Statistical analysis of small twitter data collection to identify dengue outbreaks

C. D. G. Euzebio, S. Agy (in memoriam), C. Boldorini Jr., , L. P. Porto,
J. R. Alcarás, A. S. Martinez and E. E. S. Ruiz

Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto
Universidade de São Paulo (USP)
Av. Bandeirantes, 3900. Monte Alegre, 14040-901, Ribeirão Preto, SP – Brazil
[carlos.danilo, boldorini, jose.alcaras, asmartinez, evandro]@usp.br

Abstract. This study presents an algorithmic strategy to analyze a small set of social network information to monitor the dengue disease. Previous studies have achieved similar results based on large datasets of Twitter microblogs. In this study, we successfully map dengue cases using a small data collection of tweets from a medium-size city. A set of modules were constructed to collect, categorize, and display dengue-related tweets. We compared the collected tweets with real data from confirmed dengue cases. We showed a significant correlation between the number of confirmed dengue cases and the number of dengue-related tweets, even considering such a small dataset. The results of this approach may be relevant in public health policies

CCS Concepts: • **Computing methodologies** → **Information extraction**.

Keywords: Aedes aegypti, Dengue, Social Network, Public health

1. INTRODUCTION

Nowadays, social media is frequently used to share information during natural disasters, environmental tragedies and other health care concerns [Finch et al. 2016]. Twitter is an unique social media channel, characterized by a microblog platform, used to discuss an immense diversity of topics, including the share of health conditions. A cornerstone on health surveillance techniques using social media is the largely cited paper of Chew and Eysenbach [Chew and Eysenbach 2010], in which they suggest and evaluate a complementary "infoveillance" approach to measure public perceptions during the 2009 H1N1 pandemic. Social media has also been used for dengue surveillance [Gomide et al. 2011]. At the same time, Signorini and colleagues [Signorini et al. 2011] also use Twitter to track the 2009 H1N1 influenza pandemic activity levels in the USA. They show that estimates of this flu derived from Twitter chatter accurately follow reported disease levels. They also acknowledge the benefit of the sheer volume of messages, which can differ from large cities to modest size geographic regions. Since this study, health surveillance has also moved to a combination of data sources, as seen in Santillana and co-authors [Santillana et al. 2015]. They use Google searches, Twitter microblogs, hospital visit records and data from a participatory surveillance procedure to feed a machine learning-based system capable of providing estimates of influenza activity in the USA.

Tropical weather countries with long rainy periods have ideal conditions for the emergence and proliferation of mosquito-borne diseases, such as yellow fever, zika, chikungunya, leishmaniasis and

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. ASM acknowledges the support of CNPq (309851/2018-1), Instituto de Ciência e Tecnologia de Sistemas Complexos (INCT–SC/CNPq) e ao Núcleo de Apoio da Física Médica (NAP–FisMed) USP. JRA acknowledges the support of FAPESP (2018/21694-4)

dengue. The outbursts of these diseases, especially during summer periods when rains are more often, worry sanitary agencies and health-care systems. This same research group has also tracked the Zika virus during the 2015–2016 Latin American outbreak [McGough et al. 2017]. The spread of tropical viruses, such as Zika, in the southern hemisphere has been well established [Zhang et al. 2017]. People's reaction to the Zika outbreak was also analyzed using social media (Twitter) [Fu et al. 2016].

Brazil was also in focus for surveillance during the preparations for the 2016 Brazil Olympic Games [Petersen et al. 2016]. The Brazilian research team led by Marques-Toledo [de Almeida Marques-Toledo et al. 2017] also took advantage of the large number of texts in social media mentioning a disease and its correlation with physician visits by patients to estimate disease activity. They demonstrated the potential of modeling dengue estimation and forecast from tweets, in comparison with other available web-based data such as Google Trends and Wikipedia access logs. Recently, an infomediology study investigated the rapidly spreading of the SARS-CoV-2 (severe acute respiratory coronavirus 2) in South Korea at the end of February 2020 following its initial outbreak in China [Park et al. 2020]. They studied how COVID-19-related issues have circulated on Twitter through network analysis.

