COMPUTING
SCIENCE

Title: VazaDengue: An Information System for Preventing and Combating Mosquito-Borne Diseases with Social Networks

Names: Leonardo Sousa - is a PhD student with the Informatics Department, PUC Rio, Brazil
Rafael de Mello - is a research associate with the Informatics Department, PUC Rio, Brazil
Diego Cedrim – a PhD student with the Informatics Department, PUC Rio, Brazil
Alessandro Garcia - is an Associate Professor with the Informatics Department, PUC-Rio, Brazil. He leads the *Opus Research Group*
Paolo Missier is a Reader in Large Scale Inform Management, School of Computing Science, Newcastle University, UK
Anderson Uchoa - is a PhD student with the Informatics Department, PUC Rio, Brazil
Anderson Oliveira - is a PhD student with the Informatics Department, PUC Rio, Brazil
Alexander Romanovsky – Alexander (Sascha) Romanovsky is a Professor in the Centre for Software Reliability. He is the leader of the Secure and Resilient Systems Group at the School of Computing Science, Newcastle University, UK

TECHNICAL REPORT SERIES

No. CS-TR- 1511-2017

TECHNICAL REPORT SERIES

No. CS-TR- 1511 Date 2017

Title: VazaDengue: An Information System for Preventing and Combating Mosquito-Borne Diseases with Social Networks

Authors: Leonardo Sousa, Rafael de Mello, Diego Cedrim, Alessandro Garcia, Paolo Missier, Anderson Uchoa, Anderson Oliveira, Alexander Romanovsky

Abstract:

Dengue is a disease transmitted by the Aedes Aegypti mosquito, which also transmits the Zika virus and Chikungunya. Unfortunately, the population of different countries has been suffering from the diseases transmitted by the mosquito. The communities can play an important role in combatting and preventing the mosquito-borne diseases. However, due to the limited engagement of the population, new methods need to be used to strengthen the mosquito surveillance. VazaDengue is one of these solutions that provides services that stand out from the others solutions. Generally speaking VazaDengue is a system that offers the users a platform for preventing and combating mosquito-borne diseases. The system relies on social actions of reporting mosquito breeding sites and dengue cases, in which the reports are made available to the citizens and health agencies. In addition, the system monitors social media network Twitter to enrich the information provided. It processes the natural language text from the network to classify the tweets according to a set of the predefined categories. After the classification, the relevant tweets are provided to the users as reports.

In this paper, we describe the VazaDengue features including its ability to harvest and classify tweets. Since the VazaDengue system aims at providing a dynamic and efficient environment to support rapid interventions of health agents, we present here two studies evaluating the potential contributions of the classified tweets in preventing and combating mosquito-borne diseases. The first evaluation uses a survey conducted by the Brazilian community of health agents. The goal is to evaluate the relevance of the classified tweets. The second study compares the official reports of the 2015-2016 epidemic waves in Brazil with the concentration of mosquito-related tweets found by VazaDengue. The goal is to verify if the concentration of tweets can be used for monitoring big cities. The results of these two evaluations are encouraging. We have found that the health agents tend to agree with the relevance of the classified tweets. Moreover, the concentration of tweets is likely to be effective

for monitoring big cities. The results of these evaluations are helping us to further improve the VazaDengue system to make it more useful for combating and preventing the mosquito-borne diseases.
VazaDengue: An Information System for Preventing and Combating Mosquito-Borne Diseases with Social Networks:

Keywords: dengue; mosquito; social media; surveillance; tweets

Printed and published by Newcastle University,
Computing Science, Claremont Tower, Claremont Road,
Newcastle upon Tyne, NE1 7RU, England.

