BMJ Open

Real-Time Identification of Influenza Vaccination Behavior from Online Self Reports

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2018-024018
Article Type:	Research
Date Submitted by the Author:	04-May-2018
Complete List of Authors:	Huang, Xiaolei; University of Colorado Boulder, Department of Information Science Smith, Michael; George Washington University, Department of Engineering Management & Systems Engineering Jamison, Amelia; University of Maryland, Center for Health Equity Broniatowski, David; George Washington University, Department of Engineering Management & Systems Engineering Dredze, Mark; Johns Hopkins University, Department of Computer Science Quinn, Sandra; University of Maryland , Department of Family Science; University of Maryland , Center for Health Equity Cai, Justin; University of Colorado, Department of Computer Science Paul, Michael; University of Colorado Boulder, Department of Information Science; University of Colorado, Department of Computer Science
Keywords:	Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, World Wide Web technology < BIOTECHNOLOGY & BIOINFORMATICS, PUBLIC HEALTH, EPIDEMIOLOGY

Title:
Real-Time Identification of Influenza Vaccination Behavior from Online Self Reports

Authors:
Xiaolei Huang[1], Michael C. Smith[2], Amelia M. Jamison[3], David A. Broniatowski[2], Mark Dredze[4], Sandra C. Quinn[3,5], Justin Cai[6], Michael J. Paul[1,6]

[1] Department of Information Science, University of Colorado, Boulder, CO 80309, USA
[2] Department of Engineering Management & Systems Engineering, George Washington University, Washington, DC 20052, USA
[3] Center for Health Equity, School of Public Health, University of Maryland, College Park, MD 20742, USA
[4] Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA
[5] Department of Family Science, School of Public Health, University of Maryland, College Park, MD 20742, USA
[6] Department of Computer Science, University of Colorado, Boulder, CO 80309, USA

Corresponding Author:
Michael J. Paul
Assistant Professor, Department of Information Science
315 UCB, Boulder, CO 80309, USA
1-217-552-3605
mpaul@colorado.edu

Word Count: 2,912

ABSTRACT

Introduction: The Centers for Disease Control and Prevention (CDC) spend significant resources to track influenza (flu) vaccination coverage each flu season. Emerging data from social media provide an alternative solution to surveillance at both national and local levels of flu vaccination coverage in near real-time.

Objectives: This study aimed to characterize and analyze the vaccinated population from temporal, demographic, and geographical perspectives using a new methodology: automatic classification of vaccination-related Twitter data.

Methods: We continuously collected tweets containing both flu-related terms and vaccine-related terms covering four consecutive flu seasons from 2013 to 2017. We created a machine learning classifier to identify relevant tweets, then evaluated our approach by comparing to data from the CDC.

Results: We found strong correlations of .80 between monthly Twitter predictions and CDC, with correlations as high as .95 in individual flu seasons. We also found that our approach obtained geographic correlations of .39 at the state level and .47 the regional level. Finally, we found a higher level of flu vaccine tweets among female users than male users, also consistent with the results of CDC surveys.

Conclusion: Significant correlations between our approach and CDC show the potential of using social media for vaccination surveillance. Temporal variability is captured better than geographic and demographic variability. We discuss potential paths forward for leveraging this approach.

Keywords: vaccination, surveillance, influenza, biostatistics, time-series

ARTICLE SUMMARY

Strengths and limitations of this study

- This study shows that vaccination behaviors – specifically, receiving or intending to receive a flu vaccine – can be detected and measured through Twitter.
- The signal from Twitter, which is available in real-time, closely tracks US government data.
- The proposed approach correlates moderately with geographic and demographic trends.
- The proposed approach is most robust at broad granularities, such as the national level, and has weaker performance within finer-grained geographic and demographic groups.

INTRODUCTION

The Advisory Council for Immunization Practices (ACIP) at the Centers for Disease Control and Prevention (CDC) recommends annual influenza vaccination for all healthy adults.[1] Furthermore, CDC urges individuals to get vaccinated early in the flu season, from October through January.[2] Yet, it can be difficult for researchers and practitioners working to improve influenza vaccine uptake to get accurate information in real time. Existing influenza immunization surveillance techniques have known limitations: traditional survey-based methods are time-consuming and expensive, and newer reimbursement-based systems fail to accurately capture a representative sample of population.[3]

Two national surveillance systems enable public health professionals to access information on influenza vaccine uptake. The most accessible of these systems is the CDC's FluVaxView, which aggregates uptake data from several national surveys.[4] The CDC data provide accurate estimates of vaccine uptake, although with some time lag. The earliest reports are only available after flu seasons typically peak, and final estimates are generally published at the open of the following flu season in September or October. Additionally, the panel surveys that inform the reports are expensive, take months to administer and process, and may undersample populations without a landline phone, particularly minority populations, young adults, and adults living in urban areas.[5, 6] A second system,[7] provided by the National Vaccine Program Office, uses an online tool to "live-track" influenza vaccination insurance claims from Medicare beneficiaries. While this system reduces lag time between vaccination and reporting, it only captures the population enrolled in Medicare, adults over age 65 and those under 65 living with disabilities.[7]

Social media data have revolutionized infectious disease surveillance, particularly for seasonal and pandemic influenza.[8-10] Utilizing data from social media platforms (like Twitter or Facebook), search engines (like Google), and other internet-based resources (like blogs), researchers have been able to track the spread of disease in real time with relatively high accuracy.[9] A recent meta-analysis of social media influenza surveillance efforts found that in a comparison to national health statistics (primarily from the CDC), correlation between social media data and national statistics ranged from 0.55 to 0.95,[11, 12] and the majority of projects were able to predict outbreaks more quickly than traditional surveillance methods.[10] Of these studies, the most accurate systems have harnessed natural language processing methods to identify relevant tweets.

With the development of new tools and techniques, social media data have the potential to similarly reshape the practice of influenza immunization surveillance. However, to our knowledge, no studies have attempted to utilize social media data to track influenza vaccine intentions and behaviors at the national level. To date, efforts to track influenza vaccination through social media have been much less frequent than efforts to track disease. Researchers are more likely to focus on the use of social media as a health communication tool than to explore the potential for immunization surveillance.[13] Some studies have been able to use social media data to track vaccine sentiment and general attitudes towards vaccines.[14–16] Others have focused on the spread of vaccine sentiment across online social networks.[17, 18] Some vaccine-specific studies have also attempted to use social media to identify geographic differences in vaccine uptake.[19, 20] The possibility of efficiently tracking influenza immunization in real-time is promising, but the true value of any new data source is limited without validation against known metrics.[14, 21, 22] To successfully use social media data in immunization surveillance efforts, an important first step is to validate observed trends against national survey data. In this study, we sought to validate observed patterns from Twitter, using tweets expressing either intention to seek immunization or receipt of influenza immunization, against influenza immunization data from the CDC for four consecutive flu seasons from 2013-2017.

METHODS

Patient and Public Involvement

This study did not involve patients.

Data

Twitter Data

We continuously collected tweets containing the terms “flu” or “influenza” since 2012 using the Twitter streaming Application Programming Interface (API), as part of data described in our prior work.[23] For this study, we filtered influenza-related tweets containing at least one vaccine-related term (“shot(s)”, “vaccine(s)”, and “vaccination”). We then inferred the US state for tweets using the Carmen geolocation system,[24] and the gender of each Twitter user of the dataset using the Demographer tool.[25] The Carmen tool infers locations of tweets by three main sources, coordinates of tweets, places name of tweets and locations in user profiles. The Demographer tool infers genders of Twitter users by the names of their profiles. We removed retweets, non-English tweets and the tweets not located in US. We obtained 1,124,839 tweets from

742,802 Twitter users covering four consecutive flu seasons from 2013 to 2017. More details can be found in the supplementary material (A1 and A2).

In addition to tweets about influenza vaccine, we also collected a random sample of tweets from all of Twitter. This was used to adjust the vaccine counts by time, location, and demographics, described below. The random sample includes approximately 4 million tweets per day since 2011.

CDC Data

We utilized CDC data on influenza vaccination of the four flu seasons for validating our approaches. The CDC data were downloaded from the CDC's FluVaxView system.[4] These data include vaccination coverage by month, by states, and by geographic regions as defined by the US Department of Health and Human Services (HHS). The CDC's estimates are based on several national surveys: the Behavioral Risk Factor Surveillance System (BRFSS, which targets adults), the National Health Interview Survey (NHIS), and the National Immunization Surveys (NIS, which focuses on children). In this study, we use the CDC data for adults (≥18 years old) across all racial/ethnic groups.

Automated Classification

In our study, we used natural language processing techniques to preprocess and encode tweets into feature vectors, we fed the vectors to build machine learning classifiers to automatically categorize the Twitter data that express vaccination behavior. Tweets were classified into yes or no labels in response to the question, "Does this message indicate that someone received, or intended to receive, a flu vaccine?" Specifically, we randomly sampled 10,000 tweets from our collected data starting from 2012 to 2016 and then used a crowdsourcing platform to annotate the 10,000 tweets,[26] using quality control measures to ensure accurate annotations. The classifiers were trained by the annotated tweets.

The best-performing classification model was a convolutional neural network (CNN), which had a precision (the proportion of tweets classified as vaccine intention/receipt that were correctly classified) of 89% and recall (the proportion of vaccine intention/receipt tweets that were identified by the classifier) of 80%, measured using nested five-fold cross-validation. This classifier was applied to the full dataset of 1.1 million tweets, of which 366,698 were classified as expressing that someone received or intended to receive an influenza vaccine. More details of preprocessing and encoding tweets, building and selecting machine learning models can be found in supplementary materials (A.2).

Trend Extraction and Validation

To evaluate the reliability of our Twitter classification model as a source for vaccination surveillance, we compared the Twitter data to CDC data along three dimensions: time (by month), location (by US state and region), and demographics (by gender). Specifically, CDC FluVaxView provides the monthly percentage of American adults who received an influenza vaccination in a given month in each state, as well as the percentage of Americans who report vaccination in different demographic groups each flu season.

To extract trends over time, we computed the number of vaccine intention/receipt tweets in each month per season, excluding June (the CDC does not report data for June). We only included tweets geolocated to the US. To adjust for variations in Twitter over time, we normalized the monthly counts by the number of tweets in the same month from a large random sample of tweets.[8] In addition to monthly rates for direct comparison to CDC, we also calculated weekly tweet rates, providing estimates at a finer time granularity than reported by the CDC. For monthly time series data, we applied an autoregressive integrated moving average (ARIMA) model and linear regression to predict the CDC data from the Twitter data.[27]

To extract trends by location, we computed the number of intention/receipt tweets in each of the 10 HHS regions and each of the 50 US states. We created per-capita estimates by dividing each count by the number of tweets from the same region or state from a random sample of tweets.

To extract trends by gender, we computed the number of intention/receipt tweets identified as male or female, divided by the corresponding counts from a random sample. We computed this proportion within each US state before aggregating the counts from all states, to additionally adjust for gender variation across location (we provided detailed validation steps and additional experiments in supplementary material A.3).

RESULTS

Activity by Time

Table 1 shows the correlation between the classified tweets and CDC data from the ARIMA results. The correlations are significant ($p<.01$) for all seasons. Figure 1 shows the values from both data sources over time, standardized with z-scores. While the CDC data are only available by month, we show Twitter counts by week (Sunday to Saturday), to illustrate the finer temporal granularity that is possible. In both data sets, there are seasonal peaks every October, when influenza vaccines are distributed in the

US. While the overall shapes are very similar, the Twitter data sometimes shows rises later in the flu season that do not correspond to a similar rise in the CDC data, especially in the 2013-14 season, which results in the lowest correlation.

Table 1. Pearson correlations by month in each flu season. * = p<0.05, ** = p<0.01, *** = p<0.001.

	All seasons	2013-14	2014-15	2015-16	2016-17
Monthly	.80 ***	.64 **	.95 ***	.91 ***	.91 ***

Activity by Location

The prevalence of tweets mentioning vaccine intention/receipt is shown in Figure 2, where darker color indicates more frequent vaccine mentions. We observe that states in the northwest, especially Washington and Oregon have higher rates than southeastern states, such as Florida and Alabama. There is a moderate correlation between the geographic patterns in the Twitter data compared to the CDC data, with a higher correlation at the HHS region level than at the state level (Table 2). The strength of the correlations varies by season, with much stronger correlations in the first two seasons than the latter two seasons.

Table 2. Pearson correlation by geography in each season. * = p<0.05, ** = p<0.01, *** = p<0.001.

	All seasons	2013-14	2014-15	2015-16	2016-17
State	.39**	.30*	.21	.05	.03
HHS Region	.47	.69*	.57	.14	.24

Activity by Gender

Female users are much more likely to tweet about vaccine intention/receipt than male users on Twitter. The female-to-male ratios in each of the four seasons are, respectively: 1.967, 1.727, 1.586, 1.468. This ratio is higher than in the CDC data (1.184, 1.172, 1.186, 1.196). However, the two data sources are in qualitative agreement: the vaccination rate is higher among females than males. For example, in the 2016-17 flu season, the CDC reported that among American adults, 47.0% of women were vaccinated for influenza, compared to 39.3% of men.

We visualized the gender weekly trends and gender ratio of vaccine coverage for male and female in the Figure 3. The plot of gender weekly trends shows the volume of vaccine intention/receipt tweets over time. The gender ratio has also decreased steadily over time in the Twitter data, while it has stayed fairly constant in the CDC data. The plot of gender ratio shows the female-to-male ratio of vaccine intention/receipt tweets within each US state, with darker color indicating a higher ratio. For example, the figure shows that West Virginia has more females mentioning influenza vaccine behavior than males. (We provided additional analyses in the supplementary material A.4.)

DISCUSSION

This study demonstrates that, by utilizing natural language processing techniques, Twitter data can be efficiently analyzed to identify meaningful information about influenza vaccination intentions and behaviors. When validated against CDC data, we observed a very strong correlation between the monthly Twitter-based predictions and official CDC uptake estimates. Furthermore, the consistency of our observations over the entire four-year period suggests that our classifiers are working well to reduce the noise in the Twitter dataset and hone in on the very specific set of tweets related to vaccine behaviors. These findings alone are very promising, suggesting that Twitter data can be incorporated as new resources for public health practitioners and researchers interested in accessing vaccine uptake data in real time.

This is one of the first studies to have utilized Twitter data to track vaccination behavior, and many of our analyses were exploratory. In addition to validating our observations of vaccine behavior, we conducted analyses to explore patterns temporally, geographically, and across demographic groups. Traditionally, surveillance efforts have focused on monthly or yearly data. Our Twitter dataset allows for greater flexibility and specificity when assessing temporal trends in vaccination. In addition to monthly estimates, we were able to assess tweets weekly. Although we are unable to compare our weekly counts to a validated national metric, we observed high week-to-week variability in general flu vaccine tweets before applying a classifier to filter out irrelevant tweets, but a relatively consistent and predictable pattern in week-to-week tweets indicating vaccine intention and receipt.

Our early results show that it is possible to capture geographic variability in Twitter data. These results suggest some similarities with the CDC FluVaxView maps, but the associations are not strong enough to make definitive conclusions based on geography. There may be local level trends that contribute to these observed patterns, for instance, both Washington and Oregon have higher-than-average rates of childhood vaccine exemptions, and users may feel the need to be more vocal about their vaccine choices.[28] Or during the 2016-17 flu season, several news outlets focused on the

severity of the flu in North Dakota, which may account for the increased twitter activity in that period.[29] While the value of this information is limited, it does demonstrate the potential for more detailed geographic analysis in the future, especially as the number of Twitter users continues to climb. As this capacity increases, it could be a useful tool for local and state health departments, enabling them to monitor coverage and increase efforts to promote uptake as necessary.

We are also currently working on new tools to enhance our understanding of demographic groups on Twitter. In this study, we were able to utilize the Demographer tool to identify the gender of the person tweeting. Our results suggest there are significantly more tweets signifying intention to vaccinate coming from females. CDC data suggest that this may be accurate, with significantly more females reporting vaccination than males (FluVaxView). However, the gender gap in Twitter narrowed over the course of the four seasons in our study period, despite staying constant according to the CDC. As we continue to refine our tools, we will work on developing additional demographic classifiers to explore other areas including race and age.

One of the great advantages of utilizing Twitter is the ability to capture behaviors from a broader range of adults, especially from groups that may be difficult to reach using traditional surveys, including young adults and members of minority groups such as African Americans and Hispanics.[30, 31] These populations are also the least likely to be immunized against seasonal influenza. For example, in the past flu season, influenza vaccine uptake rates for young people (age 18-49) were much lower (34%), when compared to the 65% uptake rates for adults over 65.[32] Racial disparities also continue to be a problem with influenza uptake at 37% for both African Americans and for Hispanics, compared to 46% for White adults.[32] But all groups could benefit from increased influenza vaccination, as all groups fail to reach the Healthy People 2020 recommended goal of 70% uptake.[33]

While social media is considered “big data,” we nevertheless ran into challenges with sample size. While the full dataset is indeed large, with over one million tweets, only 33.75% of those tweets can be resolved to the United States, and each experiment further filters down the data into smaller groups. For example, if tweets are counted by month within each US state, then the data needs to be split into 600 partitions (12 months times 50 states) within each year. This has an observable effect of the validity of the results: the correlations between Twitter and CDC are very strong at the national level, but weaker at the regional level, and weaker still at the state level. Sample size may also explain why the geographic correlations between Twitter and CDC (Table 2) were strong in 2013-14 and 2014-15 than in 2015-16 and 2016-17: the first two seasons contain 53% more geolocated tweets than the latter two seasons.

Our hope is that these new tools can enrich the practice of influenza immunization surveillance and inform influenza vaccination campaigns. To date, the majority of social media surveillance research has been conducted without the involvement of local, state, or governmental agencies,[10] and most efforts to include public health practitioners in social media research have focused on concentrated health communications efforts.[34, 35] These new resources allow researchers and practitioners to respond to emerging health issues in new and innovative ways, but the progress depends on the ability to integrate novel methods into existing frameworks and to validate new data streams against reliable metrics. True success will depend on the use of novel techniques to measure positive changes in population health.[36]

COMPETING INTEREST STATEMENT

MD and MJP hold equity in Sickweather Inc. MD has received consulting fees from Bloomberg LP, and holds equity in Good Analytics Inc. These organizations did not have any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. All other authors declare no competing interests.

