nEmesis: Which Restaurants Should You Avoid Today?

Adam Sadilek*
Google
Mountain View, CA
sadilekadam@google.com

Sean Brennan
University of Rochester
Rochester, NY
sbrennan@u.rochester.edu

Henry Kautz
University of Rochester
Rochester, NY
kautz@cs.rochester.edu

Vincent Silenzio
University of Rochester
Rochester, NY
v.m.silenzio@rochester.edu

Abstract

Computational approaches to health monitoring and epidemiology continue to evolve rapidly. We present an end-to-end system, *nEmesis*, that automatically identifies restaurants posing public health risks. Leveraging a language model of Twitter users' online communication, nEmesis finds individuals who are likely suffering from a foodborne illness. People's visits to restaurants are modeled by matching GPS data embedded in the messages with restaurant addresses. As a result, we can assign each venue a "health score" based on the proportion of customers that fell ill shortly after visiting it. Statistical analysis reveals that our inferred health score correlates ($r = 0.30$) with the official inspection data from the Department of Health and Mental Hygiene (DOHMH). We investigate the joint associations of multiple factors mined from online data with the DOHMH violation scores and find that over 23% of variance can be explained by our factors. We demonstrate that readily accessible online data can be used to detect cases of foodborne illness in a timely manner. This approach offers an inexpensive way to enhance current methods to monitor food safety (*e.g.*, adaptive inspections) and identify potentially problematic venues in near-real time.

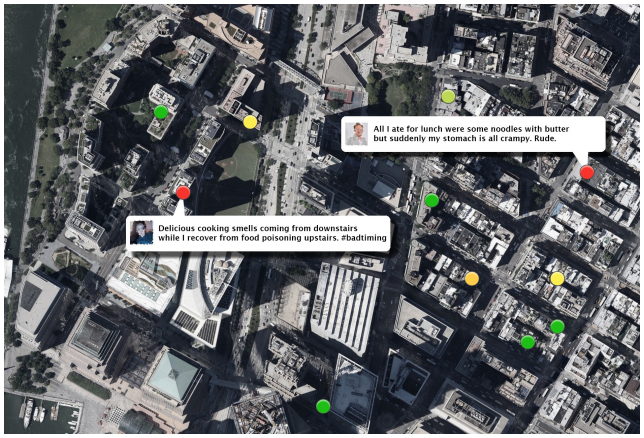

Figure 1: nEmesis analyses people's online messages and reveals individuals who may be suffering from a foodborne disease. Precise geo coordinates embedded in the messages enable us to detect specific restaurants a user had visited prior to falling ill. This figure shows a sample of users in New York City. Their most recent location is shown on the map and their likelihood of suffering from a foodborne illness is color-coded from low (green) to high (red). nEmesis enables tracking of possible health risks in a timely and scalable fashion.

Introduction

Every day, many people fall ill due to foodborne disease. Annually, three thousand of these patients die from the infection in the United States alone (CDC 2013). We argue in this paper that many of these occurrences are *preventable*. We present and validate nEmesis—a scalable approach to data-driven epidemiology that captures a large population with fine granularity and in near-real time. We are able to do this by leveraging vast sensor networks composed of users of online social media, who report—explicitly as well as implicitly—on their activities from their smart phones. We accept the inherent noise and ambiguity in people's online communication and develop statistical techniques that overcome some of the challenges in this space. As a result, nEmesis extracts important signals that enable individuals to make informed decisions (*e.g.*, "What is the probability that I will get sick if I eat lunch here?") and opens new opportunities for public health management (*e.g.*, "Given a limited budget, which restaurants should we inspect today?").

Recent work in computational epidemiology and machine learning has demonstrated that online social media enable novel surveillance and modeling tools (Lampos, De Bie, and Cristianini 2010; Paul and Dredze 2011a; Sadilek and Kautz 2013). Most research to date has focused on estimating aggregate "flu trends" in a large geographical area, typically at the national level. Researchers have shown that Internet data can be used to compute estimates of flu prevalence that correlate with the official Centers for Disease Control (CDC) statistics, but can be obtained in a more timely manner (Ginsberg et al. 2008; Signorini, Segre, and Polgreen 2011; Achrekar et al. 2012; Sadilek, Kautz, and Silenzio 2012b). Flu outbreaks can in some cases be even *predicted* by modeling the flow of infected airline passengers through their tweets (Brennan,

*Adam performed this work at the University of Rochester.

Sadilek, and Kautz 2013). This paper extends prior work beyond influenza-like disease, focusing on foodborne illness that afflicts *specific* individuals at *specific* venues.

The field of human computation (also referred to as crowdsourcing) has made significant progress in recent years (Kamar, Hacker, and Horvitz 2012). Along the way, it has been shown in a number of domains that the crowd can often act more effectively and accurately than even the best individual (*i.e.*, the "expert"). Successes with leveraging the crowd have influenced thinking within a wide range of disciplines, from psychology to machine learning, and include work on crowdsourcing diverse tasks such as text editing (Bernstein et al. 2010), image labeling (Von Ahn and Dabbish 2004), speech transcription (Lasecki et al. 2012), language translation (Shahaf and Horvitz 2010), software development (Little and Miller 2006), protein folding (Khatib et al. 2011), and providing new forms of accessibility for the disabled (Bigham et al. 2010).

This paper explores the intersection of three fields: human computation, machine learning, and computational epidemiology. We focus on real-time modeling of foodborne illness—a significant health challenge in the developing and developed world. Harnessing human and machine intelligence in a unified way, we develop an automated language model that detects individuals who likely suffer from a foodborne disease, on the basis of their online Twitter communication. By leveraging the global positioning system (GPS) data of each Twitter user and known locations of every restaurant in New York City (NYC), we detect users' restaurant visits preceding the onset of a foodborne illness. As a result, we can assign each restaurant a "health score" based on the proportion of Twitter customers that fell ill shortly after visiting the restaurant.

