Social Media Monitoring to Inform Public Health Programming

Processes and recommendations from an example Zika-related study to detect rumors and misinformation in Four Central and Latin American Countries

USAID FROM THE AMERICAN PEOPLE

HEALTH COMMUNICATION CAPACITY COLLABORATIVE

K4Health™ Knowledge for Health

Table of Contents

Acknowledgements 3

Introduction 4

Processes followed and challenges encountered 4

- Deciding the appropriateness of a social media monitoring approach 4
- Selecting a social media monitoring platform 5
- Privacy and Institutional Review Board ethics approval 5
- Constructing an appropriate search terms set 7
- Designing the codebook 7

Data analysis processes and challenges 8

- Machine learning option 8
- Training the coders 8
- Too many data! 9
- Challenges with analyzing historical data 9

Lessons learned from the research findings 10

- Over-representation of Twitter™ in the publicly available posts 10
- Over-representation of institutions and organizations in the publicly available posts 11
- Excessive repeating of certain posts 11
- Types of rumors and misinformation circulated on social media about Zika 11

Summary of recommendations for social media monitoring projects 12

Acknowledgements

Support for this report was made possible by the U.S. Agency for International Development, Bureau for Global Health, through the Health Communication Capacity Collaborative (HC3) Project, Cooperative Agreement No. AID-OAA-A-12-00058 and the Knowledge for Health (K4Health) Project, Cooperative Agreement No. AID-OAA-1300068.

In addition, the authors wish to thank Tilly Gurman and Alice Payne Merritt for helping to conceptualize this activity; Sean Maloney, Elisa Vidal, Saifra Khan and Diana Kumar for their crucial contributions to data management, coding and analysis; and Anne Ballard Sara for review and copy-editing.

Suggested citation: Leontsini, E., Parikh, P., Hunter, GC. *Social Media Monitoring to Inform Public Health Programming: Processes and recommendations from an example Zika-related study to detect rumors and misinformation in Four Central and Latin American Countries.* 2019. Baltimore: Johns Hopkins Center for Communication Programs.

Introduction

Digital forums on emerging or urgent health topics are a new domain through which public health officials can monitor public sentiment with the purpose to inform programing, by tailoring and fine tuning their efforts according to the concerns expressed by digital platform users. Content analyses of social media posts have previously been used to examine discourse around H1N1 (swine flu) and emergency contraception, as these are increasingly being discussed in the social media sphere. The immediacy of social media spurs health communication professionals to strive for quicker response times to craft messages that target key or influencing audiences and promote appropriate behaviors tailored to the concerns in social media posts. This potentially rapid feedback loop is especially appealing in emergency situations where widespread fear often propagates rumors, misinformation, and myths that do little to increase self-efficacy for disease prevention.

As part of the United States Agency for International Development (USAID) Zika response, the Health Communication Collaborative (HC3) – based at the Johns Hopkins Center for Communication Programs (CCP) – conducted social media monitoring of publicly available social media posts in 2016-17. Monthly social media monitoring reports were published to identify rumors and misinformation about Zika in El Salvador, Dominican Republic, Honduras and Guatemala with the intent to inform public health programming to address these issues directly.

