Characteristics of Zika Behavior Discourse on Twitter

Ashlynn R. Daughton, MPH[1,2], Dasha Pruss[1], Brad Arnot[1], Danielle Albers Szafir, Ph.D.[1], Michael J. Paul, Ph.D.[1]
[1]University of Colorado, Boulder, Boulder, CO; [2]Los Alamos National Laboratory, Los Alamos, NM

Abstract

Zika is an important emerging illness that has been linked to neurological syndromes in adult patients, and birth defects in patients infected in-utero. Here, we use Twitter to explore the discourse of individuals tweeting about Zika. We labeled a sample of 500 English tweets to identify common themes, used keywords to track two themes, reproduction and travel, and identify spatial patterns in tweets based on the language of the tweet. We also observed tweets made in the first person that might be indicative of reflections of personal behavior on Twitter. Future work will delve further into first person tweets and continue to examine spatial and temporal trends in the themes temporally observed.

Introduction

Mosquito borne arboviruses (most notably dengue, chikungunya, and, now, Zika) cause massive outbreaks worldwide. In 2015 Zika emerged in South America, and caused a large outbreak throughout many countries in South, Central and North America.[1–3] Although the majority of Zika cases are asymptomatic[4] or cause mild disease, evidence emerged that there was an association between infection and neurological syndromes (e.g., Guillian-Barré[5] or birth defects in fetuses infected in-utero.[6]) Specifically, women who were infected with Zika during pregnancy had fetuses with much higher rates of microcephaly, a condition where the head circumference of the fetus is very small (below the second percentile for gestation[7]), and has accompanying brain defects.[6] The intersection between this outbreak and the impacts on human behavior are of substantial public health relevance. Social media data provide one way to gain insight into human behavior with respect to well publicized public health events.

Social media data now has a rich history of use understanding health-related human behavior. Previous studies have used Twitter as a general way to understand patterns in human behavior with respect to possible risk factors for disease. Paul and Dredze used Twitter data to identify types of ailments discussed on Twitter.[8] Others identified tweets related to human behaviors designed to control infection. For example, Signorini et al. found that an important minority of persons talking about H1N1 on Twitter, also talked about possible control measures (hand hygiene and mask wearing).[9]

As Zika emerged, researchers again used Twitter to find relevant patterns. McGough et al. used a combination of official incidence reports and Internet data streams (including Twitter) to build Zika forecasts for several Central and South America countries.[10] Stefanidis et al. used data from the first 3 months of the outbreak to look at spatial clusters in discussions of Zika on Twitter.[11] They also found tweets referencing pregnancy and abortion.[11]

This study builds upon previous work by looking at trends in discourse. We present preliminary evidence that examines temporal, spatial and keyword trends in a larger Zika Twitter corpus, and analyzes the demographics of Twitter users that talked about Zika. These analyses provide a discussion of Zika tweet content at a finer resolution, and broader spatial context than previous work.

Methods

Data collection

Tweets were collected from Gnip, based on the keywords "zika", "zica" (a common Portuguese spelling of the virus[1], or 'zikv' (a common abbreviation of 'Zika Virus'). Tweets were collected from March 1, 2015 until October 31, 2016. This resulted in just under 15.5 million tweets, 7 million of which are in English. For this initial work, we only coded English tweets. We identified the likely gender of person tweeting using Demographer[12], and identified the likely location of the tweets using Carmen.[13]

[1] https://en.wiktionary.org/wiki/zica

Data labeling

This initial work has focused on the relationship between Zika and behavioral decisions (e.g., those relating to reproduction, travel, mosquito interventions etc.), as well as to important global events (e.g., 2016 Olympics). We identified a list of keywords related to these behaviors (see Table 1). We then filtered the dataset for tweets that included the keywords. Where applicable we included different plural forms (`-ies` rather than `-s` endings). For words that might match longer words or strings (e.g., 'birth' would match 'birthday'), we included `\b`, a Python regular expression marker for a word boundary.

Table 1: Behavior keywords. Keywords were identified by reading 500 hundred random English tweets and identifying words that were commonly associated with behaviors. Keywords were translated into Spanish and Portuguese using Google translate (translations not shown for brevity).

Keywords							
pregnancy	pregnancies	microcephaly	microcephalies	brain	babies	baby	\bbirth\b
defect	pregnant	condom	\bsti\b	std	sex\b	travel	female
abortion	syndrome	guillain	barre	repellent	spray	olympic	fetal
carnival	autoimmune	born	pope	catholic	ultrasound	transmission	transmit
abnormality	abnormalities	reconsider	cancel	protect	paralyze	paralyzing	church
conceive	spring break	baby-moon	paralysis	terminate			

Table 2: Label frequency and example tweets. Tweets were included in frequency if both annotators used the label.