FUNDING STATEMENT

Preparation of this manuscript was supported in part by the National Institute of General Medical Sciences under award number R01GM114771 to DAB and SCQ.

CONTRIBUTORSHIP STATEMENT

XH, MCS, DAB, MD, SCQ, and MJP contributed to the design of the study. XH, JC, MD, and MJP contributed to data collection. XH, MCS, JC, DAB, and MJP performed data analysis. XH, AMJ, DAB, SCQ, and MJP interpreted the results. All authors contributed to the editing of this manuscript.

DATA SHARING STATEMENT

All Twitter data used in this study is available in the following repository: https://figshare.com/account/projects/31742/articles/6213878

This contains the annotations for training the classifiers, as well as the classifier predictions on the full dataset. This also contains the extracted metadata, including demographics and location. In accordance with the Twitter terms of service, raw tweets are not shared, but identifiers are shared which can be used to download the tweets.

REFERENCES

[1] Grohskopf LA, Sokolow LZ, Broder KR, et al. Prevention and Control of Seasonal Influenza With Vaccines: Recommendations of the Advisory Committee on Immunization Practices—United States, 2017–18 Influenza Season. *Am J Transplant*2017;17(11):2970–82. doi:10.1111/ajt.14511 [published Online First: 30 October 2017].

[2] CDC. Morbidity and Mortality Weekly Report (MMWR) [Internet]. 2017 [cited 2018 Mar 8]. Available from: https://www.cdc.gov/mmwr/volumes/66/rr/rr6602a1.htm

[3] Santibanez T, Zhai Y, O'Halloran A, et al. Flu Vaccination Coverage, United States, 2016-17 Influenza Season [Internet]. 2017 [cited 2018 Mar 9]. Available from: https://www.cdc.gov/flu/fluvaxview/coverage-1617estimates.htm

[4] CDC. Influenza Vaccination Coverage | FluVaxView | Seasonal Influenza | CDC [Internet]. Centers for Disease Control and Prevention. Centers for Disease Control and Prevention; 2017 [cited 2018 Mar 9]. Available from: https://www.cdc.gov/flu/fluvaxview/index.htm

[5] Iachan R, Pierannunzi C, Healey K, et al. National weighting of data from the Behavioral Risk Factor Surveillance System (BRFSS). *BMC Med Res Methodol*2016;16(1):155 doi:10.1186/s12874-016-0255-7 [published Online First: 15 November 2016].

[6] Keeter S. The Impact of Cell Phone Noncoverage Bias on Polling in the 2004 Presidential Election. *Public Opin Q*2006;70(1):88–98 doi:10.1093/poq/nfj008 [published Online First: 1 January 2006].

[7] National Vaccine Program Office. Flu Vaccination Trends [Internet]. US Department of Health and Human Services. 2017 [cited 2018 Mar 9]. Available from: https://www.hhs.gov/nvpo/resources/flu/index.html

[8] Broniatowski DA, Paul MJ, Dredze M. National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic. *PLoS One*2013;8(12) doi:10.1371/journal.pone.0083672 [published Online First: 9 December 2013].

[9] VELASCO E, AGHENEZA T, DENECKE K, et al. Social Media and Internet-Based Data in Global Systems for Public Health Surveillance: A Systematic Review. *Milbank Q*2014;92(1):7–33 doi:10.1111/1468-0009.12038 [published Online First: 6 March 2014].

[10] Charles-Smith LE, Reynolds TL, Cameron MA, et al. Using social media for actionable disease surveillance and outbreak management: a systematic literature review. *PLoS One*2015;10(10):e0139701 doi: 10.137/journal.pone.0139701 [published Online First: 5 October 2015].

[11] Corley C, Cook D, Mikler A, et al. Text and Structural Data Mining of Influenza Mentions in Web and Social Media. *Int J Environ Res Public Health*2010;7(12):596–615 doi:10.3390/ijerph7020596 [published Online First: 22 February 2010].

[12] Collier N, Son N, Nguyen N. OMG U got flu? Analysis of shared health messages for bio-surveillance. J Biomed Semantics [Internet]. *Journal of Biomedical Semantics* 2011;2(Suppl 5):S9. doi:10.1186/2041-1480-2-S5-S9 [published Online First: 6 October 2011].

[13] Odone A, Ferrari A, Spagnoli F, et al. Effectiveness of interventions that apply new media to improve vaccine uptake and vaccine coverage. *Hum Vaccin Immunother*2015;11(1):72–82 doi:10.4161/hv.34313 [published Online First: 1 November 2014].
[14] Dredze M, Broniatowski DA, Hilyard KM. Zika vaccine misconceptions: A social media analysis. *Vaccine*2016;34(30):3441–2 doi:10.1016/j.vaccine.2016.05.008 [published Online First: 20 May 2016].
[15] Powell GA, Zinszer K, Verma A, et al. Media content about vaccines in the United States and Canada, 2012–2014: An analysis using data from the Vaccine Sentimeter. *Vaccine*2016;34(50):6229–35 doi:10.1016/j.vaccine.2016.10.067 [published Online First: 3 November 2016].
[16] Kang GJ, Ewing-Nelson SR, Mackey L, et al. Semantic network analysis of vaccine sentiment in online social media. *Vaccine*2017;35(29):3621–38 doi:10.1016/j.vaccine.2017.05.052 [published Online First: 27 May 2017].
[17] Salathé M, Khandelwal S. Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control. *PLOS Comput Biol*2011;7(10):e1002199 doi:10.1371/journal.pcbi.1002199 [published Online First: 13 October 2011].
[18] Salathé M, Vu DQ, Khandelwal S, et al. The dynamics of health behavior sentiments on a large online social network. *EPJ Data Sci*2013;2(1):4 doi:10.1140/epjds16 [published Online First: 4 April 2013].
[19] Nelson EJ, Hughes J, Oakes JM, et al. Estimation of Geographic Variation in Human Papillomavirus Vaccine Uptake in Men and Women: An Online Survey Using Facebook Recruitment. *J Med Internet Res*2014;16(9):e198 doi:10.2196/jmir.3506 [published Online First: 1 September 2014].
[20] Dunn AG, Surian D, Leask J, et al. Mapping information exposure on social media to explain differences in HPV vaccine coverage in the United States. *Vaccine*2017;35(23):3033–40 doi:10.1016/j.vaccine.2017.04.060 [published Online First: 29 Apri 2017].
[21] Tufekci Z. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. In: ICWSM. 2014. p. 505–14.
[22] Cohen R, Ruths D. Classifying Political Orientation on Twitter: It's Not Easy! In: Seventh International AAAI Conference on Weblogs and Social Media. 2013.
[23] Paul MJ, Dredze M. Discovering Health Topics in Social Media Using Topic Models. *PLoS One*2014;9(8):e103408 doi:10.1371/journal.pone.0103408 [published Online First: 1 August 2013].
[24] Dredze M, Paul MJ, Bergsma S, et al. Carmen: A twitter geolocation system with applications to public health. In: AAAI workshop on expanding the boundaries of health informatics using AI (HIAI). 2013. p. 45.
[25] Knowles R, Carroll J, Dredze M. Demographer: Extremely simple name demographics. In: Proceedings of the First Workshop on NLP and Computational Social Science. 2016. p. 108–13.
[26] Callison-Burch C, Dredze M. Creating speech and language data with Amazon's Mechanical Turk. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. 2010. p. 1–12.
[27] Brockwell PJ, Davis RA. Introduction to Time Series and Forecasting. New York,

NY: Springer New York; 2002.
[28] Samuel L. Vaccine exemptions are on the rise in a number of US states [Internet]. STAT. 2018 [cited 2018 Mar 9]. Available from: https://www.statnews.com/2017/01/20/vaccine-exemptions-states/.
[29] Tribune B. Record number of flu cases reported this year [Internet]. Bismarck Tribune. 2017 [cited 2018 Mar 9]. Available from: http://bismarcktribune.com/news/local/health/record-number-of-flu-cases-reported-this-year/article_90c1c917-f140-5aab-ae65-ea505d6adc65.html.
[30] Duggan M, Brenner J. The Demographics of Social Media Users - 2012 [Internet]. Pew Research Center: Internet, Science & Tech. 2013 [cited 2018 Mar 8]. Available from: http://www.pewinternet.org/2013/02/14/the-demographics-of-social-media-users-2012/
[31] Krogstad JM. Social media preferences vary by race and ethnicity [Internet]. Pew Research Center. 2015 [cited 2018 Mar 8]. Available from: http://www.pewresearch.org/fact-tank/2015/02/03/social-media-preferences-vary-by-race-and-ethnicity/
[32] CDC. Flu Vaccination Coverage, United States, 2016-17 Influenza Season [Internet]. 2017 [cited 2018 Mar 8]. Available from: https://www.cdc.gov/flu/fluvaxview/coverage-1617estimates.htm#age-group-adults
[33] HealthyPeople. Immunization and Infectious Diseases [Internet]. HealthyPeople.gov. [cited 2018 Mar 9]. Available from: https://www.healthypeople.gov/2020/topics-objectives/topic/immunization-and-infectious-diseases
[34] Moorhead SA, Hazlett DE, Harrison L, et al. A New Dimension of Health Care: Systematic Review of the Uses, Benefits, and Limitations of Social Media for Health Communication. Eysenbach G, editor. *J Med Internet Res*2013;15(4):e85 doi:10.2196/jmir.1933 [published Online First: 13 September 2011].
[35] Thackeray R, Neiger BL, Smith AK, et al. Adoption and use of social media among public health departments. *BMC Public Health*2012;12(1):242 doi:10.1186/1471-2458-12-242 [published Online First: 26 March 2012].
[36] Mooney SJ, Westreich DJ, El-Sayed AM. Epidemiology in the Era of Big Data. *Epidemiology*2016;26(3):390–4 doi:10.1097/EDE.0000000000000274 [published Online First: 1 May 2016].

Figure 1. Flu vaccination by time.

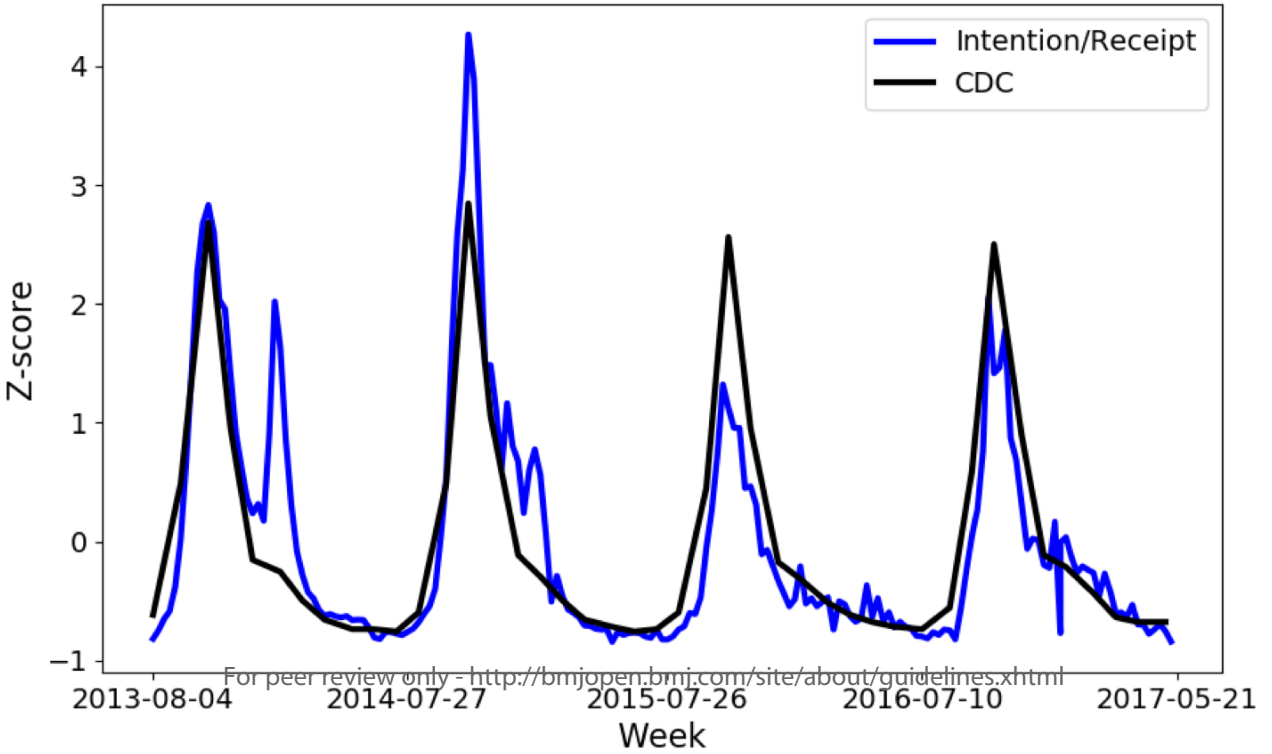


Figure 2. Flu vaccination by US state.

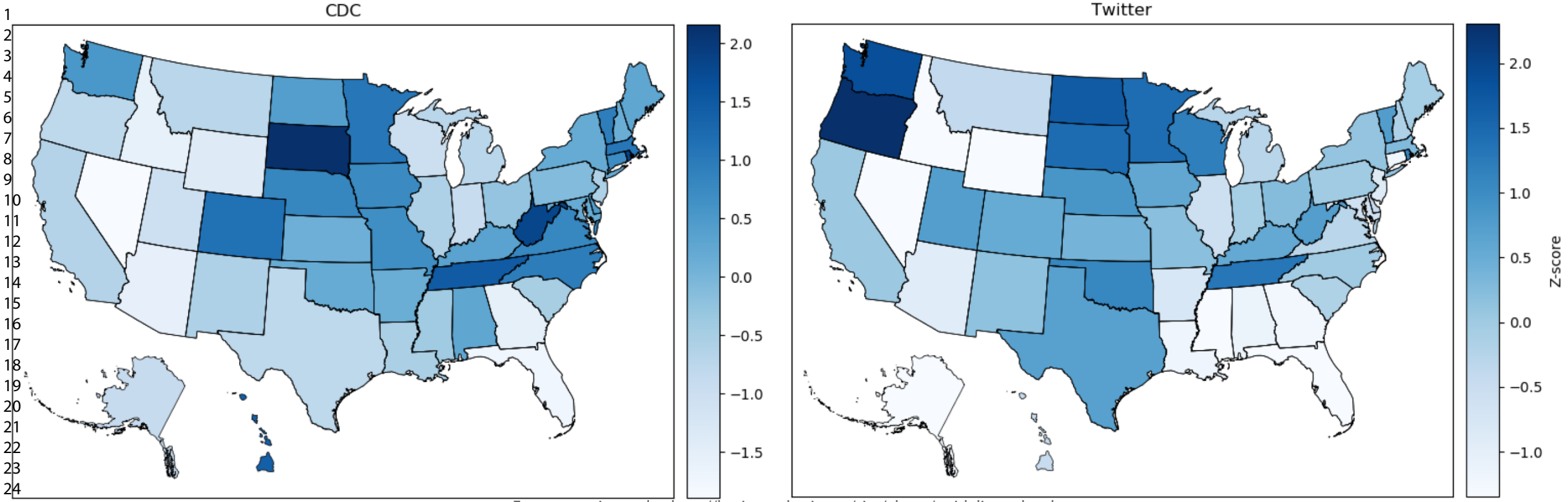
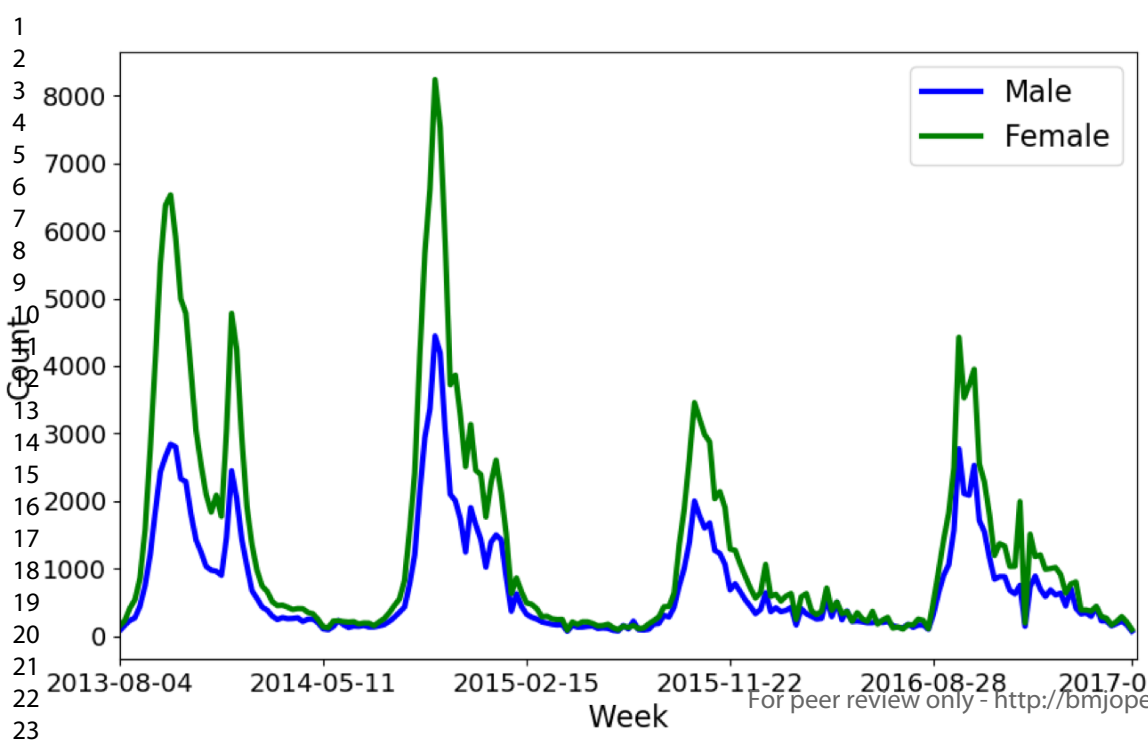
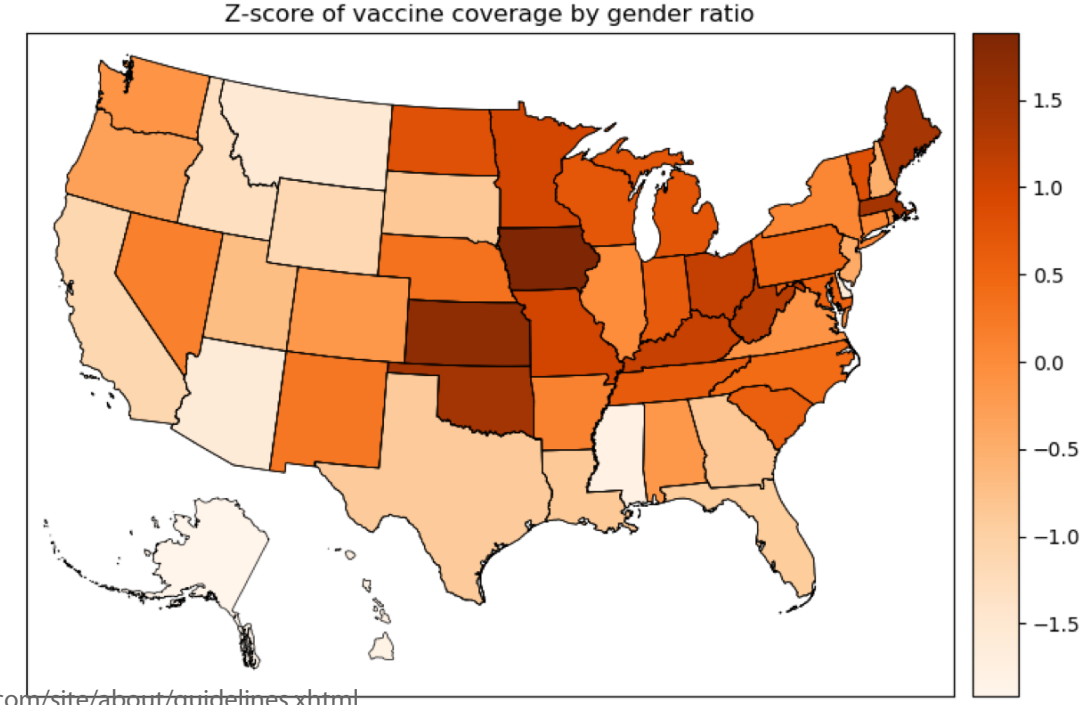


Figure 3. Flu vaccination by gender.