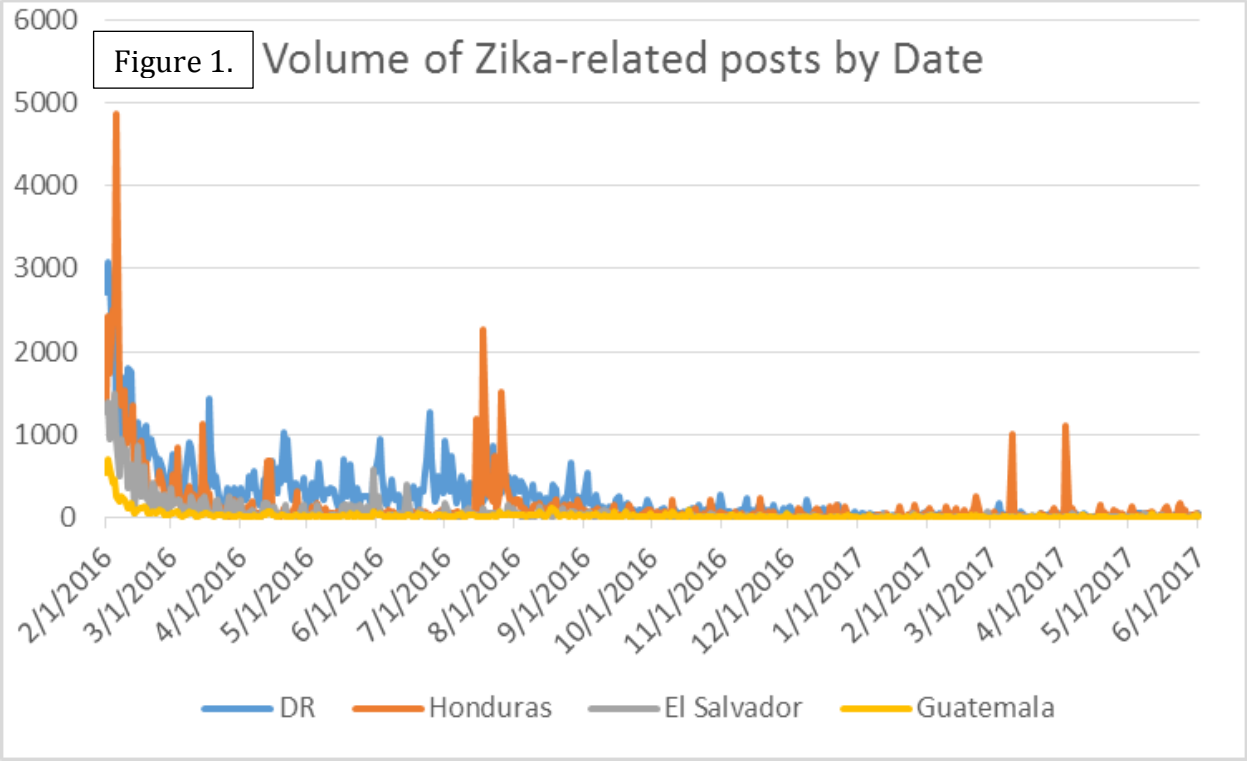
This report describes the process that was followed, challenges encountered, lessons learned, and how these shaped the activity and the findings. The social media listening findings resulting from this effort can be accessed on the Zika Communication Network. Here we offer a series of lessons learned and recommendations in an effort to inform future social media monitoring projects, especially if they have not conducted this type of research before.

Processes followed and challenges encountered

Deciding the appropriateness of a social media monitoring approach

We began the activity with a review of existing literature on the use of social media for Zika communication purposes, and the types of social media websites in the four countries of interest. We found that all four had active social media participation primarily on Facebook and Twitter, but there were differences in the volume of social media posts by country. Guatemala had the lowest volume of Zika-related posts while the Dominican Republic had the highest (Fig 1). Nevertheless all four countries had large enough volumes to justify a social media monitoring approach.

Recommendation #1: Begin with an assessment of the volume of social media use in the country of interest in order to decide if social media monitoring is appropriate.



Selecting a social media monitoring platform

We investigated several available platforms for social media monitoring. We selected Crimson Hexagon because this platform offered a reliable origin identification system for identification of posts from the four countries of interest using geo-inferencing; granted nearly unlimited historical data and no quota for number of posts available to the research team; permitted a large number of queries running in parallel at any one time; had machine learning capabilities for potential content coding by the platform; and was a good value for the membership fee. In addition, the platform allowed for filtering of posts written in Spanish, the majority language spoken by the residents in our four countries of interest. Historical data was of interest so that we could capture the entirety of public posts going back to the beginning of the Zika epidemic in 2015. In retrospect, because of the sheer volume of posts and the limitations of the machine learning, we were not able to take full advantage of all of the platform's capabilities. Practical and logistic considerations will also be important in selecting a platform.

