

Deploying nEmesis: Preventing Foodborne Illness by Data Mining Social Media

Adam Sadilek,^{*1} Henry Kautz,¹ Lauren DiPrete,² Brian Labus,²
Eric Portman,¹ Jack Teitel,¹ and Vincent Silenzio³

¹Department of Computer Science, University of Rochester, Rochester, NY

²Southern Nevada Health District, Las Vegas, NV

³School of Medicine and Dentistry, University of Rochester, Rochester, NY

Abstract

Foodborne illness afflicts 48 million people annually in the U.S. alone. Over 128,000 are hospitalized and 3,000 die from the infection. While preventable with proper food safety practices, the traditional restaurant inspection process has limited impact given the predictability and low frequency of inspections, and the dynamic nature of the kitchen environment. Despite this reality, the inspection process has remained largely unchanged for decades. We apply machine learning to Twitter data and develop a system that automatically detects venues likely to pose a public health hazard. Health professionals subsequently inspect individual flagged venues in a double blind experiment spanning the entire Las Vegas metropolitan area over three months. By contrast, previous research in this domain has been limited to indirect correlative validation using only aggregate statistics. We show that adaptive inspection process is 63% more effective at identifying problematic venues than the current state of the art. The live deployment shows that if every inspection in Las Vegas became adaptive, we can prevent over 9,000 cases of foodborne illness and 557 hospitalizations annually. Additionally, adaptive inspections result in unexpected benefits, including the identification of venues lacking permits, contagious kitchen staff, and fewer customer complaints filed with the Las Vegas health department.

Introduction

The fight against foodborne illness is complicated by the fact that many cases are not diagnosed or traced back to specific sources of contaminated food. In a typical U.S. city, if a food establishment passes their routine inspection, they may not see the health department again for up to a year. Food establishments can roughly predict the timing of their next inspection and prepare for it. Furthermore, the kitchen environment is dynamic, and ordinary inspections merely provide a snapshot view. For example, the day after an inspection, a contagious cook or server could come to work or a refrigerator could break, either of which can lead to a food poisoning. Unless the outbreak is massive, the illness is unlikely to be traced back to the venue.

We present a novel method for detecting problematic venues quickly—before many people fall ill. We use the

^{*}Adam is now at Google.

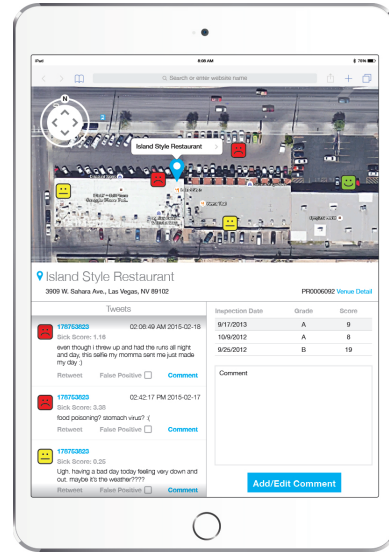


Figure 1: nEmesis web interface. The top window shows a portion of the list of food venues ranked by the number of tweeted illness self-reports by patrons. The bottom window provides a map of the selected venue, and allows the user to view the specific tweets that were classified as illness self-reports.

phrase *adaptive inspections* for prioritizing venues for inspection based on evidence mined from social media. Our system, called nEmesis, applies machine learning to real-time Twitter data — a popular micro-blogging service where people post message updates (tweets) that are at most 140 characters long. A tweet sent from a smartphone is usually tagged with the user’s precise GPS location. We infer the food venues each user visited by “snapping” his or her tweets to nearby establishments (Fig. 1). We develop and apply an automated language model that identifies Twitter users who indicate they suffer from foodborne illness in the text of their public online communication. As a result, for each venue, we can estimate the number of patrons who fell ill shortly after eating there. In this paper, we build on our prior work, where we showed a correlation between the number of “sick tweets” attributable to a restaurant and its

historic health inspection score (Sadilek et al. 2013). In this paper, however, we deploy an improved version of the model and validate its predictions in a *controlled experiment*.

