88 (0.85-0.91)
SVM	Silver	0.96	0.74	0.83 (0.81-0.85)	0.93 (0.92-0.95)

The underlined value represents the final selected model from among the variants. This is the model we further analyze in the error analysis. Because the bootstrap distribution of some test statistics exhibited non-normal behavior, their corresponding confidence intervals are wider.

RESULTS

We found that the best classifiers achieved bias-adjusted F1-scores of 87% and 66% on the "Sick" and "Multiple" classification tasks, respectively.

Classification evaluation

The performance of the classifier variations for the "Sick" and "Multiple" tasks is presented in Tables 1 and 2, respectively. All models were evaluated on the test data from after January 1, 2017.

For the "Sick" task, we found that the logistic regression model trained using the "Silver" regime achieved the highest F1-score, 87%. With the addition of 10 000 silver-labeled complement-sampled reviews, this model gained 77% in bias-adjusted F1-score over its "Biased" counterpart, a significant increase. The low bias-adjusted F1-score of 10% for the "Biased" "Sick" logistic regression is due to the misrepresentation of the full Yelp dataset by the "Biased" training, which causes the model to highly over-predict "Yes" on the complement-sampled test data. This behavior is heavily penalized by the bias-adjustment because each false positive in the small complement-sampled test data is representative of many more false positives in the full Yelp dataset.

For the "Multiple" task, we found that the "Sick-Pipelined" logistic regression model trained using the "Silver" regime achieved the highest F1-score, 66%. The use of pipelined training and prediction caused a gain of 5% for the "Silver" "Sick-Pipelined" logistic regression over its single-step counterpart.

Precision-recall trade-off

Given the rarity of reviews discussing foodborne illness, it is desirable to explore settings of the "Sick" classifiers that favor recall over precision, since DOHMH epidemiologists are willing to accept some extra false positives to reduce the risk of missing an important positive "Sick" review. We analyzed this trade-off by examining the precision-recall curves of the "Sick" logistic regression classifiers, presented in Figure 1. From the plot, we can see that "Gold" and "Silver" models begin to experience an approximately equal trade-off of precision for recall in the region of 80%–90% recall, illustrated by the slope of the curves being close to 1 point of precision lost per point of recall gained. In the 90%–100% recall region, the "Gold" model begins to experience a steep drop in precision at a recall of 92% while the "Silver" model does not experience a steep drop in precision until a recall of 98%. At this point, the precision of the "Silver" logistic regression is still 69%, 21% higher than the prototype classifier which has 48% precision at 99% recall. This indicates that even in a high-recall setting the "Silver" "Sick" classifier should provide better performance over the "Sick" prototype.

Error analysis of best "Sick" classifier

Of the 2975 reviews in the test dataset, there are 949 positive examples and 2026 negative examples for the "Sick" task. The best "Sick" classifier, "Silver" trained logistic regression, achieved an F1-score of 87%, a statistically significant 22% absolute increase over the prototype classifier, with an F1-score score of 65%. On this test dataset, the best "Sick" classifier correctly classified many reviews containing major sources of false positives for the prototype classifier. These gains are not surprising, given that this model uses 40 times more data and better document representations (TF-IDF and trigrams rather than vanilla BOW). This large performance increase will qualitatively change the efficacy of the system for DOHMH epidemiologists.

Examination of the 144 false positives identified various causes. Many of these false positives cannot be identified by a classifier only using *n*-grams up to $n = 3$. For example, one reviewer wrote, "I didn't get food poisoning," which would require 4-grams for the classifier to capture the negation. This example illustrates a major shortcoming of *n*-gram models: important dependencies or relationships between words often span large distances across a sentence. Another major source of false positives are reviews that do talk

