HHS Public Access

Author manuscript

J Public Health Manag Pract. Author manuscript; available in PMC 2024 January 01.

Published in final edited form as:

J Public Health Manag Pract. 2023 ; 29(2): 174–177. doi:10.1097/PHH.0000000000001677.

Analyzing Twitter for Community-Level Public Health Messaging

Kalie M. Wertz, MPA,

Jonathan P. Scaccia, PhD,

Mary Louise Mitsdarffer, PhD

Dawn Chorus Group,University of Delaware, Harvard University

Abstract

Funded in 2021 by the CDC, Communities RISE Together (RISE) aims to increase the reach and effectiveness of messages to address vaccine hesitancy to further health equity. Twitter is a predominant social media source used by communities to share messaging and factual local information with constituents. We looked at the Twitter accounts of the organizations in ten regional communities to examine social media communication patterns to guide how to increase messaging engagement. Specifically, we focused on Twitter post content, likes, and retweets. Our findings identified certain words- such as *food, older adults, equity,* and *covid* - that were most associated with increased likes and retweets on the platform. However, the strongest predictor of receiving likes and retweets is the number of followers. Sentiment was a significant, but not meaningful, predictor of tweet engagement.

Keywords

Social Media; COVID; communication; Twitter

Introduction

In the summer of 2021, the Center for Disease Control and Prevention (CDC) awarded five convening organizations across the United States to provide funding, management support, and technical assistance (TA) to communities to improve their capacity for addressing vaccine hesitancy and health equity. The initiative's goal was to implement and evaluate the evidence-based methods used by each community to alleviate vaccine hesitancy and apprehension. One of these convening organizations, Communities RISE (Reach, Immunizations, System Change for Equity) partnership, identified ten distinct communities across the United States to execute this work. Each of the communities formed unique community-based coalitions, composed of between two and six partner organizations, in their geographic areas. These coalitions received intensive TA and evaluation support through the partnership. Each coalition worked on measurement and implementation plans together leveraging their collective strengths.

Corresponding Authors: Jonathan P. Scaccia, PhD, 401 Penn Street, Suite 147F, Reading PA, 19601.

Human Participant Compliance Statement: None Needed

Conflicts of Interest: We have none to disclose at this time

A key aspect of the CDC's logic model was to improve the reach and effectiveness of public health messaging. Many communities rely on "microblogging" social media accounts to share news, inform the public about events and local happenings, and create dialogue with their community members (Sundstrom & Levenshus, 2017; Wang & Yang, 2020). Twitter is one of the most used social media platforms for nonprofit and community-based organizations, due to its ease of use and focus on quick, digestible information (Guo & Saxton, 2014).

To this end, we focused on scraping and analyzing Twitter posts from the accounts of each organization affiliated with each community. We then asked two central questions to better inform how organizations used Twitter to combat misinformation and better inform their communities:

1. How has Twitter been used by these communities in the past, especially since the beginning of the COVID-19 pandemic?
2. Based on these findings, is there an opportunity to predict community engagement and knowledge transmission through Twitter content?

Methods

Twitter data is one of the most easily accessible and available social media platforms for analysis. Our data collection began by identifying all Twitter accounts for each organization affiliated with one of the ten community coalitions within the RISE collaboration, resulting in 48 accounts, about 75% of the total number of organizations (n=64). We identified no Twitter accounts for any of the organizations affiliated with the community based in Cook County, Georgia, leaving a total 9 different communities to examine through our analysis.

On February 28, 2022, using the Twitter application programming interface (API) and the R package rtweet, data containing the Twitter timelines for each organization were pulled (Kearney, 2019). Our final dataset contained over 83,000 unique tweets across 48 different organizations, with an average of 8,236 (sd = 5,550) tweets per coalition and 1,915 (sd=1,209) tweets per organization. More precise metrics, consistent with common Twitter analysis (Wang & Yang, 2020). While our focus was on vaccine messaging, we chose to look at how RISE partners were using Twitter historically to understand both how they could improve vaccine messaging specifically and their messaging and engagement with tweets overall.

Data were then analyzed using R and a simple natural language processing (NLP) model to examine the textual content of tweets to predict two outcomes: likes and retweets. Following guidance from Hveitfeldt & Silge (2021), along with Silge & Robinson (2017), we then ran a basic machine learning model to examine which words were most associated with these outcomes. To do this, we divided our data into training and test sets. We used the training set to "teach" the model which words were associated with our outcomes of interest, then applied the model to unseen data (the test set). We then plotted the most significant words on a Variable Importance Plot (VIP) to see what terms were most associated with our outcomes of interest.

