Information Extraction from Social Media for Public Health

Noémie Elhadad  
Columbia University  
noemie.elhadad@columbia.edu

Luis Gravano  
Columbia University  
gravano@cs.columbia.edu

Daniel Hsu  
Columbia University  
djhsu@cs.columbia.edu

Sharon Balter  
New York City Dept. of Health and Mental Hygiene  
sbalter@health.nyc.gov

Vasudha Reddy  
New York City Dept. of Health and Mental Hygiene  
VReddy@health.nyc.gov

HaeNa Waechter  
New York City Dept. of Health and Mental Hygiene  
hwaechte@health.nyc.gov

ABSTRACT

Social media sites are a major source for non-curated, user-generated feedback on virtually all products and services. Users increasingly rely on social media to disclose sometimes serious real-life incidents rather than visiting official communication channels. This valuable, actionable, user-generated information, if extracted reliably and robustly from the social media sites, has the potential to have a positive impact on critical applications related to public health and safety, and beyond. Unfortunately, the extraction and presentation of actionable information from social media—where the output of the extraction process is used to take concrete actions in the real world—are not well supported by existing technology. Traditional information extraction approaches do not work well over the highly informal, noisy, and ungrammatical text common in social media, and they do not handle the extraction and aggregation of the rare content that important applications need to extract from high-volume streaming sources. In our ongoing collaborative project between Columbia University and the New York City Department of Health and Mental Hygiene (DOHMH), we aim to address these gaps in research and technology for one important public health application, namely, detecting and acting on foodborne outbreaks in New York City restaurants. Thus far, we have been able to address these issues successfully and have used one social media site to identify and follow up on several foodborne outbreaks that had not been reported through conventional channels.

1. DETECTING AND ACTING ON FOODBORNE OUTBREAKS

The Centers for Disease Control and Prevention (CDC) estimates that 1 in 6 Americans, or 48 million people, get sick from a foodborne disease each year. Of confirmed foodborne outbreaks investigated nationally, 45% are restaurant-related. The New York City DOHMH is the agency with primary responsibility for foodborne disease detection and outbreak investigations in New York City. New York City hosts approximately 24,000 restaurants and 15,000 food retail establishments (e.g., grocery stores, delis). Citizens can report illnesses associated with such venues through the official 311 telephone line and its associated website and app. Once an outbreak is identified, DOHMH launches an investigation that includes a restaurant inspection, testing food and clinical specimens, and collecting symptom and food exposure data from restaurant patrons, as well as conducting statistical analysis to implicate a food item; finally, DOHMH takes action to stop the outbreak (e.g., by removing a contaminated food item) and prevent its recurrence (e.g., through education). Unfortunately, only about 30 restaurant-associated foodborne outbreaks are reported annually, so almost certainly many such outbreaks are not reported to the government through the official channels, which has potentially serious public health repercussions. In fact, investigated outbreaks are thought to represent only a small fraction of all foodborne outbreaks. As we reported in [10], in a study of 294,000 New York City restaurant reviews on the Yelp website we discovered that only 3% of the illnesses that we identified had been reported through the official New York City channels, which highlights the importance of extracting reliable, otherwise-unreported outbreak-related information from social media, so that the government can analyze it and launch investigations when appropriate. Figure 4(a) shows Yelp and Twitter posts about a potential food poisoning incident at a New York City restaurant.

Since January 2012, we have had a fruitful, ongoing collaboration between the Computer Science Department at Columbia University and the New York City DOHMH, hence combining Columbia’s expertise in Computer Science with the DOHMH’s domain knowledge and infrastructure for the important application at hand. Overall, the goal of this collaboration is to identify and analyze the unprecedented volumes of user-contributed opinions and comments about restaurants on social media sites, to extract reliable indicators of disease outbreaks associated with the restaurants, an important public health task. We have already produced a proof-of-concept, operational prototype over Yelp data. (Yelp has been graciously producing now-daily feeds of New York City restaurant reviews for our use in our project.) Our prototype has been used by the DOHMH since July 2012. Our system processes each Yelp review with multiple classifiers, developed through supervised machine learning to detect (1) whether the review discusses a potential food poisoning incident; (2) whether the review hints at an ince-
bation period that is consistent with foodborne illness; and
(3) whether the review mentions that multiple people were
affected in the incident. The classifier output is aggregated
and integrated, in a time-sensitive way, for each New York
City restaurant, to produce a ranked list of restaurants, with
their associated reviews. A DOHMH epidemiologist then
analyses each case that is output by our prototype and de-
cides whether it likely corresponds to a case of true food
poisoning and whether further action by DOHMH is appro-
priate. The prototype has so far yielded highly encourag-
ing results—including helping unveil several previously un-
known outbreaks—that DOHMH has been analyzing and
acting upon as appropriate, as reported in [10].

2. NEW CHALLENGES

The existing prototype is an exciting proof of concept, hinting at the feasibility and usefulness of leveraging social
media into actionable information to support public health
practitioners. We have identified several research directions
to enable robust, comprehensive, and accurate information
extraction and inference all the while ensuring that the ex-
tracted information is presented in a useful, contextualized
fashion to facilitate the investigation of outbreaks.