Bibliographical details

Title and Authors

NEWCASTLE UNIVERSITY
Computing Science. Technical Report Series. CS-TR- 1511

Abstract:
Dengue is a disease transmitted by the Aedes Aegypti mosquito, which also transmits the Zika virus and Chikungunya. Unfortunately, the population of different countries has been suffering from the diseases transmitted by the mosquito. The communities can play an important role in combatting and preventing the mosquito-borne diseases. However, due to the limited engagement of the population, new methods need to be used to strengthen the mosquito surveillance. VazaDengue is one of these solutions that provides services that stand out from the others solutions. Generally speaking VazaDengue is a system that offers the users a platform for preventing and combating mosquito-borne diseases. The system relies on social actions of reporting mosquito breeding sites and dengue cases, in which the reports are made available to the citizens and health agencies. In addition, the system monitors social media network Twitter to enrich the information provided. It processes the natural language text from the network to classify the tweets according to a set of the predefined categories. After the classification, the relevant tweets are provided to the users as reports.
In this paper, we describe the VazaDengue features including its ability to harvest and classify tweets. Since the VazaDengue system aims at providing a dynamic and efficient environment to support rapid interventions of health agents, we present here two studies evaluating the potential contributions of the classified tweets in preventing and combating mosquito-borne diseases. The first evaluation uses a survey conducted by the Brazilian community of health agents. The goal is to evaluate the relevance of the classified tweets. The second study compares the official reports of the 2015-2016 epidemic waves in Brazil with the concentration of mosquito-related tweets found by VazaDengue. The goal is to verify if the concentration of tweets can be used for monitoring big cities. The results of these two evaluations are encouraging. We have found that the health agents tend to agree with the relevance of the classified tweets. Moreover, the concentration of tweets is likely to be effective for monitoring big cities. The results of these evaluations are helping us to further improve the VazaDengue system to make it more useful for combating and preventing the mosquito-borne diseases.
VazaDengue: An Information System for Preventing and Combating Mosquito-Borne Diseases with Social Networks:

About the authors:

Leonardo Sousa - is a PhD student with the Informatics Department, PUC Rio, Brazil.

Rafael de Mello - is a research associate with the Informatics Department, PUC Rio, Brazil.

Diego Cedrim – a PhD student with the Informatics Department, PUC Rio, Brazil.

Alessandro Garcia - is an Associate Professor with the Informatics Department, PUC-Rio, Brazil. He leads the *Opus Research Group*.

Paolo Missier – Paolo Missier is a Reader in Large Scale Inform Management, School of Computing Science, Newcastle University, UK.

Anderson Uchoa - is a PhD student with the Informatics Department, PUC Rio, Brazil.

Anderson Oliveira - is a PhD student with the Informatics Department, PUC Rio, Brazil.

Alexander Romanovsky – Alexander (Sascha) Romanovsky is a Professor in the Centre for Software Reliability. He is the leader of the Secure and Resilient Systems Group at the School of Computing Science, Newcastle University, UK

Suggested keywords:

Keywords: dengue; mosquito; social media; surveillance; tweets

RESEARCH

VazaDengue: An Information System for Preventing and Combating Mosquito-Borne Diseases with Social Networks

Leonardo Sousa[1*†], Rafael de Mello[1†], Diego Cedrim[1], Alessandro Garcia[1], Paolo Missier[2], Anderson Uchôa[1], Anderson Oliveira[1] and Alexander Romanovsky[2]

*Correspondence:
lsousa@inf.puc-rio.br
rmaiani@inf.puc-rio.br
dcgrego@inf.puc-rio.br
afgarcia@inf.puc-rio.br
paolo.missier@newcastle.ac.uk
auchoa@inf.puc-rio.br
aoliveira@inf.puc-rio.br alexander.romanovsky@newcastle.ac.uk
[1]Department of Informatics, PUC-Rio, Rio de Janeiro, Brazil
Full list of author information is available at the end of the article
†Equal contributor

Abstract

Dengue is a disease transmitted by the Aedes Aegypti mosquito, which also transmits the Zika virus and Chikungunya. Unfortunately, the population of different countries has been suffering from the diseases transmitted by the mosquito. The communities can play an important role in combatting and preventing the mosquito-borne diseases. However, due to the limited engagement of the population, new methods need to be used to strengthen the mosquito surveillance. VazaDengue is one of these solutions that provides services that stand out from the others solutions. Generally speaking VazaDengue is a system that offers the users a platform for preventing and combating mosquito-borne diseases. The system is relies on social actions of reporting mosquito breeding sites and dengue cases, in which the reports are made available to the citizens and health agencies. In addition, the system monitors social media network Twitter to enrich the information provided. It processes the natural language text from the network to classify the tweets according to a set of the predefined categories. After the classification, the relevant tweets are provided to the users as reports. In this paper, we describe the VazaDengue features including its ability to harvest and classify tweets. Since the VazaDengue system aims at providing a dynamic and efficient environment to support rapid interventions of health agents, we present here two studies evaluating the potential contributions of the classified tweets in preventing and combating mosquito-borne diseases. The first evaluation uses a survey conducted the Brazilian community of health agents. The goal is to evaluate the relevance of the classified tweets. The second study compares the official reports of the 2015-2016 epidemic waves in Brazil with the concentration of mosquito-related tweets found by VazaDengue. The goal is to verify if the concentration of tweets can be used for monitoring big cities. The results of these two evaluations are encouraging. We have found that the health agents tend to agree with the relevance of the classified tweets. Moreover, the concentration of tweets is likely to be effective for monitoring big cities. The results of these evaluations are helping us to further improve the VazaDengue system to make it more useful for combating and preventing the mosquito-borne diseases.