Recommendation #2: Weigh the features of various monitoring platforms in order to select the one that does a satisfactory job within the time and resource limitations of the project; while large data capacities and unrestricted historical data access are attractive features, they may become less important if the project is limited in scope, time or if manual coding will be necessary.

Privacy and Institutional Review Board ethics approval

Privacy and ethics are important considerations for social media monitoring and institutions vary in their determinations and guidance on the use of social media data for

research. A review of ethical considerations in using social media for public health research purposes returned mixed results. In cases where researchers use social media or blogs to deliberately interact with social media users, there is a clear need for Institutional Review Board (IRB) approval to ensure protection of human subjects during research activities. However, when the research intent was a secondary analysis of existing public information without interaction with the user, the level of IRB review needed was unclear. Some IRBs had deemed social media monitoring as non-human subjects research, while others had not.

While social media users made their posts public and no part of the user experience was manipulated by this monitoring activity, researchers may need to consider the expectations of social media users of privacy and consent. For this study, we submitted the protocol to the Johns Hopkins Bloomberg School of Public Health IRB and an IRB in each of the four countries. Exempt status was granted by all five IRBs for human subjects research based on secondary data analysis of existing public information. However, the Johns Hopkins Bloomberg School of Public Health IRB saw a qualitative difference in privacy and confidentiality between personal identifiers accessible online versus personal identifiers stored in a database by researchers without the subject's informed consent, despite the fact that the social media user publicly disclosed those personal identifiers. As a result, no personal identifiers were to be stored in our databases, and posts from minors (less than 18 years of age) had to be excluded from the analyses.

While this study did not require storage of personal identifiers, some identifiers were needed for initial coding of the posts. Furthermore, it was impossible to download data from the Crimson Hexagon platform without identifiers, such as account user name, post URL, and any other account users mentioned within the post. These nuances required an additional IRB amendment, and the development of a cumbersome and highly time-intensive procedure for hand-coding each individual post. Coders had to view the identifying information for each post through the online platform to then return to the database to code for user type (individual or institutional), gender, age (rarely disclosed by users, and when disclosed was used to exclude minors), and whether the post was original or repeated. Coders also had to manually search for and remove user names from stored data, and at the time of report writing, refrain from listing full quotes from posts so that content could not be linked back to the original author.

In the case of this monitoring activity, the lengthy IRB review process extended the start-up time for the project, and the additional privacy safeguarding procedures significantly slowed down the data analysis timeline, prohibiting large sample sizes, and ultimately limiting the range of Zika-related themes to study and inform public health programs.

Recommendation #3: Public health research on social media is an evolving realm with complex privacy considerations; respective projects may need to take extra measures to ensure that social media data is not identifiable. Allow for the extra time and effort required to de-identify the data, as well as to conduct background research on the relevant privacy considerations.

Recommendation #4: If the project will require country-specific human subjects research ethics approval, plan to submit applications to country IRBs as soon as possible to avoid program delays.

Constructing an appropriate search terms set

At the initial stages of the project, the team was interested in what social media users discussed on a broad range of Zika related themes, including rumors and misinformation. Search terms were constructed to capture posts referencing Zika as well relevant data in posts that may not include the obvious terms of Zika, dengue or chikungunya in the post content. An example of an initial, highly inclusive search term set is shown below:

zika OR zica OR dengue OR chi?ung* OR SGB OR ((guill?n OR guill??n) AND (barr? OR bare OR baré)) OR microcefalia OR ((malformaci?n* OR trastorno*) AND (cong?nit* OR neurol?gic*)) OR zancud* OR mosquito* OR Aedes OR aeg* OR vector* OR larva* OR pupa* OR insecticida* OR adulticida* OR fumig* OR fumíg* OR (roci* AND -Rocio*) OR rociá* OR (rocí* AND -Rocío*) OR Naled OR permetrina OR deltametrina OR larvicida* OR ((abat* AND -abatid*) AND -abatible*) OR pesticida* OR repelente* OR mosquitero* OR malla* OR escrin* OR escrín* OR cedrezo* OR criadero* OR recipiente* OR llanta* OR llanter* OR (pila AND -alcalina) OR (((pilas AND -pon*) AND -puestas) AND -alcalinas) OR ((barril* AND -barrilete*) AND agua) OR (tanque* AND agua) OR charco* OR zanj* OR cuneta* OR condón OR (condon* AND -condonacion*) OR preservativo* OR anticoncep* OR (transmisi?n* AND sexual*) OR ((via OR vía OR vias OR vías) AND sexual*) OR ((preven* OR prevén* OR previene* OR evit* OR evít* OR control* OR contról* OR prot?g* OR prot?j* OR protecci?n*) AND (embaraz* OR embaráz* OR embarac* OR embarác* OR natalidad)) OR (planificaci?n AND familiar)

However, due to the large amount of on-topic and off-topic data that such queries yielded, combined with the significantly increased time required to comply with all privacy safeguards at the time of coding, the research team conducted several rounds of refinement of search terms, eventually settling on a less inclusive query:

zika OR zica OR dengue OR chi?ung* OR ((guill?n OR guill??n) AND (barr? OR bare OR baré)) OR (malformaci?n* AND cong?nita*) OR microcefalia OR fumig* OR fumíg* OR repelente*

Recommendation #5: It may take several iterations to come up with a set of search terms that yield on-topic data within a volume that the research team can successfully code and analyze in their allotted time. More inclusive queries yield a broader range of relevant content and less inclusive queries can be more limiting. It is important to strike the right balance for the purposes of the specific media listening project.

Designing the codebook

Codebook construction paralleled the iterations of query construction. Many more codes were initially included, ranging from disease manifestations to mosquito control measures to sexual transmission; whether a rumor or misinformation was contained in the post; the tone and sentiment of the post, as well as its intended purpose. As the query shrank so did the codebook, primarily due to the feasibility challenges of applying a long list of codes. Only codes on rumors, myth, and misinformation were retained.

Recommendation #6: Construct a codebook for data analysis commensurate with the project's research question and with the query. As with the query, expect to modify the codebook a few times to best address the specific research questions, as well as the research team's coding capabilities. Conducting "test" runs of the query and the codebook are recommended.

Data analysis processes and challenges

Machine learning option

We experimented with the platform's machine learning feature with the intention of teaching the platform to apply some of the coding. Under this feature, the researcher assigns a specific category to approximately 30 or 40 posts that the researcher deems match that category. Eventually, the platform learns to recognize patterns of terms contained in the researcher-categorized posts and begins to assign the same category to new posts with similar patterns. There were three limitations to machine coding at the time we undertook this work: 1) The platform could only assign a single code to a post, in contrast to manual coding where multiple codes are allowed and desired; 2) With regard to tone and sentiment in particular, such as frustration, anger, or joy, the platform could only code posts in English; algorithms for Spanish, our primary language of analysis, were not available at the time; and 3) The platform was not able to effectively "learn" how to automatically filter rumors, myths or misinformation about Zika; the primary code of interest. We therefore required a content analysis team to review the database of all posts and make these determinations and code the data appropriately.

Recommendation #7: Although it did not help in this particular study, coding using machine learning may become a viable option in future social media monitoring activities, permitting a much larger volume of data to be processed.

Training the coders

To train the four-person coding team and ensure uniformity within coders, the primary investigator created a random sample of posts from each country; all coders performed content analysis on the same data set and the results were compared across coders. Initially, the agreement among coders was partial, i.e. all coders had applied the same subset of codes to the same post, however not every coder applied all applicable codes to that post. It was a surprise for all to realize how many different codes could be applied to a single relatively short and laconic social media post. Agreement improved after a few rounds of repeat coding followed by a discussion of the discrepancies after each round, with one exception: tone and sentiment.

Recommendation #8: Given the amount of time that it can take to finalize a codebook for content analysis, start the training of the coding team and initial analyses as soon as possible to determine which social media users and thematic codes to prioritize. This prioritization of codes will also determine the amount of time spent on content analysis per post, and enable an assessment of resources needed to get the job done.

