# Using Twitter data to identify geographic clustering of anti-vaccination sentiments

Benjamin Brooks

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Public Health

University of Washington
2014

Committee:
Abraham Flaxman
Andrew Whitaker
Michael Hanlon

Program Authorized to Offer Degree:
Global Health

University of Washington

## Abstract

Using Twitter data to identify geographic clustering of anti-vaccination sentiments

Benjamin Brooks

Chair of the Supervisory Committee:
Abraham Flaxman, Assistant Professor
Department of Global Health

**Introduction:** Public opinion concerning vaccination is of interest since the publication of a study in 1999 (since retracted) linking the measles, mumps, and rubella (MMR) vaccine to autism; in its wake, parental fear of vaccination has risen, vaccination rates have decreased, and occurrence of outbreaks of vaccine-preventable diseases have increased. I examined vaccination-related opinions using data collected on the social networking site Twitter to determine whether particular geographic areas in the United States expressed more negative sentiment towards vaccination than others.

**Methods:** I tested this hypothesis by combining vaccination-related Twitter data with data published through the National Notifiable Disease Surveillance System, which provides weekly counts of newly diagnosed cases of vaccine-preventable diseases for each state. In the process of working towards this goal, I tested several different sentiment classification methods, collected a new body of vaccination-related Twitter data from 2014, and examined whether the average

sentiment expressed on Twitter in 2009 during the H1N1 pandemic was similar to the average sentiment in the same geographic areas in 2014.

**Results:** I was unable to find any meaningful correlation between the average opinion expressed in small geographic units of the United States in 2009 and 2014 or between the average opinion expressed by state and the mumps incidence rate observed over the period 2009-2013. I did note, however, that the proportion of tweets containing negative sentiment (between 5-10%) remained relatively stable in the data collected in 2014, which offers some hope that there is meaningful vaccination-related opinion expressed on Twitter that persists over time.

**Conclusion:** I believe that the lack of correlation observed in this study is a product of the aggregated nature of our outbreak data and differences in the content of negative opinion expressed in 2009 during the H1N1 pandemic and in 2014. Further research on this topic should focus on improving sentiment classification of tweets published when there is not an active pandemic and identifying data sources to validate the use of social media to monitor opinions around vaccination that contain vaccination rate or outbreak data at a more localized geographic level.

## Introduction

As users of the Web have evolved from passive consumers of content produced by a small number of providers to creators and distributors of information to others in their social networks, there is now an opportunity to use this information to monitor opinions and behaviors relevant to public health practitioners and researchers in real time. The potential for search engine data in this context is well known in relation to influenza, as Google Flu Trends tracks search volumes of flu-related keywords in order to estimate flu transmission activity [1]. While search queries are useful in tracking volume of web activity, social networking sites (SNSs), such as Twitter and Facebook, offer individual-level information on opinions and behaviors that previously might have been much more difficult and expensive to collect.

The potential of these platforms as both mediums of communication for outreach services and data sources for surveillance and research in medicine and public health has been widely discussed; for example, the *Journal of Medical Internet Research* (JMIR) was founded in 1999 by Dr. Gunther Eysenbach, a professor at the University of Toronto, with an initial three issues per year; as of 2013, it is published monthly. JMIR offers a recent e-collection of published articles titled "Medicine 2.0: Social Media, Open, Participatory, Collaborative Medicine" [2]. One of the articles included in this collection offers a systematic review of the use of social networking sites (SNS) in public health [3]. Capurro et. al. note that, while SNSs have been used since the late 1990s, they were never used as a platform for public health research until 2007. Some of the earliest studies used MySpace to study health risk behaviors among teenagers [4,5,6]. As MySpace plummeted from the United States' most popular website in 2006 to its current place outside of the top 500 websites by overall traffic, researchers have turned to other social media platforms, such as Twitter and Facebook, to examine public health issues [7,8]. The

acceleration of research of this type is evident after repeating the PubMed search query used in that systematic review, which yielded 431 results published at any time before March 31, 2012 (the time of its publication); the volume of research more than doubled with 551 articles published in the two years since (Figure 1) [9].