VazaDengue: An Information System for Preventing and Combating Mosquito-Borne Diseases with Social Networks:

Keywords: dengue; mosquito; social media; surveillance; tweets

1 Introduction

Dengue is a tropical febrile illness that affects individuals of all ages. The disease is not transmitted directly from person-to-person but by the bite of a mosquito (typically the Aedes aegypti) infected with one of the four Dengue virus serotypes. Unfortunately, there is no vaccine or specific medicine to treat Dengue. To make the situation worse, its more severe version, known as dengue hemorrhagic fever, is a potentially lethal complication, affecting mainly children [1]. In spite of the risks involved, Dengue is in the list of the World Health Organization's (WHO) Neglected Tropical Diseases [1].

In the last decades, the population of different endemic countries has been continually affected by Dengue outbreaks. In this scenario, Brazil is historically one of the countries with higher incidence of the disease [1]. One of the reasons is that the country offers an appropriate environment for the mosquito and the disease proliferation. Dengue rapidly flourishes in poor urban areas, suburbs and the countryside. However, it also affects more affluent neighborhoods in tropical and subtropical countries. The burden of dengue is considered higher among the poorest citizens who grow up in communities with an inadequate water supply and without a solid waste infrastructure, and where conditions are most favorable for the proliferation of the mosquito. The immature stages of the mosquito can be found in water-filled habitats, mostly in artificial containers. A notable example includes tires containing rainwater, in which female mosquitoes may deposit their eggs. Other examples include discarded food and beverage containers and buildings under construction.

Identification of mosquito vector breeding sites is fundamental to prevent Dengue and other mosquito-borne diseases, such as Zika [2] and Chikungunya [3]. In this sense, community participation is a key factor in preventing and controlling arboviruses, i.e. viruses that are transmitted by arthropod vectors. Citizens should follow and may help authorities on monitoring the correct application of prevention practices. For instance, citizens may report the incidence of mosquito breeding sites in their neighborhoods. The concern on diagnosing possible cases of the disease in time is also important, not only to the treatment but also for generating statistical data regarding the incidence of the mosquito-borne diseases. However, public health researchers in Brazil have been reported the unsatisfactory communities' responsiveness to the prevention programs in Brazil, maintaining practices that contribute to proliferating illnesses transmitted by mosquitoes. This is even most worrisome in the context of poor communities, where settlement and sanitation features contribute to such proliferation.

Another important issue is that citizens are typically unable to follow the community health agents' work, who end up feeling vulnerable despite the prevention efforts. For instance, when a citizen reports a potential mosquito breeding site by using a common channel such as a telephone number, he and the citizens of the same community stay unaware of the actions taken by the health agents addressing the issues reported. The Brazilian Health System (SUS) requires that each confirmed case of dengue should be reported by health professionals, information about the incidence of Dengue cases may take months be processed and published for the population. In addition, we did not identify in the SUS any features to associate the locations of the Aedes mosquito focuses detected by health agents with the locations of confirmed cases of diseases transmitted by the mosquito.

On the other hand, the concern with mosquito-borne diseases and its potential consequences led citizens around the globe to share relevant online information related to the disease in social networks, such as Twitter and Instagram. Such information typically includes reporting (suspected) cases of the disease and denouncing locations with the incidence of mosquito focus. Besides, news related to the illness is also shared by the users. In this sense, it is worth mentioning the increasingly facilitated access to smartphones and the Internet in the last years, leading to the dissemination of such technologies even in poor communities, especially those located in urban zones. Thus, the current situation calls for a reflection on how public health policies could explore the collective knowledge regarding Dengue, intentionally generated or not by each citizen, towards increasing the Dengue prevention and combat.

Automated solutions for supporting the detection of cases of mosquito-borne diseases and outbreaks typically require the direct contribution of the citizens (Section 2.1) through filling up forms [4, 5]. Consequently, the coverage of the intended support becomes restricted to the willing of the applications' users on providing eventual contributions. In this sense, it is important to note that individuals largely share content through social networks. Therefore, mining and classification of the geolocated content from social networks such as Twitter represent an interesting alternative for supporting an earlier identification of potential epidemic waves of mosquito-borne diseases. The mining and classification of social network content have been used to support prevention and control activities related to natural phenomena [6] and prevention of crimes [7, 8]. However, we are not aware of any previous work exploring such technology in the context of mosquito-borne diseases.