Table 2. Model performance on "Multiple" task

Model	Training Regime	Precision	Recall	F1-Score 95% CI	AUPR 95% CI
J4.8	Prototype	< 0.01	0.69	0.01 (0.01, 0.01)	< 0.01 (< 0.01, < 0.01)
Logistic regression	Biased	0.08	0.56	0.15 (0.09-0.26)	0.25 (0.19-0.40)
Logistic regression	Gold	0.42	0.58	0.48 (0.30-0.67)	0.56 (0.49-0.67)
Logistic regression	Silver	0.64	0.58	0.61 (0.56-0.66)	0.58 (0.52-0.65)
Sick-Pipelined logistic regression	Biased	0.07	0.61	0.13 (0.09-0.23)	0.18 (0.13, 0.43)
Sick-Pipelined logistic regression	Gold	0.77	0.56	0.65 (0.60-0.70)	0.65 (0.59-0.70)
Sick-Pipelined logistic regression	Silver	0.75	0.59	0.66 (0.61-0.70)	0.71 (0.65-0.76)
Random forest	Biased	0.04	0.37	0.07 (0.05-0.12)	0.03 (0.02-0.18)
Random forest	Gold	0.75	0.24	0.36 (0.29-0.42)	0.31 (0.23-0.45)
Random forest	Silver	0.74	0.25	0.37 (0.31-0.43)	0.40 (0.34-0.49)
SVM	Biased	0.07	0.65	0.12 (0.08-0.20)	0.18 (0.12-0.48)
SVM	Gold	0.35	0.34	0.35 (0.21-0.54)	0.29 (0.21-0.57)
SVM	Silver	0.20	0.30	0.24 (0.13-0.47)	0.39 (0.30-0.64)

The underlined value represents the final selected model from among the variants. This is the model we further analyze in the error analysis.

Table 3. Confusion matrices of best classifiers

Actual Class	Predicted Class			
	No		Yes	
	Count	Rate (%)	Count	Rate (%)
Sick				
No	1882 (true negatives)	93	144 (false positives)	7
Yes	112 (false negatives)	12	837 (true positives)	88
Multiple				
No	2643 (true negatives)	98	55 (false positives)	2
Yes	114 (false negatives)	42	163 (true positives)	58

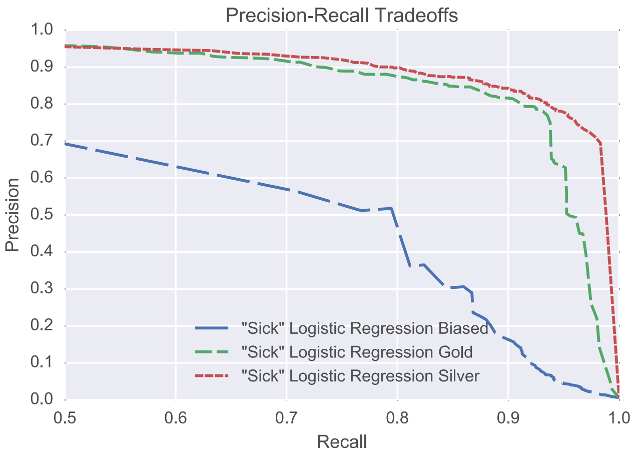

Figure 1. Precision-recall curves of "Sick" logistic regression models in the high-recall region. While the "Biased" logistic regression performance lags below, the "Gold" and "Silver" models show relatively mild losses in precision per point of recall gained until the 90-100% recall region. After 92% recall the "Gold" model begins to experience a steep drop in precision while the "Silver" model does not experience a steep drop in precision until a recall of 98%.

about food poisoning but are not current enough to meet the DOHMH criteria for follow-up, and thus are labeled "No." A third type of false positive occurs when a review talks about food poisoning in a hypothetical or future sense. For example, one reviewer reported that the food "had a weird chunky consistency…hopefully we won't get sick tonight."

Multiple causes of the 112 false negatives were also identified. One notable cause is misspellings of key words related to food poisoning in the review, such as "diherrea." Another major cause is grave references to food poisoning but the classifier predicts "No" because of a prevalence of negatively weighted *n*-grams, such as "almost threw up." A final source of false negatives is human error in the labeling of reviews for the test data. For example, one review's only reference to illness was "she began to feel sick" while at the restaurant, yet the review was labeled positive. Many of the reviews contained negation, which the best "Sick" classifier can detect due to the use of *n*-grams. *N*-grams also allow the classifier to identify that the pattern "sick of," as in "sick of the pizza," does not typically refer to actual food poisoning, compared to "got sick," which typically does.

Finally, we examined the highest-weighted *n*-grams of the best "Sick" classifier. The most highly positive-weighted features were phrases indicative of foodborne illness, such as “diarrhea,” “food poisoning,” and “got sick,” while the most highly negative features were either very positive phrases or indicative of false positives, such as “amazing” and “sick of.” These top features are encouraging, as they show the model has identified features that epidemiologists would also deem important.

Error analysis of best “Multiple” classifier

Of the 2975 reviews in the test dataset, there are 277 positive examples and 2698 negative examples for the “Multiple” task. The best “Multiple” classifier, “Silver” trained “Sick-Pipelined” logistic regression, achieved an F1-score of 66%.