We also included the overall follower count in the model, as a higher follower count is often related to more engagement simply because those accounts have a larger reach (Guo & Saxton, 2014), and because one of the coalitions was so much larger than the others.

Finally, we were interested in examining whether emotionally-charged tweet content is more likely to provoke engagement from users. We first looked at the average sentiment of the tweets by month using the bing lexicon, which codes words as either positive or negative (Liu, 2010). Then we used the best practice VADER model (Hutto & Gilbert, 2014), which was normed for microblogging settings, to see whether the emotional content of the tweet might yield these outcomes of liking or retweeting.

Results

Our results mirror prior research on the effect of engagement on Twitter. However, we also elucidate potential predictors with a goal of sharing information about social services and general news about the COVID-19 pandemic in specific communities. Since we were looking at the entire history of tweets from these organizations, we used a long baseline that in some cases went back to the inception of the accounts to understand how messaging has changed over time. We were able to identify some trends that appeared to be the result of COVID, such as more COVID-specific messages and an upward turn in sentiment during the pandemic.

Building on past research from Wang & Yang among others, we found the strongest predictor of receiving likes and retweets as engagement was the number of followers any account has at the time of posting. The more followers an account has, the stronger likelihood that they will receive more engagement, due simply to the potential “reach” of the tweet.

Further, our analysis shows that including the use of different phrases such as ‘*food*’, ‘*older adults*’, ‘*equity*’, and ‘*covid*’ within the body of the tweet was associated with increased likes and retweets on the platform. These phrases were identified through a VIP, and demonstrate that the use of these phrases are associated with an increase in either liking or retweeting the tweet.This is likely due to the nature of the tweet content, in which organizations are sharing local community news and information about social services.

To evaluate this *favorite or “like”* model with and without Texas (Figure 1), we used resampling to create a 10-fold cross validation set. From this, we can get the overall performance of the model. The R^2 was 0.05, with a standard error of 0.01. The RMSE, which is in the original units of the model, was 13.62, with a SE of 3.16, concluding our model predicted ~13 likes per tweet.

The tweet sentiment was a significant, but not necessarily meaningful, predictor of tweets. While positive messages trended upward in engagement during the pandemic, the VADER model performed poorly, leading to a lack of conclusion regarding if sentiment could lead to prediction of which tweets would receive more engagement. While overall emotionality of the tweets was statistically significant, they only predicted between 5–6% of the variance, respectively, leaving roughly 95% unknown. Essentially, how emotional the tweet’s content

had no impact on how far of a "reach" the tweet received. Prior research has demonstrated that typically more "negative" framed sentiment had a stronger correlation to likes and retweets; however, through this sample set, we did not find this same conclusion.

Discussion & Conclusion

Twitter is an important tool communities can utilize to further their communication to a wider audience. However, often community organizations are limited in resources to invest in learning more about how to be strategic on social media platforms. Building upon Paul & Dredze (2011), we aimed to find one simple solution for organizations to analyze their Twitter data for maximumization. Using Twitter data from RISE communities throughout the United States, we were able to further expand knowledge around how public health can use the practice of microblogging on Twitter to expand the reach and efficacy of community health messaging. In particular, we found that specific words or phrases could be used to predict higher engagement with likes and retweets on the platform, thus spreading the potential reach of the message further.

Such findings can be of particular use to organizations within public health, as we found that phrases that spoke to health justice, resources, and sensitive populations – i.e. *food*, *equity*, *Covid,* and *older adults* – were linked to greater engagement on the platform, leading us to infer that such phrases positively affect user engagement and potential for messaging. Organizations interested in conducting a similar analysis of their Twitter data can follow a similar methodology and use of R packages to run models and produce their own unique results. This ability to reproduce findings for other samples of data demonstrates how effective analysis of a simple data set - in this case, tweets - can help organizations make their social media usage more efficient with their own limited resources and therefore further social support. We conclude that microblogging efficacy on Twitter may be enhanced by targeting a greater follower base, focusing on sensitive populations and health equity-based language, and not advocating to target audience sentiment but instead concise facts about resources and factual information.

Funding:

Centers for Disease Control Award #: 6 NU21IP000596-01-01

Financial Disclosure:

This work is funded, in part, by CDC grant Award #: 6 NU21IP000596-01-01. These results do not necessarily represent the views of the U.S. Department of Health and Human Services (HHS), or the Centers for Disease Control and Prevention (CDC), nor does it imply endorsement of the material's methods or findings.