In this context, we launched in 2015 an interactive platform named VazaDengue ("vaza" is a slang in Portuguese for telling somebody or something to disappear), which offers for the users a system for supporting the prevention and control of mosquito-borne diseases. Its main goals are: to contribute to detecting the potential development of Dengue, Zika, and Chikungunya in specific cities before the spreading of the epidemics, and; to identify useful geolocated content to detect mosquito breeding sites in certain communities. Two different sources are used to feed the system with the relevant mosquito-related content. The first one is the (traditional) direct report performed by the users. The second (and main) source is based on filtering and harvesting the content from social networks, including Twitter and Instagram. In the case of Twitter content, we went beyond by implementing an supervised algorithm for classifying filtered tweets in Portuguese. The relevant content from both sources is then published in dynamic maps in the VazaDengue system [9].

The first version of the supervised algorithm [10] classified the filtered content about mosquito-borne diseases in one of the following categories: *suspected cases of the disease*, *mosquito focus*, *news*, and *jokes*. The last category (jokes) was included due to the traditional use of terms such as "dengue" and "mosquito" for jokes in Brazil. Indeed, the major challenge of the classification algorithm is to distinguish the relevant content from noise. After 12 months from its launching, we observed a significant change in the epidemic and tweeting scenario, especially due to the 2016' Zika outbreak in Brazil. Once Zika was not an issue in Brazil before that, it

had drastically impacted the social network content. Consequently, the number of noisy tweets had considerably grown and the main terms had changed, impacting on the accuracy of the original classifier. This new and challenging scenario led us to evolve the classifier in 2016, resulting in a new version [11] working with a new set of content categories.

Publishing geolocated content in VazaDengue offers an opportunity to explore whether such content could be useful to support the work of different categories of health professionals on preventing and controlling mosquito-borne diseases. For instance, community health agents may benefit from such data to support the identification of mosquito breeding sites. Researchers may use classified data to investigate behaviors associated with the incidence of suspected cases, confirmed cases, and mosquito breeding sites. Medical doctors may follow the incidence of the reports in their working region to warn their patients. As part of our work, we conducted an empirical study in which community health agents evaluated a sample of tweets annotated as relevant by the new classifier. As a result, we could identify in more detail some patterns of tweets that such professionals tend to annotate as relevant and other patterns of tweets in major annotated by them as non-relevant. Such findings are helping us to improve the precision of the classification algorithm. Another evaluation was conducted through comparing the geographic distribution of tweets during the two more recently concluded epidemic cycles (2015 and 2016) and reports from the Brazilian Government regarding the geographic distribution of mosquito-borne diseases in these cycles. The results indicate that mining and classifying geolocated tweets can be useful to identify potentially critical cities.

This paper introduces the VazaDengue system and presents the research steps performed to develop and evolve the classifier. Therefore, the main contributions of the presented paper are the following:

- It introduces VazaDengue, a web platform (website + mobile application) that allows the visualization in large scale of relevant content regarding the prevention and combat of mosquito-borne diseases;
- It presents a successful repurposing and retraining of the classifier to track concept drift (from Dengue to Zika);
- It presents a qualitative assessment of the relevance of VazaDengue' mined content by community health agents;
- It presents a comparison between the VazaDengue published content with official reports of recent epidemic cycles.

Section 2 presents the background and related work. Section 3 describes the VazaDengue system architecture. Sections 4 and 5 present the first and the second version of the content classifier, respectively. Finally, Section 7 describes the evaluations of the proposed technology, discussing opportunities for improvement.

2 Background and Related Work

As previously mentioned, Brazil has an appropriate environment for the mosquito and the disease proliferation. Hence, the identification of mosquito vector breeding sites is fundamental to prevent Dengue and other mosquito-borne diseases. In this sense, community participation plays a key factor. Unfortunately, the Brazilian communities have not been involved in the combat of the mosquito neither they have

been involved in prevention programs. Due to low community adherence, a number of systems were created and made available to the public with the goal of supporting citizens either in the combat of the mosquito or adherence of prevention campaigns. The purpose of this section is to present an overview of solutions in the context of Dengue surveillance. Section 2.1 describes other information systems available in Brazil that support the prevention and combat of Dengue Fever, Chikungunya, and Zika virus. Section 2.2 introduces the area of mining content from social media.