As we will see, our inferred health score correlates ($r = 0.30$, p-value of 6×10^{-4}) with the official inspection data from the Department of Health and Mental Hygiene (DOHMH). Additionally, we investigate the joint effect of multiple factors mined from online data on the DOHMH violation scores and find that over 23% of variance in the official statistics can be explained by factors inferred from online social media.

Achieving these encouraging results would be difficult without *joint* human and machine effort. Humans could not keep up with the average rate of 9,100 tweets per second that are produced globally,[1] resulting in very sparsely labeled data. Since foodborne illness is (fortunately) rare, even 99% coverage would not be enough to get a reliable signal. At the same time, the complexity of natural language would prevent machines from making sense of the data. While machines can easily provide full coverage, the signal to noise ratio would be too low to maintain adequate sensitivity and specificity. We show in this paper that including human workers and machines in a common loop cancels each others' weaknesses and results in a reliable model of foodborne disease.

[1] http://www.statisticbrain.com/twitter-statistics/

Significance of Results

We harness human computation on two different levels. One is the aforementioned *explicit* crowdsourcing of data labeling by online workers. The second—more subtle—level leverages the *implicit* human computation performed by hundreds of millions of users of online social media every day. These users make up an "organic" sensor network—a dynamic mesh of sensors interconnected with people facilitated by Internet-enabled phones. A single status update often contains not only the text of the message itself, but also location, a photo just taken, relationships to other people, and other information. The text contains a nugget of human computation as well—describing what the person thought or saw.

This paper concentrates on extracting useful and dependable signals from snippets of human computation that users perform every time they post a message. We do this via ambient tracking and inference over online data. The inference itself is in part enabled by explicit crowdsourcing.

It is essential to capture the organic sensor network computationally. A single user complaining about acute food poisoning has a small impact on the behavior of others. Even messages from very popular individuals (barring celebrities) reach relatively few followers. However, an automated system like nEmesis that tracks a large online population can find important patterns, even when they require stitching together subtle signals from low-profile users. By placing the signal in context (*e.g.*, by matching the message with a relevant restaurant), a seemingly random collection of online rants suddenly becomes an actionable alert.

We believe the pervasiveness of Internet-enabled mobile devices has reached a critical point that enables novel applications that help people make more *informed* decisions. nEmesis is one specific example of such an application.

In the remainder of the paper, we will discuss the broader context of this research, describe in detail our methodology and models, report key findings, and discuss the results.

Background and Related Work

Twitter is a widely used online social network and a particularly popular source of data for its real-time nature and open access (Smith 2011). Twitter users post message updates (tweets) up to 140 characters long. Twitter launched in 2006 and has been experiencing an explosive growth since then. As of April 2012, over 500 million accounts were registered on Twitter.

Researchers have shown that Twitter data can be used not only for flu tracking, but also for modeling mental health (Golder and Macy 2011; De Choudhury et al. 2013), and general public health (Paul and Dredze 2011b). Much work has been done outside the medical domain as well. Twitter data has been leveraged to predict movie box office revenues (Asur and Huberman 2010), election outcomes (Tumasjan et al. 2010), and other phenomena. Globally, the prevalence of social media usage is significant, and is increasing: 13% of online adults use Twitter, most of them daily and often via a phone (Smith 2011). These mobile users often attach their current GPS location to each tweet, thereby creating rich datasets of human mobility and interactions.

Foodborne illness, also known colloquially as food poisoning, is any illness resulting from the consumption of pathogenic bacteria, viruses, or parasites that contaminate food, as well as the consumption of chemical or natural toxins, such as poisonous mushrooms. The most common symptoms include vomiting, diarrhea, abdominal pain, fever, and chills. These symptoms can be mild to serious, and may last from hours to several days. Typically, symptoms appear within hours, but may also occur days or even weeks after exposure to the pathogen (J Glenn Morris and Potter 2013). Some pathogens can also cause symptoms of the nervous system, including headache, numbness or tingling, blurry vision, weakness, dizziness, and even paralysis. According to the U.S. Food and Drug Administration (FDA), the vast majority of these symptoms will occur within three days (FDA 2012).

The CDC estimates that 47.8 million Americans (roughly 1 in 6 people) are sickened by foodborne disease every year. Of that total, nearly 128,000 people are hospitalized, while just over 3,000 die of foodborne diseases (CDC 2013). The CDC classifies cases of foodborne illness according to whether they are caused by one of 31 *known foodborne illness pathogens* or by *unspecified agents*. The known pathogens account for 9.4 million (20% of the total) cases of food poisoning each year, while the remaining 38.4 million cases (80% of the total) are caused by unspecified agents. Of the 31 known pathogens, the top five (Norovirus, *Salmonella, Clostridium perfringens, Campylobacter* species, and *Staphylococcus aureus*) account for 91% of the cases (CDC 2013). The economic burden of health losses resulting from foodborne illness are staggering—$78 billion annually in the U.S. alone (Scharff 2012).

Public health authorities use an array of surveillance systems to monitor foodborne illness. The CDC relies heavily on data from state and local health agencies, as well as more recent systems such as sentinel surveillance systems and national laboratory networks, which help improve the quality and timeliness of data (CDC 2013). The NYC Department of Health carries out unannounced sanitary inspections. Each restaurant in NYC is inspected at least once a year and receives a *violation score* (higher score means more problems recorded by the inspector) (Farley 2011).

An example of the many systems in use by CDC would include the Foodborne Diseases Active Surveillance Network, referred to as FoodNet. FoodNet is a sentinel surveillance system using information provided from sites in 10 states, covering about 15% of the US population, to monitor illnesses caused by seven bacteria or two parasites commonly transmitted through food. Other systems include the National Antimicrobial Resistance Monitoring Systementeric bacteria (NARMS), the National Electronic Norovirus Outbreak Network (CaliciNet), and the National Molecular Subtyping Network for Foodborne Disease Surveillance (PulseNet), among many others.

