**Language metrics as a measure of global reach and local impact: The case of papers about Zika on Facebook and Twitter**

*Germana Barata, visiting scholar at Simon Fraser University (SFU) and researcher at the State University of Campinas (Unicamp) email: gfbarata@sfu.ca*

*Juan PabloAlperin, Assistant Professor at Simon Fraser University email: juan@alperin.ca*

There are some seemingly arbitrary decisions about what metrics are derived from each platform—all retweets on Twitter count, but re-shares on Facebook only count if they are done to another public group; quote tweets, a type of reply, count, but Facebook comments never do. These choices, likely made out of convenience, limit the value we can derive from altmetrics. For altmetrics to be of value, we need to look beyond what is currently being counted, and begin to identify other aspects of social media activity that can serve as useful indicators. What is useful, of course, depends on what it is we want the indicators for. This paper focuses in identifying indicators that could serve to signal global reach, local impact, and national relevant. Therefore, we turned to the use of language, an aspect not currently reported on by Altmetric providers.

We focused on activity around the Zika outbreak in Brazil in the beginning of 2016, believing it to be a great case study to track how scientific information was shared on social media at a moment when accurate information was crucial to boost scientific research, inform society, and lead public health authorities' actions. Our analysis investigated differences in language used on social media posts about Zika because Zika was both nationally and internationally relevant, and much Zika-related research is conducted in collaboration with non-English speaking countries, like Brazil, the second more prolific country on papers publication.

## Methodology

We therefore conducted an analysis of articles tracked by Altmetric about Zika LLC that contained the keyword "Zika" in the title and had been shared at least once on Facebook or Twitter during the outbreak's peak, between January to June 2016 – period when research efforts, reported cases, and public interest were at their highest.

We identified 844 documents, from which we have excluded 126 that were published in venues not consider places sources of original research, which were a non-traditional scholarly output (e.g., a podcast), or not actually about Zika. This resulted in a sample of 718 documents, including research articles, editorials and news items published in scholarly journals, reports from Centre of Disease Control and Prevention-CDCs, and preprints from repositories such as arXiv and biorXiv.

We subsequently extracted the text of the social media posts and removed URLs, hashtags, and @mentions using the python module twitter-preprocessor before running the python module langdetect to identify the language of the post.

We also manually coded the language of publication and the country of affiliation of each of the authors to correlate these data with the language used on posts and tweets.

## Results

The majority of documents were published exclusively in English (90.25%), with an additional 1.25% in English as well as a second language. While Portuguese was the second most popular language with a total of only 10 papers (1.4%), only 4 were exclusively in Portuguese.

Among authors, most documents (50.7%) are affiliated uniquely to English-speaking countries (ESC), although over a third (35.2%) are authored exclusively by those from non-ESC (the remaining 14.1% were co-authored by individuals from both an English speaking and a non-English speaking country). Authors from Brazil were present on 17.0% of all documents.

Even though we see a prevalence of English in the scholarly documents, the social media shows a greater diversity of languages. While English was almost exclusively used on Twitter (89.8% of posts), Facebook shows greater linguistic diversity (24.1% of posts were not in English, with Portuguese and Spanish making up 13.6% of posts).

By comparing the authors' country affiliations with the language of posts, we see that authors affiliated with institutions from non-ESC have their work shared more on non-English posts, while the contrary is also true. On Twitter, papers have a bigger chance to be shared in languages other than English when there is at least one author from non-ESC. In some cases, we also see a local-language effect, with increased proportions of posts in the mother tongue of the author's country affiliation. Documents with authors from Brazil, for instance, were shared more in Portuguese than other languages (after English).

**Discussion and Conclusion**

Although language analysis of posts shows that scholarly documents are more often shared in English on both social media, we see significant differences in the languages used to discuss research on both platforms. Facebook has a much higher percentage (24.1%) on non-English languages than on Twitter (10.2%), and we see a significant difference depending on whether the document being discussed are from English-speaking countries or not, and again, we see differences depending on the author's country of affiliation. In the case of Brazil, the country most affected by Zika, we also see a local-language effect that is especially pronounced on Facebook.

All of these differences suggest language use is a promising avenue for exploration of new indicators that may signal global reach, national relevance, and local impact. Our results suggest that Facebook is used to reach more targeted groups (that use non-English languages) while Twitter seems to be used as a global communication tool.

More generally, the broad differences in language use is a strong indication that both platforms are used in very different ways. As a result, it seems that more attention needs to be given to the relevance of Facebook as a media that can be used to spread scientific content and news at a national level. The Zika case also makes it clear that journals should consider a multi-language strategy that targets each social media platform differently to encourage wider dissemination of relevant information.

Finally, although language metrics can help to highlight local and global communication, the relatively small amount of Facebook activity (despite the much larger userbase than Twitter) indicates there are still limitations to the way that Altmetric tracks activity on Facebook. By being limited to posts on public pages, Altmetric undercounts sharing of research between users. Given that Facebook is used in more nationally oriented groups, there needs to be further investigation to better understand and track the social use of scientific information on Facebook.