A.1 Data

A.1.1 Data Collection

We collected Twitter data beginning in 2012. However, the tweets collected during 2012-13 flu season were removed in this study, because the data did not cover the whole flu season. We discarded retweets and non-English tweets.[1] For the CDC's data, we collected the data from the 2013 to 2017 flu seasons, where each flu season starts in July and ends in May in the following year. To match CDC data, we removed tweets posted in June. The statistical description of our final data is listed in Table 1.

Table 1. Overview of Twitter data in this study

Flu Season	Tweet count	Unique user count
2013 July - 2014 May	264,171	199,733
2014 July - 2015 May	336,644	219,012
2015 July - 2016 May	232,591	147,564
2016 July - 2017 May	263,535	175,770
Total	1,124,839	742,079

A.1.2 Data Preprocessing

Tweets have some unique characteristics that do not exist in traditional text, such as hashtags, hyperlinks, and colloquial language. To make the text more appropriate for natural language processing tools, we preprocessed each tweet according to the following steps:

1. Hyperlinks, hashtags, user mentions in each tweet were replaced with "<url>", "<hashtag>", and "<user>," respectively.
2. Repeated punctuation was replaced with "[punctuation] <repeat>".
3. Each tweet was lowercased and tokenized using NLTK.[2]

A.1.3 Data Annotation

To build training data, we collected annotations for a random sample of 10,000 tweets from the full collection. Annotations were obtained from Amazon Mechanical Turk,[3] with three independent annotations per tweet. Tweets were labeled with the following:

- Does this message indicate that someone received, or intended to receive, a flu vaccine? (yes or no)
    - If yes: has the person already received a vaccine, or do they intend to receive the vaccine in the future.

We refer to tweets labeled "yes" as "intention/receipt" and tweets labeled "no" as "other".

We rejected annotators whose agreement was anomalously low (percentage agreement was ≤ 60%). Three bad annotators were removed from our final dataset. We took a majority vote on the remaining 29,970 annotations to obtain the final labels. If there was not a majority label, then we defaulted to the "other" label. The dataset contained 10,000 tweets, with 32.8% labeled as positive for "intention/receipt", with a kappa score of 0.793, using Fleiss' kappa to measure the inter-annotator agreement.[4] Then we manually corrected 168 labels of the dataset and finally obtained 31.1% labeled as positive for "intention/receipt".

A.2 Automatic Assessment Methods

To automatically identify tweets expressing vaccination intention/receipt, we used the labeled data to train two machine learning classifiers: Logistic Regression (LR) and Convolutional Neural Network (CNN). The LR model achieved the best performance among other classifiers in our previous study.[5] We implemented Logistic Regression (LR) classifier using the scikit-learn toolkit.[6] CNN has been drawn significant attention in recent years because of its impressive performance on text classification tasks.[7] We trained the two models on the annotated Twitter data. After optimizing the model parameters and hyperparameters, we compared the two models. We finally chose the model that achieved the best performance in the validation experiments.

A.2.1 Logistic Regression

We fed the LR model with TF-IDF weighted n-gram (uni-, bi- and tri-gram) features, as well as part-of-speech (POS) counts from TweeboParser,[8] and emoji and emoticon features derived from two open lexicons.[9, 10] Feature counts were normalized to sum to 1 within each tweet. The list of features we used in this study are shown in Table 2.

Table 2 Details of the feature set for Logistic Regression classifier

Feature name	Feature attributes
N-gram	TF-IDF scores of unigrams, bigrams, trigrams
Part-of-Speech	Counts of POS tags, normalized by the total tags in the tweet
Emoji	Counts of negative and positive emojis, normalized by total counts.
Emoticon	Counts of negative and positive emoticons, normalized by total counts.

We balanced the weight of each label by adjusting weights inversely proportional to class frequencies in the training dataset. We adopted cross entropy as the loss function with $l2$norm penalty for weight regularization.

A.2.2 Convolutional Neural Network

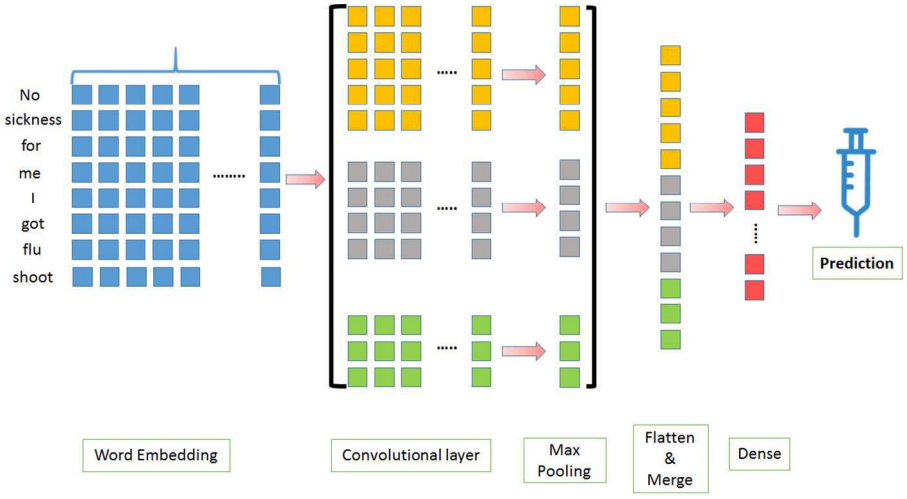


Figure 1. The architecture of the CNN model.

The embedding layer converts processed tweets into an embedding matrix of floating point values, where each row is a vector representation of a word. The embedding matrix is then fed into the convolutional layer, where the matrix will be screened and sampled by the filters. We set 150 filters in this layer. Each filter is a square sliding window and we defined three different sizes of filters: 3*3, 4*4, 5*5. We set the filter stride to 1 and padding mode to "VALID". We obtained the squares by sliding the filters over the matrix. Those captured squares will be fed into the next layer, the pooling layer. We adopt 1-max pooling as the strategy to extract a max scalar value from each square, which outputs the maximum value. We stack another convolutional layer and pooling layer following the first pooling layer, for which the operation steps are the same.

Outputs from the stacked convolutional and pooling layers are flattened, concatenated and fed to the next layer, the dense layer, where it learns and generates a fixed representation for each tweet. We set the activation function as rectified linear unit (ReLU).[11] We set the output dimension of this dense layer to 150. A dropout was applied in the layer, where dropout is a standard method to prevent overfitting by randomly set a proportion of values to zero during training.[12]

We fed the outputs from the dense layer to the sigmoid function to predict the final binary label, "intention/receipt" or "other". We adopted the binary cross entropy function with $l2$penalty to calculate the loss of predictions. Adam with a learning rate of 0.001 and decay of 0.003 was adopted to optimize the parameters.[13]

A.2.3 Experiment Settings

We randomly sliced the dataset into three pieces: 80% as training set, 10% as development set and 10% as testing set. We trained our two methods, LR and CNN, on the training set, tuned parameters on the development set, and evaluated the methods on the testing set. We

balanced weights of predicted labels in the two models. The models' parameters were selected by accuracy on the development set. The CNN model was trained by 10 epochs, batch size was set by 64, and the dropout rate was set to 0.2. We fixed the length of inputs by either padding sentence to 40 words or slicing the first 40 words. Outputs of the classifiers are probabilities of "intention/receipt", which consider true only if the values are equal to or larger than 0.5 and vice versa. "Precision", "recall", "f1-score" were used to evaluate the performance of each method on the testing set. We focused on the performance of "intention/receipt", not "other" label, which consistently keeps the same evaluation metrics with our previous work.[5]

A.2.4 Selecting Word Embeddings

Word embedding is a language modeling technique that maps words into a set of word vectors.[14] The CNN model in our study was fed with the word vectors. There are two popular frameworks to generate the vectors, Word2vec and GloVe.[14, 15] We selected the best embedding model from the following options:

1. We obtained pre-trained word embedding by running word2vec from Gensim over our collected tweet dataset.[16] We set the tool's default settings except for changing minimum count of words to 1 and number of iterations to 15. We finally obtained 100 dimensional embedding for each word (denoted as *word2vec*).
2. We obtained an embedding model by GloVe with its default parameter settings from its official website (denoted as *glovec*).
3. Google provides pretrained word2vec embeddings on its news dataset,[14] and Stanford provides pretrained GloVe embeddings on its Twitter dataset (denoted as *pre-word2vec* and *pre-glovec* respectively).[15]
4. Character-level embeddings have recently been shown to perform well on text classification.[17] We built word embeddings using a one-hot encoding of characters (denoted as *character*).

We fed the different embedding models to the same CNN model with the fixed parameters. We evaluated the performance by precision, recall and F1-score. The performance is shown in Table 3.

Table 3 Performance of different word embeddings on our dataset.

Word Embeddings	Precision	Recall	F1-score
word2vec	0.84	0.80	**0.84**
glovec	0.82	0.75	0.78
pre-glovec	0.79	0.80	0.79
pre-word2vec	0.90	0.77	0.82
character	0.86	0.73	0.79

Finally, we chose the *word2vec* model trained on the collected data in this study, because it achieves the best performance. We also trained embeddings with 50 and 200 dimensions for

both Word2vec and GloVe, but their performance was worse than with 100 dimensions. The word embedding trained on our collected data outperformed pre-trained models from Google and Stanford. Thus, we chose this embedding model for our experiments.

A.2.5 Test Performance of Classifiers

Table 4 Classification performance on test data.

Method	Precision	Recall	F1-score
LR-ngram*	0.84	0.80	0.82
CNN-embedding	0.89	0.80	**0.84**
LR-embedding-average	0.83	0.65	0.73

We used the precision, recall, and F1-score to measure the performance of the two classifiers. We selected the classifier for our analysis tasks based on the best F1-score. We show the test performance in Table 4, where embedding refers to the word vectors from the selected word2vec model, and embedding-average means the trained features of LR are word vectors created by averaging the word vectors of all words in each tweet. Compared to the other two models, the CNN-embedding has better precision and F1-score. We finally selected CNN-embedding for categorizing all the tweets we collected.

A.3 Validation Experiments

In this section, we provide additional details and experiments on the validation process of comparing the Twitter data to the CDC data.

A.3.1 Experimental Steps

We ran both classifiers (LR and CNN) on all tweets from the 2013 to 2017 seasons to obtain labeled tweets. We restricted the analysis to tweets from the United States. We validated our approach across three dimensions: time, geography, and demography.

- Time:
    a. We counted both the weekly and monthly number of tweets classified as “intention/receipt”. To be consistent with CDC’s week definitions, we used the epidemiological week instead of the ISO week to calculate the counts. The data from Twitter and CDC were normalized by z-score separately.
    b. Because the types of data were time-series, we ran the time series model, “autoregressive integrated moving average” (ARIMA), to obtain relationship Twitter and CDC, which was (p, d, q) = (0, 1, 0). The result suggested a linear relationship between the trends of CDC and Twitter. We then fitted the time

series data by a linear regression model using Twitter trends to predict CDC trends.

c. We additionally calculated Pearson correlation and Spearman correlation scores on the Twitter counts and CDC data.

- Geography:
  a. For geographic regions (referred to as “Region”), we aggregated the total counts of “intention/receipt” tweets for the 10 HHS regions separately. In the “Region-year” experiment, we treated the regional tweets in each flu season as a separate point. We normalized the counts of “Region” and “Region-year” by dividing the number of tweets from that region, using the random sample of tweets from the Twitter streaming API.
  b. For “State” and “State-year”, we excluded five locations, Northern Mariana Islands, US Virgin Islands, Puerto Rico, Guam, and District of Columbia. These experiments follow the same process as the region experiments, but within individual US states.
  c. All the values were normalized by z-scores.
  d. We validated the geographic data by measuring Pearson and Spearman correlations.
- Demography:
  a. For “Gender”, we first counted positive tweets separately for males and females for each flu season. We divided the female counts by male counts of each flu season to generate gender ratios for the Twitter data. Finally, the ratios were normalized by z-score.

A.3.2 Correlation Results

Table 5.1 shows the Pearson correlations over time for both the CNN and LR models. Table 5.2 shows the correlations over geography for the LR model.

Table 5.1 Validation by Pearson correlation for time. * = p<0.05, ** = p<0.01, *** = p<0.001.

Validation model	All	2013-14 season	2014-15 season	2015-16 season	2016-17 season
CNN	89.85%***	89.71%***	98.45%***	98.48%***	96.65%***
LR	89.68%***	92.67%***	99.22%***	98.53%***	98.41%***

Table 5.2 Validation of LR by Pearson correlation for geography. * = p<0.05, ** = p<0.01, *** = p<0.001.

Validation model	State	State year	Region	Region year
LR	43.28%**	21.20%**	45.61%	-12.07%

Table 6.1 shows the Spearman correlation by time, and Table 6.2 shows the Spearman correlation by geography.

As the data is split into finer granularities, such as State or State-year, the correlation scores tend to decrease. This might be caused by a smaller sample size of tweets in the smaller bins. This suggests that if we could obtain more data, this approach will be more accurate.

Table 6.1 Validation by Spearman correlation for time, * = p<0.05, ** = p<0.01, *** = p<0.001.

Validation model	All	2013-14 season	2014-15 season	2015-16 season	2016-17 season
CNN	92.88%***	94.76%***	97.04%***	90.00%***	94.31%***
LR	93.43%***	95.67%***	97.49%***	93.63%***	94.31%***

Table 6.2 Validation by Spearman correlation score for geography. * = p<0.05, ** = p<0.01, *** = p<0.001.

Validation model	State	State year	Region	Region year
CNN	40.15%**	23.57%***	55.15%	-8.80%
LR	44.62%**	20.76%**	45.45%	-13.27%

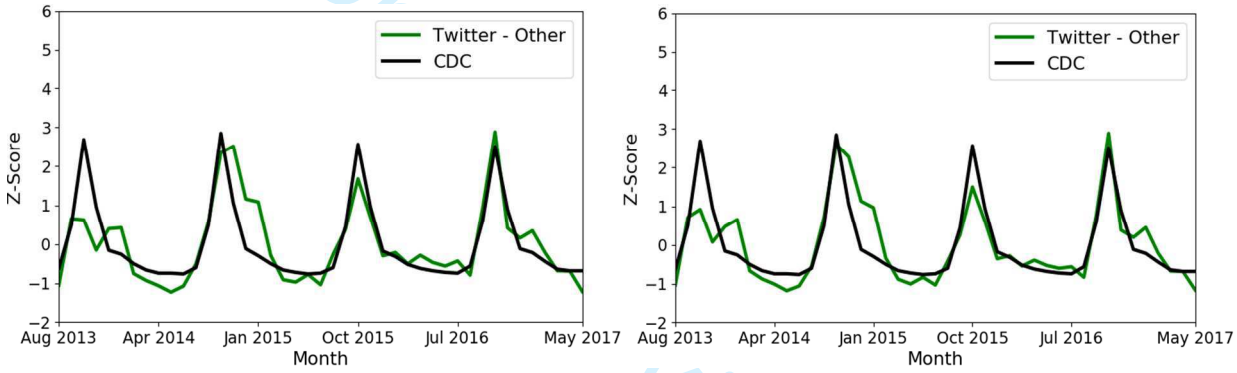
A.3.4 Validation of "Other" Tweets

We have focused on the "intention/receipt" tweets under the assumption that they will be more meaningful than the tweets classified as "other", i.e., tweets that contain vaccine-related phrases but do not explicitly state that someone received or intends to receive a vaccine. In this section, we measured the predictive value of the "other" tweets, which might also correlate with CDC data, and we compare the correlations to the correlations of the "intention/receipt" tweets.