2.1 Information Systems Supporting Dengue Prevention and Combat

There are a number of mobile applications and websites in Brazil supporting Dengue prevention and combat. Most of these services only provide information about the Dengue Fever and the Aedes mosquito. For example, the *UNA-SUS Dengue* [12] is an Android application (app for short) developed by the Federal University of Health Sciences of Porto Alegre (UFCSPA). Its main goal is to provide useful information for individuals infected with the Dengue Fever. Based on the patient characteristics (age, gender, weight, among others) and his symptoms, the system provides information about a aproppriate treatment. Based on such data, the application classifies the patient in a particular risk group and provides, for the patient needs of fluid replacement. Moreover, the UNA-SUS Dengue app also provides tips related to the treatment and prevention of dengue.

Dengue Brazil [13] is another app designed to provide information regarding the Dengue. Its primary goal is to provide information about dengue prevention actions, treatments, and news relevant to the disease. Informative videos and public health advertisements addressing the Dengue fever are also available. The app allows users to share news by e-mail. *Dengue Brazil* also lists other Internet sources of information providing information about Dengue prevention and combat.

Radar Dengue [4] is an mobile app developed by Unicesumar. Its main goal is to inform the population of Maringá city (Paraná State - Brazil) about mosquito breeding sites around the city. The users can use the app to reports the breeding sites, they also con include a picture in the report before sending it. Such content is used to update a map indicating potential outbreaks of dengue fever.

Among the information systems that support the dengue prevention and combat in Brazil, there are two systems that are similar to the VazaDengue. The fist one is the *Observatorio do Aedes Aegypti* [5]. It is a more comprehensive information system than the aforementioned systems. It was launched in May 2014 and is composed of an Android application and a web portal. The system was developed by Innovation Lab in Health (LAIS) of the University Hospital Onofre Lopes (UFRN), in partnership with the city and state administration. Through using Georeferenced location, the system allows citizens to denounce mosquito breeding sites and suspected cases of dengue, zika, and chikungunya. Public health agents can also use information provided by citizens to plan their prevention and combat activities. Despite of providing similar features to VazaDengue, the *Observatorio do Aedes Aegypti* does not explore social media as Twitter and Instagram.

The *InfoDengue* [14] is the second information system that is similar to the VazaDengue. It was developed in partnership between the Oswaldo Cruz Foundation, Getulio Vargas Foundation and the Health Department of the city of Rio

de Janeiro. The system is based on a preliminary study that the authors conducted using historical series from 2011 to 2014 (provided by the Federal University of Minas Gerais - UFMG) and data from January to December 2015. Based on the preliminary study, the InfoDengue captures climate time series, dengue case reporting and activity on Twitter at the beginning of each week. It uses this data to find indicators of dengue transmission for the states of Espírito Santo, Paraná, Rio de Janeiro and Minas Gerais. Then, it uses these indicators to classify the cities from these states into some categories of risk. Thus, the system is able to show a risk map to inform the public about the week's level of attention and the evolution of the disease incidence. A report is also sent automatically to the health agencies. Although this system is similar to VazaDengue regarding to the monitoring of Twitter, it is completely different from the VazaDengue. First, the system does not classify the tweets according to their content. Second, the InfoDengue was not developed to the citizens report mosquito breeding sites and diseases cases. Third, the system only covers few states instead of the entire county. Fourth, the system is based on probabilistic models to create a risk map at the beginning of each week Finally, the system does not have a mobile app.

2.2 Mining Content from Social Media

Our goal in developing the VazaDengue system is to provide a dynamic and efficient environment to support rapid interventions from health agents. Considering the already mentioned limitation of citizen's engagement in the direct contributions, we need to find new ways to acquire the relevant content from alternative sources, for instance, social media networks. In this context, Twitter[1] and Instagram[2] are natural choices due to their popularity – they have a broad coverage of active users posting content everyday, especially in Brazil. For instance, Twitter has more than 313 million of active users per month [15]. Facebook is another suitable social network for our context. Unfortunately, Facebook [3] does not provide free means to obtain social media data. On the other hand, Twitter and Instagram allow acquisition of content through the use of free APIs.