The Southern Nevada Health District started a controlled experiment with nEmesis on January 2, 2015. Venues with the highest predicted risk on any given day are flagged and subsequently verified by a thorough inspection by an environmental health specialist. For each adaptive inspection, we perform a paired control inspection independent of the online data to ensure full annual coverage required by law and to compensate for the geographic bias of Twitter data. During the first 3 months, the environmental health specialists inspected 142 venues, half using nEmesis and half following the standard protocol. The latter set of inspections constitutes our control group. The inspectors were not told whether the venue comes from nEmesis or control.

nEmesis downloads and analyzes all tweets that originate from Las Vegas in real-time. To estimate visits to restaurants, each tweet that is within 50 meters of a food venue is automatically “snapped” to the nearest one as determined by the Google Places API. We used Google Places to determine the locations of establishments because it includes latitude/longitude data that is more precise than the street address of licensed food venues. As we will see, this decision allowed nEmesis to find problems at unlicensed venues.

For this snapping process, we only consider tweets that include GPS coordinates. Cell phones determine their location through a combination of satellite GPS, WiFi access point fingerprinting, and cell-tower triangulation (Lane et al. 2010). Location accuracy typically ranges from 9 meters to 50 meters and is highest in areas with many cell towers and WiFi access points. In such cases, even indoor localization (e.g., within a mall) is accurate.

Once nEmesis snaps a user to a restaurant, it collects all of his or her tweets for the next five days, including tweets with no geo-tag and tweets sent from outside of Las Vegas. This is important because most restaurant patrons in Las Vegas are tourists, who may not show symptoms of illness until after they leave the city. nEmesis then analyses the text of these tweets to estimate the probability that the user is suffering from foodborne illness.

Determining if a tweet indicates foodborne illness of the user is more complex than simply scanning for a short list of keywords. By its nature, Twitter data is noisy. Even a seemingly explicit message, such as “I just threw up,” is incomplete evidence that the author of the tweet has a foodborne illness. By using a language model rather than relying on individual keywords, our method is able to better model the meaning behind the tweet and is therefore able to capture even subtle messages, such as “have to skip work tomorrow” or “I need to go to a pharmacy.” Fig. 2 lists the 20 most significant positive and negative language features that contribute to the score.

nEmesis then associates the individual sickness scores to the food venues from which the users originally tweeted. Each snapped twitter user is a proxy for an unknown number of patrons that visited but did not tweet. Since contracting foodborne illness and tweeting at the right times and places is a relatively rare occurrence, even a single ill individual can

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	f-k i'm	-0.6393
tummy	1.0718	@ID 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
naucious	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
@ID 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	@ID sick	-0.4704
poisoning	0.5879	if	-0.4695
feel better tomorrow	0.5643	i feel better	-0.4670

Figure 2: The top 20 most significant negatively and positively weighted features in our language model.

be a strong evidence of a problem. The web interface (Fig. 1) is used by the the managing health specialist sort venues by the number of sick users, and dispatches inspectors.

Fig. 3 illustrates the full nEmesis process. On a typical day we collect approximately 15,900 geo-tagged tweets from 3,600 users in the Las Vegas area. Approximately 1,000 of these tweets, written by 600 unique users, snap to a food venue. nEmesis then tracks these 600 users and downloads all their subsequent tweets for the following five days. These subsequent tracked tweets are then scored by the language model. Finally, venues are ranked based on the number of tweets with sickness score exceeding the threshold of 1.0 determined on a withheld validation set. During the experiment, nEmesis identified on average 12 new tweets per day that were strongly indicative of foodborne illness.

To the best of our knowledge, this is the first study that directly tests the hypothesis that social media provides a signal for identifying specific sources of *any* disease through a controlled, double-blind experiment during a real-world deployment.