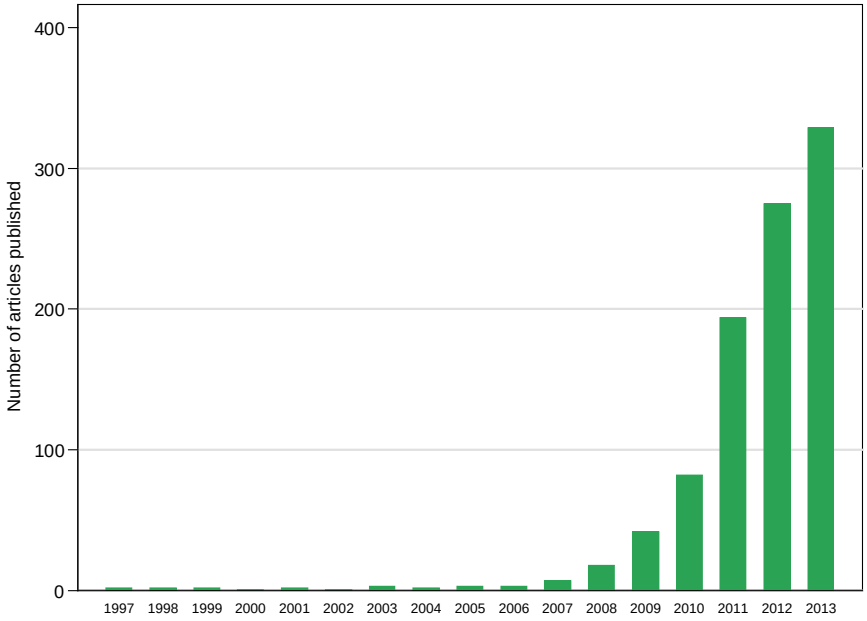


***Figure 1*:** Volume of academic work published using social networking sites for public health research, 1997-2013.

SNS data has been used in a wide array of applications, ranging from studies of the use of SNS during natural disasters to data from Twitter used to monitor misuse of antibiotics [10,11]. Although most of this research has been conducted in high-income countries, there is some work that has been done in low- and middle-income countries; for example, research conducted in South Africa used social network data to understand factors associated with low HIV/AIDS knowledge [12]. The emergence of this sort of research in middle-income countries in particular may continue as Twitter adds users, with large countries like Indonesia, Brazil, and Mexico entering the top ten in active Twitter users in 2013 [13].

***Use of social networking sites during H1N1 infuenza pandemic***

In 2009, the H1N1 influenza, also known as "swine flu", pandemic marked an important milestone in public health research based on social network data; as described in a subsequent paper studying the contents of tweets published about H1N1, it marked "the first instance in which a global pandemic has occurred in the age of Web 2.0 and presents a unique opportunity to investigate the potential role of these technologies in public health emergencies" [14]. A number of academic papers have been published based on data collected during the H1N1 pandemic and subsequent vaccine rollout. One of those studies, published in 2011, used a series of keywords to identify and collect Twitter data related to vaccination over a six month period after the H1N1 vaccine became available to the public [15]. The researchers developed a classifier by compiling a training dataset where students tagged tweets as containing positive, negative, neutral, or irrelevant sentiment toward the vaccine for about 10% of their data; this classifier was then used to categorize the rest of the tweets into one of the three bins.

While this study demonstrated that users with anti-vaccination opinions tended to cluster within the social network, the authors were primarily interested in analyzing the spread of opinions within a Twitter user's social network. They only used a basic measure to validate whether those opinion manifested themselves in measurable outcomes of public health concern (i.e., vaccination rates or disease outbreak). They used geographic information associated with individual Twitter accounts to compare the average "sentiment ratings" of different regions of the US to H1N1 vaccination rates and found a reasonably strong positive correlation (i.e., more positive sentiment, higher vaccination coverage).

### *Study aims*

My goal at the outset of this project was to extend this work by examining whether these clusters can be linked to particular geographic areas at the state or, preferably, sub-state level and whether those areas have experienced outbreaks of vaccine-preventable disease since the original link between autism and the MMR vaccine was published. Public opinion concerning vaccination is of interest since the publication of a study in 1999 (since retracted) linking the measles, mumps, and rubella (MMR) vaccine to autism [16]; in its wake, parental fear of vaccination has risen, vaccination rates have decreased, and occurrence of outbreaks of vaccine-preventable diseases have increased (for more information, see the *British Medical Journal*'s series describing the worldwide scare over the MMR vaccine) [17].

We tested this hypothesis by combining vaccination-related Twitter data with data published through the National Notifiable Disease Surveillance System, which provides weekly case counts of newly diagnosed cases of key infectious diseases (including those that are preventable through vaccine) for each state [18]. In the process of working towards this goal, we tested several different sentiment classification methods, collected a new body of vaccination-related Twitter data from 2014, and examined whether the average sentiment expressed on Twitter in 2009 during the H1N1 pandemic was similar to the average sentiment in the same geographic areas in 2014.

## Data and Methods

### *Twitter data retrieval and prospective collection*

Our first step was to extract data and code from the public GitHub repository made available by the previous study's authors [19]. We were able to extract 477,468 tweets containing one or more of the following vaccination-related keywords: *vaccination* OR *vaccine*

OR *vaccinated* OR *vaccinate* OR *vaccinating* OR *immunized* OR *immunize* OR *immunization* OR *immunizing*. These tweets were collected between August 2009 and January 2010 from 150,093 unique users. A portion of this data was used as a training dataset, made up of 39,297 tweets that had been labeled as containing positive, negative, neutral, or irrelevant sentiment towards vaccination by students at Penn State.