Data Availability Statement.

All data are publicly available, though a cleaned dataset is available from the corresponding author upon request.

References

Guo C, & Saxton GD (2014). Tweeting Social Change: How Social Media Are Changing Nonprofit Advocacy. Nonprofit and Voluntary Sector Quarterly, 43(1), 57–79. 10.1177/0899764012471585

Hutto C, & Gilbert E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Proceedings of the International AAAI Conference on Web and Social Media, 8(1), 216–225. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14550

Hvitfeldt E, & Silge J. (2021). Supervised machine learning for text analysis in R. Chapman and Hall/CRC.

Kearney MW (2019). rtweet: Collecting and analyzing Twitter data. Journal of open source software, 4(42), 1829.

Liu B. (2010). Sentiment analysis and subjectivity. Handbook of natural language processing, 2(2010), 627–666.

Paul M, & Dredze M. (2011, July). You are what you tweet: Analyzing twitter for public health. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 5, No. 1).

Silge J, & Robinson D. (2017). Text mining with R: A tidy approach. " O'Reilly Media, Inc.".

Sundstrom B. and Levenshus AB (2017), "The art of engagement: dialogic strategies on Twitter", Journal of Communication Management, Vol. 21 No. 1, pp. 17–33. 10.1108/JCOM-07-2015-0057

Wang Y, & Yang Y. (2020). Dialogic communication on social media: How organizations use Twitter to build dialogic relationships with their publics. Computers in Human Behavior, 104, 106183.

Implications for Policy & Practice

- Partner with organizations with many followers! By leveraging their reach and coordinating messaging, you can more effectively get your specific message out.
- Use specific keywords targeting certain populations and/or outcomes in order to receive more engagement from those audiences.
- More evaluation from social media accounts can let us gauge more real-time reactions and engagements that are hard to pick up in other messaging opportunities (flyers, posters, tabling events, etc). While this is unlikely to be the only source of data used to understand trends, it can supplement other practices and be used to generate more comprehensive policy

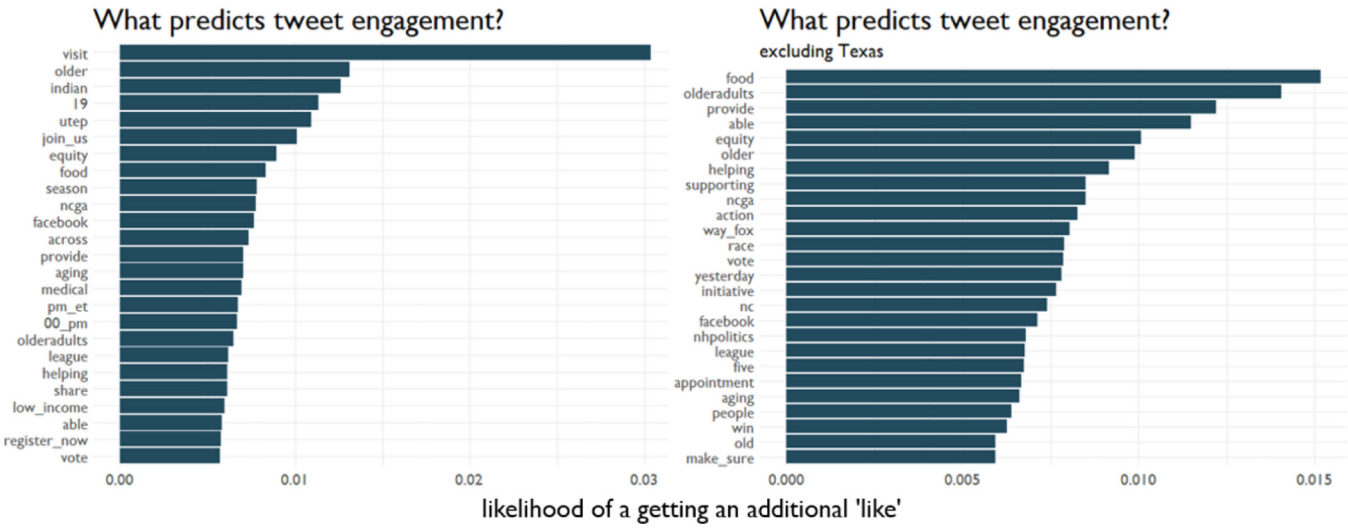

Figure 1:

words associated with “likes” both with and without the Texas-based community (with largest number of followers)