Related work

Since the famous cholera study by John Snow (1855), much work has been done in capturing the mechanisms of epidemics. There is ample previous work in computational epidemiology on building relatively coarse-grained models of disease spread via differential equations and graph theory (Anderson and May 1979; Newman 2002), by harnessing simulated populations (Eubank et al. 2004), and by analysis of official statistics (Grenfell, Bjornstad, and Kappey 2001). Such models are typically developed for the purposes of assessing the impact a particular combination of an outbreak and a containment strategy would have on humanity or ecol-

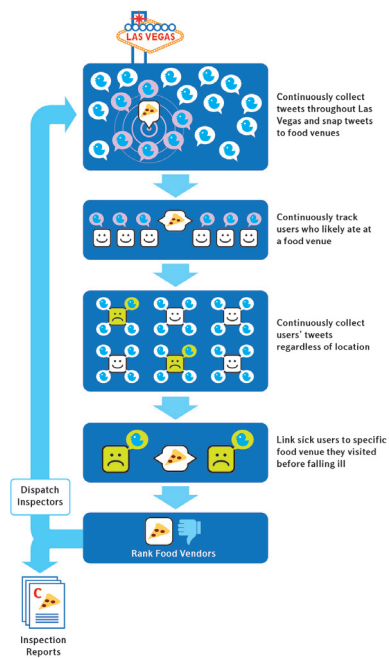


Figure 3: Adaptive inspection process. Starting from the top: all tweets geotagged in the Las Vegas area are collected. Tweets that geotagged within 50 meters of a food venue are snapped to that venue, and the Twitter IDs of the users are added to a database of users to be tracked. All tweets of tracked users are collected for the next five days, whether or not the users remain in Las Vegas. These tweets are evaluated by the language model to determine which are self-reports of symptoms of foodborne illness. Venues are ranked according to the number of patrons who later reported symptoms. Health department officials use the nEmesis web interface to select restaurants for inspection. Inspectors are dispatched to the chosen restaurants, and findings reported.

ogy (Chen, David, and Kempe 2010).

Most prior work on using data about users' online behavior has estimated aggregate disease trends in a large geographical area, typically at the level of a state or large city. Researchers have examined influenza tracking (Culotta 2010; Achrekar et al. 2012; Sadilek and Kautz 2013; Broniatowski and Dredze 2013; Brennan, Sadilek, and Kautz 2013), mental health and depression (Golder and Macy 2011; De Choudhury et al. 2013), and well as general public health across a broad range of diseases (Brownstein, Freifeld, and Madoff 2009; Paul and Dredze 2011b).

Some researchers have begun modeling health and contagion of *specific individuals* by leveraging fine-grained online social and web search data (Ugander et al. 2012; White and Horvitz 2008; De Choudhury et al. 2013). For example, in (Sadilek, Kautz, and Silenzio 2012) we showed that Twitter users exhibiting symptoms of influenza can be accurately detected using a model of language of Twitter posts. A detailed epidemiological model can be subsequently built by following the interactions between sick and

healthy individuals in a population, where physical encounters estimated by spatio-temporal co-located tweets.

Our earlier work on nEmesis (Sadilek et al. 2013) scored restaurants in New York City by their number of sick tweets using an initial version of the language model described here. We showed a weak but significant correlation between the scores and published NYC Department of Health inspection scores. Although the data came from the same year, many months typically separated the inspections and the tweets.

Other researchers have recently tried to use Yelp restaurant reviews to identify restaurants that should be inspected (Harrison et al. 2014). Keywords were used to filter 294,000 Yelp reviews for New York City to 893 possible reports of illness. These were manually screened and resulted in the identification of 3 problematic restaurants.

Background: Foodborne Illness

Foodborne illness, known colloquially as *food poisoning*, is any illness resulting from the consumption of contaminated food, pathogenic bacteria, viruses, or parasites that contaminate food, as well as the consumption of chemical or natural toxins such as poisonous mushrooms. The US Centers for Disease Control and Prevention (CDC) estimates that 47.8 million Americans (roughly 1 in 6 people) are sickened each year by foodborne disease. Of that total, nearly 128,000 people are hospitalized, while just over 3,000 die of foodborne diseases (CDC 2013).

CDC classifies cases of foodborne illness according to whether they are caused by one of 31 *known foodborne illness pathogens* or by *unspecified agents*. These 31 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. Food poisoning episodes associated with these 31 known pathogens account for an estimated 44% of all hospitalizations resulting from foodborne illness, as well as 44% of the deaths. The economic burden of health losses resulting from foodborne illness are staggering. One recent study estimated the aggregated costs in the US alone to be \$77.7 billion annually (Scharff 2012).