A major challenge in monitoring foodborne illness is in capturing actionable data in real time. Like all disease surveillance programs, each of the systems currently in use by CDC to monitor foodborne illness entails significant costs and time lags between when cases are identified and the data is analyzed and reported.

Support vector machine (SVM) is an established model of data in machine learning (Cortes and Vapnik 1995). We learn an SVM for linear binary classification to accurately distinguish between tweets indicating the author is afflicted by foodborne disease and all other tweets. Linear binary SVMs are trained by finding a hyperplane defined by a normal vector with the maximal margin separating it from the positive and negative datapoints.

Finding such a hyperplane is inherently a quadratic optimization problem given by the following objective function that can be solved efficiently and in a parallel fashion using stochastic gradient descent methods (Shalev-Shwartz, Singer, and Srebro 2007).

$$\min_w \frac{\lambda}{2}||w||^2 + \mathcal{L}(w, D) \tag{1}$$

where λ is a regularization parameter controlling model complexity, and $\mathcal{L}(w, D)$ is the hinge-loss over all training data D given by

$$\mathcal{L}(w, D) = \sum_i \max\left(0, 1 - y_i w^T x_i\right) \tag{2}$$

Class imbalance, where the number of examples in one class is dramatically larger than in the other class, complicates virtually all machine learning. For SVMs, prior work has shown that transforming the optimization problem from the space of individual datapoints $\langle x_i, y_i \rangle$ in matrix D to one over *pairs* of examples $\left\langle x_i^+ - x_j^-, 1 \right\rangle$ yields significantly more robust results (Joachims 2005).

Active learning is a machine learning approach, where the training data is provided *adaptively*. The model we are inducing typically ranks unlabeled data according to the expected information gain and requests labels for top-k examples, given budget constraints (Settles 2010). The labels are typically provided by a single human expert. In a number of domains, active learning has been repeatedly shown to achieve the same level of model quality while requiring only a fraction of (often exponentially less) labeled data, as compared to nonadaptive ("label all") learning approaches (Cohn, Atlas, and Ladner 1994).

Methods

This section describes in detail our method of leveraging human and machine computation to learn an accurate language model of foodborne disease, which is subsequently used to detect restaurants that could pose health risks. We begin by describing our data collection system, then turn to our active data labeling framework that leverages human as well as machine intelligence, and finally concentrate on the induction and application of the language model itself.

Data Collection

We have obtained a database of all restaurant inspections conducted by the Department of Health and Mental Hygiene in New York City. A total of 24,904 restaurants have been recently inspected at least once and appear in the database.

As each inspection record contains the name and address of the restaurant, we used Google Maps[2] to obtain exact GPS

[2]https://developers.google.com/maps/documentation/geocoding/

coordinates for each venue. We then use the location to tie together users and restaurants in order to estimate *visits*. We say that a user visited a restaurant if he or she appeared within 25 meters of the venue at a time the restaurant was likely open, considering typical operating hours for different types of food establishments.

Since foodborne disease is not necessarily contracted at a venue already recorded in the DOHMH database, future work could explore the interesting problem of finding *undocumented* venues that pose health hazards. This could be done by analyzing visits that appear to be—at first sight—false negatives. As the food industry is becoming increasingly mobile (*e.g.*, food trucks and hot dog stands), its health implications are more difficult to capture. We believe online systems based on methods presented in this paper will be an important component of future public health management.

Using the Twitter Search API[3], we collected a sample of public tweets that originated from the New York City metropolitan area. The collection period ran from December 26, 2012 to April 25, 2013. We periodically queried Twitter for all recent tweets within 100 kilometers of the NYC city center in a distributed fashion.

Twitter users may alternate between devices, not necessarily publishing their location every time. Whenever nEmesis detects a person visiting a restaurant it spawns a separate data collection process that listens for new tweets from that person. This captures scenarios where someone tweets from a restaurant using a mobile device, goes home, and several hours later tweets from a desktop (without GPS) about feeling ill.

The GPS noise could lead to false positive as well as false negative visits. We validate our visit detector by analyzing data for restaurants that have been closed by DOHMH because of severe health violations. A significant drop in visits occurs in each venue after its closure. Furthermore, some users explicitly "check-in" to a restaurant using services such as FourSquare that are often tied to a user's Twitter account. As each check-in tweet contains venue name and a GPS tag, we use them to validate our visit detector. 97.2% of the explicit 4,108 restaurant check-ins are assigned to the correct restaurant based on GPS alone.

Altogether, we have logged over 3.8 million tweets authored by more than 94 thousand unique users who produced at least one GPS-tagged message. Out of these users, over 23 thousand visited at least one restaurant during the data collection period. We did not consider users who did not share any location information as we cannot assign them to restaurants. To put these statistics in context, the entire NYC metropolitan area has an estimated population of 19 million people.[4] Table 1 summarizes our dataset.

Restaurants in DOHMH inspection database	24,904
Restaurants with at least one Twitter visit	17,012
Restaurants with at least one *sick* Twitter visit	120
Number of tweets	3,843,486
Number of detected *sick* tweets	1,509
***Sick* tweets associated with a restaurant**	479
Number of unique users	94,937
Users who visited at least one restaurant	23,459

Table 1: Summary statistics of the data collected from NYC. Note that nearly a third of the messages indicating foodborne disease can be traced to a restaurant.

Labeling Data at Scale

To scale the laborious process of labeling training data for our language model, we turn to Amazon's Mechanical Turk.[5] Mechanical Turk allows requesters to harness the power of the crowd in order to complete a set of human intelligence tasks (HITs). These HITs are then completed online by hired workers (Mason and Suri 2012).

We formulated the task as a series of short surveys, each 25 tweets in length. For each tweet, we ask "Do you think the author of this tweet has an upset stomach today?". There are three possible responses ("Yes", "No", "Can't tell"), out of which a worker has to choose exactly one.