We examined the reason behind the 114 false negative reviews. Many false negatives were due to incorrect predictions made by the pipelined “Sick” classifier. Most other false negatives were caused by the inability of trigram models to capture longer phrases. Phrases indicating multiple illnesses, such as “we both got really sick,” typically span more than 3 contiguous words, leaving no way for a classifier using trigrams to detect them directly.

Of the 277 true positives, 163 were correctly classified. Reviews containing phrases clearly indicating multiple illnesses in a bigram or trigram, such as “both got sick,” scored highest; however, such concise *n*-grams are rare. The classifier’s highly weighted features are *n*-grams that simply refer to multiple people without referring to food poisoning. The classifier can capture references to multiple people in a trigram, but these references are often devoid of context, making it hard to determine if multiple people simply did something together or multiple people became ill. Analysis of the true positive test reviews with respect to these feature weights suggests that the classifier tends to select reviews that contain an abundance of *n*-grams about multiple people. Examination of these features shows that the *n*-gram model class is not sufficient for the “Multiple” task, indicated by its low performance relative to the “Sick” task and the need for detection of long phrases, which it cannot do. While it is tempting to simply extend the *n*-gram range to longer sequences, this approach fails due to a well-known statistical issue called “sparsity”: specific longer phrases become extremely rare in the data and are not seen in enough quantity for models to learn from them.

DISCUSSION

In this study, we have presented an automated text-classification system for the surveillance and detection of foodborne illness in online NYC restaurant reviews from Yelp. Using this system, NYC DOHMH epidemiologists are able to monitor millions of reviews, a previously impossible task, to aid in the identification and investigation of foodborne illness outbreaks in NYC. As of May 21, 2017, this system has been instrumental in the identification of 10 outbreaks and 8523 reports of foodborne illness associated with NYC restaurants since July 2012.

Aided by simple prototype classifiers, DOHMH epidemiologists have evaluated and labeled 13 526 Yelp reviews for 2 key indicators of foodborne illness since July 2012. Although these data are biased by the prototype classifier’s selection criterion, we showed how these biased data and additional complement-sampled data could be combined in a bias-adjusted training regime to build significantly higher-performing classifiers, an issue that commonly plagues deployed needle-in-a-haystack systems.

We evaluated the performance of our prototype classifiers and several other well-known classification models on 2 tasks, namely “Sick” and “Multiple.” We found that logistic regression trained with the “Silver” regime performed best for the “Sick” task and that the “Silver” “Sick-Pipelined” logistic regression performed best on the “Multiple” task, with bias-adjusted F1-scores of 87% and 66%, respectively.

As future work, we are currently exploring the use of modern deep learning techniques to further improve upon the classifiers by using soft measures of word similarity and models that are not limited to short contiguous spans of text, the key limitation found in the error analysis. We also intend to examine the performance of our system in locations outside of NYC.

This study is granted institutional review board exempt status under National Science Foundation grant IIS-15-63785, titled “III: Medium: Adaptive Information Extraction from Social Media for Actionable Inferences in Public Health.” Although the raw Yelp data are not publicly available, all code used to reproduce the final experiments in this manuscript can be found at https://github.com/teffland/FoodborneNYC/tree/master/jamia_2017/.

CONCLUSION

The importance of effective information extraction regarding foodborne illness from social media sites is increasing with the rising popularity of online restaurant review sites and the decreasing likelihood that younger people will report food poisoning via official government channels. In this investigation, we described details of the DOHMH system for foodborne illness surveillance in online restaurant reviews from Yelp. Our system has been instrumental in the identification of 10 outbreaks and 8523 reports of foodborne illness associated with NYC restaurants since July 2012. Our evaluation has identified strong classifiers for both tasks, whose deployment will allow DOHMH epidemiologists to more effectively monitor Yelp for improved foodborne illness investigations.

FUNDING

This work was supported by National Science Foundation grant IIS-15-63785, Google Research Award “Information Extraction from Social Media: Detecting Disease Outbreaks,” Alfred P. Sloan Foundation grant G-2015-14017, the Centers for Disease Control and Prevention PHEP grant NU90TP000546, and CDC/ELC grant NU50CK000407-03.

COMPETING INTERESTS

This material is based on work supported in part by a Google Research Award. In accordance with Columbia University reporting requirements, LG acknowledges ownership of Google stock as of the writing of this paper.