Although the majority of the studies regarding disease tracking work at the level of large territory monitoring, we are not aware in the literature about the effectiveness of health monitoring in small areas using social media. On one hand, Mousset and his team have analyzed the spatio-temporal peculiarities of small-scale events [Mousset et al. 2018], but they have not considered any health issue on it. On the other hand, Wang and colleagues [Wang et al. 2020] used a large volume of geotagged Twitter streaming data to predict trends of flu cases based on a specific partial differential equation.

For the past decade, the city of Ribeirão Preto, located in the state of São Paulo, in the southeastern part of Brazil, has had health related issues related to dengue outbursts during the summer. In this paper, we show that Twitter microblogs content can be used to track dengue disease in a relatively small geographic region based only on the word mentions to the mosquito and the disease itself. Even for a small data collection, we show that a correlation analysis based on the count itself and difference per unit time interval of word mentions in time (derivative) compared to officially confirmed cases present significant correlation coefficients.

This paper is organized as follows: Section 2 provides the technical details on the methods used for data mining on dengue-infected individuals using Twitter, specifically discussing the database and the analyzer built to categorize the tweets; Section 3 presents a discussion for Ribeirão Preto using data from 2015 to 2017, comparing the real infected patients with the obtained data; Section 4 concludes this study, showing a strong correlation between the obtained Twitter data and the confirmed cases, which motivates validation of this technique and its propagation to other diseases and locations on future studies.

2. MATERIALS AND METHODS

The Secretary of Health of Ribeirão Preto releases data concerning suspected and confirmed cases in the city. This data set is small when compared to the state or national scale, nevertheless they present high correlation as shown in the counts and difference of the counts. See Fig. 1. Counts of cases and their difference is the basic measure we use along this study. Only a small fraction of the population of Ribeirão Preto makes use of a social network. In social networks, particularly Twitter, we are interested in the cases that consist of a few words related to dengue disease during a time interval, making the set even smaller. The suspected cases of the released data plays a similar role to the tweet counts with the specific feature. Thus, these cases are compared to the confirmed one, serving a basis of comparison for a much smaller data set. As we show, the relatively small number of tweet counts is not a flaw to infer a possible outbreak to public health measures to be taken beforehand.

Since Ribeirão Preto is a relatively large, but very country oriented city (differing from São Paulo

or Rio de Janeiro, for example, which are much larger and much more urbanized cities). A region identification of the outbursts would be difficult by the tweets collected, especially considering the fact that Twitter users are mostly located on regions with higher economical statuses and, thus, not reflecting an accurate picture of the population profile of some neighborhoods.