Despite the variability in the underlying etiology of foodborne illness, the signs and symptoms of disease overlap considerably. The most common symptoms include vomiting, diarrhea (occasionally bloody), abdominal pain, fever, and chills. These symptoms can be mild to serious, and may last from hours to several days. Some pathogens can also cause symptoms of the nervous system, including headache, numbness or tingling, blurry vision, weakness, dizziness, and even paralysis. occur days to even weeks after exposure to the pathogen (J Glenn Morris and Potter 2013). According to the US Food and Drug Administration (FDA), the vast majority of these symptoms will occur within three days (FDA 2012).

Public health authorities use an array of surveillance systems to monitor foodborne illness. In the US, 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). 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.

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 can entail significant time lags between when cases are identified and the data is analyzed and reported. Whereas this is not as important a limitation in terms of epidemiological surveillance, using surveillance data to actively intervene in outbreaks of foodborne illnesses can be challenging when surveillance data may not infrequently identify cases after the window of opportunity needed to prevent additional cases (Heymann 2004).

Methods

Inspection Protocols

Clark County, Nevada is home to over 2 million people and hosts over 41 million annual visitors to the Las Vegas metropolitan area. The Southern Nevada Health District (SNHD) is the governmental agency responsible for all public health matters within the county and is among the largest local health departments in the United States by population served. In 2014, SNHD conducted 35,855 food inspections in nearly 16,000 permitted facilities.

At the Southern Nevada Health District, food establishments are required by law to be inspected once per calendar year. A routine inspection is a risk-based process addressing the food establishments control over the five areas of risk for foodborne illness: personal hygiene, approved food source, proper cooking temperatures, proper holding times and temperatures, and sources of contamination. Violations are weighted based on their likelihood to directly cause a foodborne illness and are divided into critical violations at 5 demerits each (e.g., food handlers not washing hands between handling raw food and ready to eat food), to major violations at 3 demerits each (hand sink not stocked with soap), to good food management practices with no demerit value (leak at the hand sink).

Demerits are converted to letter grades, where 0-10 is an A, 11-20 is a B, 21-39 is a C, and 40+ is an F (immediate closure). A repeated violation of a critical or major item causes the letter grade to drop to the next lower rank. A grade of C or F represents a serious health hazard.

Controlled Experiment: Adaptive Inspections

During experiment, when a food establishment was flagged by nEmesis in an inspector's area, he was instructed to conduct a standard, routine inspection on both the flagged facility (adaptive inspection) and also a provided control facility

(routine inspection). Control facilities were selected according to their location, size, cuisine, and their permit type to pair the facilities as closely as possible. The inspector was blind as to which facility was which, and each facility received the same risk-based inspection as the other.

Labeling Data at Scale

To scale the laborious process of labeling training data for our language model, we turn to Amazon's Mechanical Turk.¹ 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). Responses from workers who exhibit consistently low annotator agreement with the majority were eliminated.

Workers were paid for their efforts only after we were reasonably sure their responses were sincere based on inter-annotator agreement. 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.

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

¹<https://www.mturk.com/>

leverage *human guided machine learning*—a novel learning method that considerably reduces the amount of human effort required to reach any given level of model quality, even when the number of negatives is many orders of magnitude larger than the number of positives (Sadilek et al. 2013). 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.

In a postmortem, we have manually verified submitted tweets and 97% were correct sick tweets. This verification step could also be crowdsourced. We note that searching for relevant tweets is significantly more time consuming than simply deciding if a given tweet contains a good example of sickness. Future work could explore multi-tiered architecture, where a small number of workers acting as “supervisors” verify data provided by a larger population of “assistants”. Supervisors as well as assistants would collaborate with an automated model, such as the SVM classifier described in this paper, to perform search and verification tasks.

Language Model

Support vector machines (SVMs) are an established method for classifying high-dimensional data (Cortes and Vapnik 1995). We train a linear binary SVM by finding a hyperplane (defined by a normal vector w) with the maximal margin separating 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).

