Zika Virus Twitter Exploration

Carolyn Clayton

May 16, 2016

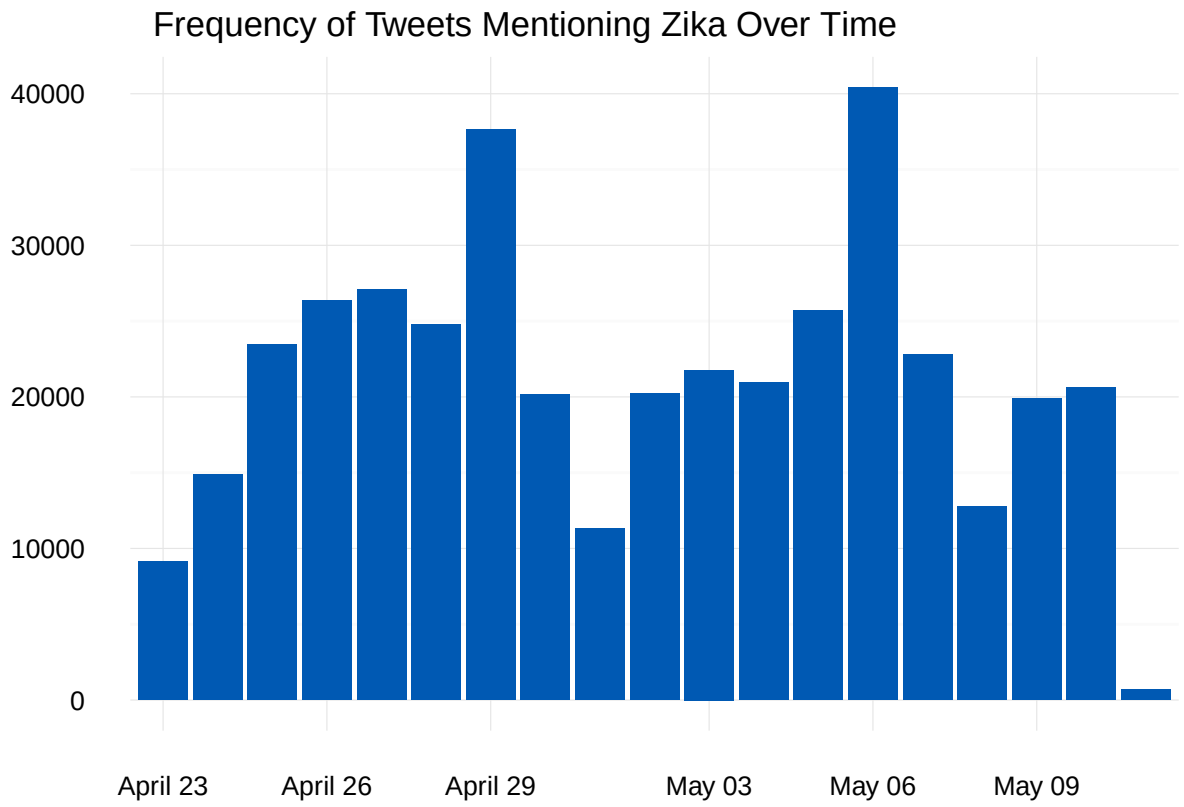
Study Background

It has been proposed that Twitter data could be used as a proxy to monitor disease outbreaks in semi-real time across the globe. It is assumed that those locations that are experiencing the outbreak are more likely to tweet about it. In this case, we wanted to analyze tweets containing mentions of Zika virus, compare frequency across time and geolocation, and conduct a sentiment analysis to determine whether tweets were more positive or more negative over time.

Data was pulled from Twitter on 5/2/2016 and 5/11/2016 using the twitteR package (v. 1.1.9). Tweets were searched on the word "Zika" ignoring casing and any attached symbols (e.g. ZIKA, #zika, and @zika were also pulled). The data represented tweets from April 23 to May 11 and contained 400,683 tweets by 199,072 users. For a word-by-word analysis, the tidytext package (v. 0.1.0) was used to separate tweets into individual words.

Findings

An average of 21,089 tweets mentioning Zika were tweeted per day; an average of 41.2% of which were retweets. The highest number of tweets occurred on Fridays, with a contrasting lower volume of tweets on Saturdays and Sundays.



For more meaningful results, common English, Spanish, and Portuguese words (e.g. “the”, “el”, and “o”), and common Twitter words (e.g. “rt” and “http”) were removed. There was a great disparity in frequency of most common words and hashtags. By far the most commonly used word was “Zika,” as expected given that “Zika” was the search keyword used when pulling tweets. However, it was surprising that “virus” was much less commonly used, appearing only 25.3% as often as “Zika.”

Common hashtags are mostly as expected, with “#zika” appearing 8 times as often as the next most-frequent hashtag. Unexpectedly, the hashtag “#6” appears frequently. This appears to be largely due to “likes” from YouTube sharing to Twitter from users who liked a Portuguese cartoon, episode 6 of which mentions Zika Virus.