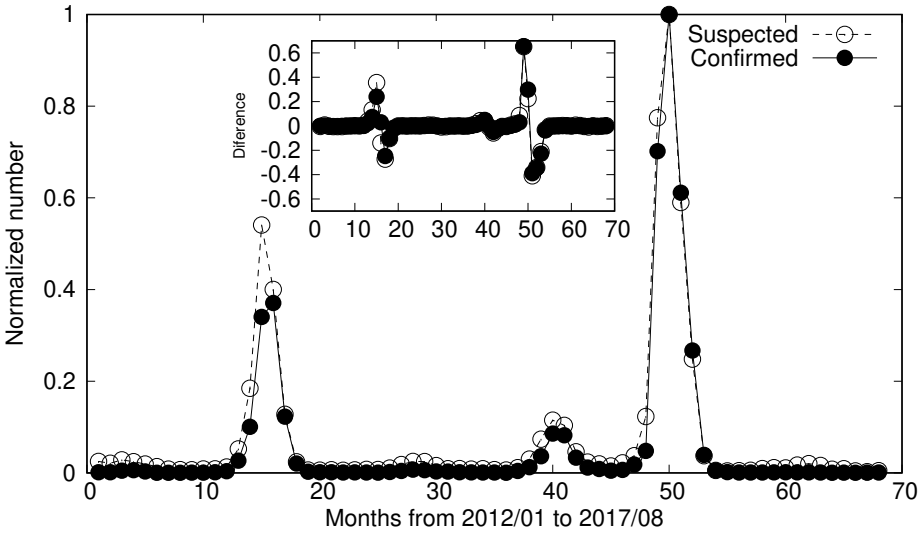

Fig. 1. Epidemiological data on dengue of Ribeirão Preto Health Secretary for suspected and confirmed cases. From 12/2012 to 8/2017, there have been registered 108,414 suspected dengue cases and 53,702 have been confirmed. The plot depicts the comparison of suspected and confirmed cases with data being normalized by their maximum values, 21,227 for the suspected and 13,319 for the confirmed cases. Correlation $r = 0.99$, 95% confidence interval $[0.98, 0.99]$. **Inset:** Difference of successive points (derivative) of data of suspected and confirmed cases. The correlation is $r = 0.97$, with 95% confidence interval $[0.95, 0.98]$.

Since 2014, the Laboratory of Complex Computational Systems (LSCC) of the Computing and Mathematics Department of the Faculty of Philosophy, Sciences and Letters of Ribeirão Preto (University of São Paulo – USP) automatically collects and process tweets on topics related to health issues. The terms used on its retrieval system include (but are not limited to) H1N1, Dengue and Zika. For that reason, this database was chosen for this study. This retrieval and processing system has two main modules: the tweet collector and the analysis tool, as illustrated in Fig. 2.

2.1 Collected tweets database

The LSCC system automatically collects full tweets, including their geographic location (latitude and longitude), if available. For this study, we retrieve only the tweets restricted to the geographic coordinates of Ribeirão Preto, using the *Tweepy*[1] library.

Only tweets related to dengue were incorporated to the database. The subject of the tweet is determined by the presence of some terms in the tweet content. Machado [Machado et al. 2017] proposed a combination of terms that can indicate if a tweet refers to dengue, these are: "dengue", "mosquito (da) dengue", "combate (da) dengue" and "casos (de) dengue" (in English, respectively, "dengue", "dengue mosquito", "fighting dengue" and "dengue cases"). Based on this reference, we