First and foremost, our current prototype has focused on
a single website, namely, Yelp, and only recently have we
started to integrate Twitter, as well as the user reports sub-
mitted through the official 311 channels, into the prototype.
(Figure 1(b) shows the initial interface of our prototype, sup-
porting the labeling and analysis by the DOHMH of Yelp,
Twitter, and 311 entries identified by the prototype.) To im-
prove the coverage of our system, and drastically expand its
“outbreak recall,” we must vastly expand the social media
content with general context about restaurants (e.g., the
overall recent sentiment about attributes such as food and
service), and select additional content when suggested by the
inference about potential outbreaks, thus following “leads”
and directions that merit further investigation. This expan-
sion will require (1) identifying and selecting sources, so that
we can capture as much user-provided evidence on the web
as possible; (2) capturing context, so that we can identify
which restaurant is being referred to in specific social me-
dia posts; and (3) selecting content adaptively, so that we
can follow “leads” and directions that merit further investi-
gation as the information extraction and inference process
progresses.

To handle this expansion in content and to minimize the
number of “false positives” that we produce—false positives
unnecessarily consume human resources at DOHMH—we
must fundamentally improve how we extract and derive in-
ferrred attributes for the restaurants. The (time-sensitive)
attributes of restaurants that we extract and infer can be
expanded and aggregated. The expanded set may include
general context and adaptive content selected earlier, as
well as specific attributes related to foodborne disease as
extracted from the text streams: user demographic infor-
mation, counts of individuals showing symptoms, recency of
exposure, indicators for displayed symptom types, and the
likelihood of cooperation with a subsequent investigation.
Attributes that account for potential confounding factors
such as seasonal trends (e.g., stomach virus seasons) can also
be used. The most relevant cases of interest—the restau-
rant that indeed must be investigated—are a comparably
small fraction of the set of all potential cases, and many
(though not all) of the inference tasks require annotations
from highly trained public health experts. New methods for
active learning are needed here to handle both severe class
imbalance in our source document types, as well as heterogeneity in the inference tasks requiring human annotations.

To support DOHMH practitioners in deciding effectively and efficiently whether further investigation on a restaurant is warranted, the extracted and inferred information for each restaurant must be consolidated and presented in a way that reveals the lineage of the extracted information and our confidence in its correctness.

Overall, we expect our collaboration to have substantial impacts on public health: it has potential to help prevent or curtail foodborne illness outbreaks, reduce public health expenditures, and increase public awareness and involvement in public health issues. As mentioned above, a prototype of our application is already in use at the NYC DOHMH, and addressing the new challenges above has the potential to vastly enhance its effectiveness. Following the best practices established at public health agencies across the country, we will report the findings about restaurant foodborne disease outbreaks detected in New York City across agencies [10]. Because such activity is relevant to other cities in the country (and typically carried out through spontaneous reporting and 311-type calls only rather than by online tracking), we hope to share our system with other cities.

Besides foodborne illness outbreaks, our general approach is also well suited for other public health and safety issues, including product recall investigations, rodent control, indoor air quality, noise pollution, and post-disaster health concerns. Our project may then have a strong, direct impact on society by supporting public health practitioners through next-generation, data-driven solutions.

3. RELATED WORK

Many research efforts have analyzed social media, search engine query logs, and beyond to analyze public health phenomena. In biosurveillance, diseases of interest are analyzed by monitoring over time specific information sources, whether emergency department visits, search engine query logs, or Twitter messages [4] [8] [9] [13] [15]. These methods generally operate through simple content analysis techniques, which might ignore potential confounders [5] [12]. Other interesting, recent work analyzes social media content to correlate it with relevant official health statistics and ratings [11] [19]. In contrast, our work has a strong focus on extracting and presenting actionable information from social media: through our focus on restaurant of interest, we propose the extraction of concrete attributes (e.g., number of people affected in a potential foodborne outbreak, incubation period associated with each case) that directly support decision-making for a public health application (e.g., to decide whether an expensive investigation should be launched). Our proposed research is in line with efforts such as the City of Chicago Department of Public Health’s “FoodBorne Chicago” system [21], which tracks foodborne illness on Twitter, with a follow-up, web-based form that users can complete and which might lead to action by the City. Beyond public health, there has also been much work on monitoring trending events and discussion topics in social media, as well as identifying real-world event content on social media (e.g., [1] [2] [3] [7] [13] [14] [15] [17] [22]).

This body of work generally focuses on events—interpreted broadly—involving large volumes of information, where temporal analysis and clustering can be meaningfully performed. In contrast, our focus is on extracting actionable, “needle in a haystack” content associated with rare occurrences (e.g., just a handful of reliable social media posts on food poisoning from a particular restaurant might be sufficient to merit further investigation by a governmental agency).

Acknowledgments

We thank Mohip Jorder, Fotis Psallidas, Alden Quimby, and Henri Stern for designing and developing the proof-of-concept system for our project, which has been running at the New York City Department of Health and Mental Hygiene. We also thank Cassandra Harrison and Kenya Murray for their invaluable contributions to our project on the DOHMH side. Finally, we are grateful to Yelp for providing our project with a (now-daily) feed of user reviews for all New York City restaurants.

References