The top five most common mentions were more uniformly frequent. However, 3,026 tweets mentioned @YouTube, @fecastanhari, and the hashtag #6 indicating these mentions were due to synced YouTube likes from the Portuguese cartoon. The majority of remaining mentions are likely due to retweets rather than direct conversation, given that the dataset contained 41.2% retweets and the original poster is automatically mentioned in a retweet.

Table 1: Five Most Common Words, Hashtags, and Mentions

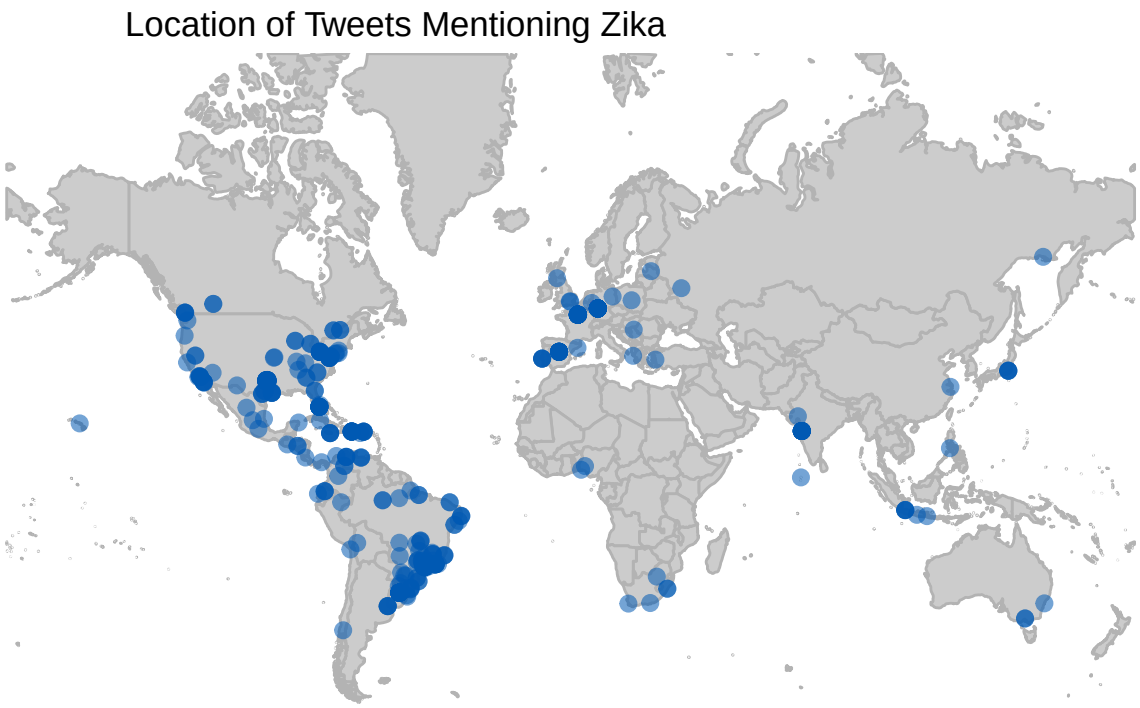
Type	Word	Frequency
Word	zika	384,723
	virus	97,359
	puerto	32,125
	rico	31,350
	dengue	19,126
Hashtag	#zika	60,200
	#zikavirus	7,614
	#health	3,734
	#6	3,140
	#news	2,647
Mention	@youtube	5,534
	@fecastanhari	3,193
	@who	2,987
	@hillaryclinton	2,909
	@zika_news	2,425

40 Most Common Hashtags

Note: Approximately sized by frequency

Geospatial data was available for 469 tweets. This represents 0.117% of total tweets as geolocation is an opt-in service and is only available for original tweets (i.e. not for retweets). Twitter data appears to be a poor indicator of location of Zika virus outbreaks as the majority of current cases appear in countries from Brazil to Mexico, including the Caribbean (Centers for Disease Control and Prevention, 2016). Most tweets

with geospatial data were located in the Western hemisphere with a large number originating from North America, although there was also a concentration in Europe. This may be due to media coverage in North America and Europe, or a higher percentage of the population with access to or engagement with Twitter.