[1] `https://www.tweepy.org/`

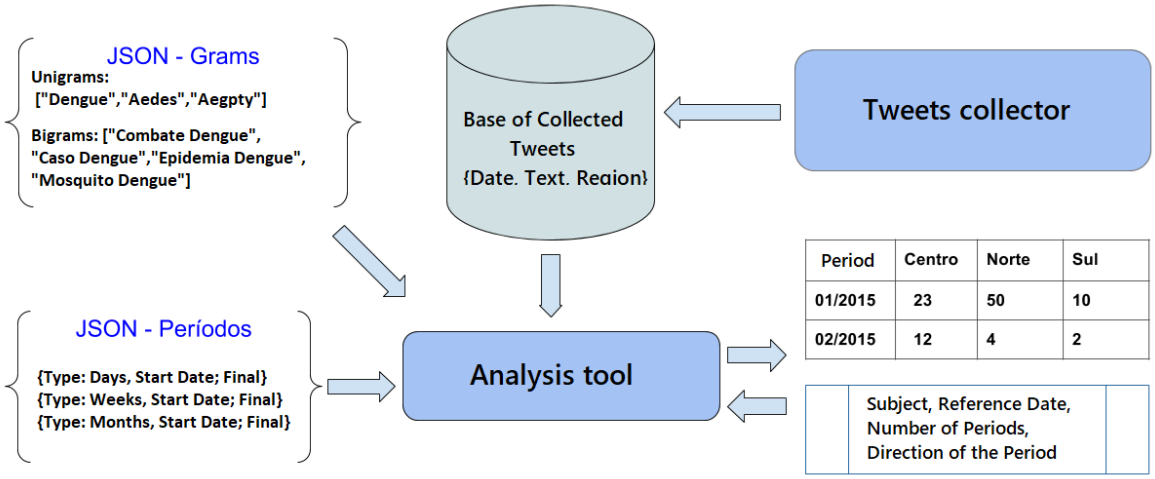

Fig. 2. The system for collecting, storing and analyzing tweets. Unigrams and/or bigrams related to dengue and its symptoms are categorized according to their region and time. From the tweets automatically collected in a given time interval, the ones from Ribeirão Preto are selected for further analysis. Out of these tweets, the ones containing two of the following words "dengue", "febre" (fever) or "aedes" are selected to have joint distribution of dates and location (region of the city). From the tweets automatically collected in a given time interval, the ones from Ribeirão Preto are selected for further analysis. Out of these tweets, the ones containing two of the following words "dengue", "febre" (fever) or "aedes" are selected to have joint distribution of dates and location (region of the city) as the table on the right hand side. In our studies we have added up the columns referring to the region of the city considering only the date (marginal) distribution. Python 3.5.2 has been used as the programming language.

defined a heuristic to identify the subject treated in the text of a tweet. A tweet was defined to be related to "dengue" if

—Two of the (unigrams) terms from the list "dengue", "febre" (fever) and "aedes" appear in the tweet text, but not necessarily contiguously, or;

—One of the (bigrams) terms from the list "mosquito dengue", "combate dengue" and "caso dengue" appears in the text of the tweet.

This categorization allowed us to further filter the tweets to focus only on dengue-related ones. This data retrieval system could be further analyzed and categorized according to other interests.

2.2 Tweet analyzer

This module allowed several kinds of comparison. It was built to provide useful data according to user entries: subject (as "dengue" for this study); geographic region of the city (general, central, east, and so on); and time interval of interest (for example, last 10 days, months between August 2018 and February 2019, etc.). Once these parameters were provided by the user, the following operations were conducted for each selected tweet: normalization of the tweet text through. The NLTK tool was used, with the corpus in Portuguese. Then there was the separation of the sentences of each tweet, and then of the words. We did the removal of special characters and changed from uppercase to lowercase letters. Repeated words were removed and there was no need for stemming. After analyzing all the tweets, the analyzer returns, as a result, a cvs file, in which the lines correspond to the analysis intervals, the columns correspond to the regions, and each cell content is the number of tweets with the subject identified as "dengue".

3. RESULTS AND DISCUSSION

The data from Ribeirão Preto Health Secretary[2] ranges from December 2012 to August 2017. During this period, they registered 108,414 suspected dengue cases, from which 53,702 have been confirmed. The plots for Fig. 1 show the values of the monthly counts divided by their respective maxima values (21,227 for suspected and 13,319 for confirmed cases). One is able to see an agreement of these data, which is corroborated by the correlation coefficient $r = 0.99$, and 95% confidence interval [0.98,1.0]. One may also see that a peak occurs between March and April of 2013, a smaller one in April–May of 2015 and the highest one in February of 2016. The two strongest peaks characterize the strong dengue outbreaks in Ribeirão Preto during the summers of 2013 and 2016.

From January 2015 to June 2017, our proposed system collected 8,080 tweets in Ribeirão Preto which used the unigrams described above, from which 384 have been considered to be tweet estimators for the dengue outbreak. There have been three distinct (different intensities) dengue outbreaks in Ribeirão Preto in this chosen interval. Here, our aim is to corroborate the idea that even with small data sets one is still able to obtain significant statistical results. In Fig. 3, one sees the agreement of these two data set: the small one with 384 tweets and the larger one, with 39,787 confirmed dengue cases from the Healthy Secretary of Ribeirão Preto. The tweets and confirmed cases data have been normalized by their highest value, in the time interval, 56 and 13,319, respectively. Both dengue outbreaks, which occurred in this time interval, have been detected by the tweet data sets. The correlation between the data set is $r = 0.88$, with 95% confidence interval $[0.76, 0.94]$. This confidence interval validates the significance of the calculated Pearson correlation (Kolmogorov-Smirnov test indicates the normality of the data). This way, we demonstrate the viability of using tweets as an indicator of dengue outbreaks, even though the number of tweets is not large, even if compared to the confirmed ones. The fluctuation of tweet data after the 20th month is due random circumstances, once weaker outbreaks have not been confirmed by the health authorities.