Perceptions of the tone and sentiment of posts varied considerably among members of the coding team. Compared to qualitative research coding of interview transcripts, for example, reading a post to discern the author's intended emotions was more limiting. Obtaining high inter-coder agreement for tone and sentiment would have required a considerable time investment in further training, and was consequently dropped from the codebook.

The idioms used in Spanish by the social media users varied slightly from country to country. Each coder was therefore assigned a country with which they were most familiar and whenever new idioms were encountered the whole team was trained to recognize them.

Recommendation #9: Do not underestimate the time and effort required for thematic analyses of short posts.

Too many data!

The team soon realized the large volume of posts (several thousand per month per country) that the carefully crafted queries returned. To minimize the potential of missing any on-topic posts, the queries cast a wide net, resulting in a large number of off-topic posts included in each data set. To help mitigate this, the search terms in the query were limited (see above). It was necessary to further limit coding to a manageable number of posts, based on the weekly rate of coding by the team. The platform returned the posts by month in a random display order, 100 at a time, and each coder coded the first three hundred posts per month per country, a sample of about 5-10% of relevant posts.

Recommendation #10: As institutions, organizations, interest groups and the media increasingly turn to public social media websites to communicate with their audiences, and as the repeating of original posts becomes commonplace, assume large data yields from social media monitoring activities and plan accordingly.

Challenges with analyzing historical data

Due to the various delays described above, we rarely analyzed data in real time. Moreover, analysis started after the volume of posts had peaked (Fig 1), and it was necessary to review preceding months. A caveat with analyzing historical data is that while the platform downloaded the post as it was published, some of the social media users' parameters could have changed since the time of that posting. When looking up each historical post, we found that the user could have deleted it, the account could have been closed, and that links in the post may no longer be valid. The limitation posed by these challenges in analyzing historical data is that the coder, or the coder's supervisor conducting quality checks, did not always have access to the full linked information in the post. It also limited the ability to share real time feedback with public health officials.

Recommendation #11: If projects are interested in providing real-time feedback based on social media monitoring, several "tests" of content analysis are needed beforehand to troubleshoot the problems that will arise in the coding process, and to train the right

numbers of exclusively assigned analysts. As such, real-time feedback is more realistic if the team has done this type of coding and analysis before.

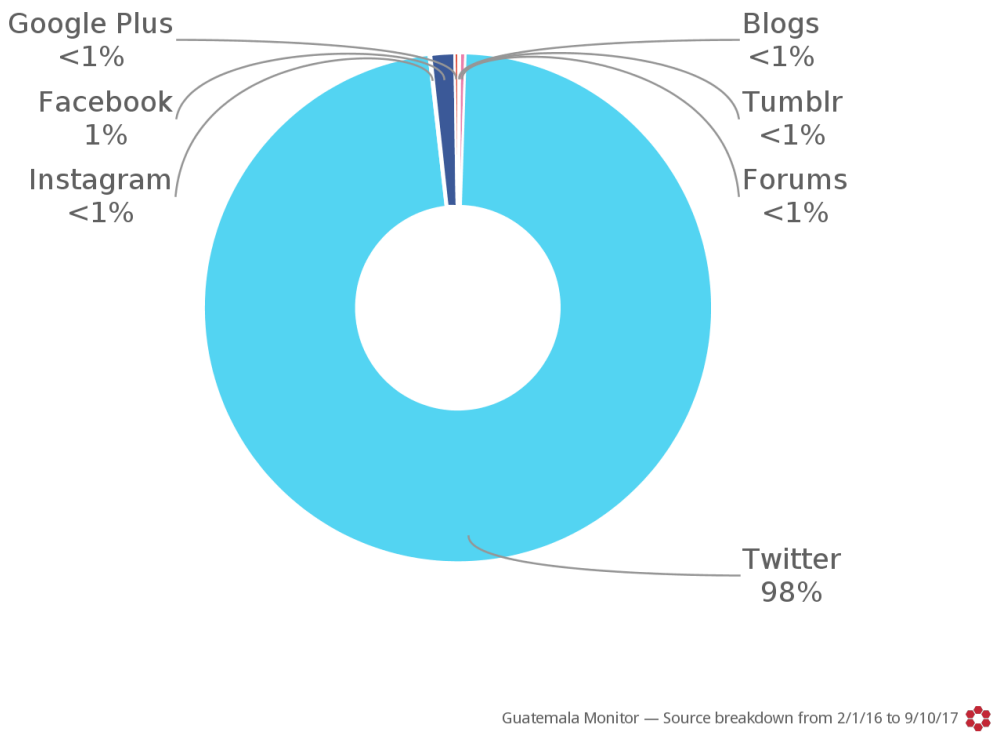
Lessons learned from the research findings