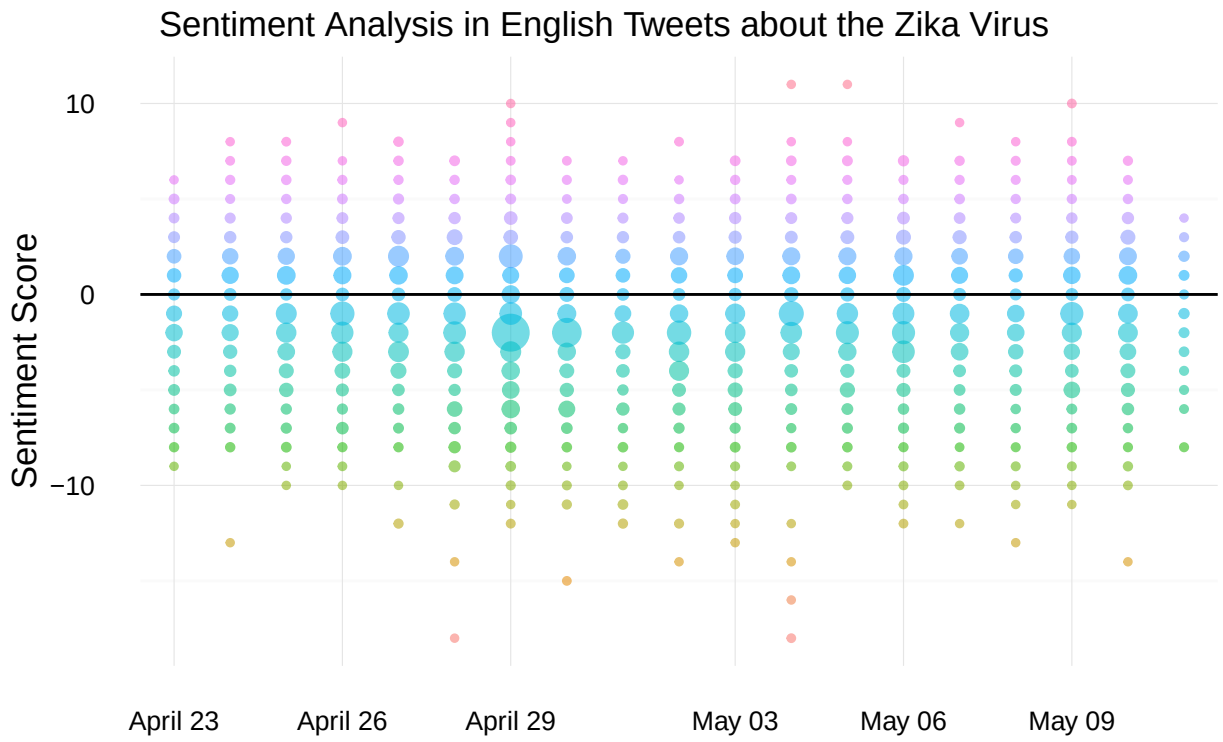


Note: Data is from 469 tweets created between April 23 and May 11.

A sentiment analysis was performed by using scoring developed by Finn, Arup, and Nielsen (AFINN, 2011). This lexicon contains English words assigned a score from -5 to 5; most words with a score of -4 or -5 are expletives. Sentiment score for a tweet is calculated as a sum of positively and negatively scored words (e.g. agonised = -3, benefit = 2) within a tweet. As such, it is possible to have a 0 score if there is an even weighting of positive and negative sentiments. For example, a tweet with a score of -10 was "Damn I just heard about that Zika shit that's crazy!" A tweet with a score of 8 was "South Korean Olympic Committee Unveils Athletes' Uniforms Designed to Protect Against Zika #funny #LOL https://t.co/v4xhqJIdQ4". Analysis was performed word-by-word so negations were un-accounted for; e.g. "Zika is not good for babies" would be reported as a positive tweet.

Sentiment scores were plotted by day, with larger points indicating a larger number of tweets at a given score. The average score of tweets overall was -1.2 (SD = 2.2). Tweets had a fairly constant average score near -1.0, but dipped down for 6 days from April 28 to May 3. The average score was lowest on May 2 (-1.8, SD = 2.1)

and highest on April 24 (-0.6, SD = 2.3).



Note: Data is from 117553 tweets created between April 23 and May 11.
Size of points represents number of tweets with a given score on a given day.

Conclusions

The large number of tweets mentioning Zika serves as a rich dataset for exploration of public perception about the Zika virus. English tweets analyzed with the AFINN lexicon generally had a slightly negative score, although there was a dip in score for 6 days. Of the tweets where geolocation data was available a large number were from North America and Europe, indicating a confounding factor such as media coverage; however the remaining tweets often originated from countries where the outbreak is manifested.

There are some limitations to this study. The data only spans 0 days, so inference on trends over time is limited. The source of data may also be biased as the population of tweeters may not reflect the population as a whole. In addition, the sentiment analysis could only be performed on English words due to the available lexicons. Geospatial analysis was limited as location data was only available for a small percentage of tweets. The relatively high number of tweets located in North America and Europe may indicate that this data should not be used as the sole tool when assessing the spread of Zika virus.

Despite this, with such a large sample size there is a wide variety of sentiments and opinions available to

delve into. The ease of obtaining tweets and the near real-time data available make it an exciting arena for further development in public health awareness and response.

References

AFINN. 2011. Accessed May 10, 2016 via http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010

Centers for Disease Control and Prevention. May 12, 2016. "All Countries and Territories with Active Zika Virus Transmission." Accessed May 15, 2016 via http://www.cdc.gov/zika/geo/active-countries.html

Julia Slige. April 29, 2016. "The Life-Changing Magic of Tidying Text." Accessed April 29, 2016 via http://juliasilge.com/blog/Life-Changing-Magic/