The Twitter Streaming API is a free API provided by Twitter that allows anyone to retrieve at most a 1 percent sample of all Twitter data by providing some filtering parameters. It means that, once the number of tweets matching the given parameters reaches 1 percent of all the Twitter tweets, Twitter will begin to sample the data returned to the user. The Twitter Streaming API has been used to support several types of research [16, 17, 18]. The Instagram-API is a free API provided by Instagram that allows anyone to retrieve data about users, relationships, media, comments, likes, and locations. The defined API terms are [19] that users own their media and that it is not allowed to use the Instagram API to crawl or store media without the express consent of the owner. Since, there are restrictions to retrieving large amounts of data in a short period of time, we concentrate our preliminary analyses in Twitter.

[1] www.twitter.com
[2] www.instagram.com
[3] www.facebook.com

Twitter has been used as a source of epidemic information, in which allows public health systems to perform real-time surveillance. For instance, Mampos and Cristianini [20] developed a monitoring tool for Twitter. The tool analyzed tweets in order to find statements of disease's symptoms in the tweets' content. The authors used these statements to generate statistical information about flu epidemic in the United Kingdom. The goal was to verify if they machine learning algorithm could measure the prevalence of diseases in a population. Using the tweets retrieved by their tool, they calculate the score for the diffusion of ILI (Influenza-like Illness) in various regions of the country. They compared their score with official data from the Health Protection Agency, and they obtained on average a statistically significant linear correlation greater than 95%. Similarly, Achrekar et al. [21] developed an architecture to monitors tweets with mention of flu indicators. Their goal was to track and predict the emergence and spread of an influenza epidemic in a population. They collected tweets from 2009 until 2010 and compared with data provided by the CDC (Center for Disease Control and Prevention). The authors found that the tweets were highly correlated with ILI activity with the CDC data. Based on this result, they build auto-regression models to predict a number of ILI cases in a population. They tested the models with the historic CDC data, and they realized that the Twitter data considerably improved the models' accuracy in predicting ILI cases.

Twitter has also been used for Dengue Surveillance. Gomide et al. [22] investigate if Dengue epidemic is reflected on Twitter. They proposed a methodology that is based on four dimensions: volume, location, time and content. The methodology allowed them to investigate to what extend the Twitter content can be used to support surveillance. First, the authors performed a sentimental analysis of the public perception in order to focus on tweets that expressed personal experience about the dengue disease. The analysis allowed them to remove irrelevant content. Then they compared the number of tweets posted from 2009 to 2011 with official statistics. They also constructed a correlated linear regression model for predicting the number of dengue cases using the proportion of tweets expressing personal experience. Their results indicate that the Twitter data can be used to predict, spatially and temporally, dengue epidemics by means of clustering.

Although these previous studies have focused on tracking epidemic information, they differ from VazaDengue due to their limited or insufficient solutions for rapid combat of epidemic waves. Firstly, these studies have relied on disease-related posts from previous epidemic waves. However, the epidemic waves change constantly due to different reasons e.g., changes on environment and ecological factors. Therefore, exploring disease-related posts from previous epidemic waves tends to be ineffective in each epidemic wave. Secondly, these previous work do not focus on identify mosquito breeding sites through the analysis of social media content.

The analysis of tweet content has been applied in other context as well. For example, Gerber et al. [7] investigated the use of spatiotemporally tagged tweets for crime prediction. The authors used linguistic analysis and statistical topic modeling to analyze tweets from Chicago City, Illinois. This allowed them to automatically identify relevant discussion topics, incorporating them into a crime prediction model. As a

result, it was observed that adding Twitter-based topics led to improving the performance of crime prediction in 19 of the 25 crime types analyzed. These results indicate that analyzing tweet content can help in enriching crime prediction models.

Similarly, Chen et al. [8] have also used Twitter to support the prediction of crimes. However, they have improved a crime prediction model by adding sentiment analysis mined from Twitter and weather predictors. According to them, weather factors, especially temperature, may influence the incidence of crimes. Based on such perspective, the authors built a logistic regression to predict crime in the Chicago area. The authors compared their prediction model with the actual theft incidents that occurred in Chicago, Illinois, between December 25, 2013 and January 30, 2014. The developed model was able to successfully predict future crime in each area of the city, surpassing the benchmark model used in the study.

Twitter can also be used for situation awareness, i.e., tweets can assist in providing processes and strategies for users who seek awareness in emergencies. In this context, Vieweg et al. [6] investigate two concurrent emergency events in North America via Twitter. During the two analyzed events, the authors identified features of information generated during emergencies. These features can be used to support software systems that employ data extraction strategies. Aware that Twitter can provide useful information to increase the disaster readiness of the general public, Zhu et al. [23] investigated the factors that affect Twitter users' retweet decision. Their goal was understanding these factors in order to optimize the communications of disaster messages. The authors identified factors that may have an impact on a user's decision to retweet a certain tweet.