We were able to collect a supplemental 467,673 tweets with the same vaccine-related keywords between April 15 and July 25, 2014 from a total of 247,869 users, with a gap in data observed between April 22 and May 8 due to a technical problem. Figure 2 shows the volume of tweets collected over the period of data collection in 2009 and 2014.

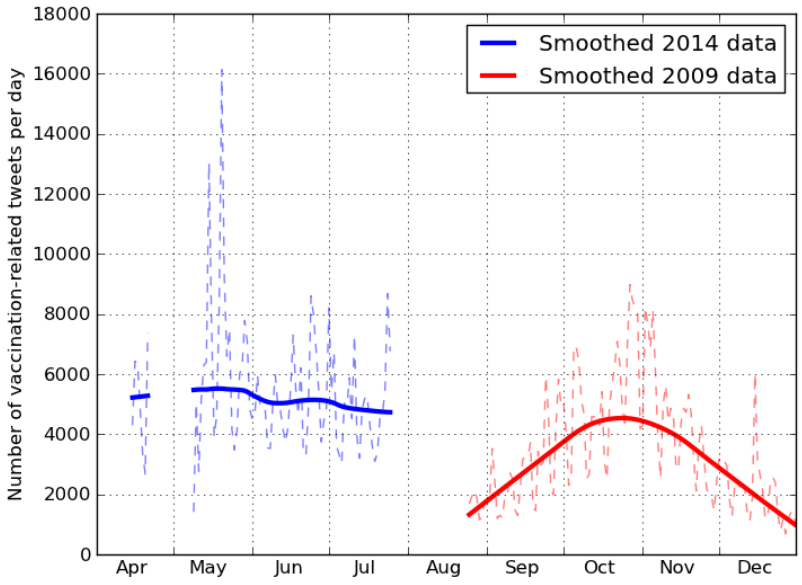


***Figure 2*:** Number of vaccination-related tweets observed per day in 2009 and 2014 data. Dotted lines show daily tweet volumes and solid lines show a smoothed trend that ignores weekly variation and days of unusually high discussion. Tweets collected in January 2010 are excluded from this figure for ease of interpretability. Even without an active pandemic of vaccine-preventable disease, increased overall Twitter usage has resulted in an increase in the overall discussion of vaccination on Twitter in 2014 relative to 2009.

### *Geolocation of Twitter users*

Twitter users are offered two options for sharing information about their geographic location. First, they can enable a geotagging option, which shares the GPS coordinates from which a user is tweeting. This option is used by only 0.6% of users in our dataset in 2014; the approach used by 70% of users in the same time period is a free-form text box where they can input information about their location. These locations can be very specific (e.g., "Capitol Hill, Seattle, WA") or may contain information that is practically useless in determining a user's actual location (e.g., "Planet Earth"). We used Python's *yql* package, which abstracts the API used to access Yahoo!'s PlaceFinder services, to discern user locations based on these free-form text responses. One advantage of using PlaceFinder is that it limits the depth of geographic information returned based on the type of information in the query (e.g., PlaceFinder will return a missing value for city if a user put "Texas" as their location). This feature is useful in this context, as it allows us to easily test hypotheses at varying depths of geographic granularity without assigning users to a depth of geographic information for which we are not confident of their location. We were able to geolocate 127,189 users from the data collected in 2014 to at least the country level; 61,649, or about half of those users, were identified as users located in the United States.

### *Training data for classifiers*

In restoring the data from 2009 to a relational database, we anecdotally noted some inconsistencies between the labels assigned in the training data and our own opinions of what qualified as a tweet containing negative sentiment. Indeed, the authors of the previous study acknowledge some discrepancies in the classifications applied by different students to the same tweets (i.e., interrater reliability).

In order to evaluate the reliability of the training data published in the original study, I reclassified 1,678 tweets from the training dataset and compared our classification to that found in the original training data. Rather than first separating out irrelevant tweets and then parsing the remaining corpus of data into positive, negative, and neutral, we simplified the classification scheme into only two classes: negative versus non-negative. We theorized that the real Twitter users of interest in our study are those who have negative opinions and, thus, might be less likely to vaccinate their children or be vaccinated themselves. We also expected that a simpler classification might result in improved detection of negative sentiment, rather than straining our classifier to distinguish between four classes of possible sentiment. This hypothesis is not well-studied in previous research, but some work published on the topic notes variation in the relative performance of different classification algorithms depending on the number of classes [20].

***Measuring algorithmic performance***

The measure of algorithmic performance presented in previous work on this topic is accuracy, which is the percentage of total tweets in a training dataset that were labeled correctly by the classifier. Given issues reproducing the exact classification mechanism used in the previous study and a new focus on detecting negative tweets, we decided to test several algorithms for detecting negative sentiment in tweets. The metrics presented here are precision and recall (also known as positive predictive value and sensitivity, respectively, in health literature). We chose the algorithm that produced the highest F1 score, calculated as the harmonic mean of precision and recall (Figure 3). We chose not to present any measure of pure accuracy due to the overwhelming majority of tweets containing non-negative sentiment; as a result, a classifier which simply labeled all tweets as non-negative would have a relatively high accuracy. From a public health perspective, where the idea of sensitivity is most well known in

the context of diagnostic tests, it's important to note that a "positive test result" here is akin to the algorithm detecting negative sentiment.