We paid the workers 1 cent for every tweet evaluated, making each survey 25 cents in total. Each worker was allowed to label a given tweet only once. The order of tweets was randomized. Each survey was completed by exactly five workers independently. This redundancy was added to reduce the effect of workers who might give erroneous or outright malicious responses. Inter-annotator agreement measured by Cohen's κ is 0.6, considered a moderate to substantial agreement in the literature (Landis and Koch 1977).

For each tweet, we calculate the final label by adding up the five constituent labels provided by the workers (Yes$=1$, No$=-1$, Can't tell$=0$). In the event of a tie (0 score), we consider the tweet healthy in order to obtain a high-precision dataset.

Human Guided Machine Learning. Given that tweets indicating foodborne illness are relatively rare, learning a robust language model poses considerable challenges (Japkowicz and others 2000; Chawla, Japkowicz, and Kotcz 2004). This problem is called *class imbalance* and complicates virtually all machine learning. In the world of classification, models induced in a skewed setting tend to simply label all data as members of the majority class. The problem is compounded by the fact that the minority class (sick tweets) are often of greater interest than the majority class.

We overcome class imbalance faced by nEmesis through a combination of two techniques: human guided active learning, and learning a language model that is robust under class imbalance. We cover the first technique in this section and discuss the language model induction in the following section.

Previous research has shown that under extreme class imbalance, simply *finding* examples of the minority class and providing them to the model at learning time significantly improves the resulting model quality and reduces human labeling cost (Attenberg and Provost 2010). In this work, we present a novel, scalable, and fully automated learning method—called *human guided machine learning*—that considerably reduces the amount of human effort required to reach any given level of model quality, even when the num-

[3] http://search.twitter.com/api/

[4] http://www.census.gov/popest/metro/

[5] https://www.mturk.com/

ber of negatives is many orders of magnitude larger than the number of positives. In our domain, the ratio of sick to healthy tweets is roughly 1:2,500.

In each human guided learning iteration, nEmesis samples representative and informative examples to be sent for human review. As the focus is on the minority class examples, we sample 90% of tweets for a given labeling batch from the top 10% of the most likely sick tweets (as predicted by our language model). The remaining 10% is sampled uniformly at random to increase diversity. We use the HITs described above to obtain the labeled data.

In parallel with this automated process, we hire workers to actively find examples of tweets in which the author indicates he or she has an upset stomach. We asked them to paste a direct link to each tweet they find into a text box. Workers received a base pay of 10 cents for accepting the task, and were motivated by a bonus of 10 cents for each unique relevant tweet they provided. Each wrong tweet resulted in a 10 cent deduction from the current bonus balance of a worker. Tweets judged to be too ambiguous were neither penalized nor rewarded. Overall, we have posted 50 HITs that resulted in 1,971 submitted tweets (mean of 39.4 per worker). Removing duplicates yielded 1,176 unique tweets.

As a result, we employ human workers that "guide" the classifier induction by correcting the system when it makes erroneous predictions, and proactively seeking and labeling examples of the minority classes. Thus, people and machines work together to create better models faster.

In the following section, we will see how a combination of human guided learning and active learning in a loop with a machine model leads to significantly improved model quality.

Learning Language Model of Foodborne Illness

As a first step in modeling potentially risky restaurants, we need to identify Twitter messages that indicate the author is afflicted with a foodborne disease at the time of posting the message. Recall that these messages are rare within the massive stream of tweets.

We formulate a semi-supervised cascade-based approach to learning a robust support vector machine (SVM) classifier with a large area under the ROC curve (*i.e.*, consistently high precision and high recall). We learn an SVM for linear binary classification to accurately distinguish between tweets indicating the author is afflicted by foodborne illness (we call such tweets "sick"), and all other tweets (called "other" or "normal").

In order to learn such a classifier, we ultimately need to effortlessly obtain a high-quality set of labeled training data. We achieve this via the following "bootstrapping" process, shown in Fig. 2.

We begin by creating a simple keyword-matching model in order to obtain a large corpus of tweets that are potentially relevant to foodborne illness. The motivation is to produce an initial dataset with relatively high recall, but low precision that can be subsequently refined by a combination of human and machine computation. The keyword model contains 27 regular expressions matching patterns such as "stomach ache", "throw up", "Mylanta", or "Pepto Bismol". Each regular expression matches many variations on a given phrase, accounting for typos and common misspellings, capitalization, punctuation, and word boundaries. We created the list of patterns in consultation with a medical expert, and referring to online medical ontologies, such as WebMD.com, that curate information on diagnosis, symptoms, treatments, and other aspects of foodborne illness.

Each tweet in our corpus C containing 3.8 million collected tweets is ranked based on how many regular expressions match it (step 1 in Fig. 2). We then take the top 5,800 tweets along with a uniform sample of 200 tweets and submit a HIT to label them, as described in the previous section. This yields a high-quality corpus of 6,000 labeled tweets (step 2).

We proceed by training two different binary SVM classifiers, M_s and M_o, using the SVM^{light} package (step 3).[6] M_s is highly penalized for inducing false positives (mistakenly labeling a normal tweet as one about sickness), whereas M_o is heavily penalized for creating false negatives (labeling symptomatic tweets as normal). We train M_s and M_o using the dataset of 6,000 tweets, each labeled as either "sick" or "other". We then select the bottom 10% of the scores predicted by M_o (*i.e.*, tweets that are normal with high probability), and the top 10% of scores predicted by M_s (*i.e.*, likely "sick" tweets).

The intuition behind this cascading process is to extract tweets that are with high confidence about sickness with M_s, and tweets that are almost certainly about other topics with M_o from the corpus C. We further supplement the final corpus with messages from a sample of 200 million tweets (disjoint from C) that M_o classified as "other" with high probability. We apply thresholding on the classification scores to reduce the noise in the cascade.