3 The VazaDengue System

According to a representative of PAHO[4] in Brazil, we should take into account three basic premises to get into control of the dengue epidemic [24]. The first one is to contact affected communities. The second one is to encourage the population to identify Aedes Aegypti mosquito and eliminate them. Finally, representative highlights the importance of an active surveillance process. Given these premises, we have created the VazaDengue system.

VazaDengue is a system that offers for the users a platform for preventing and combating mosquito-borne diseases. Its main goal is to strengthen the entomological surveillance of the mosquito that transmits Dengue, Zika, and Chikunguny by providing a geolocated reports addressing the mosquito-borne disease. The system is based on social actions for reporting mosquito bleeding sites and dengue cases. Thus, it allows users to manipulate geolocated data obtained either from the system's users or from social media. These two different sources are used to feed the system with mosquito relevant data. The first source is the traditional one, in which the users directly report cases of mosquito breeding site or disease cases. The second (and main) source is based on filtering and harvesting the content from social media, including Twitter and Instagram. The data collected from these two sources are classified according to categories, and then they are published in dynamic maps in the VazaDengue system [9]. The users already send the data classified according

[4]The Pan American Health Organization: http://www.paho.org/hq/

to the categories. In the case of Twitter content, we use a supervised algorithm for classifying tweets in Portuguese.

We launched the VazaDengue system in 2015, and it includes two implementation versions: an implementation of the system as a web portal and an application for mobile devices – the Android app is available for downloading and the iOS version is under development. Both versions provide the same functionality, in which allows users to visualize geolocated data in dynamic maps as well as report occurrences of the mosquito breeding or cases of sick people. The VazaDengue system, its components, functionality and architecture are explained in the following subsections.

3.1 VazaDengue Functionality

VazaDengue system offers the users three main services, (1) allowing them to report mosquito breeding sites and dengue cases, (2) monitor social media for updated information and (3) visualize existing reports. The system combines the data of the two first services to ensure a real-time surveillance activity through dynamic maps, which is available to population and government agencies.

3.1.1 Reporting mosquito breeding sites and dengue cases

The first service offered by the VazaDengue is a communication channel to report mosquito breeding sites and dengue cases. Thus, the uses can act as health agents in order to notify occurrences related to Dengue, Zika or Chikungunya. The users can use this service to collect three types of reports:

- Mosquito Breeding Site: This type of report allows the user to send to the system a possible dengue mosquito breeding. Alongside with the report, the users can inform where the breeding site is located, and if the location comprises a public or a private area.
- Sick Person: This type of report allows the users to send cases of an individual who is sick due to the dengue mosquito. The users can specify between three types of diseases: Dengue, Zika or Chikungunya. Alongside with the report, the users can inform the age of the patient and if the health agents have visited the region where the patient lives.
- Illness Suspicion: This type of report allows the user to send the case of a person who is only suspected of Dengue fever, Zika or Chikungunya. Alongside with the report, the user can tell what symptoms the person is feeling. This type of report can be useful to health agents provide a first diagnostic.

All these three types of reports must have attached information about the user's location and the date he is reporting. The users can attach a photo optionally.

3.1.2 Monitoring social media

The VazaDengue also monitors social media as Twitter and Instagram. This comprises the second service offered by the system. In this service, tweets and Instagram posts that are related to mosquito content are retrieved and treated as a report.

In the case of tweets, we have a tweet classifier that processes natural language to classifier the tweets (Section 4). After the classification, tweets are provided to the users as reports. They are plotted on the map according to their classification. Yellow markers represent Mosquito Focus tweets, and red markers represent Sick

Person or Suspected Disease tweets. In addition, green markers represent tweets that mention the news. Tweets that represent jokes are not displayed. All posts retrieved from Instagram are plotted with blue markers since mosquito-related posts have not been classified yet.

This service can be used to monitor epidemic waves. Since Twitter creation (2006), As discussed in Section 2.2, the analysis of content published on social media like Twitter has the potential to reveal useful data [7]. Moreover, exploring social media can include users who do not have mobile devices.