We kept the same experiment settings for the tweets of the "other" label as the "intention/receipt" tweets. We calculated the Pearson correlation with the CDC data. The results are shown in Table 7. We plot the monthly flu vaccine prevalence between "other" (denote as Twitter-Other) and the CDC and weekly prevalence of Twitter data in Figure 2. The "other" tweets have lower Pearson correlation than "intention/receipt" tweets overall with the CDC data. In Figure 2.2, the other tweets in the dataset have very high week-to-week variability, with numerous spikes that do not fit the seasonal trends. This suggests that our classifier is reducing the noise and improving our identification of vaccine behaviors

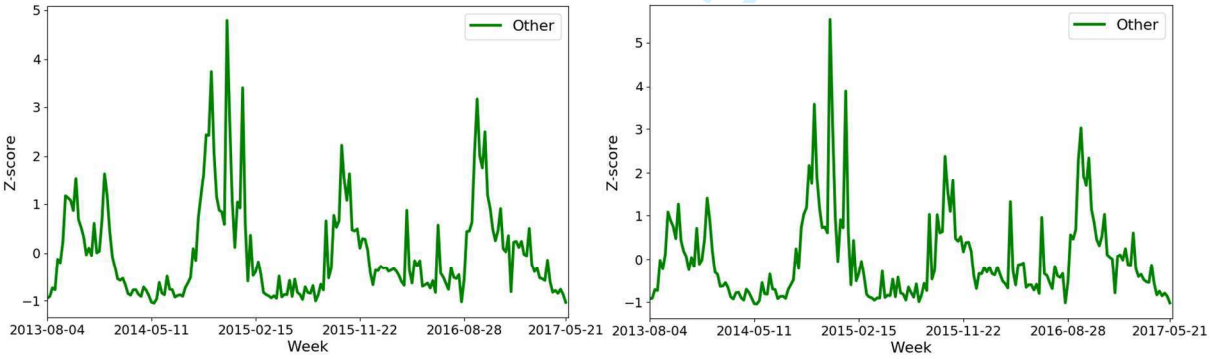
Table 7 Validation Results of CNN and LR by "other" label. * = p<0.05, ** = p<0.01, *** = p<0.001.

Validation Task	CNN	LR
All seasons	81.95%***	84.42%***
State	17.33%	20.01%
State-year	11.11%	13.40%
Region	58.74%	58.87%
Region-year	45.05%**	50.03%**



(a) LR (b) CNN

Figure 2.1 Monthly prevalence of "Other" trends from Twitter compared to the CDC.

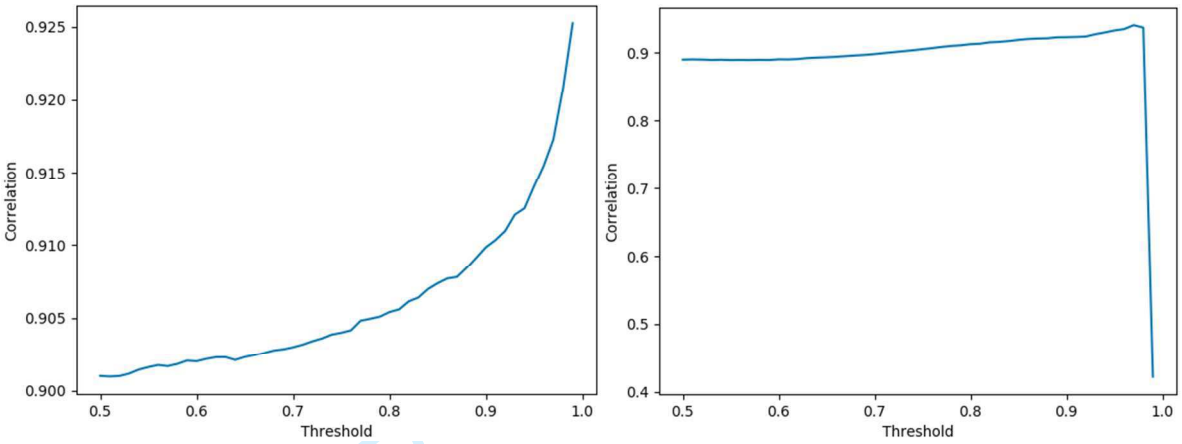


(a) LR (b) CNN

Figure 2.2. Weekly time series of tweets classified as "Other" by LR (a) and CNN (b).

A.4 Additional Analyses

A.4.1 Sensitivity of the Classification Threshold

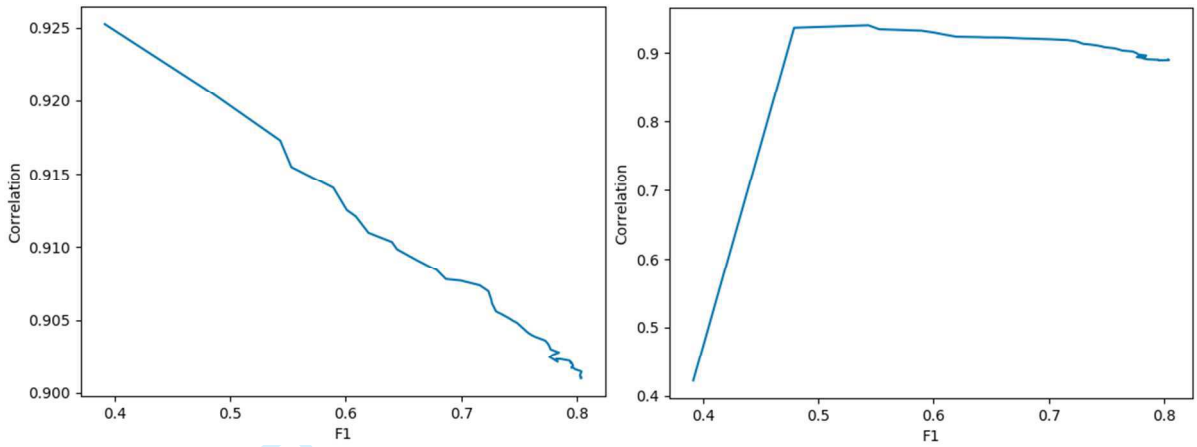


(a) LR (b) CNN

Figure 3. The relationship between the prediction threshold and correlation coefficient.

In this section, we explore how the threshold of classifiers impacts the Pearson correlation. Specifically, the threshold of how the probability of a tweet being positive before it is actually positive. By default, anything with probability greater than or equal to 0.5 is classified as positive, but this threshold can be raised to increase precision (at the expense of recall).

In Figure 3(a) and 3(b), we plotted the relationship between Pearson correlation and prediction threshold for both LR and CNN. Both approaches show that increasing the predicting threshold can improve the correlation coefficient. Increasing the threshold indicates higher confidence of the classifier, that is to say, a tweet will only be considered as “intention/receipt” when the classifier has high confidence. In the view of the classifier, only the tweets have enough evidence to indicate vaccination will be classified as “intention/receipt”. Additionally, we could find that when the threshold of CNN is set to near 0.95, the correlation score decreases rapidly, so raising the threshold does not always improve performance monotonically.



(a) LR (b) CNN

Figure 4 The relationship between the F1 score and correlation coefficient.

.

In Figure 4(a) and 4(b), we explore the relationship between the F1-score and Pearson correlation, because our criteria for selecting the best classifier was by F1-score. The CNN classifier reaches the highest correlation coefficient at around an F1-score of 0.5. Under both models, the correlation drops when the F1 score is too high, likely because the optimal balance is high precision and low recall, even if that drops the F1 score.

For the LR model, while the correlation varies with F1 score, the correlation values are all very similar, and all are above .90. However, the CNN model is not very stable with respect to the correlation coefficient, which might indicate the LR is more robust. We also combined the two approaches to see if we could achieve better performance in the next section.

A.4.2 An Ensemble Perspective of the Two Models

We combined the two models using two linear combination approaches: combining monthly counts of tweets from the LR and CNN (weighted-counts), and combining the prediction probabilities of each approach (weighted-prob). We calculated the combination by the formulas below:

$$Weighted - output = \sum_{i=1}^{2} W_i * X_i \quad (1)$$

$$W_i = \frac{F1_i}{\sum_{i=1}^{2} F1_i} \quad (2)$$

, where F1 is the F1-score of each classifier achieved on the test data, and X_i is the count number of each classifier for “weighted-counts” or the predicted probability of “intention/receipt” of each tweet by i-th classifier. Specifically, the weighted-count is weighted sum of weighted counts from the LR and CNN approaches; the weighted_prob, instead of counts, we calculated the predicting probability of each tweet by weighted sum of the probabilities from each classifier. The F1-score of each method was used as the weight in the Equation (1). The weight were normalized by the sum of weights to ensure they are within 0 and 1, as shown in Equation (2).

For the validation, we evaluated the performance of the tweets classified as “intention/receipt” and “other”. We validated the two ensemble approaches by calculating Pearson correlation with

the CDC data. The results are shown in Table 8. We find that the weighted-counts performs slightly better than the weighted-prob on the tweets classified as “intention/receipt”. The ensemble ways show promising results, outperforming a single classifier.

Table 8. Validation Results of CNN and LR. * = p<0.05, ** = p<0.01, *** = p<0.001.

Validation Task	Intention/receipt		Other	
	Weighted-Counts	Weighted-Prob	Weighted-Counts	Weighted-Prob
All seasons	89.93%***	89.48%***	83.48%***	83.98%***
State	40.61%*	43.74%**	18.76%	19.15%
State-year	29.55%*	28.14%*	9.17%	11.50%
Region	47.50%	43.16%	58.84%	59.13%
Region-year	32.46%*	26.37%	48.03%**	49.68%**

A.4.3 Simpson’s Paradox

In our previous work,[5] LR achieved a .90 correlation on the three consecutive flu seasons (2013-14, 2014-15, 2015-16). In this work, we added a fourth flu season, and LR received a lower correlation score after adding the 2016-17 season. To explore why the correlation dropped, we calculated the correlation on the 2016-17 by itself, to see if this season had a lower correlation that caused the overall correlation to drop. The results are shown in Table 9, comparing the first three seasons (2013-16), the fourth season (2016-17), and all four seasons.

Surprisingly, we discovered that the CNN achieves lower correlation scores than LR on both Seasons 2013-16 and Season 2016-17, even though it exceeds LR on all seasons. This behavior could be explained by “Simpson’s paradox”, a common paradoxical phenomenon in data analysis.[18]

Table 9 Pearson correlation of two different time periods.

Validation Task	Intention/receipt	
	CNN	LR
Seasons 2013-16	89.20%	**90.27%**
Season 2016-17	96.65%	**98.41%**
All seasons	**89.85%**	89.68%

A.4.4 Additional Trend Figures

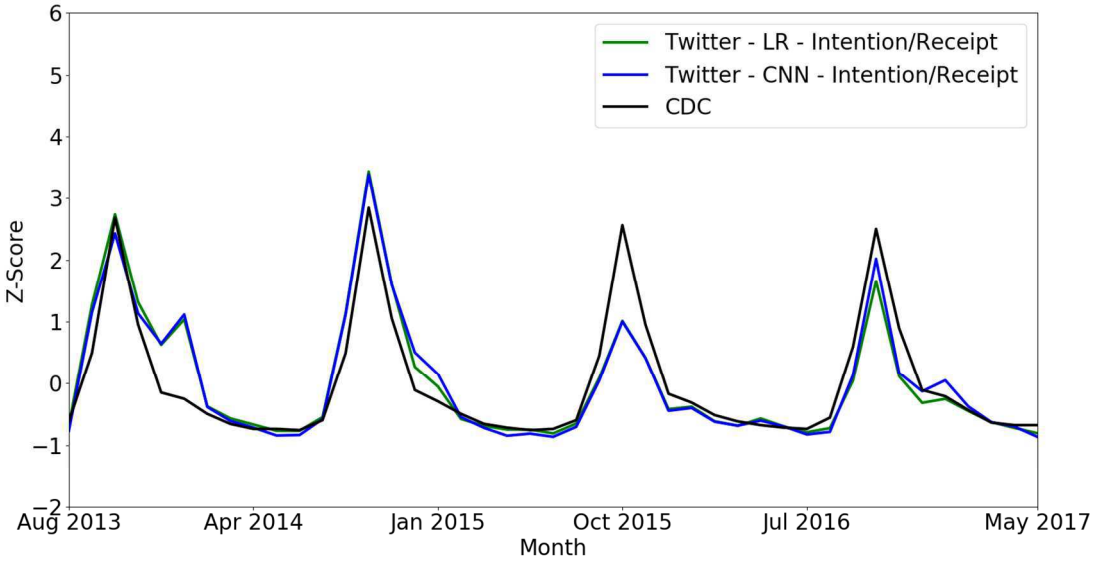
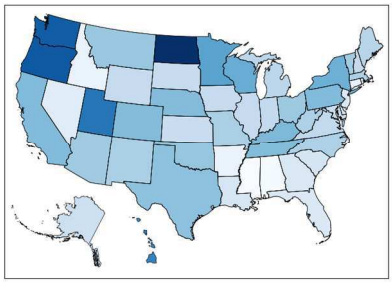
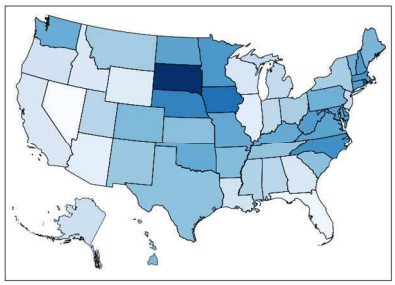
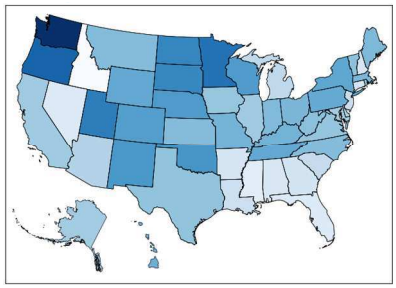
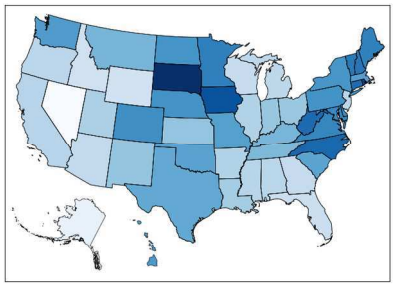
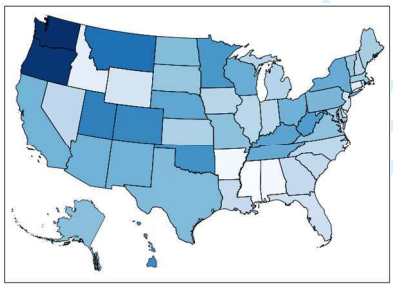
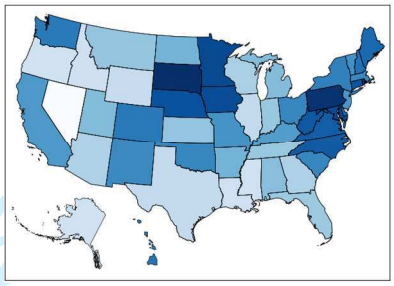
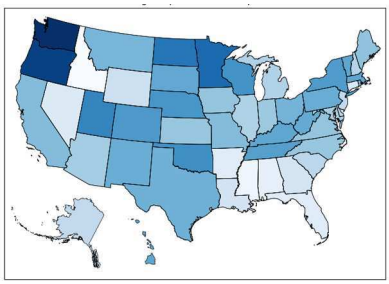
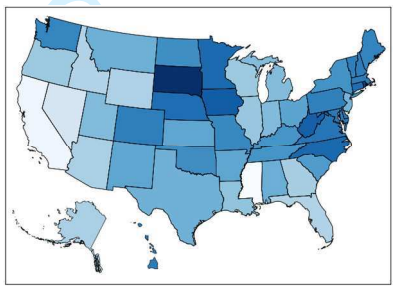


Figure 5. Monthly prevalence of vaccination trends from Twitter and CDC.

Figure 5 shows both the CNN time series (blue) alongside the LR time series (green) and CDC data. There are only minor differences in the trends of the two models. Notice that each peak of the plots is usually in October of the flu season. Yet, there is a distinct peak between Jan. 2014 and Feb. 2014, which might indicate many people also talked about taking flu vaccination shots during that time.

We visualized vaccine coverage in the 50 states each flu season in the Figure 6.[19] We find there are some similar patterns between the Twitter and CDC that the states in the northeast of US show high vaccine coverage and southeast of the US show the lower vaccine coverage, while there are also some clear differences, for example, in the Twitter data, Washington and Oregon show consistently very dark colors.

Figure 6. Flu vaccine trends of both the Twitter and CDC in the U.S.

Flu Season	Twitter	CDC
2014-15		
2015-16		
2016-17		
All		

References

[1] Lui M. saffsd/langid.py [Internet]. GitHub. 2017 [cited 2018 Mar 9]. Available from: https://github.com/saffsd/langid.py.

[2] Bird S, Loper E. NLTK: the natural language toolkit. In: Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. Association for Computational Linguistics; 2004. p. 31.

[3] Callison-Burch C, Dredze M. Creating speech and language data with Amazon's Mechanical Turk. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. 2010. p. 1–12.

[4] Fleiss JL, Cohen J. The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. *Educ Psychol Meas*1973;33(3):613–9 doi:10.1177/001316447303300309 [published First: 1 October 1973].

[5] Huang X, Smith MC, Paul MJ, et al. Examining patterns of influenza vaccination in social media. In: Proceedings of the AAAI Joint Workshop on Health Intelligence (W3PHIAI), San Francisco, CA, USA. 2017. p. 4–5.

[6] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res* 2011;12(Oct):2825–30.

[7] Kim Y. Convolutional Neural Networks for Sentence Classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014. p. 1746–51.

[8] Gimpel K, Schneider N, O'Connor B, Das D, Mills D, Eisenstein J, et al. Part-of-speech Tagging for Twitter: Annotation, Features, and Experiments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA, USA: Association for Computational Linguistics; 2011. p. 42–7. (HLT '11).