To validate this approach, we compare the data obtained by the difference of successive values (derivative) along the months. This method can be validated using the suspected and confirmed cases from the Healthy Secretary (Fig. 1). We show this data in the inset of Fig. 1, and one is able to identify the three dengue outbreaks that occurred during this time interval. The correlation between these data is $r = 0.97$, with 95% confidence interval [0.95,0.98]. The difference for the normalized number of tweets compared to the confirmed case, although one can visually infer the outbreaks as depicted in the inset of Fig. 3, the correlation between the data $r = 0.53$ is not significant, since the 95% confidence interval is $[0.20, 0.75]$. We believe the noise associated with the small amount of data is enough to decorrelate the data.

One way to check this belief is to consider the moving average, since it softens the noise. Indeed, when considering 2 and 3 points moving averages, one retrieves the correlations between the tweets and confirmed dengue cases. For 2 points moving average, the data are depicted in Fig. 4 and one is able to identify the two dengue outbreaks. The correlation between the tweets and confirmed data is $r = 0.82$ with 95% confidence interval $[0.64, 0, 91]$, showing its significance. For 3 points moving average, data are smoothed further, also with significant correlation $r = 0.86$ and 95% confidence interval $[0.71, 0.93]$, as shown in Fig. 5.

In this way, we have shown that considering a city that may produce a small amount of tweets (compared to the cases confirmed by its health authorities), these tweets can be used as indicators of dengue outbreaks. An automated system can be created to monitor the number of unigrams and bigrams per week (for instance) in the regions of the city. As these numbers increase, specific health actions can be taken in these dengue cases in suspicious regions.. This may be an inexpensive alternative to explore the disease propagation.

[2] `https://www.ribeiraopreto.sp.gov.br/ssaude/pdf/boletim_dengue.pdf`

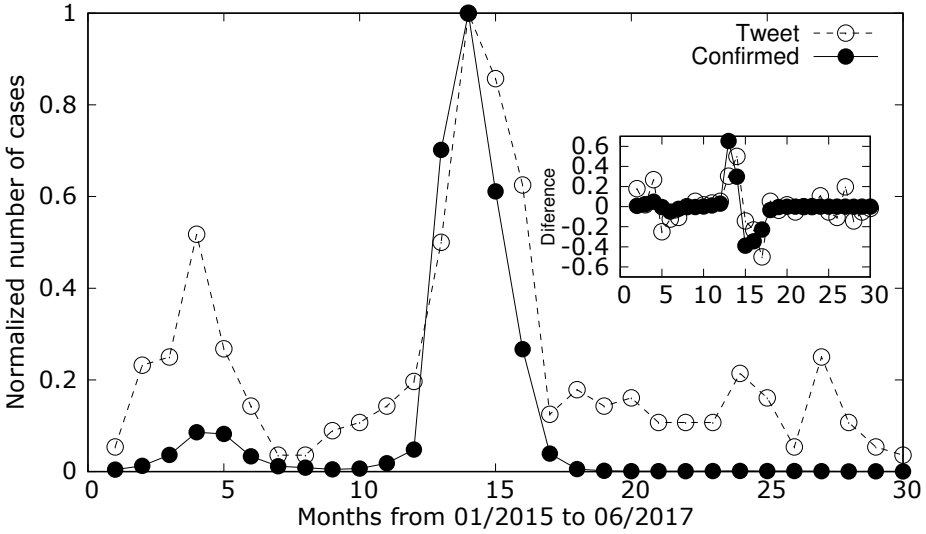

Fig. 3. Comparison between 384 tweets (out of 8,080) from a database of dengue related tweets and confirmed cases of dengue from Ribeirão Preto Health Secretary. The plot depicts the comparison of suspected and confirmed cases with curves being normalized by their maximum values, 56 for the tweets and 13,319 for the confirmed cases. Correlation $r = 0.88$, 95% confidence interval $[0.76, 0.94]$. **Inset:** Difference of successive points (derivative) of tweet and confirmed data with correlation coefficient $r = 0.53$, the 95% confidence interval is $[0.20, 0.75]$.