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

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

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

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 $\langle x_i^+ - x_j^-, 1 \rangle$ yields significantly more robust results (Joachims 2005).

We use the trained SVM language model to predict how likely each tweet indicates foodborne illness. The model is trained on 8,000 tweets, each independently labeled by five human annotators as described above. As features, the SVM uses all uni-gram, bi-gram, and tri-gram word tokens that appear in the training data at least twice. For example, a tweet “*My tummy hurts.*” is represented by the following feature vector:

$\{my, tummy, hurts, my\ tummy, tummy\ hurts, my\ tummy\ hurts\}$

Prior to tokenization, we convert all text to lower case and strip punctuation. Additionally, we replace mentions of user identifiers (the “@” tag) with a special @ID 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.

Training the model associates a real-valued weight to each feature. The score the model assigns to a new tweet is the sum of the weights of the features that appear in its text. There are more than one million features; Fig. 2 lists the 20 most significant positive and negative features. While tweets indicating illness are sparse and our feature space has a very high dimensionality, with many 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). Evaluation of the language on a held-out test set of 10,000 tweets shows 0.75 precision and 0.96 recall. The high recall is critical because evidence of illness is very scarce.

System Architecture

nEmesis consists of several modules that are depicted at a high-level in Fig. 3. Here we describe the architecture in more detail. We implemented the entire system in Python, with NoSQL data store running on Google Cloud Platform. Most of the code base implements data download, cleanup, filtering, snapping (e.g., “at a restaurant”), and labeling (“sick” or “healthy”). There is also a considerable model learning component described in the previous two sections.

Downloader: This module runs continuously and asynchronously with other modules, downloading all geo-coded tweets based upon the bounding box defined for the Las Vegas Metro area. These tweets are then persisted to local database in JSON format.

Tracker: For each unique Twitter User that tweets within the bounding box, this module continues to download all of their tweets for 2 weeks, independent of location (also using the official Twitter API). These tweets are also persisted to local storage in JSON format.

Snapper: The responsibility of this module is to identify Las Vegas area tweets that are geocoded within 50 meters of a food establishment. It leverages Google Places API,

which serves precise location for any given venue. We built an in memory spatial index that included each of those locations (with a square boundary based on the target distance we were looking for). For each tweet, nEmesis identifies a list of Google Places in the index that overlapped with the Tweet based on its lat/long. If a given tweet had one or more location matches, the matching venues are added as an array attribute to the Tweet.

Labeler: Each tweet in the data store is piped through our SVM model that assigns it an estimate of probability of foodborne illness. All tweets are annotated and saved back into the data store.

Aggregation Pipelines: We use Map Reduce framework on Google App Engine to support custom aggregation pipeline. It updates statistics about each venue (number of sick tweets associated with that venue, *etc.*).

Web Interface. The health professionals interact with nEmesis through a web application shown in Fig. 1. All modules described above work together to produce a unified view that lists most likely offending venues along with supporting evidence. This allows inspectors to make informed decisions how to allocate their resources. The application was written using a combination of Python for the data access layer and AngularJS for the front-end.

Developing the SVM model took 3 engineer-months. The backend modules above (Downloader through Labeler) took 2 engineer-months, and the Web Interface took an additional engineer-month.

Lessons Learned

A major challenge was implementing the SVM language model and calibrating its output. This included research work to arrive at a robust model, as well as engineering work to scale it to the size and real-time nature of the data.

The initial format of our mTurk HITS for labeling training data used a payout of 3 cents per tweet with 10 tweets per survey. We discovered that we could reduce payouts to 1 cent per tweet and increase tweets per survey to 25 without increasing worker attrition. Our initial surveys also had additional “Yes” options for other illness types, *e.g.*, cold and allergies. The original hope with these options was that they would help the classifier more easily discriminate between general sickness and food-related sickness. However, it became clear that these options were confusing workers, resulting in low inter-annotator agreement, so we abandoned them.

nEmesis is deployed on Google Cloud with automatic deployment whenever there was a check-in to the codebase. Because the data store is schema-less, there is not a need for any schema deployment (table creation, indexes, *etc.*) that are traditionally a part of a SQL database deployment process.