[9] Kralj Novak P, Smailović J, Sluban B, Mozetič I. Sentiment of Emojis. *PLoS One*2015 Dec 7;10(12):e0144296 doi:10.1371/journal.pone.0144296 [published Online First: 7 December 2015].

[10] Mohammad SM, Turney PD. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text. Stroudsburg, PA, USA: Association for Computational Linguistics; 2010. p. 26–34.

[11] Nair V, Hinton GE. Rectified Linear Units Improve Restricted Boltzmann Machines. In: Proceedings of the 27th International Conference on International Conference on Machine Learning. USA: Omnipress; 2010. p. 807–14. (ICML'10).

[12] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J Mach Learn Res* 2014;15:1929–58.

[13] Kingma DP, Ba J. Adam: A method for stochastic optimization [Internet]. arXiv preprint arXiv:1412.6980. 2014. Available from: https://arxiv.org/pdf/1412.6980.pdf

[14] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada: Curran Associates Inc.; 2013. p. 3111–9.

[15] Pennington J, Socher R, Manning CD. GloVe: Global Vectors for Word Representation. In: Empirical Methods in Natural Language Processing (EMNLP) [Internet]. 2014. p. 1532–43. Available from: http://www.aclweb.org/anthology/D14-1162

[16] Rehurek R, Sojka P. Software framework for topic modelling with large corpora. In: In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. 2010. p. 45–50.

[17] Kim Y, Jernite Y, Sontag D, et al. Character-Aware Neural Language Models. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Phoenix, Arizona; 2016. p. 2741–9.
[18] Pearl J. Comment: Understanding Simpson's Paradox. *Am Stat* 2014;68(1):8–13 doi:10.1080/00031305.2014.876829 [published Online First: 21 February 2014].
[19] Root B. matplotlib/basemap [Internet]. GitHub. 2018 [cited 2018 Mar 9]. Available from: https://github.com/matplotlib/basemap

BMJ Open

Can Online Self-Reports Assist in Real-Time Identification of Influenza Vaccination Uptake? A Cross-Sectional Study of Influenza Vaccine-Related Tweets in the US, 2013-2017

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2018-024018.R1
Article Type:	Research
Date Submitted by the Author:	07-Nov-2018
Complete List of Authors:	Huang, Xiaolei; University of Colorado Boulder, Department of Information Science Smith, Michael; George Washington University, Department of Engineering Management & Systems Engineering Jamison, Amelia; University of Maryland, Center for Health Equity Broniatowski, David; George Washington University, Department of Engineering Management & Systems Engineering Dredze, Mark; Johns Hopkins University, Department of Computer Science Quinn, Sandra; University of Maryland , Department of Family Science; University of Maryland , Center for Health Equity Cai, Justin; University of Colorado, Department of Computer Science Paul, Michael; University of Colorado Boulder, Department of Information Science; University of Colorado, Department of Computer Science
Primary Subject Heading:	Health informatics
Secondary Subject Heading:	Epidemiology, Public health
Keywords:	Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, World Wide Web technology < BIOTECHNOLOGY & BIOINFORMATICS, PUBLIC HEALTH, EPIDEMIOLOGY

Title:

Can Online Self-Reports Assist in Real-Time Identification of Influenza Vaccination Uptake? A Cross-Sectional Study of Influenza Vaccine-Related Tweets in the US, 2013-2017

Authors:

Xiaolei Huang[1], Michael C. Smith[2], Amelia M. Jamison[3], David A. Broniatowski[2], Mark Dredze[4], Sandra C. Quinn[3,5], Justin Cai[6], Michael J. Paul[1,6]

[1] Department of Information Science, University of Colorado, Boulder, CO 80309, USA
[2] Department of Engineering Management & Systems Engineering, George Washington University, Washington, DC 20052, USA
[3] Center for Health Equity, School of Public Health, University of Maryland, College Park, MD 20742, USA
[4] Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA
[5] Department of Family Science, School of Public Health, University of Maryland, College Park, MD 20742, USA
[6] Department of Computer Science, University of Colorado, Boulder, CO 80309, USA

Corresponding Author:

Michael J. Paul
Assistant Professor, Department of Information Science
315 UCB, Boulder, CO 80309, USA
1-217-552-3605
mpaul@colorado.edu

Word Count: 3,303

ABSTRACT

Introduction: The Centers for Disease Control and Prevention (CDC) spend significant time and resources to track influenza (flu) vaccination coverage each flu season using national surveys. Emerging data from social media provide an alternative solution to surveillance at both national and local levels of flu vaccination coverage in near real-time.

Objectives: This study aimed to characterize and analyze the vaccinated population from temporal, demographic, and geographical perspectives using automatic classification of vaccination-related Twitter data.

Methods: In this cross-sectional study, we continuously collected tweets containing both flu-related terms and vaccine-related terms covering four consecutive flu seasons from 2013 to 2017.We created a machine learning classifier to identify relevant tweets, then evaluated the approach by comparing to data from the CDC's FluVaxView. We limited our analysis to tweets geolocated within the US.

Results: We assessed 1,124,839 tweets. We found strong correlations of .799 between monthly Twitter estimates and CDC, with correlations as high as .950 in individual flu seasons. We also found that our approach obtained geographic correlations of .387 at the US state level and .467 at the regional level. Finally, we found a higher level of flu vaccine tweets among female users than male users, also consistent with the results of CDC surveys on vaccine uptake.

Conclusion: Significant correlations between Twitter data and CDC data show the potential of using social media for vaccination surveillance. Temporal variability is captured better than geographic and demographic variability. We discuss potential paths forward for leveraging this approach.

Keywords: vaccination, surveillance, influenza, biostatistics, time-series

ARTICLE SUMMARY

Strengths and limitations of this study

- This study shows how to measure influenza vaccination uptake through Twitter, which has advantages and disadvantages compared to traditional survey methods.
- The signal from Twitter is available in real-time and has potential to be localized to specific geographic locations.
- While Twitter can be considered “big data”, the sample size is more limited when narrowed to specific populations.
- Certain vulnerable populations, including children and older adults, are underrepresented in Twitter data.

INTRODUCTION

The Advisory Council for Immunization Practices (ACIP) at the Centers for Disease Control and Prevention (CDC) recommends annual influenza vaccination for all healthy adults.[1] Furthermore, CDC urges individuals to get vaccinated early in the flu season, from October through January.[2] Yet, it can be difficult for researchers and practitioners working to improve influenza vaccine uptake to get accurate information in real time. Existing influenza immunization surveillance techniques have known limitations: traditional survey-based methods are time-consuming and expensive, and newer reimbursement-based systems fail to accurately capture a representative sample of population.[3]

Two national surveillance systems enable public health professionals to access information on influenza vaccine uptake in the United States (US). The most accessible of these systems is the CDC's FluVaxView, which aggregates uptake data from several national surveys.[4] The CDC data provide accurate estimates of vaccine uptake, although with some time lag. The earliest reports are only available after flu seasons typically peak, and final estimates are generally published at the start of the following flu season in September or October. Additionally, the panel surveys that inform the reports are expensive, take months to administer and process, and may undersample populations without a landline phone, particularly minority populations, young adults, and adults living in urban areas.[5, 6] A second system,[7] provided by the National Vaccine Program Office, uses an online tool to "live-track" influenza vaccination insurance claims from Medicare beneficiaries. While this system reduces lag time between vaccination and reporting, it only captures the population enrolled in Medicare, adults over age 65 and those under 65 living with disabilities.[7]

Social media data have been utilized in new tools for infectious disease surveillance, particularly for seasonal and pandemic influenza.[8-10] Utilizing data from social media platforms (like Twitter or Facebook), search engines (like Google), and other internet-based resources (like blogs), researchers have been able to track the spread of disease in real time with relatively high accuracy.[9] A recent meta-analysis of social media influenza surveillance efforts found that in a comparison to national health statistics (primarily from the CDC), correlation between social media data and national statistics ranged from 0.55 to 0.95,[11, 12] and the majority of projects were able to predict outbreaks more quickly than traditional surveillance methods.[10] Of these studies, the most accurate systems have harnessed natural language processing methods to identify relevant tweets. However, few of these tools have been fully integrated into public health practice.

With the development of new tools and techniques, social media data have the potential to similarly inform the practice of influenza immunization surveillance. However, to our knowledge, no studies have attempted to utilize social media data to track influenza vaccine intentions and uptake at the national level. To date, efforts to track influenza vaccination through social media have been much less frequent than efforts to track disease. Researchers are more likely to focus on the use of social media as a health communication tool than to explore the potential for immunization surveillance.[13] Some studies have been able to use social media data to track vaccine sentiment and general attitudes towards vaccines.[14–16] Others have focused on the spread of vaccine sentiment across online social networks.[17, 18] Some vaccine-specific studies have also attempted to use social media to identify geographic differences in vaccine uptake.[19, 20] The possibility of efficiently tracking influenza immunization in real-time is promising, but the true value of any new data source is limited without validation against known metrics.[14, 21, 22] To successfully use social media data in immunization surveillance efforts, an important first step is to validate observed trends against national survey data. In this study, we sought to validate observed patterns from Twitter, using tweets expressing either intention to seek immunization or receipt of influenza immunization, against influenza immunization data from the CDC for four consecutive flu seasons from 2013-2017.

METHODS

Patient and Public Involvement

This study did not involve patients.

Ethics approval

This work was conducted under Johns Hopkins University Homewood IRB No. 2011123: "Mining Information from Social Media", which qualified for an exemption under category 4.

Data

Twitter Data

We continuously collected tweets containing the terms “flu” or “influenza” since 2012 using the Twitter streaming Application Programming Interface (API), as part of data described in our team’s prior work on Twitter-based health surveillance.[23] For this study, we filtered influenza-related tweets containing at least one vaccine-related term (“shot(s)”, “vaccine(s)”, and “vaccination”). We then inferred the US state for tweets using the Carmen geolocation system,[24] and the gender of each Twitter user of the dataset using the Demographer tool.[25] The Carmen tool infers locations of tweets by

three main sources, coordinates of tweets, places name of tweets and locations in user profiles, and most often represents the home location of the user rather than their location while tweeting. The Demographer tool infers binary genders of Twitter users by the names of their profiles. We removed retweets, non-English tweets and tweets not located in the US. We obtained 1,124,839 tweets from 742,802 Twitter users covering four consecutive flu seasons from 2013 to 2017. More details can be found in the supplementary material (A1 and A2).

In addition to tweets about influenza vaccination, we also collected a random sample of tweets from all of Twitter. This was used to adjust the vaccine counts by time, location, and demographics, as described below. The random sample includes approximately 4 million tweets per day since 2011.

CDC Data

We utilized CDC data on influenza vaccination of the four flu seasons for validating our approaches. The CDC data were downloaded from the CDC's FluVaxView system.[4] These data include vaccination coverage by month, by states, and by geographic regions as defined by the US Department of Health and Human Services (HHS). The CDC's estimates are based on several national surveys: the Behavioral Risk Factor Surveillance System (BRFSS, which targets adults), the National Health Interview Survey (NHIS), and the National Immunization Surveys (NIS, which focuses on children). In this study, we use the CDC data for adults (≥18 years old) across all racial/ethnic groups. The CDC reports the "sex" of the respondents, although the underlying surveys ask for "gender" rather than sex,[26, 27] making this variable comparable to our definition of gender in Twitter.

Automated Classification

In our study, we used natural language processing techniques to preprocess and encode tweets into feature vectors, then used the vectors to build machine learning classifiers to automatically categorize Twitter messages that express vaccination behavior. Tweets were classified into yes or no labels in response to the question, "Does this message indicate that someone received, or intended to receive, a flu vaccine?" Specifically, we randomly sampled 10,000 tweets from our collected data from 2012 to 2016 and then used a crowdsourcing platform to annotate the 10,000 tweets,[28] using quality control measures to ensure accurate annotations. The classifiers were trained by the annotated tweets.

The best-performing classification model was a convolutional neural network (CNN), which had a precision (the proportion of tweets classified as vaccine intention/receipt that were correctly classified) of 89.4% and recall (the proportion of vaccine

intention/receipt tweets that were identified by the classifier) of 80.0%, measured using nested five-fold cross-validation. This classifier was applied to the full dataset of 1,124,839 tweets, of which 366,698 were classified as expressing that someone received or intended to receive an influenza vaccine. More details of preprocessing and encoding tweets, and building and selecting machine learning models, can be found in the supplementary materials (A.2) as well as in our prior preliminary work using simpler models.[29]

Trend Extraction and Validation

To evaluate the reliability of the Twitter classification model as a source for vaccination surveillance, we compared the Twitter data to CDC data along three dimensions: time (by month), location (by US state and region), and demographics (by gender). Specifically, CDC FluVaxView provides the monthly percentage of American adults who received an influenza vaccination in a given month in each state, as well as the percentage of Americans who report vaccination in different demographic groups each flu season.

To extract trends over time, we computed the number of vaccine intention/receipt tweets in each month per season, excluding June (the CDC does not report data for June). We only included tweets geolocated to the US. To adjust for variations in Twitter over time, we divided the monthly counts by the number of tweets in the same month from the large random sample of tweets.[8] In addition to monthly rates for direct comparison to CDC, we also calculated weekly tweet rates, providing estimates at a finer time granularity than reported by the CDC. For monthly time series data, we applied an autoregressive integrated moving average (ARIMA) model and linear regression to estimate the CDC data from the Twitter data.[30]

To extract trends by location, we computed the number of intention/receipt tweets in each of the 10 HHS regions and each of the 50 US states. We created per-capita estimates by dividing each count by the number of tweets from the same region or state from the random sample of tweets.

To extract trends by gender, we computed the number of intention/receipt tweets identified as male or female, divided by the corresponding counts from the random sample. We computed this proportion within each US state before aggregating the counts from all states, to additionally adjust for gender variation across location. We provided detailed validation steps and additional experiments in supplementary material A.3.

Confidence Intervals

We present 95% confidence intervals for all results. There are two sources of variability we must account for when constructing confidence intervals. One source is the set of points included in the correlation. The other is the set of tweets used to estimate the level of vaccine intention in each group. When estimating values within fine-grained groups, such as specific US states, the number of tweets can be small, leading to high variability in the estimates that propagates to the estimate of the correlation.

To address these issues, we construct confidence intervals using bootstrap resampling.[31] We perform sampling at two levels. First, we sample the set of tweets used to calculate the estimate in each group (e.g., the tweets in a specific month or location). We then sample the set of points that are included in the calculation of the correlation (e.g., the set of months). The confidence intervals are constructed from 100 bootstrap samples.

RESULTS

Activity by Time

Table 1 shows the correlation between the classified tweets and CDC data from the ARIMA results along with 95% confidence intervals. Figure 1 shows the values from both data sources over time, standardized with z-scores. While the CDC data are only available by month, we show Twitter counts by week (Sunday to Saturday) to illustrate the finer temporal granularity that is possible. In both data sets, there are seasonal peaks every October, when influenza vaccines are distributed in the US. While the overall shapes are very similar, the Twitter data sometimes shows rises later in the flu season that do not correspond to a similar rise in the CDC data, especially in the 2013-14 season, which results in the lowest correlation.

Table 1. Pearson correlations (95% CI) by month in each flu season.

	All seasons	2013-14	2014-15	2015-16	2016-17
Monthly	.799 (.797 - .801)	.644 (.639 - .647)	.950 (.948 - .951)	.909 (.905 - .913)	.910 (.909 - .912)

Activity by Location

The prevalence of tweets mentioning vaccine intention/receipt in each location is shown in Figure 2, where darker color indicates more frequent vaccine mentions. We observe that states in the northwest, especially Washington and Oregon, have higher rates than southeastern states, such as Florida and Alabama. There is a moderate correlation

between the geographic patterns in the Twitter data compared to the CDC data, with a higher correlation at the HHS region level than at the state level (Table 2). The strength of the correlations varies by season, with much stronger correlations in the first two seasons than the latter two seasons.

Table 2. Pearson correlations (95% CI) by geography in each season.

	All seasons	2013-14	2014-15	2015-16	2016-17
State	.387 (.362 - .394)	.300 (.261 - .308)	.214 (.193 - .243)	.051 (.015 - .057)	.025 (.002 - .040)
HHS Region	.467 (.445 - .483)	.690 (.650 - .714)	.573 (.539 - .600)	.137 (.090 - .179)	.244 (.213 - .272)

Activity by Gender

Female users are much more likely to tweet about vaccine intention/receipt than male users on Twitter. The female-to-male ratios in each of the four seasons are (with 95% CIs), respectively: 1.97 (1.96 - 1.98), 1.73 (1.72 - 1.74), 1.59 (1.58 - 1.59), 1.47 (1.46 - 1.48). This ratio is higher than in the CDC data (1.18, 1.17, 1.19, 1.20). However, the two data sources are in relative agreement: the vaccination rate is higher among females than males. For example, in the 2016-17 flu season, the CDC reported that among American adults, 47.0% of women were vaccinated for influenza, compared to 39.3% of men.

We visualized the gender weekly trends and gender ratio of vaccine coverage across locations in Figure 3. The plot of gender weekly trends shows the volume of vaccine intention/receipt tweets over time. The gender ratio has also decreased steadily over time in the Twitter data, while it has stayed fairly constant in the CDC data. The plot of gender ratio shows the female-to-male ratio of vaccine intention/receipt tweets within each US state, with darker color indicating a higher ratio. For example, the figure shows that West Virginia has more females mentioning influenza vaccine behavior than males. We provided additional analyses in the supplementary material A.4.

DISCUSSION

By utilizing natural language processing techniques, Twitter data can be effectively analyzed to identify meaningful information about influenza vaccination intentions and behaviors at the population level. Our key finding is the strong correlation between monthly Twitter-based estimates of vaccination uptake and official CDC uptake estimates. Additionally, exploratory analysis suggests that natural language processing

tools can be developed to further investigate significant patterns in self-reported vaccine uptake by time, location, and demographics.

Traditionally, surveillance efforts have focused on monthly or yearly data. Twitter data allows for greater flexibility and specificity when assessing temporal trends in vaccination. For example, this study shows that it is possible to extract weekly data in addition to monthly estimates. Although we are unable to compare our weekly counts to a validated national metric, we observed high week-to-week variability in general flu vaccine tweets before applying a classifier to filter out irrelevant tweets, but a relatively consistent and predictable pattern in week-to-week tweets indicating vaccine intention and receipt, suggesting that the classifiers are reducing noise at this granularity.