At this point, we begin to iterate the human guided active learning loop shown in the gray box in Fig. 2. The cycle consists of learning an updated model M from available training data (step 4), labeling new examples, and finally using our active learning strategy described above to obtain labeled tweets from human workers (steps 5 and 6). This process is repeated until sufficient model quality is obtained, as measured on an independent evaluation set.

As features, the SVM models use all uni-gram, bi-gram, and tri-gram word tokens that appear in the training data. For example, a tweet *"My tummy hurts."* is represented by the following feature vector:

$$\Big(\textit{my}, \textit{tummy}, \textit{hurts}, \textit{my tummy}, \textit{tummy hurts}, \textit{my tummy hurts}\Big).$$

Prior to tokenization, we convert all text to lower case and strip punctuation. Additionally, we replace mentions of user names (the "@" tag) with a special `@MENTION` token, and all web links with a `@LINK` token. We do keep hashtags (such as #upsetstomach), as those are often relevant to the author's health state, and are particularly useful for disambiguation of short or ill-formed messages. When learning the final SVM M, we only consider tokens that appear at least three times in the training set. Table 2 lists the most significant positive and negative features M found.

While our feature space has a very high dimensionality (M operates in more than one million dimensions), with many

[6] http://svmlight.joachims.org/

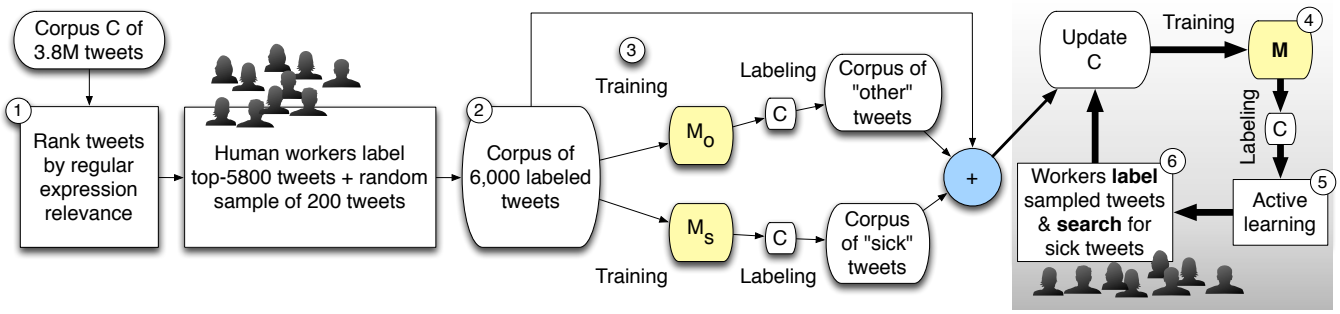

Figure 2: A diagram of our cascade learning of SVMs. Human computation components are highlighted with crowds of people. All other steps involve machine computation exclusively. The dataset C contains our 3.8 million tweets from NYC that are relevant to restaurants.

Positive Features		Negative Features	
Feature	**Weight**	**Feature**	**Weight**
stomach	1.7633	think i'm sick	−0.8411
stomachache	1.2447	i feel soooo	−0.7156
nausea	1.0935	fuck i'm	−0.6393
tummy	1.0718	@MENTION sick to	−0.6212
#upsetstomach	0.9423	sick of being	−0.6022
nauseated	0.8702	ughhh cramps	−0.5909
upset	0.8213	cramp	−0.5867
nautious	0.7024	so sick omg	−0.5749
ache	0.7006	tired of	−0.5410
being sick man	0.6859	cold	−0.5122
diarrhea	0.6789	burn sucks	−0.5085
vomit	0.6719	course i'm sick	−0.5014
@MENTION i'm getting	0.6424	if i'm	−0.4988
#tummyache	0.6422	is sick	−0.4934
#stomachache	0.6408	so sick and	−0.4904
i've never been	0.6353	omg i am	−0.4862
threw up	0.6291	@LINK	−0.4744
i'm sick great	0.6204	@MENTION sick	−0.4704
poisoning	0.5879	if	−0.4695
feel better tomorrow	0.5643	i feel better	−0.4670

Table 2: Top twenty most significant negatively and positively weighted features of our SVM model M.

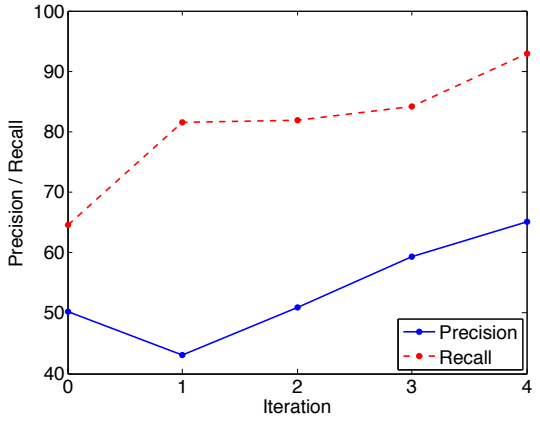

Figure 3: Precision and recall curves as we increase the number of iterations of the SVM pipeline shown in Fig. 2. Iteration 0 shows the performance of M trained with only the initial set of 6,000 tweets. In iteration 1, M is additionally trained with a sample of "other" tweets. We see that recall improves dramatically as the model experienced a wide variety of examples, but precision drops. Subsequent iterations (2-4) of the human guided machine learning loop yield significant improvement in both recall and precision, as workers search for novel examples and validate tweets suggested by the machine model.

possibly irrelevant features, support vector machines with a linear kernel have been shown to perform very well under such circumstances (Joachims 2006; Sculley et al. 2011; Paul and Dredze 2011a).

In the following section, we discuss how we apply the language model M to independently score restaurants in terms of the health risks they pose, and compare our results to the official DOHMH inspection records.