Code & Definition	Example (Paraphrased to preserve privacy)	Freq (%)	Cohen's κ
Travel discussion changing plans, sharing information about travel advisories	CDC says pregnant women should consider delaying travel	34 (6.8%)	0.47
Sex and/or safe sex behaviors: contraceptive methods or comments about sexual transmission.	Men, if you're partner is pregnant, abstain or use condoms to prevent zika	13 (2.6%)	0.64
Delaying pregnancy: advisories to delay pregnancy, or use methods to prevent pregnancy.	Delay pregnancy if you live in a place with zika	17 (5%)	0.65
Reproductive rights: legal and personal opinions about abortion or birth control use.	Americans weigh in on late-term abortions in cases of zika-linked defects	14 (3.4%)	0.41
Birth defects: birth defects associated with Zika	Baby born with zika-related birth defect	56 (11.2%)	0.65
Paralyzing syndromes: Guillian-Barré or other paralyzing syndromes	Guillain-Barré cases associated with #Zika infection rise	17 (3.4%)	1.0
Intervention strategies ways to interrupt transmission (e.g., mosquito spraying)	Country to spray for mosquitos to prevent zika	37 (7.4%)	0.56
Olympics or large group gatherings information about very large events (e.g., Olympic games, Carnival etc.)	Olympic athlete decides to not compete over zika fears	68 (13.6%)	0.78
Misleading jokes, references to conspiracy theories that are factually inaccurate	Brazilian government admits microcephaly not caused by zika	8 (1.6%)	0.38

The resulting subset included 3,572,320 tweets. Of these, we randomly sampled 500 for coding. Labels were decided iteratively as a group, with the final labeling schema available in Table 2. Tweets were coded independently by two team members. Multiple labels per tweet were allowed. Labels were assigned based solely on the content of the tweet; we did not attempt to look at the context of any URLs linked, or find additional context from that user's other tweets. Inter-rater reliability is presented in Table 2 (see Cohen's κ).

Behavior threads

Lastly, we performed initial explorations of two specific types of tweets. Because our initial labeled corpus of 500 is too small to build a robust classifier system, we identified a few keywords that related (1) reproduction/ birth control (keywords: 'birth control' 'abortion') and (2) travel ('travel' 'cancel' 'plane\b' 'airline'). Keywords were used in lieu of more sophisticated methods as a crude approximation of topics of public health relevance to begin to identify temporal trends in topics. These topics were chosen because they relate to the two most common types of advisories observed - recommendations about limiting travel and delaying pregnancy.

Results

Demographics and label frequencies

Within the behavior Twitter corpus, we have a somewhat higher proportion of male tweeters (61%) compared to female tweeters (37%) (the remaining 2% could not be identified). This is in contrast to surveys that suggest that men and women use Twitter at roughly the same frequencies.[14] We were able to predict a user's location for 61% of tweets. The majority came from the United States (20%). Venezuela, United Kingdom, Brazil, Canada, and India contributed 1-3% each. 179 additional countries had at least 1 tweet present in the dataset. Additional spatial analyses are available in Figure 2.

Labels, descriptions and example tweets are listed in Table 2. The most common label in our random sample was 'birth defects', followed by discussions of the Olympics, travel, 'intervention strategies', safe sex behaviors and 'reproductive rights'.

We also noticed a rare, but interesting handful of tweets that were written in first person. These are an interesting subset because they provide insight into individual's opinions on policy decisions, or reactions to guidelines on health behaviors, beyond what might be inferred by noticing what kinds of informational things an individual retweets or links to. A paraphrased example is 'Canceling our baby-moon trip due to Zika concerns'.

Trends in behavior themes

Figure 1 shows the fraction of behavior tweets each day that have keywords matching reproduction or travel keywords (see section Behavior threads).