Over-representation of Twitter™ in the publicly available posts

The dominant social media websites – Facebook™, Twitter™, and Instagram™ – were all included in the platform's search system, although the platform only had access to publicly available posts. In addition, public institutions and organizations comprised most of the public posts on these websites compared to individual social media users. The latter tended to post under privacy protections, particularly on Facebook, while the former tended to post on the entirely public Twitter™ platform. Only posts from public Facebook pages – those that do not require an account to be viewed were available on the platform. Twitter™ was therefore the dominant source of posts in this study (Fig 2).

Recommendation #12: Social media monitoring, particularly on Twitter, may be better utilized to examine general trends in Zika-related topics and to identify sources that people trust and engage with on social media, rather than looking for actual views and opinions, rumors and/or misinformation. The latter was similar to seeking a needle in a haystack due to the high volume of posts.

Figure 2. Breakdown of post by source (example from Guatemala)



Over-representation of institutions and organizations in the publicly available posts

Since institutions and organizations comprised the majority of public posts on these social media platforms, as compared to individual members, it quickly became apparent that posts from the former constituted the majority of our study sample. This was a limitation, as individual social media posts might have been much more informative of perceptions around Zika, however, this would have required access to private websites, and unique IRB approvals. It is possible, however, that social media monitoring for other topics might yield more conversation amongst individuals on publicly available platforms.

The majority of authors of Zika-related posts were ministries of health, municipalities, political parties, politicians, international and national NGOs, and the press – often 'digital only' press. Other authors included interest groups and public personalities. Though individuals were not the dominant source of conversation regarding Zika, the high number of posts from news media, interest groups, and local governments regarding Zika-related issues demonstrated that it was a topic that drew the interest and engagement of individual users on social media. In that sense, the monitoring of public posts was a strength, as this type of social media monitoring activity can identify prominent accounts on social media that could be utilized for SBCC initiatives in specific locations (see recommendation 12, above).

Excessive repeating of certain posts

Another important lesson was that certain posts were shared repeatedly either by the original author or by other authors, often hundreds of times a day, thus inflating the volume of posts on the topic. Understanding the motives behind this phenomenon was not within the scope of the study, but likely benefits could include publicity and political gains for the authors. To account for repeated posts, we refined the codebook to categorize original and repeated posts, and counted and reported the number of times repeated posts were posted in that month, adding to the time demands for data analysis.

Recommendation #13: Flooding the internet with repeated posts is a problematic trend in social media. Plan to spend time and effort separating the repeats from the new information. Machine learning may become helpful in filtering out repeated posts which are otherwise very tedious to sort manually.

Types of rumors and misinformation circulated on social media about Zika

Rumors and misinformation were detected within the sample of relevant posts per country per month that we analyzed. Indeed, we found that over the course of July 2016-May 2017, consistent themes emerged in the rumors and misinformation regarding Zika. These fell into the following main thematic categories:

- Zika virus was created in laboratories of large organizations as an attempt to control the population, or to increase the sales of pharmaceuticals or agrochemicals.
- Non-evidence based causes of microcephaly; for instance, as a side-effect of insecticides.
- Non-evidence based home remedies and mosquito repellants, such as "natural" treatments and products.