***Sentiment classification algorithms***

We tested the four following sentiment classification algorithms, which are described below:

- AFINN
- Custom vaccination keywords
- Logistic regression
- Naïve Bayes

AFINN, named for its creator, is a list of words rated on a scale from -5 (very negative) to 5 (very positive) that was developed for sentiment analysis in microblogs (i.e., Twitter, etc.) [21]. The AFINN dictionary includes 2,477 words with an associated sentiment strength. We classified tweets with an overall negative score based on the AFINN dictionary as negative, and remaining tweets as positive. Because the AFINN dictionary is designed to detect sentiment across topics, we also produced a custom set of vaccination-related keywords that we hypothesized to be associated with negative opinions around vaccination. We used this shortened dictionary to tag tweets; any tweet that contained at least one of these words was tagged as a negative tweet (for a list of words, see Appendix A).

The logistic regression classifier parses all tweets in our training data into single words and estimates a coefficient for each word based on its association with negative tweets. Using these coefficients, the classifiers assigns a value between zero and one estimating the probability of a tweet being negative. Tweets were dichotomized into predictions of positive or negative sentiment based on whether that predicted probability exceeded (i.e., predicted negative) or was less than 0.5. Similarly, we tested a naïve Bayes classifier, which, along with the logistic regression classifier, was implemented with Python's *scikit-learn* package, which was also used to evaluate the recall, precision, and F1 score of all of our classifiers.

***Figure 3:*** Metrics used to evaluate sentiment algorithm performance.

| ***Precision (positive predictive value)*** | ***Recall (sensitivity)*** | ***F1 score*** |
|---|---|---|
| $P = \frac{\sum True\ negative\ tweets}{\sum Classified\ negative\ tweets}$ | $R = \frac{\sum Classified\ negative\ tweets}{\sum True\ negative\ tweets}$ | $F1 = \frac{2PR}{P+R}$ |

*Vaccination rate and outbreak data*

In the United States, vaccination status data is collected regularly through the National Immunization Survey and used to estimate vaccination coverage for children aged 19-35 months in each state as well as select metropolitan areas [22]. Coverage estimates are provided for each routine childhood vaccination and used to create aggregate measures such as the percentage of children who have completed the entire routine vaccination series. The most recent year for which this data is publicly available is 2012.

Outbreaks of infectious disease are reported through the National Notifiable Disease Surveillance System, which publishes a weekly report of the number of new cases observed of a variety of diseases, including those that are preventable through routine vaccination [17]. Again, this data is reported at the state level and for select large metropolitan areas. We extracted yearly counts of new cases observed for each year and state from 2009-2013 for mumps as a measure of the degree to which outbreaks of diseases directly preventable through the MMR vaccine have been observed and calculated incidence rates using 2013 state population estimates from the United States Census [23].

## Results

### *Sentiment classification and training data*

We compared the classification of the 1,678 tweets we identified as either negative or non-negative and compared it to the classification assigned in the original training dataset. Of

those tweets that we classified as negative, the original training data only identified 36% as negative, which suggests a surprising lack of agreement between our definition of a negative tweet and that found in their training data. As a result of this disagreement, we chose to train and test each algorithm using our reclassified training data.

In evaluating our sentiment algorithms, we found that logistic regression provided the best overall classification for sentiment around vaccination when using the combined measure of precision and recall as the metric of success. That said, only our custom sentiment classifier was able to identify "true" negative tweets (i.e., recall) at a reasonably high level, with a recall of 59%. Precision and recall statistics were not presented in the original study, and we were unable to reproduce the exact classification approach used and thus were unable to compare the performance of our sentiment classification algorithms directly.

| Sentiment algorithm | Precision | Recall | F1 score |
|---|---|---|---|
| *Logistic regression* | 70% | 28% | 40% |
| *AFINN* | 25% | 41% | 31% |
| *Naïve Bayes* | 79% | 19% | 31% |
| *Custom sentiment* | 19% | 59% | 29% |

***Table 1*:** Precision and recall of algorithms tested for detection of negative sentiment in vaccination-related tweets.

### *Correlation of vaccination sentiment over time*

When comparing the average sentiment expressed on Twitter from small areas of the United States over the two periods in time for which we had data, we were unable to identify any meaningful correlation. We compared the average sentiment of Twitter users in data collected during the H1N1 epidemic in 2009 and compared them to the average sentiment of users in 2014. We found 323 counties with at least 20 users in each period of data collection, and 233 cities with at least 20 users in each period. In both cases, we did not see any correlation between

average sentiment across the two time periods (Figure 4). The Pearson correlation coefficient when comparing either cities or counties did not exceed 0.11.