Results

We begin by annotating all tweets relevant to restaurant visits with an estimated likelihood of foodborne illness, using the language model M learned in the previous section. Fig. 3 shows the precision and recall of the model as we iterate through the pipeline in Fig. 2. The model is always evaluated on a static independent held-out set of 1,000 tweets. The model M achieves 63% precision and 93% recall after the final learning iteration. Only 9,743 tweets were adaptively labeled by human workers to achieve this performance: 6,000 for the initial model, 1,176 found independently by human computation, and 2,567 labeled by workers as per M's request. The total labeling cost was below $1,500. The speed with which workers completed the tasks suggests that we have been overpaying them, but our goal was not to minimize human work costs. We see in Fig. 3 that the return of investment on even small amounts of adaptively labeled examples is large in later iterations of the nEmesis pipeline.

Using Twitter data annotated by our language model and matched with restaurants, we calculate a number of features for each restaurant. The key metric for a restaurant x is the fraction of Twitter visitors that indicate foodborne illness within 100 hours after appearing at x. This threshold is selected in order to encompass the mean onset of the majority of foodborne illness symptoms (roughly 72 hours after ingestion) (FDA 2012). We denote this quantity by $f(x)$ or, in

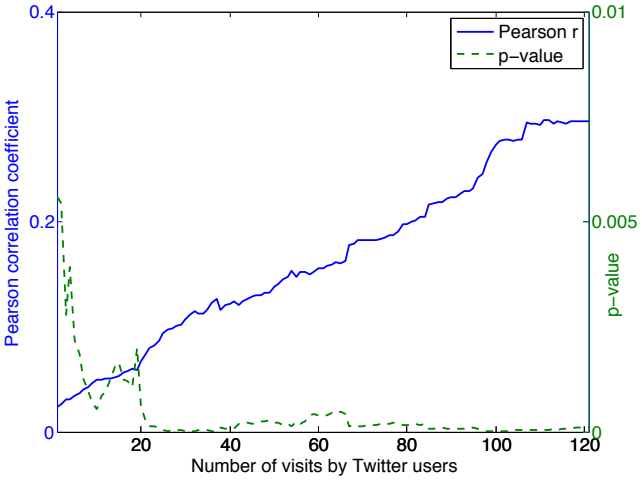

Figure 4: We obtain increasingly stronger signal as we concentrate on restaurants with larger amounts of associated Twitter data. Pearson correlation coefficient increases linearly as we consider venues with at least n visits recorded in the data (horizontal axis). At the same time, the correlation is increasingly significant in terms of p-value as we observe more data. Note that even sparsely represented restaurants (*e.g.*, with one recorded visit) exhibit weak, but significant correlation.

general, as function f when we do not refer to any specific restaurant.

As a first validation of f, we correlate it with the official inspection score s extracted from the DOHMH database. A restaurant may have been inspected multiple times during our study time period. To create a single score $s(x)$, we calculate the arithmetic mean of x's violation scores between December 2012 to April 2013. Fig. 4 shows Pearson correlation between f and s as a function of the density of available Twitter data. The horizontal axis shows the smallest number of Twitter visits a restaurant has to have in order to be included in the correlation analysis.

We see that the correlation coefficient increases from $r = 0.02$ (p-value of 5.6×10^{-3}) to $r = 0.30$ (p-value of 6×10^{-4}) when we look at restaurants with a sufficient number of visits. The signal is weak, but significant, for restaurants where we observe only a few visits. Moreover, the p-value becomes increasingly significant as we get more data.

Focusing on restaurants with more than 100 visits (there are 248 such restaurants in our dataset), we explore associations between s and additional signals mined from Twitter data (beyond f). Namely, we observe that the number of visits to a restaurant declines as s increases (*i.e.*, more violations): $r = -0.27$ (p-value of 3.1×10^{-4}). Similarly, the number of distinct visitors decreases as s increases: $r = -0.17$ (p-value of 3.0×10^{-2}). This may be a result of would-be patrons noticing a low health score that restaurants are required to post at their entrance.

We consider alternative measures to f as well. The absolute number of sick visitors is also strongly associated with s: $r = 0.19$ (p-value of 9.5×10^{-3}). Note that this association is not as strong as for f. Finally, we can count the number of consecutive *sick days* declared by Twitter users after visiting a restaurant. A sick day of a user is defined as one in which the user posted at least one sick tweet. We find similarly strong association with s here as well: $r = 0.29$ (p-value of 10^{-4}).

We do not adjust f by the number of restaurants the users visited, as most ill individuals do not appear in multiple restaurants in the same time frame. In general, however, adjusting up as well as down could be appropriate. In one interpretation, a sick patron himself contributes to the germs in the restaurants he visits (or happens to have preferences that consistently lead him to bad restaurants). Thus, his contribution should be adjusted up. In a more common scenario, there is a health hazard within the restaurant itself (such as insufficient refrigeration) that increases the likelihood of foodborne illness. If a person had visited multiple venues before falling ill, the probability mass should be spread among them, since we do not know a priori what subset of the visits caused the illness. A unified graphical model, such as a dynamic Bayesian network, over users and restaurants could capture these interactions in a principled way. The network could model uncertainty over user location as well. This is an intriguing direction for future research.

Our final validation involves comparison of two distributions of s: one for restaurants with $f > 0$ (*i.e.*, we have observed at least one user who visited the establishment and indicated sickness afterwards) and one for restaurants with $f = 0$ (no Twitter evidence of foodborne disease). We call the first multi-set of restaurant scores $S_{e=1} = \{s(x) : f(x) > 0\}$ and the second $S_{e=0} = \{s(x) : f(x) = 0\}$.

Fig. 5 shows that restaurants in set $S_{e=1}$ (where we detect sick users) have significantly worse distribution of health violation scores than places where we do not observe anybody sick ($S_{e=0}$). Nonparametric Kolmogorov-Smirnov test shows that the two distributions are significantly different (p-value of 1.5×10^{-11}). Maximum-likelihood estimate shows that both distributions are best approximated with the log-normal distribution family.