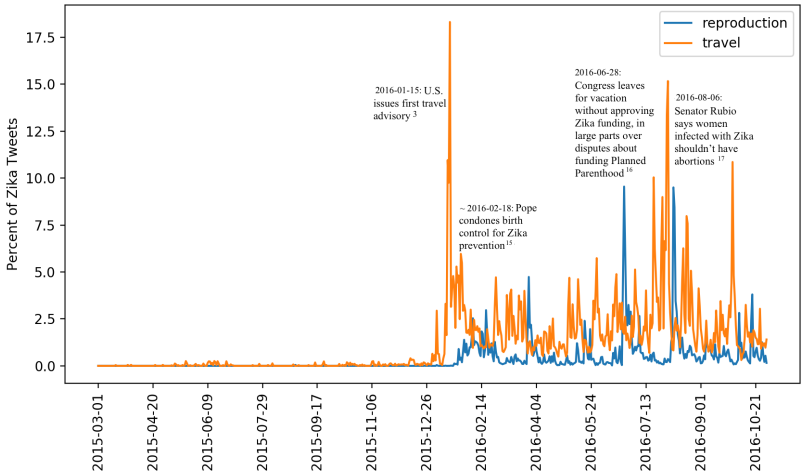

Figure 1: Initial behavior trends for the three categories of words. Blue tweets are those that talk about reproduction and birth control and orange are tweets about travel. Annotations come from articles in the popular news media.[15–17] Dates are formatted yyyy-mm-dd.

These keywords are not meant to be exhaustive representations of discourse, but are useful to identify general patterns

to explore more in future work. Several spikes within the dataset seem to correspond to policy discourse in the US and have been added as annotations onto Figure 1. Overall, travel tweets are much more common than those about reproduction. Given the intensely personal and political nature of abortion and reproductive rights, this is unsurprising. However, further investigation into the content of these tweets is warranted.

Spatial trends

Figure 2 shows the number of geotagged tweets by language in each location by month between August 2015 and June 2016. Unsurprisingly, the volume of tweets in each language strongly correlates with the primary language spoken there. In general, tweets were rare in the first few months of the outbreak, and began increasing primarily among Portuguese speakers November and December 2015. In January and February 2016 the number of tweets increased primarily among Spanish and Portuguese speakers in Central and South America, possibly because the outbreak was extremely active at that time. In contrast, there is an increase in English tweets throughout the spring and summer of 2016, possibly in accordance with increased media attention to the outbreak.