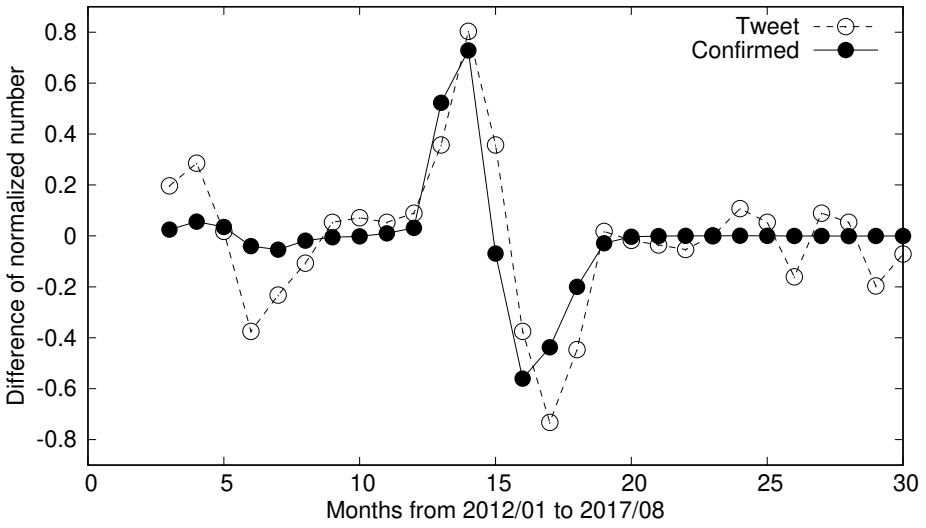

Fig. 4. Moving average (2 points) of difference of successive points (derivative) of data of inset of Fig. 1. The correlation coefficient is $r = 0.82$, with 95% confidence interval $[0.64, 0.91]$.

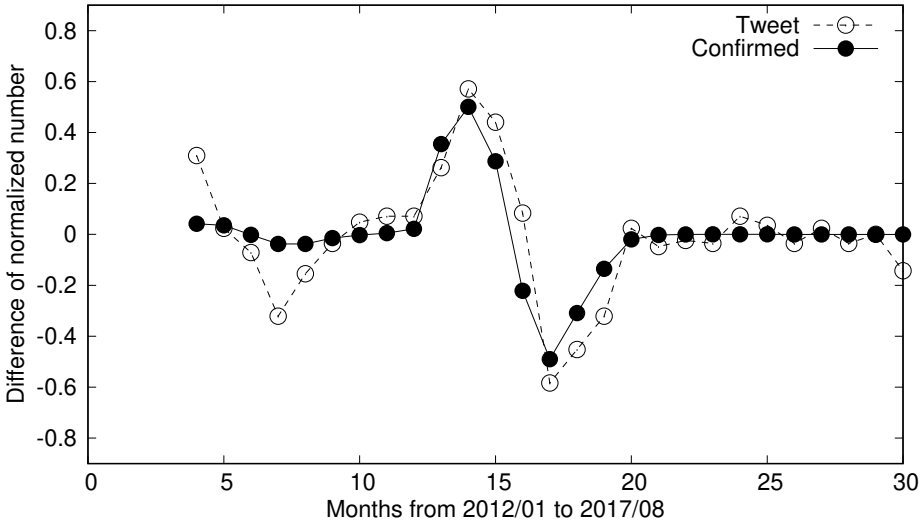

Fig. 5. Moving average (3 points) of difference of successive points (derivative) of data of inset of Fig. 1. Correlation $r = 0.86$, 95% confidence interval $[0.71, 0.93]$.