When we use a language model for tweets about influenza-like disease (*i.e.*, instead of a model specific to foodborne disease) developed in Sadilek, Kautz, and Silenzio (2012a), the signal nearly vanishes. Namely, we define a new quantity, f^I, as an analog to f. $f^I(x)$ denotes the fraction of Twitter visitors that indicate an influenza-like illness within 100 hours after appearing at a given restaurant x. Pearson correlation coefficient between f^I and s is $r = 0.002$ (p-value of 1.9×10^{-4}). This demonstrates the importance of using a language model specific to foodborne illness rather than general sickness reports.

Finally, we perform multiple linear regression analysis to model the joint effects of the features we infer from Twitter data. Specifically, we learn a model of the DOHMH violation score $s(x)$ for restaurant x as a weighted sum of our features a_i with additional constant term c and an error term ϵ: $s(x) = c + \sum_i w_i a_i(x) + \epsilon$.

Table 3 lists all features and their regression coefficient. As we would expect from our analysis of correlation coefficients above, the proportion of sick visitors (f) is the most dominant feature that contributes to an increased violation

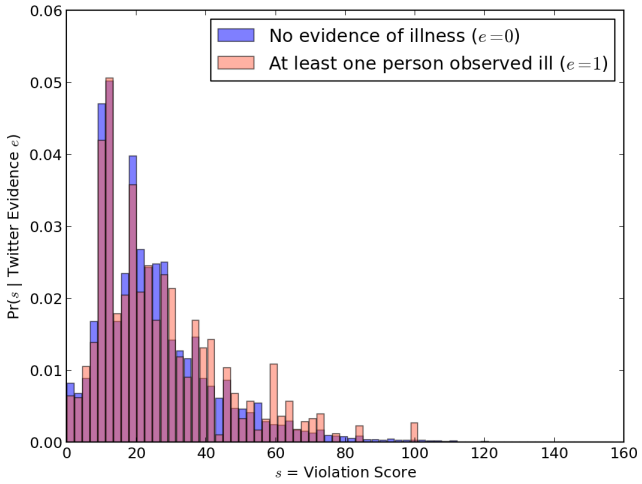

Figure 5: Probability distributions over violation scores (higher is worse) for restaurants, where we have not observed evidence of illness ($Pr(s \mid e = 0)$; blue), and restaurants in which we observed at least one individual who subsequently became ill ($Pr(s \mid e = 1)$; orange). Nonparametric Kolmogorov-Smirnov test shows that the two distributions are significantly different (p-value of 1.5×10^{-11}).

Feature	Regression Coefficient
Constant term c	$+16.1585$ ***
Number of visits	-0.0015 ***
Number of distinct visitors	-0.0014 ***
Number of *sick* visitors (f^T)	$+3.1591$ ***
Proportion of *sick* visitors (f)	$+19.3370$ ***
Number of sick days of visitors	0 ***

Table 3: Regression coefficients for predicting s, the DOHMH violation score, from Twitter data. *** denotes statistical significance with p-value less than 0.001.

score, followed by the absolute number of sick visitors (f^T). Interestingly, the number of sick days explains no additional variance in s. This may reflect the fact that typical episodes of foodborne illness commonly resolve within a single day (*e.g.*, the proverbial "24-hour bug").

The effect of the observed number of visits and the number of distinct visitors is significantly weaker in the regression model than in correlation analysis—suggesting that the health states of the visitors indeed do explain most of the signal. Overall, we find that 23.36% of variance in s is explained by our factors mined from Twitter data (shown in Table 3).

Conclusions and Future Work

We present nEmesis, an end-to-end system that "listens" for relevant public tweets, detects restaurant visits from geotagged Twitter messages, tracks user activity following a restaurant visit, infers the likelihood of the onset of foodborne illness from the text of user communication, and finally ranks restaurants via statistical analysis of the processed data.

To identify relevant posts, we learn an automated language model through a combination of machine learning and human computation. We view Twitter users as noisy sensors and leverage their *implicit* human computation via ambient tracking and inference, as well as their *explicit* computation for data exploration and labeling. Humans "guide" the learning process by correcting nEmesis when it makes erroneous predictions, and proactively seek and label examples of sick tweets. Thus, people and machines work together to create better models faster.

While nEmesis' predictions correlate well with official statistics, we believe the most promising direction for future work is to address the *discrepancy* between these two fundamentally different methodologies of public health management: analysis of noisy real-time data, and centralized inspection activity. Our hope is that the *unification* of traditional techniques and scalable data mining approaches will lead to better models and tools by mitigating each others' weaknesses.

As we have discussed throughout this paper, the most daunting challenge of online methods is data incompleteness and noise. We have presented machine learning techniques that at least partially overcome this challenge. At the same time, one of the strong aspects of systems like nEmesis is their ability to measure the signal of interest more directly and at scale. While DOHMH inspections capture a wide variety of data that is largely impossible to obtain from online social media or other sources (such as the presence of rodents in a restaurant's storage room), our Twitter signal measures a perhaps more actionable quantity: a probability estimate of *you becoming ill* if you visit a particular restaurant.

DOHMH inspections are thorough, but largely sporadic. A cook who occasionally comes to work sick and infects customers for several days at a time is unlikely to be detected by current methods. Some individuals may even be unaware they are causing harm (*e.g.*, "Typhoid Mary"). Similarly, a batch of potentially dangerous beef delivered by a truck with faulty refrigeration system could be an outlier, but nonetheless cause loss of life.

nEmesis has the potential to complement traditional methods and produce a more comprehensive model of public health. For instance, *adaptive* inspections guided, in part, by real-time systems like nEmesis now become possible.

Acknowledgments

We thank the anonymous reviewers for their insightful feedback. This research was supported by grants from ARO (W911NF-08-1-024) ONR (N00014-11-10417), NSF (IIS-1012017), NIH (1R01GM108337-01), and the Intel Science & Technology Center for Pervasive Computing.