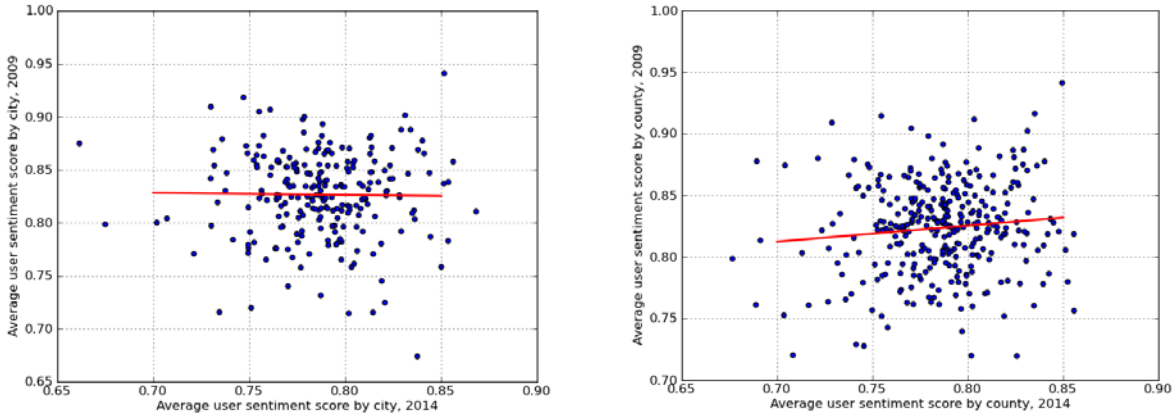


***Figure 4:*** Average user sentiment in cities (left; $R^2$ = -0.02) and counties (right; $R^2$ = 0.11) in the United States with at least 20 users tweeting about vaccination in 2009 and 2014. Red lines show the linear fit of the data. A lower average sentiment score represents more negative sentiment being expressed.

We compared the average sentiment observed by users in each state in 2014 to the mumps incidence rate per 100,000 for all cases observed over the period 2009-2013. In this case, we would expect a negative correlation; that is, as average sentiment score increases, the mumps incidence rate decreases. However, that relationship was not observed in our data; in fact, it appears that some of the states with the highest average sentiment scores also had the highest mumps incidence rates (Figure 5).

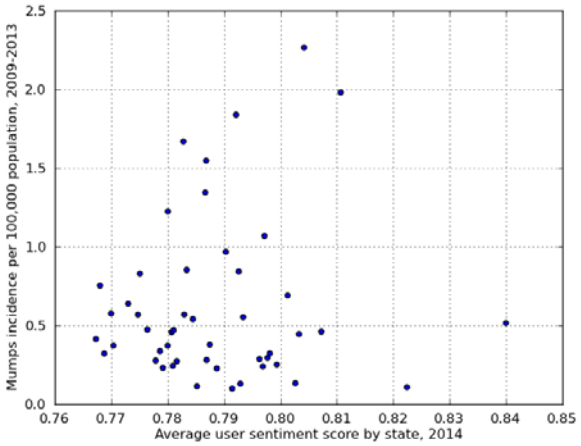


***Figure 5:*** Comparison of average vaccination sentiment expressed by users in 2014 to the mumps incidence rate observed over the period 2009-2013 by state. In this case, we would have expected a negative correlation, with areas with more negative sentiment experiencing a higher incidence rates. New York and New Jersey are excluded from this plot because of exceptionally high incidence rates of 16.7 and 7.2 per 100,000, respectively, as is Rhode Island because of its unusually low average sentiment score of 0.72.

*Stability of negative sentiment over time*

In order to evaluate the stability of negative sentiment expressed on Twitter around vaccination, we used our custom keywords algorithm to calculate the proportion of tweets observed on each day in our 2014 data containing at least one of the words we believed to be associated with negative vaccination sentiment. For this exercise, we chose to use the custom vaccination keywords algorithm because it scored highest using the recall metric, which gives the proportion of true negative tweets that an algorithm also classified as negative. We observed a reasonably stable trend over time, with the fraction of tweets containing negative keywords generally hovering between five and 10 percent of total vaccination-related tweets (Figure 6). This result is promising because it suggests that the signal we are attempting to capture persists over time and might offer some information of public health importance if combined appropriately with vaccination rate and outbreak data.

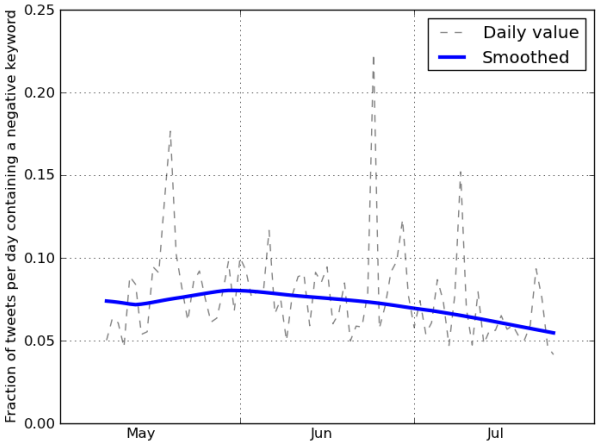


***Figure 6:*** Fraction of vaccination-related tweets containing a negative keyword per day in 2014 data collected. Data collected in April excluded for ease of interpretability.