It is possible to capture geographic variability in Twitter data using the Carmen tool. Our results suggest some similarities with the CDC FluVaxView maps, but the associations are not strong enough to make definitive conclusions based on geography. There may be local level trends that contribute to these observed patterns. While the value of this information is limited, it does demonstrate the potential for more detailed geographic analysis in the future, especially as the number of Twitter users continues to climb.

Demographic classifiers are still under development. We were able to utilize the Demographer tool to identify the gender of the person tweeting. Our results suggest there are significantly more tweets indicating intention to vaccinate coming from females. CDC data suggest that this may be accurate, with significantly more females reporting vaccination than males according to FluVaxView. However, the gender gap in Twitter narrowed over the course of the four seasons in our study period, despite staying constant according to the CDC. Other important demographic attributes, like age, are challenging to classify and therefore not considered in this study.[32] Further refinement of demographic classifiers is necessary.

There are limitations to working with social media data. While social media is considered “big data,” we nevertheless ran into challenges with sample size. While the full dataset is indeed large, with over one million tweets, only 33.8% of those tweets can be resolved to the United States, and each experiment further filters down the data into smaller groups. For example, if tweets are counted by month within each US state, then the data needs to be split into 600 partitions (12 months times 50 states) within each year. This has an observable effect of the validity of the results: the correlations between Twitter and CDC are very strong at the national level, but weaker at the regional level, and weaker still at the state level. Sample size of tweets may also explain why the geographic correlations between Twitter and CDC (Table 2) were strong in 2013-14 and 2014-15 than in 2015-16 and 2016-17: the first two seasons contain 25.8% more geolocated tweets than the latter two seasons.

Errors in the natural language classifiers also limit overall accuracy of the approach. We investigated why the correlation with CDC was substantially lower in the

2013-14 season compared to others, and while there is no single conclusive explanation, we observed that the classifiers mis-identified flu-related tweets as indicating vaccine intentions during the peak of the flu season in January 2014, such as tweets expressing regret about not being vaccinated. This type of error was common during this month, resulting in an spike in classified tweets that did not correspond with a true rise in vaccine uptake.

These data limitations affect all social media focused research. However, among studies that utilize natural language processes to study social media data, this is one of the first studies to track vaccination uptake. Our focus on messages that explicitly indicated intention or receipt of vaccination was unique. Existing research has focused on vaccine attitudes or sentiments alone, or substitutes other measures as a proxy for behavior.[33] For example, Salanthe & Khandelwal's 2011 assessment of vaccine-related Tweets during the H1N1 influenza pandemic found strong correlation between vaccine sentiment expressed in tweets and CDC vaccine uptake rates.[17] Another study by Dunn et al. mapped exposure to negative information about HPV vaccines on Twitter to state-level vaccine uptake rates.[20] A more recent study from Tangherlini et al. focused on instances of parents opting-out of immunizations by identifying narratives describing vaccine exemptions on "Mommy blogs".[34]

Our results suggest that self-report data from Twitter can enrich the practice of influenza immunization surveillance and inform influenza vaccination campaigns. To date, the majority of social media surveillance research has been conducted without the involvement of local, state, or governmental agencies.[10] Indeed, most efforts to include public health practitioners in social media research have focused on health communications efforts.[35, 36] By utilizing an adaptable machine learning technique, research questions can be tailored to suit the needs of specific projects or organizations. For example, while we focused on estimating vaccination coverage from FluVaxView, future work could use this data in a study design that is focused on supporting decision making.[37] It may also be possible to utilize social media to track the impact and effectiveness of vaccines in a community, as early work suggests.[38]

Development of demographic classifiers for factors such as age and race/ethnicity is an important next step. One advantage of utilizing Twitter is the ability to capture behaviors from a broader range of adults, especially from groups that may be difficult to reach using traditional surveys, including young adults and members of minority groups such as African Americans and Hispanics.[30, 31] While all groups fail to reach the Healthy People 2020 recommendation of 70% uptake, these same populations (young adults and racial/ethnic minorities) are also the least likely to be immunized against seasonal influenza.[39 - 41]

Incorporating self-report social media data may allow researchers and practitioners to respond to emerging health issues in new and innovative ways, but the progress depends on the ability to integrate novel methods into existing frameworks and

to validate new data streams against reliable metrics. True success will depend on the use of novel techniques to measure positive changes in population health.[42]

COMPETING INTEREST STATEMENT

MD and MJP hold equity in Sickweather Inc. MD has received consulting fees from Bloomberg LP, and holds equity in Good Analytics Inc. These organizations did not have any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. All other authors declare no competing interests.

FUNDING STATEMENT

Preparation of this manuscript was supported in part by the National Institute of General Medical Sciences under award number R01GM114771 to DAB and SCQ and by the National Science Foundation under award number IIS-1657338 to XH and MJP.

CONTRIBUTORSHIP STATEMENT

XH, MCS, DAB, MD, SCQ, and MJP contributed to the design of the study. XH, JC, MD, and MJP contributed to data collection. XH, MCS, JC, DAB, and MJP performed data analysis. XH, AMJ, DAB, SCQ, and MJP interpreted the results. All authors contributed to the editing of this manuscript.

DATA SHARING STATEMENT

All Twitter data used in this study is available in the following repository: https://figshare.com/account/projects/31742/articles/6213878

This contains the annotations for training the classifiers, as well as the classifier inferences on the full dataset. This also contains the extracted metadata, including demographics and location. In accordance with the Twitter terms of service, raw tweets are not shared, but identifiers are shared which can be used to download the tweets.

ACKNOWLEDGEMENT

An early version of this research was presented at the AAAI Joint Workshop on Health Intelligence (W3PHIAI) in February 2017.

Figure Legends

Figure 1: Monthly levels of flu vaccination activity as measured by the CDC versus Twitter.

Figure 2: Levels of flu vaccination activity per US state as measured by the CDC versus Twitter.

Figure 3: Levels of flu vaccination activity of male versus female users in Twitter across time (left) and location (right).

REFERENCES

[1] L. A. Grohskopf *et al.*, "Prevention and Control of Seasonal Influenza With Vaccines: Recommendations of the Advisory Committee on Immunization Practices—United States, 2017–18 Influenza Season," *Am. J. Transplant.*, vol. 17, no. 11, pp. 2970–2982, 2017.

[2] CDC, "Morbidity and Mortality Weekly Report (MMWR)," 2017. [Online]. Available: https://www.cdc.gov/mmwr/volumes/66/rr/rr6602a1.htm. [Accessed: 08-Mar-2018].

[3] T. Santibanez *et al.*, "Flu Vaccination Coverage, United States, 2016-17 Influenza Season," 2017. [Online]. Available: https://www.cdc.gov/flu/fluvaxview/coverage-1617estimates.htm. [Accessed: 09-Mar-2018].

[4] CDC, "Influenza Vaccination Coverage | FluVaxView | Seasonal Influenza | CDC," *Centers for Disease Control and Prevention*, Dec-2017. [Online]. Available: https://www.cdc.gov/flu/fluvaxview/index.htm. [Accessed: 09-Mar-2018].

[5] S. Keeter, "The Impact of Cell Phone Noncoverage Bias on Polling in the 2004 Presidential Election," *Public Opin. Q.*, vol. 70, no. 1, pp. 88–98, Jan. 2006.

[6] R. Iachan, C. Pierannunzi, K. Healey, K. J. Greenlund, and M. Town, "National weighting of data from the Behavioral Risk Factor Surveillance System (BRFSS)," *BMC Med. Res. Methodol.*, vol. 16, no. 1, p. 155, Nov. 2016.

[7] N. V. P. Office, "Flu Vaccination Trends," *US Department of Health and Human Services*, 2017. [Online]. Available: https://www.hhs.gov/nvpo/resources/flu/index.html.

[8] D. A. Broniatowski, M. J. Paul, and M. Dredze, "National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic," *PLoS One*, vol. 8, no. 12, 2013.

[9] E. VELASCO, T. AGHENEZA, K. DENECKE, G. KIRCHNER, and T. I. M. ECKMANNS, "Social Media and Internet-Based Data in Global Systems for Public Health Surveillance: A Systematic Review," *Milbank Q.*, vol. 92, no. 1, pp. 7–33, 2014.

[10] L. E. Charles-Smith *et al.*, "Using social media for actionable disease surveillance and outbreak management: a systematic literature review," *PLoS One*, vol. 10, no. 10, p. e0139701, 2015.

[11] C. Corley, D. Cook, A. Mikler, and K. Singh, “Text and Structural Data Mining of Influenza Mentions in Web and Social Media,” *Int. J. Environ. Res. Public Health*, vol. 7, no. 12, pp. 596–615, Feb. 2010.
[12] N. Collier, N. Son, and N. Nguyen, “OMG U got flu? Analysis of shared health messages for bio-surveillance,” *J. Biomed. Semantics*, vol. 2, no. Suppl 5, p. S9, 2011.
[13] A. Odone *et al.*, “Effectiveness of interventions that apply new media to improve vaccine uptake and vaccine coverage,” *Hum. Vaccin. Immunother.*, vol. 11, no. 1, pp. 72–82, 2015.
[14] M. Dredze, D. A. Broniatowski, and K. M. Hilyard, “Zika vaccine misconceptions: A social media analysis,” *Vaccine*, vol. 34, no. 30, pp. 3441–3442, Jun. 2016.
[15] G. A. Powell *et al.*, “Media content about vaccines in the United States and Canada, 2012–2014: An analysis using data from the Vaccine Sentimeter,” *Vaccine*, vol. 34, no. 50, pp. 6229–6235, 2016.
[16] G. J. Kang *et al.*, “Semantic network analysis of vaccine sentiment in online social media,” *Vaccine*, vol. 35, no. 29, pp. 3621–3638, 2017.
[17] M. Salathé and S. Khandelwal, “Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control,” *PLOS Comput. Biol.*, vol. 7, no. 10, p. e1002199, Oct. 2011.
[18] M. Salathé, D. Q. Vu, S. Khandelwal, and D. R. Hunter, “The dynamics of health behavior sentiments on a large online social network,” *EPJ Data Sci.*, vol. 2, no. 1, p. 4, 2013.
[19] E. J. Nelson, J. Hughes, J. M. Oakes, J. S. Pankow, and S. L. Kulasingam, “Estimation of Geographic Variation in Human Papillomavirus Vaccine Uptake in Men and Women: An Online Survey Using Facebook Recruitment,” *J. Med. Internet Res.*, vol. 16, no. 9, p. e198, Sep. 2014.
[20] A. G. Dunn, D. Surian, J. Leask, A. Dey, K. D. Mandl, and E. Coiera, “Mapping information exposure on social media to explain differences in HPV vaccine coverage in the United States,” *Vaccine*, vol. 35, no. 23, pp. 3033–3040, May 2017.
[21] Z. Tufekci, “Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls.,” in *ICWSM*, 2014, vol. 14, pp. 505–514.
[22] R. Cohen and D. Ruths, “Classifying political orientation on Twitter: It’s not easy!,” in *ICWSM*, 2013.
[23] M. J. Paul and M. Dredze, “Discovering Health Topics in Social Media Using Topic Models,” *PLoS One*, vol. 9, no. 8, p. e103408, Aug. 2014.
[24] M. Dredze, M. J. Paul, S. Bergsma, and H. Tran, “Carmen: A twitter geolocation system with applications to public health,” in *AAAI workshop on expanding the boundaries of health informatics using AI (HIAI)*, 2013, vol. 23, p. 45.
[25] R. Knowles, J. Carroll, and M. Dredze, “Demographer: Extremely simple name demographics,” in *Proceedings of the First Workshop on NLP and Computational Social Science*, 2016, pp. 108–113.
[26] National Center for Immunization and Respiratory Diseases, “National Immunization Surveys (NIS),” 2018. .
[27] National Center for Chronic Disease Prevention and Health Promotion, “Behavioral Risk Factor Surveillance System Questionaires,” 2018. .

[28] C. Callison-Burch and M. Dredze, “Creating speech and language data with Amazon’s Mechanical Turk,” in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, 2010, pp. 1–12.

[29] X. Huang *et al.*, “Examining patterns of influenza vaccination in social media,” in *AAAI Joint Workshop on Health Intelligence (W3PHIAI)*, 2017, pp. 542–546.

[30] J. Franke, W. K. Härdle, and C. M. Hafner, “ARIMA Time Series Models,” in *Statistics of Financial Markets: An Introduction*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 255–282.

[31] B. Efron and R. Tibshirani, “[Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy]: Rejoinder,” *Stat. Sci.*, vol. 1, no. 1, pp. 77–77, Feb. 1986.

[32] L. Flekova, J. Carpenter, S. Giorgi, L. Ungar, and D. Preo\ctiuc-Pietro, “Analyzing biases in human perception of user age and gender from text,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, vol. 1, pp. 843–854.

[33] J. Du, J. Xu, H.-Y. Song, and C. Tao, “Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with Twitter data.,” *BMC Med. Inform. Decis. Mak.*, vol. 17, no. Suppl 2, p. 69, 2017.

[34] T. R. Tangherlini *et al.*, “‘Mommy Blogs’ and the Vaccination Exemption Narrative: Results From A Machine-Learning Approach for Story Aggregation on Parenting Social Media Sites.,” *JMIR public Heal. Surveill.*, vol. 2, no. 2, p. e166, 2016.

[35] X. Zhou, E. W. Coiera, G. Tsafnat, D. Arachi, M.-S. Ong, and A. G. Dunn, “Using social connection information to improve opinion mining: Identifying negative sentiment about HPV vaccines on Twitter,” *Stud. Health Technol. Inform.*, vol. 216, pp. 761–765, 2015.

[36] K. A. McGregor and M. E. Whicker, “50 - Natural Language Processing Approaches to Understand HPV Vaccination Sentiment,” *J. Adolesc. Heal.*, vol. 62, no. 2, Supplement, pp. S27–S28, 2018.

[37] K. P. Alberti, J. P. Guthmann, F. Fermon, K. D. Nargaye, and R. F. Grais, “Use of Lot Quality Assurance Sampling (LQAS) to estimate vaccination coverage helps guide future vaccination efforts,” *Trans. R. Soc. Trop. Med. Hyg.*, vol. 102, no. 3, pp. 251–254, 2008.

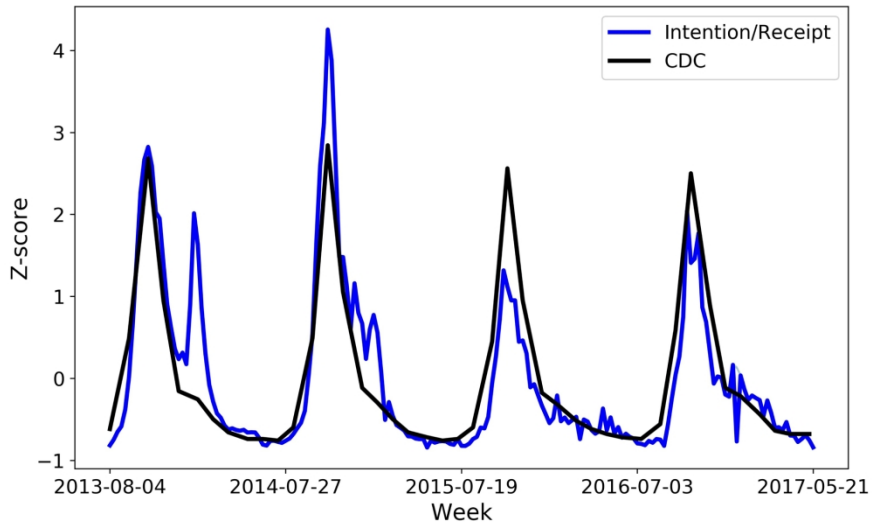
[38] M. Wagner, V. Lampos, E. Yom-Tov, R. Pebody, and I. J. Cox, “Estimating the Population Impact of a New Pediatric Influenza Vaccination Program in England Using Social Media Content,” *J. Med. Internet Res.*, vol. 19, no. 12, p. e416, Dec. 2017.

[39] J. M. Krogstad, “Social media preferences vary by race and ethnicity,” *Pew Research Center*, Feb-2015. [Online]. Available: http://www.pewresearch.org/fact-tank/2015/02/03/social-media-preferences-vary-by-race-and-ethnicity/. [Accessed: 08-Mar-2018].

[40] CDC, “Flu Vaccination Coverage, United States, 2016-17 Influenza Season,” 2017. [Online]. Available: https://www.cdc.gov/flu/fluvaxview/coverage-1617estimates.htm#age-group-adults. [Accessed: 08-Mar-2018].

[41] HealthyPeople, “Immunization and Infectious Diseases,” *HealthyPeople.gov*. [Online]. Available: https://www.healthypeople.gov/2020/topics-objectives/topic/immunization-and-infectious-diseases. [Accessed: 09-Mar-2018].
[42] S. J. Mooney, D. J. Westreich, and A. M. El-Sayed, “Epidemiology in the Era of Big Data,” *Epidemiology*, vol. 26, no. 3, pp. 390–394, May 2015.

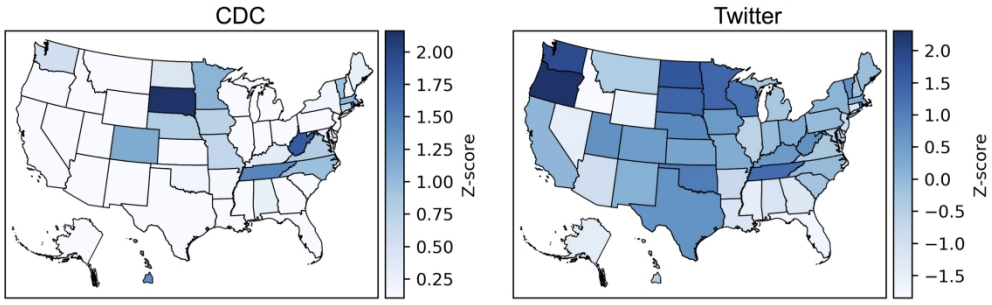
Figure 1. Flu vaccination by time.