The system is highly asynchronous, with many modules running in parallel. These modules further communicate with other systems (*e.g.*, Twitter API, Google Places API). Many of the processing steps can fail for reasons beyond our control (*e.g.*, a call to Twitter API times out because of temporary network issue). We have learned that the data pipelines need to have comprehensive exception catching

logic to detect and recover from a variety of errors. Many of the errors are non-reproducible, and occur rarely and unpredictably. Therefore, programatic testing and monitoring is essential.

The data pipelines also need to be ‘self healing’ – in case of a failure necessary steps are immediately taken to ensure that data is reprocessed and not lost. The use of Google Cloud platform for the datastore and front end ensures there is really no operational requirement for the team. The system is always available. Application servers quiesce when not in use, and automatically come back on line when needed. In addition the application server tier will automatically scale (up or down) if needed based on user volume.

Results and Discussion

We determined that adaptive inspections uncover significantly more demerits: 9 vs. 6 per inspection (p-value of 0.019). We use paired Mann-Whitney-Wilcoxon test to calculate the probability that the distribution of demerits for adaptive inspection is stochastically greater than the control distribution (Mann and Whitney 1947). This test can be used even if the shapes of the distributions are non-normal and different, which is the case here. Chi-squared test at the level of discrete letter grades shows a significant skew towards worse grades in adaptive inspections.

The most important distinction, however, is between restaurants with minor violations (grades A and B) and those posing considerable health risks (grade C and worse). nEmesis uncovers 11 venues in the latter category, whereas control finds only 7, a 64% improvement.

All of our data, suitably anonymized to satisfy Twitter’s terms of use, is available upon request to other researchers for further analysis.

CDC studies show that each outbreak averages 17.8 afflicted individuals and 1.1 hospitalizations (CDC 2013). Therefore we estimate that adaptive inspections saved 71 infections and 4.4 hospitalizations over the three month period. Since the Las Vegas health department performs over 35,000 inspections annually, nEmesis can prevent over 9,126 cases of foodborne illness and 557 hospitalizations in Las Vegas alone. This is likely an underestimate as an adaptive inspection can catch the restaurant sooner than a normal inspection. During that time, the venue continues to infect customers.

Adaptive inspections yield a number of unexpected benefits. nEmesis alerted SNHD to an unpermitted seafood establishment. This business was flagged by nEmesis because it uses a comprehensive list of food venues independent of the permit database. An adaptive inspection also discovered a food handler working while sick with an influenza-like disease. Finally, we observed a reduced amount of foodborne illness complaints from the public and subsequent investigations during the experiment. Between January 2, 2015 and March 31, 2015, SNHD performed 5 foodborne illness investigations. During the same time frame the previous year, SNHD performed 11 foodborne illness investigations. Over the last 7 years, SNHD averaged 7.3 investigations during this three month time frame. It is likely that nEmesis alerted

the health district to food safety risks faster than traditional complaint channels, prior to an outbreak.

Given the ambiguity of online data, it may appear hopeless to identify problematic restaurants fully automatically. However, we demonstrate that nEmesis uncovers significantly more problematic restaurants than current inspection processes. This work is the first to directly validate disease predictions made from social media data. To date, all research on modeling public health from online data measured accuracy by correlating aggregate estimates of the number of cases of disease based on online data and aggregate estimates based on traditional data sources (Grassly, Fraser, and Garnett 2005; Brownstein, Wolfe, and Mandl 2006; Ginsberg et al. 2008; Golder and Macy 2011; Sadilek et al. 2013). By contrast, each prediction of our model is verified by an inspection following a well-founded professional protocol. Furthermore, we evaluate nEmesis in a controlled double-blind experiment, where predictions are verified in the order of hours.

Finally, this study also showed that social-media driven inspections can discover health violations that could never be found by traditional protocols, such as unlicensed venues. This fact indicates that it may be possible to adapt the nEmesis approach for identifying food safety problems in non-commercial venues, ranging from school picnics to private parties. Identifying possible sources of foodborne illness among the public could support more targeted and effective food safety awareness campaigns.