## Discussion

Our attempts to use Twitter data to identify geographic pockets of anti-vaccination sentiment has attuned us to the challenges that must be overcome in order to derive utility for public health practitioners in this context. Broadly, those challenges fall into three categories: accurate classification of sentiment, accurate geographic classification of users, and potential sample bias in the users from which data is collected. Without outbreak data disaggregated to a more localized geographic level and a classifier that can more consistently determine whether a tweet contains negative sentiment or not, it will remain difficult to use Twitter data to monitor public sentiment around vaccination accurately. Furthermore, even if an improved classifier were produced, the ability to validate its usefulness as a predictor of potential outbreaks, either outbreak or vaccination rate data would need to be available at a more granular level.

While we did generate a new training dataset that we believe produced more plausible classification of tweets into negative and non-negative bins, we did not test for any interrater reliability or test-retest reliability. All tweets were categorized by a single person and only at a

single point in time. Our training dataset was much smaller than that provided by the previous study, and it is plausible that the degree of disagreement between our training data and that produced in the previous study is a combination of a relatively small sample size and a lack of testing for reliability of the ratings assigned in our training data.

***Reproducibility***

Given the heavy reliance of this thesis on previous work on H1N1 vaccination sentiments expressed on Twitter, it has unexpectedly proven to be a case study in reproducibility of published scientific research. We struggled at the outset to reproduce the results presented in earlier work, generally as a byproduct of issues understanding the relational data structure used to store information and the software written to produce results for the final report. Below, we offer some suggestions for best practices for publishing relational data and code used in producing scientific research.

There is no agreed upon standard for publishing relational databases used in academic work. However, a "data dump" in some delimited text format should be accompanied with a database schema that establishes variable names, each variable's contents, and relationships (i.e., foreign keys) that exist between tables. In cases where some "gold standard" dataset is divided into testing and training datasets, it should be made clear exactly what data were included in the test dataset and which were included in the training dataset.

If data from a relational database is being included in a repository, each query that was used to generate results presented in the publication should be clearly stated. Following each of these queries should be the code that produces any figure or result presented in the paper. As a rule of thumb, I think publishing code that was written at some point in the development of the final publication but did not actually produce any results presented should not be included in a

public repository. In this experience, trying to work with code that I had never seen before with no contact with the original author is hard enough; when there are scripts included that aren't even relevant, it becomes that much more difficult to parse out which produced published results and which were eliminated as part of the research process. That said, I can understand there are cases where including additional code may help prove that the authors were comprehensive in their approach, or perhaps that their unpublished work may prove useful to someone else. In these cases, I would include this sort of code in a public repository, as long as it is very clearly separated from any work that went into producing results shown in the academic work. We are producing a public branch of the GitHub repository that was used to conduct this work that adheres to these standards as an example of how we think reproducibility could be improved in this context [24].

***Possible explanations for null result***

Our approach of using a sentiment classifier developed using data collected during the H1N1 pandemic to classify tweets published in 2014 relied on the assumption that opinions around vaccination expressed today would correlate with opinions expressed around H1N1 in 2009. Given that there is not a strong correlation between H1N1 vaccination rates in 2010 and MMR vaccination rates in 2012 (the most recent year for which data has been published), it is plausible that this hypothesis was incorrect and that the overlap between the population that chose not to vaccinate against H1N1 may be different than the population that chooses to forego routine childhood vaccinations (Figure 7).

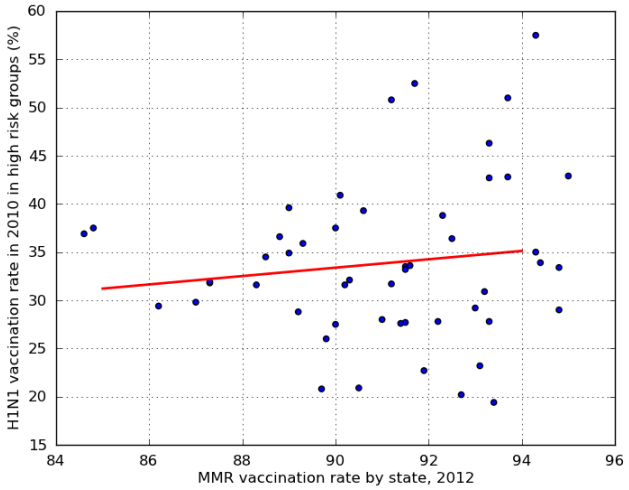


***Figure 7:*** Relationship between H1N1 vaccination rate in high risk groups in 2010 and MMR vaccination rates in 2012, by state. A weakly positive relationship suggests that it may not be appropriate to correlate opinions expressed about the H1N1 vaccine with those expressed about general childhood vaccination ($R^2 = 0.13$).

