Social Media Representation and Illness Detection on Social media: Foodborne Illness as a Case Study

Authors: Nina Cesare, Quynh Nguyen, Christan Grant, Elaine Nsoesie

Summary:

Researchers widely recognize the importance of addressing bias in digital data but identifying and unpacking the nature of this bias is a work in progress. This study seeks to build on this literature assessing the utility of detecting foodborne illness on Twitter. It acknowledges that effective disease surveillance requires considering both who uses a specific platform, as well as how they use it. It first measures how the composition of users tweeting about foodborne illness symptoms differs from the composition of individuals impacted by foodborne illness. It then characterizes how individuals within specific demographic groups vary regarding how they discuss illness within this platform. Understanding the former will allow researchers to more accurately assess the association between digitally reported symptoms and offline illness, and the latter may improve the effectiveness and comprehensiveness of digital platforms as tools for identifying disease outbreak

Introduction:

The availability of self-reported health symptoms on digital media has captured the interest of epidemiologists and public health researchers interested in identifying disease outbreak and/or tracking the spread of illness. As outlined by Fung, Tse, and Fu (2015), data from digital sources such as social media and search queries have been used to monitor the spread of official information on disease outbreaks, estimate the number of individuals impacted by an outbreak, and forecast disease spread. The benefits of using these data are many; they are less resource-intensive to gather than traditional data on disease outbreaks, and they are spatially and temporally granular. However, use of these data poses challenges – many of which are related to demographic representation.

Demographic representation is most directly related to the ability to draw useful inferences from digital samples. The demographic composition of the individuals who post on social media may be significantly different than the demographic composition of the area from which their posts originate. This has the potential to bias estimates of offline illness generated using these data and may impact the distribution of resources necessary for combatting illness. Henly et al. (2017), for example, found that counties in which instances of foodborne illness are detectable through Yelp typically exhibit stronger signifiers of affluence (e.g. higher median income, higher educational attainment) than counties in which there are no Yelp reports. The same may be true of other self-reported digital traces of foodborne illness. As a first step, this study will assess the sociodemographic composition (e.g. household income, racial composition, access to health resources) of U.S. counties in which there exists both documentation of foodborne illness and tweets about foodborne illness symptoms and compare this to the composition of counties in which there are documented outbreaks but no tweets. Moreover, it will assess how these disparities increase or decrease as we stratify the sample according to the gender and approximate age of the individuals tweeting.

In addition to statistically biasing estimates of illness prevalence, representation within social media may influence the content of the data gathered. One challenge of using digital data to detect illness is that the text is unsolicited and the vocabulary used may not match that of a researcher-designed questionnaire. Existing work has conducted content and sentiment analyses on tweets addressing

illness, noting that understanding how individuals discuss illness within this space is essential for improving the use of Twitter as health monitoring tool (Brownstein et al. 2008; Chew and Eysenbach 2010),. To date, no studies have analyzed how this content varies by age or gender. This study will fill this gap by analyzing the vocabulary and content of foodborne illness-related posts from male and female users, as well as users above and below age 30.

Data:

Data for this study come from two sources. The first data source are tweets describing foodborne illness – specifically tweets that report the following descriptors and/or symptoms: 'vomit,' 'diarrhea' and 'food poisoning.' These data contain all tweets spanning approximately one year (2013). In total there are 11,921,408 tweets posted by 5,767,028 unique users. Among these users, 62.9% are female, and 37.1% are male. The second data source are reports of foodborne illness that were distributed by the Centers for Disease Control and Prevention (CDC).[1] Outbreak estimates are aggregated at the county level.

We rely on machine learning to identify the demographic traits of individuals reporting illness. We use gender classification techniques proposed in Cesare et al. (2017) to identify users as male or female. This approach, in summary, uses a weighted ensemble model that incorporates three gender prediction approaches based on user name alone. For age, we will apply a previously developed classifier that uses a gradient boosted decision tree algorithm with selected self-presentation and activity features to predict whether users are under 18, 18-30 or over 30 with 67% accuracy. In this analysis, we identify whether users are 18 to 30 or over 30 years old.

Preliminary results:

In regard to representation, preliminary results find a lower percentage of non-Hispanic white residents in counties with both documented outbreaks and tweets in comparison with counties that have outbreaks but no tweets. This difference is stronger for males than females ($t=3.407$, 0.006 and $t=0.889$, $p=0.386$). Counties in which there are outbreaks and tweets have higher household incomes than counties in which there are outbreaks alone. This effect is also stronger for males than females, but it is significant for both ($t=-4.660$, $p<0.001$ and $t=-2.315$ $p<0.05$ for males and females, respectively). A similar pattern holds true for median county age, with outbreak and tweet counties having lower median ages than outbreak counties ($t=7.03$, $p<0.001$ and $t=3.301$, $p<0.01$ for men and women, respectively). In summary, it seems that Twitter is effective for detecting foodborne illness is counties with more diverse, younger, and wealthy populations. This effect is particularly strong when considering tweets issued by female users.