References

Achrekar, H.; Gandhe, A.; Lazarus, R.; Yu, S.; and Liu, B. 2012. Twitter improves seasonal influenza prediction. *Fifth Annual International Conference on Health Informatics*.

Asur, S., and Huberman, B. 2010. Predicting the future with social media. In *WI-IAT*, volume 1, 492–499. IEEE.

Attenberg, J., and Provost, F. 2010. Why label when you can search?: Alternatives to active learning for applying human resources to build classification models under extreme class imbalance. In *SIGKDD*, 423–432. ACM.

Bernstein, M.; Little, G.; Miller, R.; Hartmann, B.; Ackerman, M.; Karger, D.; Crowell, D.; and Panovich, K. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, 313–322. ACM.

Bigham, J.; Jayant, C.; Ji, H.; Little, G.; Miller, A.; Miller, R.; Miller, R.; Tatarowicz, A.; White, B.; White, S.; et al. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, 333–342. ACM.

Brennan, S.; Sadilek, A.; and Kautz, H. 2013. Towards understanding global spread of disease from everyday interpersonal interactions. In *Twenty-Third International Conference on Artificial Intelligence (IJCAI)*.

CDC. 2013. Estimates of Foodborne Illness in the United States.

Chawla, N.; Japkowicz, N.; and Kotcz, A. 2004. Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter* 6(1):1–6.

Cohn, D.; Atlas, L.; and Ladner, R. 1994. Improving generalization with active learning. *Machine Learning* 15(2):201–221.

Cortes, C., and Vapnik, V. 1995. Support-vector networks. *Machine learning* 20(3):273–297.

De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2013. Predicting depression via social media. *AAAI Conference on Weblogs and Social Media*.

Farley, T. 2011. Restaurant grading in New York City at 18 months. *http://www.nyc.gov*.

FDA. 2012. *Bad Bug Book*. U.S. Food and Drug Administration, 2nd edition.

Ginsberg, J.; Mohebbi, M.; Patel, R.; Brammer, L.; Smolinski, M.; and Brilliant, L. 2008. Detecting influenza epidemics using search engine query data. *Nature* 457(7232):1012–1014.

Golder, S., and Macy, M. 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* 333(6051):1878–1881.

J Glenn Morris, J., and Potter, M. 2013. *Foodborne Infections and Intoxications*. Food Science and Technology. Elsevier Science.

Japkowicz, N., et al. 2000. Learning from imbalanced data sets: a comparison of various strategies. In *AAAI workshop on learning from imbalanced data sets*, volume 68.

Joachims, T. 2005. A support vector method for multivariate performance measures. In *ICML 2005*, 377–384. ACM.

Joachims, T. 2006. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 217–226. ACM.

Kamar, E.; Hacker, S.; and Horvitz, E. 2012. Combining human and machine intelligence in large-scale crowdsourcing. In *International Conference on Autonomous Agents and Multiagent Systems*, 467–474.

Khatib, F.; Cooper, S.; Tyka, M. D.; Xu, K.; Makedon, I.; Popović, Z.; Baker, D.; and Players, F. 2011. Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences* 108(47):18949–18953.

Lampos, V.; De Bie, T.; and Cristianini, N. 2010. Flu detector-tracking epidemics on Twitter. *Machine Learning and Knowledge Discovery in Databases* 599–602.

Landis, J. R., and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *biometrics* 159–174.

Lasecki, W. S.; Miller, C. D.; Sadilek, A.; Abumoussa, A.; Borrello, D.; Kushalnagar, R.; and Bigham, J. P. 2012. Real-time captioning by groups of non-experts. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*, UIST '12.

Little, G., and Miller, R. 2006. Translating keyword commands into executable code. In *Proceedings of the 19th annual ACM symposium on User interface software and technology*, 135–144. ACM.

Mason, W., and Suri, S. 2012. Conducting behavioral research on amazons mechanical turk. *Behavior research methods* 44(1):1–23.

Paul, M., and Dredze, M. 2011a. A model for mining public health topics from Twitter. *Technical Report. Johns Hopkins University. 2011*.

Paul, M., and Dredze, M. 2011b. You are what you tweet: Analyzing Twitter for public health. In *Fifth International AAAI Conference on Weblogs and Social Media*.

Sadilek, A., and Kautz, H. 2013. Modeling the impact of lifestyle on health at scale. In *Sixth ACM International Conference on Web Search and Data Mining*.

Sadilek, A.; Kautz, H.; and Silenzio, V. 2012a. Modeling spread of disease from social interactions. In *Sixth AAAI International Conference on Weblogs and Social Media (ICWSM)*.

Sadilek, A.; Kautz, H.; and Silenzio, V. 2012b. Predicting disease transmission from geo-tagged micro-blog data. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Scharff, R. L. 2012. Economic burden from health losses due to foodborne illness in the United States. *Journal of food protection* 75(1):123–131.

Sculley, D.; Otey, M.; Pohl, M.; Spitznagel, B.; Hainsworth, J.; and Yunkai, Z. 2011. Detecting adversarial advertisements in the wild. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM.

Settles, B. 2010. Active learning literature survey. *University of Wisconsin, Madison*.

Shahaf, D., and Horvitz, E. 2010. Generalized task markets for human and machine computation. AAAI.

Shalev-Shwartz, S.; Singer, Y.; and Srebro, N. 2007. Pegasos: Primal estimated sub-gradient solver for svm. In *Proceedings of the 24th international conference on Machine learning*, 807–814. ACM.

Signorini, A.; Segre, A.; and Polgreen, P. 2011. The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PLoS One* 6(5).

Smith, A. 2011. Pew internet & american life project. *http://pewresearch.org/pubs/2007/twitter-users-cell-phone-2011-demographics*.

Tumasjan, A.; Sprenger, T.; Sandner, P.; and Welpe, I. 2010. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 178–185.

Von Ahn, L., and Dabbish, L. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 319–326. ACM.