Qualitatively, reasons for choosing not to vaccinate against H1N1 are likely different from reasons parents choose not to vaccinate their children against MMR. H1N1 was seen as threat across the population, but the lack of severity of the disease for the majority of people who contracted it and questions of the vaccine's efficacy likely influenced individual decisions to vaccinate. Alternatively, the reasons for not participating in routine childhood vaccinations such as MMR are products of an entirely different set of challenges. Published work has pointed to a number of factors influencing childhood vaccine refusal, ranging from perceived low susceptibility to the disease, fear of vaccine safety, and state laws that affect the ease with which parents can receive non-medical vaccine exemptions [25, 26].

All of these limitations and challenges aside, I do not believe that this work is necessarily entirely in vain. A classifier developed explicitly to examine opinions around general vaccination, rather than using a classifier that was developed based on data from the H1N1 pandemic, might have a much higher rate of accurately detecting negative sentiment in tweets. Combining an appropriately trained classifier with an increased effort to distinguish between

informational and sentimental tweets would likely make for dramatic improvements in performance. Some research on this topic has shown that, although again in the context of influenza, controlling for "media chatter" can improve the ability of Twitter data to identify meaningful public health information [27].

Given the state-level limit on the geographic granularity of outbreak and vaccination rate data, there are also a couple of opportunities to seek improved validation data. Among its many publications of key public health indicators at the county level, the Institute for Health Metrics and Evaluation has made some attempts to estimate vaccination rates in small areas in the United States; however, this work was never completed [28]. If county-level vaccination rates were published, it would be enormously useful in determining the potential of this approach to monitoring opinions around vaccination. Additionally, in the wake of more frequent outbreaks, the Council on Foreign Relations has used media reports to generate a map visualizing outbreaks of vaccine-preventable disease since 2008. This data is not standardized for analysis and may not be a comprehensive list of all outbreaks observed but at least provides some increased geographic granularity as to where a fraction of outbreaks have occurred [29].

The emphasis that policy-focused groups like the Council on Foreign Relations have placed on generating awareness for recent outbreaks of vaccine-preventable disease globally underscores the importance of understanding public opinion around vaccination. While social media offers only one of many avenues for analyzing these opinions, it is a source that is updated constantly and thus can provide insights to public health decision makers in real time. That said, the challenges of using Twitter in public health are many; discerning user sentiment from messages of 140 characters or less is challenging, and resolving users to a specific geographic location is not trivial. The expanding body of research in public health using social media,

however, offers evidence of its utility as a data source and medium of communication. We hope that future work continues to exploit the potential of social media to help maintain high vaccination rates, which remains one of public health's greatest success stories.

## References


[1] Ginsberg J, Mohebbi MH, et. al. (2009) “Detecting influenza epidemics using search engine query data.” *Nature.* DOI: 10.1038/nature07634.

[2] Various authors. (2011-2014) “Medicine 2.0: Social Media, Open, Participatory, Collaborative Medicine” *Journal of Medical Internet Research.* Available: http://www.jmir.org/themes/52. Accessed 20 August 2014.

[3] Capurro D, Cole K, et. al. (2014) “The Use of Social Networking Sites for Public Health Practice and Research: A Systematic Review.” *Journal of Medical Internet Research.* DOI: 10.2196/jmir.2679.

[4] Moreno MA, Parks M, Richardson LP. (2007) “What are adolescents showing the world about their health risk behavior on MySpace?” *Medscape General Medicine*.

[5] Moreno MA, Parks MR, et. al. (2009) “Display of health risk behaviors on MySpace by adolescents: prevalence and associations.” *JAMA Pediatrics (formerly Archive of Pediatrics & Adolescent Medicine).* DOI: 10.1001/archpediatrics.2008.528.

[6] Moreno MA, Vanderstoep A, et. al. (2009) “Reducing at-risk adolescents’ display of risk behavior on a social networking web site: a randomized controlled pilot intervention trial.” *JAMA Pediatrics (formerly Archive of Pediatrics & Adolescent Medicine).* DOI: 10.1001/archpediatrics.2008.502.

[7] Cashmore P. (11 July 2006) “MySpace, America’s Number One.” *Mashable*, Available: http://mashable.com/2006/07/11/myspace-americas-number-one/. Accessed 20 August 2014.

[8] “Site overview: myspace.com.” *Alexa*, Available: http://www.alexa.com/siteinfo/myspace.com. Accessed 19 August 2014.

[9] Corlan AD. (2004) “Medline trend: automated yearly statistics of PubMed results for any query.” Available: http://dan.corlan.net/medline-trend.html. Accessed 19 August 2014.