CONTRIBUTORS

Columbia author contributions

- TE: Designed and evaluated the alternative machine learning techniques for the classification of foodborne illness occurrence in Yelp restaurant reviews. Co-authored manuscript.
- AL: Designed and evaluated the alternative machine learning techniques for the classification of foodborne illness occurrence in Yelp restaurant reviews. Co-authored manuscript.
- LG: Coordinated the design and evaluation of the alternative machine learning techniques for the classification of foodborne illness

occurrence in Yelp restaurant reviews. Co-authored, critically reviewed, and provided extensive feedback on manuscript.

- DH: Coordinated the design and evaluation of the alternative machine learning techniques for the classification of foodborne illness occurrence in Yelp restaurant reviews. Co-authored, critically reviewed, and provided extensive feedback on manuscript.

DOHMH author contributions

- SB: Conceptualized and coordinated the incorporation of Yelp reviews into the DOHMH foodborne illness complaint system. Co-authored, critically reviewed, and provided extensive feedback on manuscript.
- VR: Conceptualized and coordinated the incorporation of Yelp reviews into the DOHMH foodborne illness complaint system. Oversaw the collection of feedback data and data cleaning. Co-authored, critically reviewed, and provided extensive feedback on manuscript.
- KD: Conducted literature review regarding other uses of social media to detect foodborne illness complaints and outbreaks. Co-authored, critically reviewed, and provided extensive feedback on manuscript.
- HW: Conceptualized and coordinated the incorporation of Yelp reviews into the DOHMH foodborne illness complaint system. Co-authored, critically reviewed, and provided extensive feedback on manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGEMENTS

We thank Fotis Psallidas for his contributions to the incorporation of Yelp reviews and the original machine learning classifiers into the DOHMH foodborne illness complaint system and helpful discussions. We thank Giannis Karamanolakis and Lampros Flokas for helping with data cleaning and validation of the results. We thank Yelp for providing the DOHMH with access to its raw feed of business reviews for New York City. We also thank Lan Li, Faina Stavinsky, Daniel O'Halloran, and Jazmin Fontenot for helpful discussions.

REFERENCES

1. Scallan E, Griffin PM, Angulo RV, *et al*. E. Foodborne illness acquired in the United States: unspecified agents. *Emerg Infect Dis*. 2011;17(1): 16–22.
2. Gould, LH, Walsh KA, Vieria AR, *et al*. Surveillance for foodborne disease outbreaks: United States, 1998–2008. *MMWR Surveill Summ*. 2013;62(2):1–34.
3. Santillana M, Nguyen AT, Dredze M, *et al*. Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS Comput Biol*. 2015;11(10):1–15.
4. Bahk, CY, Scales DA, Mekaru SR, *et al*. Comparing timeliness, content, and disease severity of formal and informal source outbreak reporting. *BMC Infect Dis*. 2015;15(135):1–6.
5. Freifeld CC, Mandl KD, Resi BY, *et al*. HealthMap: global infectious disease monitoring through automated classification and visualization of internet media reports. *J Am Med Inform Assoc*. 2008;15(2):150–57.
6. Harris JK, Mansour R, Choucair B, *et al*. Health department use of social media to identify foodborne illness: Chicago, Illinois, 2013–2014. *MMWR Morb Mortal Wkly Rep*. 2014;63(32):681–85.
7. Sadilek A, Kautz H, DiPrete L, *et al*. Deploying nEmesis: preventing foodborne illness by data mining social media. *Proc Conf AAAI Artif Intell*; February 12–17, 2016; Phoenix, Arizona; 3982–90.
8. Nsoesie EO, Kluberg SA, Brownstein JS. Online reports of foodborne illness capture foods implicated in official foodborne outbreak reports. *Prev Med*. 2014;67:264–69.
9. Harrison C, Jorder H, Stern F, *et al*. Using online reviews by restaurant patrons to identify unreported cases of food-borne illness: New York City, 2012–2013. *MMWR Morb Mortal Wkly Rep*. 2014;63(20):441–45.
10. Quinlan R. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers; 1993.
11. Leskovec J, Rajaraman A, Ullman JD. *Mining of Massive Datasets*. Cambridge: Cambridge University Press; 2014.
12. Cox DR. The regression analysis of binary sequences with discussion. *J R Stat Soc Series B Stat Methodol*. 1958;20:215–42.
13. Breiman L. Random forests. *Mach Learn*. 1997;45(1);5–32.
14. Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995;20(3):273–97.
15. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Boca Raton, FL: CRC Press; 1994.