The success of this study has led the Southern Nevada Health District to win a CDC grant to support the further development of nEmesis and its permanent deployment statewide.

Acknowledgements

This research was partly funded by NSF grants 1319378 and 1516340; NIH grant 5R01GM108337-02; and the Intel ISTC-PC.

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*.

Anderson, R., and May, R. 1979. Population biology of infectious diseases: Part I. *Nature* 280(5721):361.

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.

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)*.

Broniatowski, D. A., and Dredze, M. 2013. National and local influenza surveillance through twitter: An analysis of the 2012–2013 influenza epidemic. *PLoS ONE* 8(12).

Brownstein, J. S.; Freifeld, B. S.; and Madoff, L. C. 2009.

Digital disease detection - harnessing the web for public health surveillance. *N Engl J Med* 260(21):2153–2157.

Brownstein, J.; Wolfe, C.; and Mandl, K. 2006. Empirical evidence for the effect of airline travel on inter-regional influenza spread in the united states. *PLoS medicine* 3(10):e401.

CDC. 2013. Surveillance for foodborne disease outbreaks united states, 2013: Annual report. Technical report, Centers for Disease Control and Prevention National Center for Emerging and Zoonotic Infectious Diseases.

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.

Chen, P.; David, M.; and Kempe, D. 2010. Better vaccination strategies for better people. In *Proceedings of the 11th ACM conference on Electronic commerce*, 179–188. ACM.

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

Culotta, A. 2010. Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, 115–122. ACM.

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

Eubank, S.; Guclu, H.; Anil Kumar, V.; Marathe, M.; Srinivasan, A.; Toroczkai, Z.; and Wang, N. 2004. Modelling disease outbreaks in realistic urban social networks. *Nature* 429(6988):180–184.

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.

Grassly, N.; Fraser, C.; and Garnett, G. 2005. Host immunity and synchronized epidemics of syphilis across the united states. *Nature* 433(7024):417–421.

Grenfell, B.; Bjornstad, O.; and Kappey, J. 2001. Travelling waves and spatial hierarchies in measles epidemics. *Nature* 414(6865):716–723.

Harrison, C.; Jorder, M.; Stern, H.; Stavinsky, F.; Reddy, V.; Hanson, H.; Waechter, H.; Lowe, L.; Gravano, L.; and Balter, S. 2014. Using a restaurant review website to identify unreported complaints of foodborne illness. *Morb Mortal Wkly Rep* 63(20):441–445.

Heymann, D. L. 2004. *Control of communicable diseases manual : an official report of the American Public Health Association*. American Public Health Association, 18th edition.

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.
- Landis, J. R., and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *biometrics* 159–174.
- Lane, N. D.; Miluzzo, E.; Lu, H.; Peebles, D.; Choudhury, T.; and Campbell, A. T. 2010. A survey of mobile phone sensing. *Communications Magazine, IEEE* 48(9):140–150.
- Mann, H., and Whitney, D. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* 18:50–60.
- Mason, W., and Suri, S. 2012. Conducting behavioral research on amazons mechanical turk. *Behavior research methods* 44(1):1–23.
- Newman, M. 2002. Spread of epidemic disease on networks. *Physical Review E* 66(1):016128.
- 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.; Brennan, S.; Kautz, H.; and Silenzio, V. 2013. nEmesis: Which restaurants should you avoid today? In *AAAI Conference on Human Computation and Crowdsourcing*.
- Sadilek, A.; Kautz, H.; and Silenzio, V. 2012. 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.
- 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.
- Snow, J. 1855. *On the mode of communication of cholera*. John Churchill.
- Ugander, J.; Backstrom, L.; Marlow, C.; and Kleinberg, J. 2012. Structural diversity in social contagion. *Proceedings of the National Academy of Sciences* 109(16):5962–5966.
- White, R., and Horvitz, E. 2008. Cyberchondria: Studies of the escalation of medical concerns in web search. Technical Report MSR-TR-2008-177, Microsoft Research. Appearing in *ACM Transactions on Information Systems*, 27(4), Article 23, November 2009, DOI 10.1145/1629096.1629101.