[10] Huang C, Chan E, Hyder AA. (2010) “Web 2.0 and Internet Social Networking: A New tool for Disaster Management? – Lessons from Taiwan.” *BMC Medical Informatics & Decision Making*. DOI:10.1186/1472-6947-10-57.

[11] Scanfeld D, Scanfeld V, Larson EL. (2010) “Dissemination of health information through social networks: twitter and antibiotics.” *American Journal of Infection Control*. DOI: 10.1016/j.ajic.2009.11.004.

[12] Wagenaar BH, Sullivan PS, Stephenson R. (2012) ”HIV Knowledge and Associated Factors among Internet-Using Men Who Have Sex with Men (MSM) in South Africa and the United States.” *PLoS ONE.* DOI: 10.1371/journal.pone.0032915.

[13] Richter F. (20 November 2013) “Twitter’s Top 5 Markets Account for 50% of Active Users.” *Statista*. Available: http://www.statista.com/chart/1642/regional-breakdown-of-twitter-users/. Accessed 19 August 2014.

[14] Chew C, Eysenbach G. (2010) “Pandemics in the Age of Twitter: Content Analysis of Tweets during H1N1 Outbreak.” *PLoS ONE.* DOI: 10.1371/journal.pone.0014118.

[15] Salathé M, Khandelwal S. (2011) “Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control.” *PLoS Computational Biology.* DOI: 10.1371/journal.pcbi.1002199.

[16] Wakefield AJ, SH Murch, et. al. (1998) “RETRACTED: Ileal-lymphoid-nular hyperplasia, non-specific colitis, and pervasive developmental disorder in children.” *The Lancet*. DOI: 10.1016/S0140-6736(97)11096-0.
[17] Deer B. (2011) “How the case against the MMR vaccine was fixed.” *BMJ*. DOI: http://dx.doi.org/10.1136/bmj.c5347.
[18] Centers for Disease Control and Prevention. (2014) “National Notifiable Diseases Surveillance System: Morbidity and Mortality Weekly Report Tables.” Available: http://wonder.cdc.gov/mmwr/mmwrmorb.asp. Accessed 20 August 2014.
[19] Salathé M, Khandelwal S. (2011) “Data and code from ‘Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control.’” *GitHub*. Available: https://github.com/salathegroup/vaccine-sentiment. Accessed 20 August 2014.
[20] Drake A, Ringger E, Ventura D. (2008) “Sentiment Regression: Using Real-Valued Scores to Summarize Overall Document Sentiment.” *IEEE International Conference on Semantic Computing*. DOI: 10.1109/ICSC.2008.67.
[21] Nielsen FA. (2011) "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs.” *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts’*.
[22] Centers for Disease Control and Prevention. (2012) “National Immunization Survey: Coverage with Individual Vaccines and Vaccination Series by State and Local area.” Available: http://www.cdc.gov/vaccines/imz-managers/coverage/nis/child/data/tables-2012.html. Accessed 20 August 2014.
[23] United States Census Bureau. “Population estimates by state: 2013.” Available: https://www.census.gov/popest/data/index.html. Accessed 20 August 2014.
[24] Brooks BP, et. al. (2014) “Data and code accompanying ‘Using Twitter data to identify geographic clustering of anti-vaccination sentiments.’” *GitHub*. Available (in progress): https://github.com/uwescience/twittervaccine/tree/public. Accessed 20 August 2014.
[25] Omer SB, Salmon DA, et. al. (2009) “Vaccine Refusal, Mandatory Immunization, and the Risks of Vaccine-Preventable Diseases.” *New England Journal of Medicine*. DOI: 10.1056/NEJMsa0806477.
[26] Atwell JE, Van Otterloo J, et. al. (2010) “Nonmedical Vaccine Exemptions and Pertussis in California, 2010.” *Pediatrics*. DOI: 10.1542/peds.2013-0878.
[27] Broniatowski DA, Paul MJ, Dredze M. (2013) “National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic.” *PLoS ONE*. DOI: 10.1371/journal.pone.0083672.
[28] Mokdad A. Personal correspondence. May 2014.
[29] Council on Foreign Relations. “Vaccine-Preventable Outbreaks Map.” Available: http://www.cfr.org/interactives/GH_Vaccine_Map/. Accessed 20 August 2014.

***Appendix A:*** Words and phrases used to tag tweets as containing negative sentiment in custom vaccination keywords algorithm.

- “shouldn’t”
- “should not”
- “must not”
- “mustn’t”
- “don’t”
- “do not”
- “won’t”
- “will not”
- “renounce”
- “renounced”
- “renounces”
- “boycott”
- “refuse”
- “mandatory”
- “forced”
- “forces”
- “forcing”
- “coerce”
- “coerced”
- “coercing”
- “required”
- “requires”
- “have to”
- “avoid”
- “never”
- “untested”
- “poison”
- “not necessary”
- “not vaccinating”
- “not vaccinated”
- “not vaccinate”