4. CONCLUSION

Here, we propose an algorithmic strategy to analyze social network information in order to provide insights on possible outbreaks of dengue in a restricted region. We considered Twitter collected in Ribeirão Preto and used a set of modules previously developed to collect, categorize and display data on dengue related tweets and analyzed this information. We compared the collected tweet data with real data from confirmed dengue cases from the Health Secretary of the city. Our analysis shows the existence of a significant correlation between the number of confirmed dengue cases and the number of dengue related tweets. This means that tweets can be used to infer dengue outbreaks, even if the number of tweets is a much smaller fraction of the total possible cases. In turn, the number of cases is localized to a city, which is the novelty of our study. The significance of our results for small data sets encourages the use of the proposed methodology could be used to efficiently identify sick patients and take preventive measures, in smaller cities with even smaller samples. The study of the minimal sample size to gather significant results is one of our perspectives to future studies.

REFERENCES

Chew, C. and Eysenbach, G. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. *PLOS ONE* 5 (11): e14118, 2010.

de Almeida Marques-Toledo, C., Degener, C. M., Vinhal, L., Coelho, G., Meira, W., Codeço, C. T., and Teixeira, M. M. Dengue prediction by the web: Tweets are a useful tool for estimating and forecasting dengue at country and city level. *PLOS Neglected Tropical Diseases* 11 (7): e0005729, 2017.

Finch, K. C., Snook, K. R., Duke, C. H., Fu, K. W., Tse, Z. T. H., Adhikari, A., and Fung, I. C. H. Public health implications of social media use during natural disasters, environmental disasters, and other environmental concerns. *Natural Hazards* 83 (1): 729–760, 2016.

Fu, K.-W., Liang, H., Saroha, N., Tse, Z. T. H., Ip, P., and Fung, I. C.-H. How people react to Zika virus outbreaks on Twitter? A computational content analysis. *American Journal of Infection Control* 44 (12): 1700–1702, 2016.

Gomide, J., Veloso, A., Meira, W., Almeida, V., Benevenuto, F., Ferraz, F., and Teixeira, M. Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In *Proceedings of the 3rd International Web Science Conference*. WebSci '11. Association for Computing Machinery, New York, NY, USA, 2011.

Machado, M., Temporal, J. C., Pardo, T. A., and Ruiz, E. E. Mineração de tópicos e aspectos em microblogs sobre dengue, chikungunya, zika e microcefalia. In *Anais Principais do XVII Workshop de Informática Médica*. SBC, Porto Alegre, RS, Brasil, 2017.

McGough, S. F., Brownstein, J. S., Hawkins, J. B., and Santillana, M. 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, 2017.

Mousset, P., Pitarch, Y., and Tamine, L. Studying the Spatio-Temporal Dynamics of Small-Scale Events in Twitter. In *Proceedings of the 29th on Hypertext and Social Media*. pp. 73–81, 2018.

Park, H. W., Park, S., and Chong, M. Conversations and Medical News Frames on Twitter: Infodemiological Study on COVID-19 in South Korea. *Journal of Medical Internet Research* 22 (5): e18897, May, 2020.

Petersen, E., Wilson, M. E., Touch, S., McCloskey, B., Mwaba, P., Bates, M., Dar, O., Mattes, F., Kidd, M., Ippolito, G., et al. Rapid spread of Zika virus in the Americas-implications for public health preparedness for mass gatherings at the 2016 Brazil Olympic Games. *International Journal of Infectious Diseases* vol. 44, pp. 11–15, 2016.

Santillana, M., Nguyen, A. T., Dredze, M., Paul, M. J., Nsoesie, E. O., and Brownstein, J. S. Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance. *PLOS Computational Biology* 11 (10): 1–15, 10, 2015.

Signorini, A., Segre, A. M., and Polgreen, P. M. The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PLOS ONE* 6 (5): e19467, 2011.

Wang, Y., Xu, K., Kang, Y., Wang, H., Wang, F., and Avram, A. Regional influenza prediction with sampling twitter data and PDE model. *International Journal of Environmental Research and Public Health* 17 (3): 678, 2020.

Zhang, Q., Sun, K., Chinazzi, M., y Piontti, A. P., Dean, N. E., Rojas, D. P., Merler, S., Mistry, D., Poletti, P., Rossi, L., et al. Spread of Zika virus in the Americas. *Proceedings of the National Academy of Sciences* 114 (22): E4334–E4343, 2017.