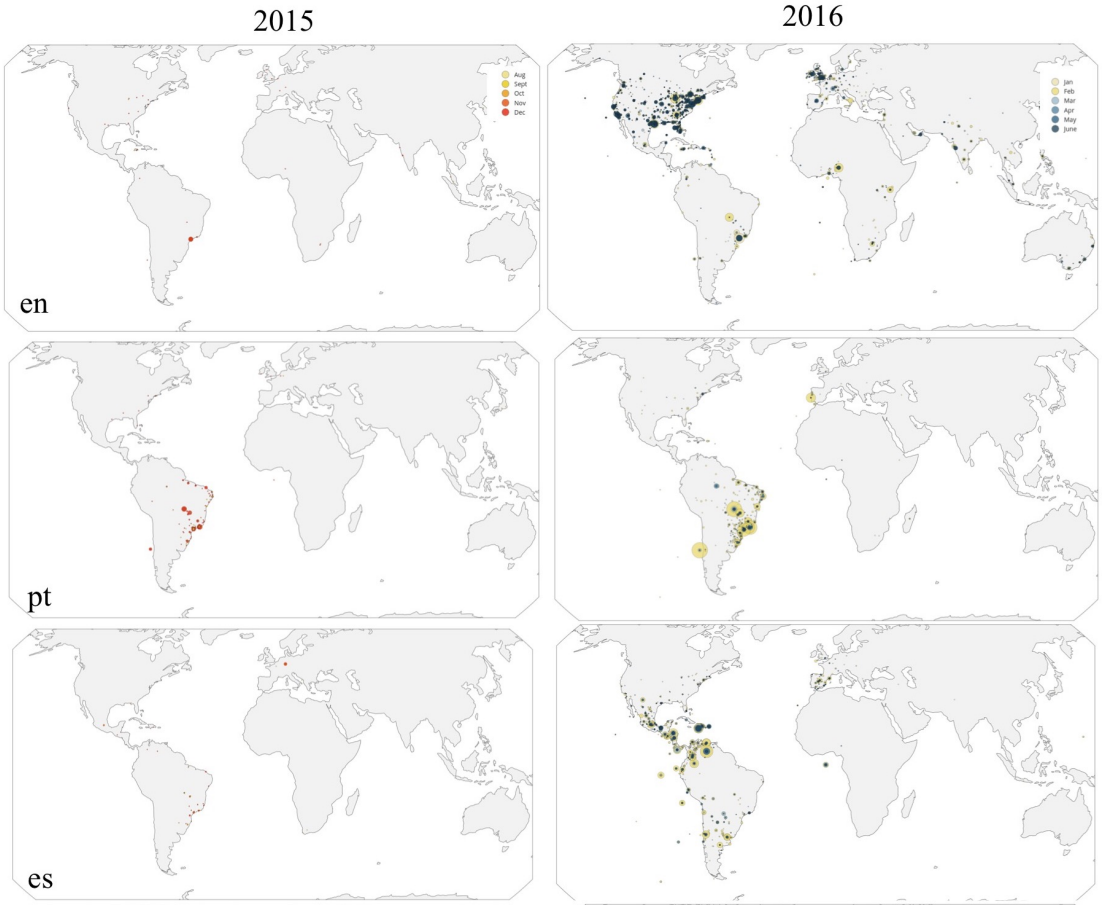

Figure 2: Spatial context, August 2015 - June 2016. The area of the bubble is directly proportional to the number of tweets. Tweets are divided by language - 'en' corresponds to English, 'pt' to Portuguese, 'es' to Spanish.

Conclusion

This work describes common themes in Zika-related tweets, as well as temporal and spatial trends in those tweets. Future work will employ other techniques, like topic modeling and named entity recognition to identify topics discussed within each theme identified. Additional focus will be on first person tweets because they provide insights into

people's individual decisions to change behavior. Although first person tweets were rare, we found tweets describing the decision to change travel plans based on advisories from public health organizations. Analysis of these tweets will allow us to understand how public health agencies can better use Twitter to communicate important public health information.

References

1. Scott C. Weaver and others. Zika virus: History, emergence, biology, and prospects for control. *Antiviral Research*, 130:69–80, June 2016.

2. Jessica Patterson, Maura Sammon, and Manish Garg. Dengue, Zika and Chikungunya: Emerging Arboviruses in the New World. *Western Journal of Emergency Medicine*, 17(6):671–679, November 2016.

3. Mary Kay Kindhauser and others. Zika: the origin and spread of a mosquito-borne virus. *Bulletin of the World Health Organization*, 94(9):675–686C, September 2016.

4. Seyed M. Moghadas and others. Asymptomatic Transmission and the Dynamics of Zika Infection. *Scientific Reports*, 7(1), December 2017.

5. Beatriz Parra and others. Guillain–Barré Syndrome Associated with Zika Virus Infection in Colombia. *New England Journal of Medicine*, 375(16):1513–1523, October 2016.

6. A. S. Oliveira Melo and others. Zika virus intrauterine infection causes fetal brain abnormality and microcephaly: tip of the iceberg?: Physician Alert. *Ultrasound in Obstetrics & Gynecology*, 47(1):6–7, January 2016.

7. Jernej Mlakar and others. Zika Virus Associated with Microcephaly. *New England Journal of Medicine*, 374(10):951–958, March 2016.

8. Michael J. Paul and Mark Dredze. You are what you tweet: Analyzing Twitter for public health. In *5th International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*, July 2011.

9. Alessio Signorini, Alberto Maria Segre, and Philip M. Polgreen. The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. *PLoS ONE*, 6(5):e19467, May 2011.

10. Sarah F. McGough and others. Forecasting Zika Incidence in the 2016 Latin America Outbreak Combining Traditional Disease Surveillance with Search, Social Media, and News Report Data. *PLOS Neglected Tropical Diseases*, 11(1):e0005295, January 2017.

11. Anthony Stefanidis and others. Zika in Twitter: Temporal Variations of Locations, Actors, and Concepts. *JMIR Public Health and Surveillance*, 3(2):e22, April 2017.

12. Rebecca Knowles, Josh Carroll, and Mark Dredze. Demographer: Extremely simple name demographics. In *EMNLP Workshop on Natural Language Processing and Computational Social Science*, pages 108–113, 2016.

13. Mark Dredze and others. Carmen: A Twitter Geolocation System with Applications to Public Health. In *Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence*. AAAI Workshops, 2013.

14. Maeve Duggan and others. Social media update 2014. Technical report, Pew Research Center, January 2015.

15. Daniel Burke and Elizabeth Cohen. Pope suggests contraceptives could be used to slow spread of Zika. *CNN.com*, February 2016.

16. Everett Burgess and Jennifer Haberkorn. Democrats block Zika funding bill, blame GOP. *Politico*, June 2016.

17. Marc Caputo. Rubio: No abortion for Zika-infected women. *Politico*, August 2016.