Monthly levels of flu vaccination activity as measured by the CDC versus Twitter.

191x114mm (300 x 300 DPI)

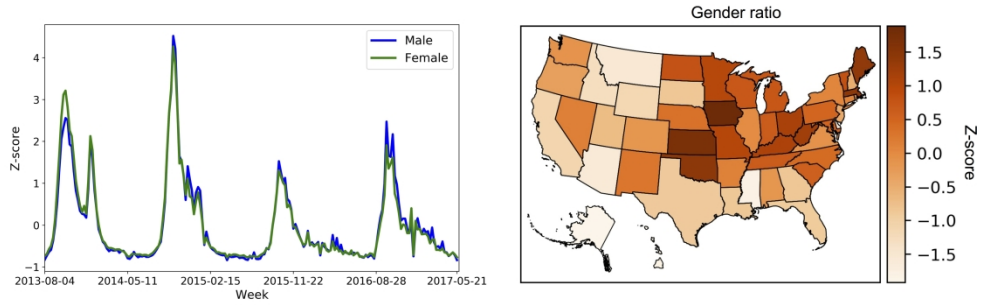
Figure 2. Flu vaccination by US state.



Levels of flu vaccination activity per US state as measured by the CDC versus Twitter.

274x107mm (300 x 300 DPI)

Figure 3. Flu vaccination by gender.



Levels of flu vaccination activity of male versus female users in Twitter across time (left) and location (right).

303x114mm (300 x 300 DPI)

A.1 Data

A.1.1 Data Collection

We collected Twitter data beginning in 2012. However, the tweets collected during 2012-13 flu season were removed in this study, because the data did not cover the complete flu season. We discarded retweets and non-English tweets.[1] For the CDC data, we collected the data from the 2013 to 2017 flu seasons, where each flu season starts in July and ends in May in the following year. To match CDC data, we removed tweets posted in June. The statistical description of our final data is listed in Table 1.

Table 1. Overview of Twitter data in this study

Flu Season	Tweet count	Unique user count
2013 July - 2014 May	264,171	199,733
2014 July - 2015 May	336,644	219,012
2015 July - 2016 May	232,591	147,564
2016 July - 2017 May	263,535	175,770
Total	1,124,839	742,079

A.1.2 Data Preprocessing

Tweets have some unique characteristics that do not exist in traditional text, such as hashtags, hyperlinks, and colloquial language. To make the text more appropriate for natural language processing tools, we preprocessed each tweet according to the following steps:

1. Hyperlinks, hashtags, user mentions in each tweet were replaced with “<url>”, “<hashtag>”, and “<user>” respectively.
2. Repeated punctuation was replaced with “[punctuation] <repeat>”.
3. Each tweet was lowercased and tokenized using NLTK.[2]

A.1.3 Data Annotation

To build training data, we collected annotations for a random sample of 10,000 tweets from the full collection. Annotations were obtained from Amazon Mechanical Turk,[3] with three independent annotations per tweet. Tweets were labeled with the following:

- Does this message indicate that someone received, or intended to receive, a flu vaccine? (yes or no)
    - If yes: has the person already received a vaccine, or do they intend to receive the vaccine in the future.

We refer to tweets labeled “yes” as “intention/receipt” and tweets labeled “no” as “other”.

We rejected annotators whose agreement was anomalously low (percentage agreement was ≤ 60%). Three bad annotators were removed from our final dataset. We took a majority vote on the remaining 29,970 annotations to obtain the final labels. If there was not a majority label, then we defaulted to the "other" label. The dataset contained 10,000 tweets, with 32.8% labeled as positive for "intention/receipt", with a kappa score of 0.79, using Fleiss' kappa to measure the inter-annotator agreement.[4] Then we manually corrected 168 labels of the dataset and finally obtained 31.1% labeled as positive for "intention/receipt".

A.2 Automatic Assessment Methods

To automatically identify tweets expressing vaccination intention/receipt, we used the labeled data to train two machine learning classifiers: Logistic Regression (LR) and Convolutional Neural Network (CNN). The LR model achieved the best performance among other classifiers in our previous study.[5] We implemented Logistic Regression (LR) classifier using the scikit-learn toolkit.[6] CNN has been drawn significant attention in recent years because of its impressive performance on text classification tasks.[7] We trained the two models on the annotated Twitter data. After optimizing the model parameters and hyperparameters, we compared the two models. We finally chose the model that achieved the best performance in the validation experiments.

A.2.1 Logistic Regression

We fed the LR model with TF-IDF weighted n-gram (uni-, bi- and tri-gram) features, as well as part-of-speech (POS) counts from TweeboParser,[8] and emoji and emoticon features derived from two open lexicons.[9, 10] Feature counts were normalized to sum to 1 within each tweet. The list of features we used in this study are shown in Table 2.

Table 2 Details of the feature set for Logistic Regression classifier

Feature name	Feature attributes
N-gram	TF-IDF scores of unigrams, bigrams, trigrams
Part-of-Speech	Counts of POS tags, normalized by the total tags in the tweet
Emoji	Counts of negative and positive emojis, normalized by total counts.
Emoticon	Counts of negative and positive emoticons, normalized by total counts.

We balanced the weight of each label by adjusting weights inversely proportional to class frequencies in the training dataset. We adopted cross entropy as the loss function with $l2$norm penalty for weight regularization.

A.2.2 Convolutional Neural Network

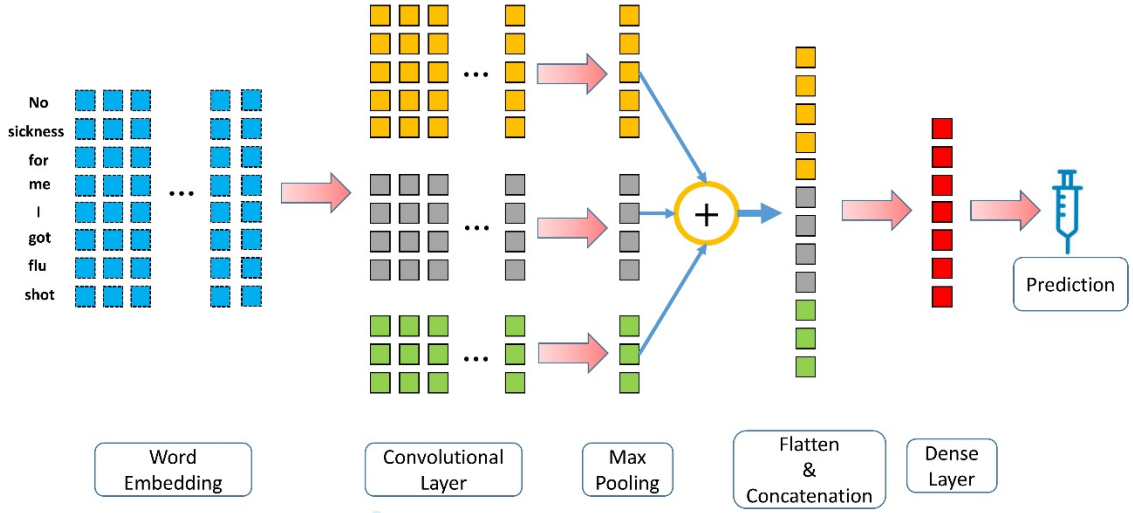


Figure 1. The architecture of the CNN model.

The embedding layer converts processed tweets into an embedding matrix of floating point values, where each row is a vector representation of a word. The embedding matrix is then fed into the convolutional layer, where the matrix will be screened and sampled by the filters. We set 150 filters in this layer. Each filter is a square sliding window and we defined three different sizes of filters: 3*3, 4*4, 5*5. We set the filter stride to 1 and padding mode to "VALID". We obtained the squares by sliding the filters over the matrix. Those captured squares will be fed into the next layer, the pooling layer. We adopt 1-max pooling as the strategy to extract a max scalar value from each square, which outputs the maximum value. We stack another convolutional layer and pooling layer following the first pooling layer, for which the operation steps are the same.

Outputs from the stacked convolutional and pooling layers are flattened, concatenated and fed to the next layer, the dense layer, where it learns and generates a fixed representation for each tweet. We set the activation function as rectified linear unit (ReLU).[11] We set the output dimension of this dense layer to 150. A dropout was applied in the layer, where dropout is a standard method to prevent overfitting by randomly set a proportion of values to zero during training.[12]

We fed the outputs from the dense layer to the sigmoid function to predict the final binary label, "intention/receipt" or "other". We adopted the binary cross entropy function with $l2$ penalty to calculate the loss of predictions. Adam with a learning rate of 0.001 and decay of 0.003 was adopted to optimize the parameters.[13]

A.2.3 Experiment Settings

We randomly sliced the dataset into three pieces: 80% as training set, 10% as development set and 10% as testing set. We trained our two methods, LR and CNN, on the training set, tuned parameters on the development set, and evaluated the methods on the testing set. We

balanced weights of predicted labels in the two models. The models' parameters were selected by accuracy on the development set. The CNN model was trained by 10 epochs, batch size was set by 64, and the dropout rate was set to 0.2. We fixed the length of inputs by either padding sentence to 40 words or slicing the first 40 words. Outputs of the classifiers are probabilities of "intention/receipt", which consider true only if the values are equal to or larger than 0.5 and vice versa. "Precision", "recall", "f1-score" were used to evaluate the performance of each method on the testing set. We focused on the performance of "intention/receipt", not "other" label, which consistently keeps the same evaluation metrics with our previous work.[5]

A.2.4 Selecting Word Embeddings

Word embedding is a language modeling technique that maps words into a set of word vectors.[14] The CNN model in our study was fed with the word vectors. There are two popular frameworks to generate the vectors, Word2vec and GloVe.[14, 15] We selected the best embedding model from the following options:

1. We obtained pre-trained word embedding by running word2vec from Gensim over our collected tweet dataset.[16] We set the tool's default settings except for changing minimum count of words to 1 and number of iterations to 15. We finally obtained 100 dimensional embedding for each word (denoted as *word2vec*).
2. We obtained an embedding model by GloVe with its default parameter settings from its official website (denoted as *glovec*).
3. Google provides pretrained word2vec embeddings on its news dataset,[14] and Stanford provides pretrained GloVe embeddings on its Twitter dataset (denoted as *pre-word2vec* and *pre-glovec* respectively).[15]
4. Character-level embeddings have recently been shown to perform well on text classification.[17] We built word embeddings using a one-hot encoding of characters (denoted as *character*).

We fed the different embedding models to the same CNN model with the fixed parameters. We evaluated the performance by precision, recall and F1-score. The performance is shown in Table 3.

Table 3 Performance of different word embeddings on our dataset.

Word Embeddings	Precision	Recall	F1-score
word2vec	0.894	0.800	**0.843**
glovec	0.820	0.751	0.784
pre-glovec	0.794	0.800	0.797
pre-word2vec	0.895	0.767	0.826
character	0.858	0.729	0.788

Finally, we chose the *word2vec* model trained on the collected data in this study, because it achieves the best performance. We also trained embeddings with 50 and 200 dimensions for both Word2vec and GloVe, but their performance was worse than with 100 dimensions. The word embedding trained on our collected data outperformed pre-trained models from Google and Stanford. Thus, we chose this embedding model for our experiments.

A.2.5 Test Performance of Classifiers

Table 4 Classification performance on test data.

Method	Precision	Recall	F1-score
LR-ngram*	0.837	0.799	0.818
CNN-embedding	0.894	0.800	**0.843**
LR-embedding-average	0.828	0.651	0.729

We used the precision, recall, and F1-score to measure the performance of the two classifiers. We selected the classifier for our analysis tasks based on the best F1-score. We show the test performance in Table 4, where embedding refers to the word vectors from the selected word2vec model, and embedding-average means the trained features of LR are word vectors created by averaging the word vectors of all words in each tweet. Compared to the other two models, the CNN-embedding has better precision and F1-score. We finally selected CNN-embedding for categorizing all the tweets we collected.

A.3 Validation Experiments

In this section, we provide additional details and experiments on the validation process of comparing the Twitter data to the CDC data.

A.3.1 Experimental Steps

We ran both classifiers (LR and CNN) on all tweets from the 2013 to 2017 seasons to obtain labeled tweets. We restricted the analysis to tweets from the United States. We validated our approach across three dimensions: time, geography, and demography.

- Time:
    a. We counted both the weekly and monthly number of tweets classified as “intention/receipt”. To be consistent with CDC’s week definitions, we used the epidemiological week instead of the ISO week to calculate the counts. The data from Twitter and CDC were normalized by z-score separately.
    b. Because the types of data were time-series, we ran the time series model, “autoregressive integrated moving average” (ARIMA), to obtain relationship Twitter and CDC, which was (p, d, q) = (0, 1, 0). The result suggested a linear

relationship between the trends of CDC and Twitter. We then fitted the time series data by a linear regression model using Twitter trends to predict CDC trends.

c. We additionally calculated Pearson correlation and Spearman correlation scores on the Twitter counts and CDC data.

- Geography:
  a. For geographic regions (referred to as "Region"), we aggregated the total counts of "intention/receipt" tweets for the 10 HHS regions separately. In the "Region-year" experiment, we treated the regional tweets in each flu season as a separate point. We normalized the counts of "Region" and "Region-year" by dividing the number of tweets from that region, using the random sample of tweets from the Twitter streaming API.
  b. For "State" and "State-year", we excluded five locations, Northern Mariana Islands, US Virgin Islands, Puerto Rico, Guam, and District of Columbia. These experiments follow the same process as the region experiments, but within individual US states.
  c. All the values were normalized by z-scores.
  d. We validated the geographic data by measuring Pearson and Spearman correlations.
- Demography:
  a. For "Gender", we first counted positive tweets separately for males and females for each flu season. We divided the female counts by male counts of each flu season to generate gender ratios for the Twitter data. Finally, the ratios were normalized by z-score.

A.3.2 Correlation Results

Table 5.1 shows the Pearson correlations over time for both the CNN and LR models. Table 5.2 shows the correlations over geography for the LR model.

Table 5.1 Validation by Pearson correlation for time.

Validation model	All	2013-14 season	2014-15 season	2015-16 season	2016-17 season
CNN	0.899	0.897	0.985	0.985	0.967
LR	0.897	0.927	0.992	0.985	0.984

Table 5.2 Validation of LR by Pearson correlation for geography.

Validation model	State	State year	Region	Region year
LR	0.433	0.212	0.456	-0.121

Table 6.1 shows the Spearman correlation by time, and Table 6.2 shows the Spearman correlation by geography.

As the data is split into finer granularities, such as State or State-year, the correlation scores tend to decrease. This might be caused by a smaller sample size of tweets in the smaller bins. This suggests that if we could obtain more data, this approach will be more accurate.

Table 6.1 Validation by Spearman correlation for time.

Validation model	All	2013-14 season	2014-15 season	2015-16 season	2016-17 season
CNN	0.929	0.948	0.970	0.900	0.943
LR	0.934	0.957	0.975	0.936	0.943

Table 6.2 Validation by Spearman correlation score for geography.

Validation model	State	State year	Region	Region year
CNN	0.402	0.236	0.552	-0.088
LR	0.446	0.208	0.455	-0.133

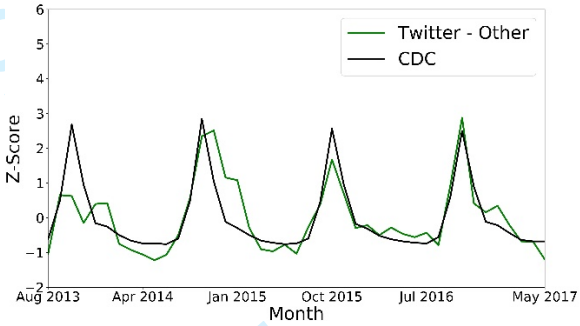
A.3.4 Validation of "Other" Tweets

We have focused on the "intention/receipt" tweets under the assumption that they will be more meaningful than the tweets classified as "other", i.e., tweets that contain vaccine-related phrases but do not explicitly state that someone received or intends to receive a vaccine. In this section, we measured the predictive value of the "other" tweets, which might also correlate with CDC data, and we compare the correlations to the correlations of the "intention/receipt" tweets.

We kept the same experiment settings for the tweets of the "other" label as the "intention/receipt" tweets. We calculated the Pearson correlation with the CDC data. The results are shown in Table 7. We plot the monthly flu vaccine prevalence between "other" (denote as Twitter-Other) and the CDC and weekly prevalence of Twitter data in Figure 2. The "other" tweets have lower Pearson correlation than "intention/receipt" tweets overall with the CDC data. In Figure 2.2, the other tweets in the dataset have very high week-to-week variability, with numerous spikes that do not fit the seasonal trends. This suggests that our classifier is reducing the noise and improving our identification of vaccine behaviors.

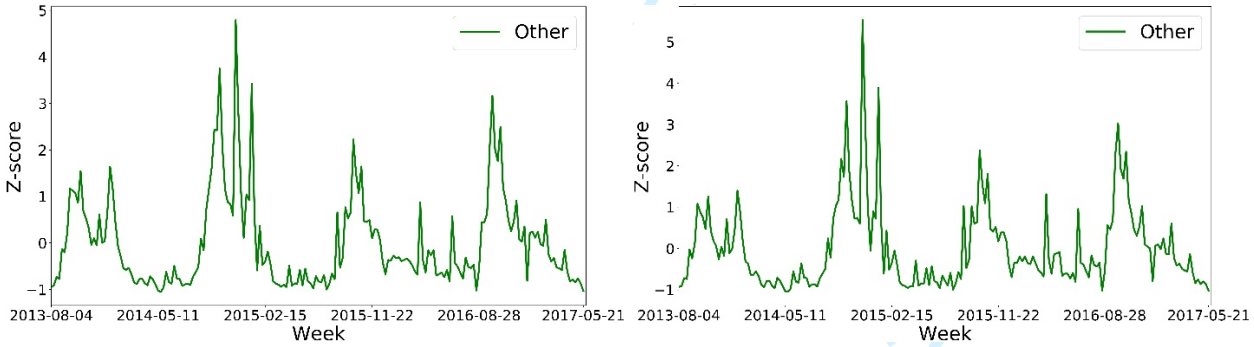
Table 7 Validation Results of CNN and LR by “other” label.