Table 1: Representation in Counties with Tweets, Outbreaks by Gender

User demographics	County category	Measure	Value
Female	Counties with outbreaks (n=493)	Income	54677.62
		Percent non-Hispanic White	75.71
		Median age	38.97
		Income	54943.40

[1] See: https://www.cdc.gov/fdoss/annual-reports/index.html

	Counties with outbreaks and tweets (n=475)	Percent non-Hispanic White	75.57
		Median age	38.80
	Counties with outbreaks, no tweets (n=18)	Income	47663.94
		Percent non-Hispanic White	80.43
		Median age	43.62
Male	Counties with outbreaks (n=493)	Income	54677.62
		Percent non-Hispanic White	75.75
		Median age	38.97
	Counties with outbreaks and tweets (n=469)	Income	55166.02
		Percent non-Hispanic White	75.2
		Median age	38.67
	Counties with outbreaks, no tweets (n=24)	Income	45133.42
		Percent non-Hispanic White	85.67
		Median age	44.90

While age predictions for the data are in-process, preliminary term frequency analyses of data gathered on the same keywords from Twitter's streaming API (n= 536,142 tweets) revealed notable differences in the vocabulary use between users of different age groups. We found that users classified as over 30 often use medical terms (e.g. norovirus, cholera, symptom) to describe or seek to identify illness, whereas users under 30 frequently reference eateries (e.g. Taco Bell, restaurant) and use slang or profanity to discuss illness. We expect to find similar trends in our working data source.

In addition to analyzing vocabulary, we will also use a machine learning approach to determine whether users within specific groups typically use Twitter to seek information about illness, or whether they use Twitter to report and discuss symptoms of illness. Similar to Powell et al. (2016)'s analysis of conversation on Twitter regarding vaccines, we will pre-designate a classification framework based on information-seeking or symptom-reporting behaviors, label a random subsample of tweets using this framework, and train a machine learning classifier to label the remaining tweets. Given previous success using a feed forward neural network (FFNN) classifier to label tweets with exercise keywords as relevant or irrelevant measures of offline fitness, we will try adapting this classifier to the new labeled dataset. We will then use these labels to characterize the conversation content of male and female users, as well as users at/over and under age 30, hypothesizing that there are significant differences in how users perceive and utilize Twitter as a tool for gathering and sharing health information.

Discussion:

The anticipated impact of this study is twofold. First, we expect that assessing the demographic composition of counties in which we see outbreaks and tweets in comparison with counties in which there are tweets alone will assist researchers who seek to improve the use of digital data as a tool for monitoring real-time disease outbreak. Second, highlighting demographic variation in how individuals discuss foodborne illness on Twitter may lead to more effective and comprehensive identification of disease outbreak. This study also raises critical ethical questions regarding the costs and benefits of using unsolicited data to identify outbreaks and disease spread. While unsolicited information on illness symptoms provides quick and valuable insight into public wellness, we may ask how best to manage studies in which research subjects may be unaware of their participation.

References

Brownstein, John S., Clark C. Freifeld, Ben Y. Reis, and Kenneth D. Mandl. 2008. "Surveillance Sans Frontières: Internet-Based Emerging Infectious Disease Intelligence and the HealthMap Project." *PLOS Medicine* 5(7):e151.

Cesare, Nina, Christan Grant, Jared B. Hawkins, John S. Brownstein, and Elaine O. Nsoesie. 2017. "Demographics in Social Media Data for Public Health Research: Does It Matter?" *ArXiv Preprint ArXiv:1710.11048*.

Chew, Cynthia and Gunther Eysenbach. 2010. "Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak." *PLOS ONE* 5(11):e14118.

Fung, Isaac Chun-Hai, Zion Tsz Ho Tse, and King-Wa Fu. 2015. "The Use of Social Media in Public Health Surveillance." *Western Pacific Surveillance and Response* 6(2).

Henly, Samuel, Gaurav Tuli, Sheryl A. Kluberg, Jared B. Hawkins, Quynh C. Nguyen, Aranka Anema, Adyasha Maharana, John S. Brownstein, and Elaine O. Nsoesie. 2017. "Disparities in Digital Reporting of Illness: A Demographic and Socioeconomic Assessment." *Preventive Medicine* 101:18–22.

Krogstad, Jens Manuel. 2015. "Social Media Preferences Vary by Race and Ethnicity." *Pew Research Center*. Retrieved September 12, 2018 (http://www.pewresearch.org/fact-tank/2015/02/03/social-media-preferences-vary-by-race-and-ethnicity/).

Powell, Guido Antonio, Kate Zinszer, Aman Verma, Chi Bahk, Lawrence Madoff, John Brownstein, and David Buckeridge. 2016. "Media Content about Vaccines in the United States and Canada, 2012–2014: An Analysis Using Data from the Vaccine Sentimeter." *Vaccine* 34(50):6229–35.