- Doubts and confusion about how the dengue mosquito that many in the region have lived with for years could transmit so many other new infections.
- Zika was at times conflated with dengue and chikungunya as a single disease.
- Skepticism surrounding the sexual transmission of Zika.
- Misinformation about vector control units demanding payment for services that should have been provided free of charge.
- Doubts that microcephaly is indeed linked to Zika.

Overall, we did not uncover a large number of rumors or misinformation surrounding Zika, and the quantity decreased significantly in March 2017. However, the misinformation trends detected in social media posts can be used to inform implementing partner social and behavior change communication initiatives.

Recommendation #14: Independently of the method utilized to identify rumors and misinformation, Zika social and behavior change communication initiatives should always strive to:

a. Address rumors and misinformation immediately and directly
b. Generate trust and use trusted messengers
c. Make a greater effort to inform communities of government vector control activities so that they are prepared when vector control units arrive
d. Listen to the community's concerns and address them in a transparent way

Summary of recommendations for social media monitoring projects

1. Begin with an assessment of the volume of social media use in the country of interest in order to decide if social media monitoring is appropriate.
2. Weigh the features of various monitoring platforms in order to select the one that does a satisfactory job within the time and resource limitations of the project; while large data capacities and unrestricted historical data access are attractive features, they may become less important if the project is limited in scope, time or if manual coding will be necessary.
3. Public health research on social media is an evolving realm with complex privacy considerations; respective projects may need to take extra measures to ensure that social media data is not identifiable. Allow for the extra time and effort required to de-identify the data, as well as to conduct background research on the relevant privacy considerations.
4. If the project will require country-specific human subjects research ethics approval, plan to submit applications to country IRBs as soon as possible to avoid program delays.
5. It may take several iterations to come up with a set of search terms that yield on-topic data within a volume that the research team can successfully code and analyze in their allotted time. More inclusive queries yield a broader range of relevant content and less inclusive queries can be more limiting. It is important to strike the right balance for the purposes of the specific media listening project.
6. Construct a codebook for data analysis commensurate with the project's research question and with the query. As with the query, expect to modify the codebook a

few times to best address the specific research questions, as well as the research team's coding capabilities. Conducting "test" runs of the query and the codebook are recommended.

7. Although it did not help in this particular study, coding using machine learning may become a viable option in future social media monitoring activities, permitting a much larger volume of data to be processed.
8. Given the amount of time that it can take to finalize a codebook for content analysis, start the training of the coding team and initial analyses as soon as possible to determine which social media users and thematic codes to prioritize. This prioritization of codes will also determine the amount of time spent on content analysis per post, and enable an assessment of resources needed to get the job done.
9. Do not underestimate the time and effort required for thematic analyses of short posts.
10. As institutions, organizations, interest groups and the media increasingly turn to public social media websites to communicate with their audiences, and as the repeating of original posts becomes commonplace, assume large data yields from social media monitoring activities and plan accordingly.
11. If projects are interested in providing real-time feedback based on social media monitoring, several "tests" of content analysis are needed beforehand to troubleshoot the problems that will arise in the coding process, and to train the right numbers of exclusively assigned analysts. As such, real-time feedback is more realistic if the team has done this type of coding and analysis before.
12. Social media monitoring, particularly on Twitter, may be better utilized to examine general trends in Zika-related topics and to identify sources that people trust and engage with on the social media, rather than looking for actual views and opinions, rumors and/or misinformation. The latter was similar to seeking a needle in a haystack due to the high volume of posts.
13. Flooding the internet with repeated posts is a problematic trend in social media. Plan to spend time and effort separating the repeats from the new information. Machine learning may become helpful in filtering out repeated posts which are otherwise very tedious to sort manually.
14. Independently of the method utilized to identify rumors and misinformation, Zika social and behavior change communication initiatives should always strive to:
    a. Address rumors and misinformation immediately and directly
    b. Generate trust and use trusted messengers
    c. Make a greater effort to inform communities of government vector control activities so that they are prepared when vector control units arrive
    d. Listen to the community's concerns and address them in a transparent way