Validation Task	CNN	LR
All seasons	0.820	0.844
State	0.173	0.200
State-year	0.111	0.134
Region	0.587	0.589
Region-year	0.451	0.500



(a) LR (b) CNN

Figure 2.1 Monthly prevalence of “Other” trends from Twitter compared to the CDC.

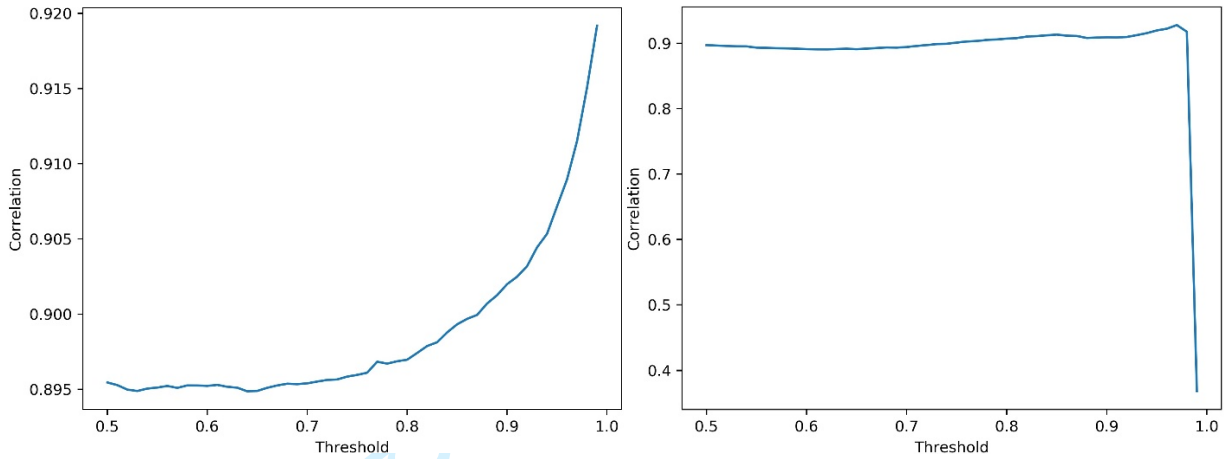


(a) LR (b) CNN

Figure 2.2. Weekly time series of tweets classified as “Other” by LR (a) and CNN (b).

A.4 Additional Analyses

A.4.1 Sensitivity of the Classification Threshold

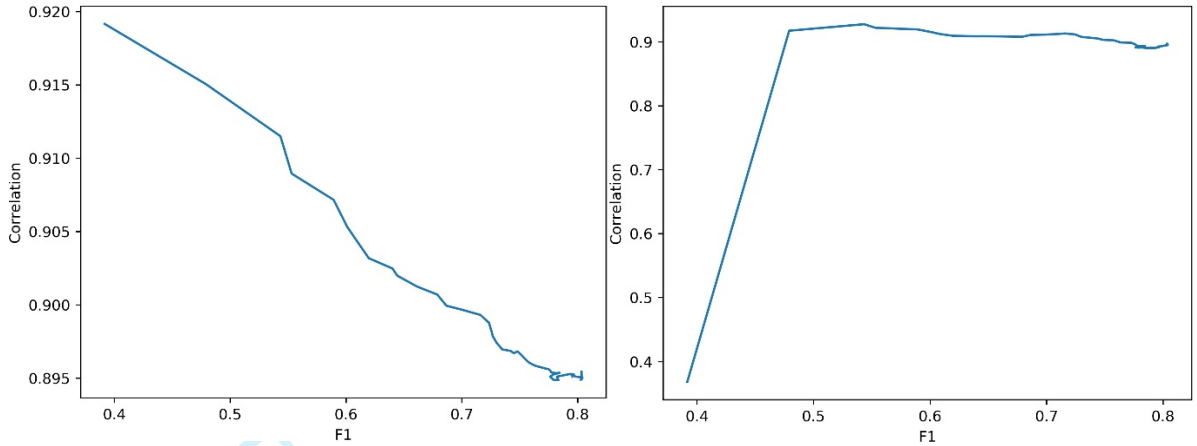


(a) LR (b) CNN

Figure 3. The relationship between the prediction threshold and correlation coefficient.

In this section, we explore how the threshold of classifiers impacts the Pearson correlation. Specifically, the threshold of how the probability of a tweet being positive before it is actually positive. By default, anything with probability greater than or equal to 0.5 is classified as positive, but this threshold can be raised to increase precision (at the expense of recall).

In Figure 3(a) and 3(b), we plotted the relationship between Pearson correlation and prediction threshold for both LR and CNN. Both approaches show that increasing the predicting threshold can improve the correlation coefficient. Increasing the threshold indicates higher confidence of the classifier, that is to say, a tweet will only be considered as “intention/receipt” when the classifier has high confidence. In the view of the classifier, only the tweets have enough evidence to indicate vaccination will be classified as “intention/receipt”. Additionally, we could find that when the threshold of CNN is set to near 0.950, the correlation score decreases rapidly, so raising the threshold does not always improve performance monotonically.



(a) LR (b) CNN

Figure 4 The relationship between the F1 score and correlation coefficient.

.

In Figure 4(a) and 4(b), we explore the relationship between the F1-score and Pearson correlation, because our criteria for selecting the best classifier was by F1-score. The CNN classifier reaches the highest correlation coefficient at around an F1-score of 0.500. Under both models, the correlation drops when the F1 score is too high, likely because the optimal balance is high precision and low recall, even if that drops the F1 score.

For the LR model, while the correlation varies with F1 score, the correlation values are all very similar, and all are above .900. However, the CNN model is not very stable with respect to the correlation coefficient, which might indicate the LR is more robust. We also combined the two approaches to see if we could achieve better performance in the next section.

A.4.2 An Ensemble Perspective of the Two Models

We combined the two models using two linear combination approaches: combining monthly counts of tweets from the LR and CNN (weighted-counts), and combining the prediction probabilities of each approach (weighted-prob). We calculated the combination by the formulas below:

$$Weighted - output = \sum_{i=1}^{2} W_i * X_i \quad (1),$$

$$W_i = \frac{F1_i}{\sum_{i=1}^{2} F1_i} \quad (2),$$

where F1 is the F1-score of each classifier achieved on the test data, and X_i is the count number of each classifier for “weighted-counts” or the predicted probability of “intention/receipt” of each tweet by i-th classifier. Specifically, the weighted-count is the weighted sum of weighted counts from the LR and CNN approaches; for weighted-prob, instead of counts, we calculated the prediction probability of each tweet by the weighted sum of the probabilities from each classifier. The F1-score of each method was used as the weight in the Equation (1). The weights were normalized by the sum of weights to ensure they are within 0 and 1, as shown in Equation (2).

For the validation, we evaluated the performance of the tweets classified as “intention/receipt” and “other”. We validated the two ensemble approaches by calculating Pearson correlation with the CDC data. The results are shown in Table 8. We find that the weighted-counts performs slightly better than the weighted-prob on the tweets classified as “intention/receipt”. The ensemble ways show promising results, outperforming a single classifier.

Table 8. Validation Results of CNN and LR.

Validation Task	Intention/receipt		Other	
	Weighted-Counts	Weighted-Prob	Weighted-Counts	Weighted-Prob
All seasons	0.899	0.895	0.835	0.840
State	0.406	0.437	0.188	0.192
State-year	0.296	0.281	0.092	0.115
Region	0.475	0.432	0.588	0.591
Region-year	0.325	0.264	0.480	0.497

A.4.3 Simpson’s Paradox

In our previous work,[5] LR achieved a .90 correlation on the three consecutive flu seasons (2013-14, 2014-15, 2015-16). In this work, we added a fourth flu season, and LR received a lower correlation score after adding the 2016-17 season. To explore why the correlation dropped, we calculated the correlation on the 2016-17 by itself, to see if this season had a lower correlation that caused the overall correlation to drop. The results are shown in Table 9, comparing the first three seasons (2013-16), the fourth season (2016-17), and all four seasons.

Surprisingly, we discovered that the CNN achieves lower correlation scores than LR on both Seasons 2013-16 and Season 2016-17, even though it exceeds LR on all seasons. This behavior could be explained by “Simpson’s paradox”, a common paradoxical phenomenon in data analysis.[18]

Table 9 Pearson correlation of two different time periods.

Validation Task	Intention/receipt	
	CNN	LR
Seasons 2013-16	0.892	**0.903**
Season 2016-17	0.967	**0.984**
All seasons	**0.899**	0.897

A.4.4 Additional Trend Figures

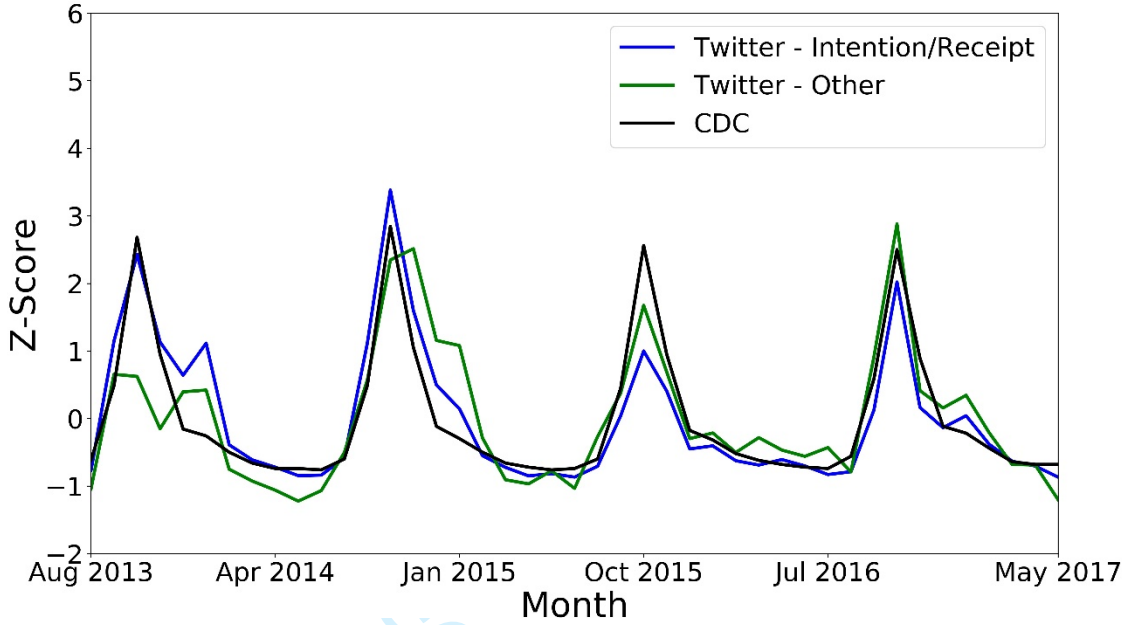
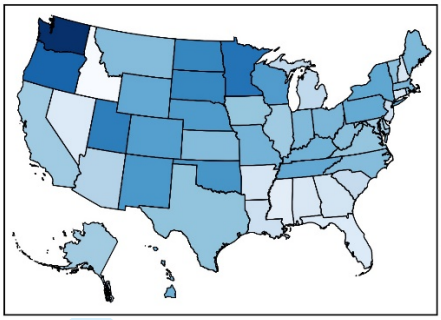
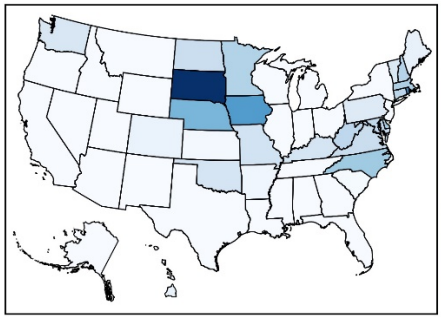
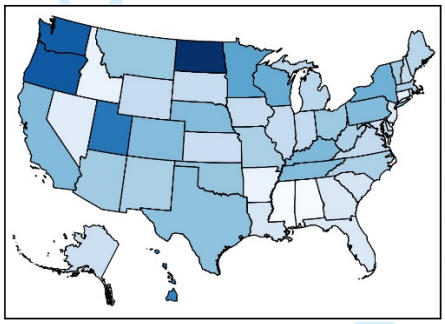
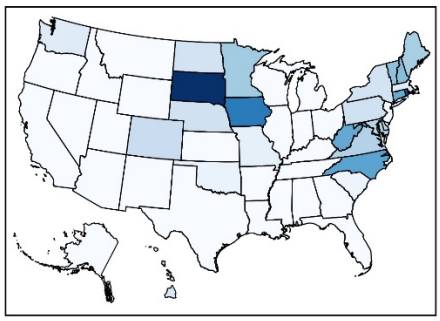
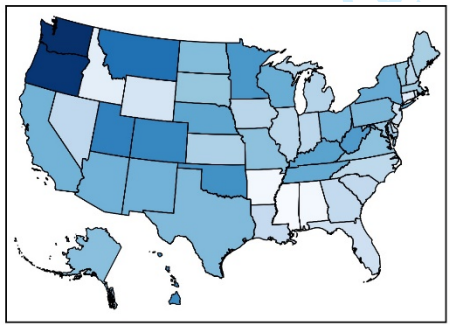
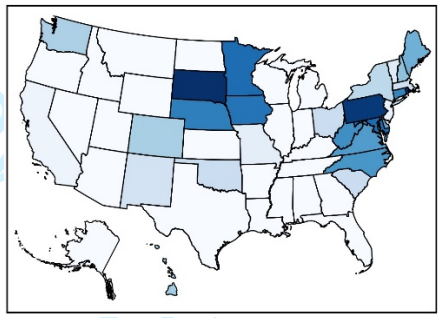
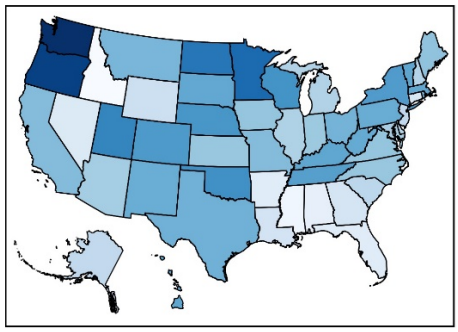
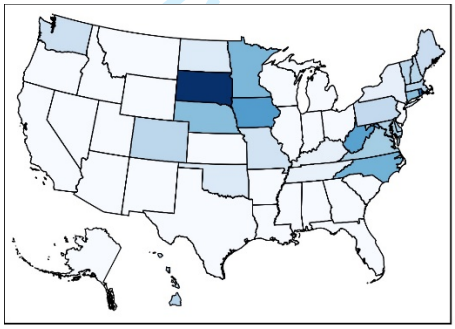


Figure 5. Monthly prevalence of vaccination trends from Twitter and CDC.

Figure 5 shows both the CNN time series (blue) alongside the LR time series (green) and CDC data. There are only minor differences in the trends of the two models. Notice that each peak of the plots is usually in October of the flu season. Yet, there is a distinct peak between Jan. 2014 and Feb. 2014, which might indicate many people also talked about taking flu vaccination shots during that time.

We visualized vaccine coverage in the 50 states each flu season in the Figure 6.[19] We find there are some similar patterns between the Twitter and CDC that the states in the northeast of US show high vaccine coverage and southeast of the US show the lower vaccine coverage, while there are also some clear differences, for example, in the Twitter data, Washington and Oregon show consistently very dark colors.

Figure 6. Flu vaccine trends of both the Twitter and CDC in the U.S.

Flu Season	Twitter	CDC
2014-15		
2015-16		
2016-17		
All		

References

1. Lui M. saffsd/langid.py. GitHub.
2. Bird S, Loper E. NLTK: the natural language toolkit. In: *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*. Association for Computational Linguistics; 2004:31. doi:10.3115/1219044.1219075
3. Callison-Burch C, Dredze M. Creating speech and language data with Amazon's Mechanical Turk. In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. ; 2010:1-12.
4. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull*. 1971;76(5):378-382. doi:10.1037/h0031619
5. Huang X, Smith MC, Paul M, et al. Examining patterns of influenza vaccination in social media. In: *AAAI Joint Workshop on Health Intelligence (W3PHIAI)*. ; 2017:542-546.
6. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res*. 2011;12(Oct):2825-2830.
7. Kim Y. Convolutional Neural Networks for Sentence Classification. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, {EMNLP} 2014, October 25-29, 2014, Doha, Qatar, {A} Meeting of SIGDAT, a Special Interest Group of the {ACL}*. ; 2014:1746-1751.
8. Gimpel K, Schneider N, O'Connor B, et al. Part-of-speech Tagging for Twitter: Annotation, Features, and Experiments. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics; 2011:42-47.
9. Kralj Novak P, Smailović J, Sluban B, et al. Sentiment of Emojis. *PLoS One*. 2015;10(12):e0144296. doi:10.1371/journal.pone.0144296
10. Mohammad SM, Turney PD. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In: *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. CAAGET '10. Stroudsburg, PA, USA: Association for Computational Linguistics; 2010:26-34.
11. Nair V, Hinton GE. Rectified Linear Units Improve Restricted Boltzmann Machines. In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML'10. USA: Omnipress; 2010:807-814.
12. Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J Mach Learn Res*. 2014;15:1929-1958.
13. Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv Prepr arXiv14126980*. 2014.
14. Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems*. Lake Tahoe, Nevada: Curran Associates Inc.; 2013:3111-3119.
15. Pennington J, Socher R, Manning CD. GloVe: Global Vectors for Word Representation. In: *Empirical Methods in Natural Language Processing (EMNLP)*. ; 2014:1532-1543. http://www.aclweb.org/anthology/D14-1162.
16. Rehurek R, Sojka P. Software framework for topic modelling with large corpora. In: *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ; 2010:45-50.
17. Kim Y, Jernite Y, Sontag D, et al.. Character-Aware Neural Language Models. In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. Phoenix, Arizona;

2016:2741-2749.
18. Pearl J. Comment: Understanding Simpson's Paradox. *Am Stat*. 2014;68(1):8-13. doi:10.1080/00031305.2014.876829
19. Root B. matplotlib/basemap. GitHub.