3.1.3 Visualization of existing reports

The third service offered by the VazaDengue system is the visualization of the 500 most recent report that have been registered by other users. We defined 500 as the default value for the reports based on the amount of tweets. As the tweets are also considered reports and due to the amount of tweets posted everyday,we could not display much more than 500 tweets, otherwise, it could impact the system performance. especially in the case of mobile applications, in which resources are scarce. In addition, the users could be an overload of tweets. However, someone could argue that 500 reports are not adequate. Thus, we intend to allow users to configure how many reports he wants to visualize.

The application clients receive these reports and plot them on the map according to their coordinates. Once the reports have been plotted on the map, the user can click on one report to access further information about it. The reports a represented by map marker. The color of the map marker varies according to the type of reports. The yellow markers represent "Mosquito Breeding Site" reports, green markers are "Informational" reports, and red markers represent disease-related reports: "Sick Person" and "Illness Suspicion."

3.2 VazaDengue Main Components

The VazaDengue architecture contains three main components: **Application Server**, **Data Crawler**, and **System Client**. Application Server is the core of VazaDengue system. It is the responsible for providing to the user the VazaDengue services. Data Crawler manages the social media services. It is in charge of retrieving social media content. System Client is the interface between users and the services provided by the VazaDengue system. It consumes the services provided by the System Server. The Figure 1 presents all the components that comprise the VazaDengue system. The main components are highlighted in a dark gray color.

The **Application Server** is the back-end of the system, in which is responsible for exposing an interface to the services associated with the application domain. The Application Server is the component that responds to HTTP requests through the architecture and REST object transfer pattern. This component is responsible for implementing business rules, authentication, and publishable data creation. It is the back-end of the system in which focuses on answering HTTP requests, implements the business rules, authenticates users and processes the received data.

The **Data Crawler** is responsible for monitoring Twitter and Instagram. Data Crawler retrieves mosquito-related data from social media and stores them in the VazaDengue database. All request are sent to Apache Web server and, then, it

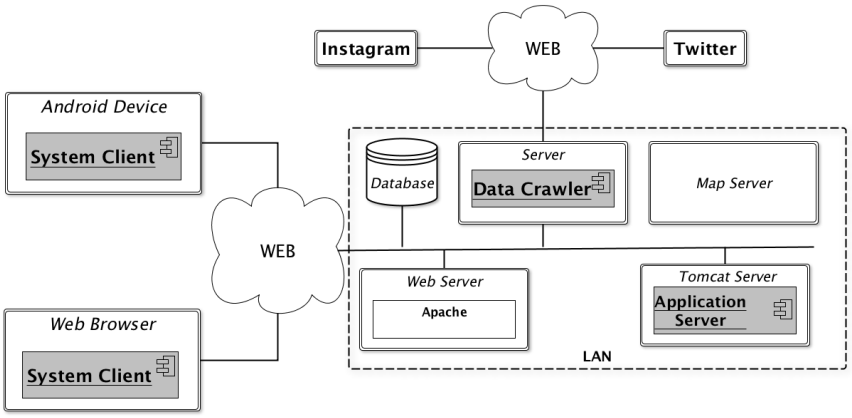

Figure 1 VazaDengue architecture

redirects to the Application Server. Regarding Twitter, tweets and user profiles are stored in the database as dengue reports. This allows us to process these data (filtering and classification) and to offer a new layer to users visualize the retrieved tweets. Therefore, the database has a structure to store the tweet information, like date, text, the number of retweets and favorite, and the location if available. It also stores information about the tweet's owner: id, name, screen name, location, description and his profile image. The VazaDengue database contains a similar structure to store Instagram posts.

The **System Client** is the component that the users interact with the VazaDengue system. Each client is an access point for the services provided by the Application Services. The users can interact with the system via two clients: Web Interface and the Mobile Interface. The *Web Interface* is the web access point for the uses. The purpose of this interface is to allow users, who do not use mobile devices, interact with the collection and visualization of mosquito-related data. *Mobile Interface*, in its turn, provides the same functions available on Web Interface. Nevertheless, the main goal is to create a mobile application to users notify dengue focus by taking advantage of mobile features, like GPS and camera. This interface is currently available for Android devices, but an iOS version is under development. Both clients communicate with the server, through API REST, to retrieve data and display them in map layers.

3.3 Architectural Decisions

During the design of the VazaDengue, our main concern was related to the communication among the three most important components, especially the communication with the Application Server since it is responsible for implement the business rules. Thus, we have taken into account mainly the interoperability, scalability, and performance to build the VazaDengue architecture. For the communication sake, we designed the architecture following a REST web service that uses JSON to transmit data among the components. Thus, any browser can read and write data without technological difficulties.

Since we are developing a system that provides services for several mobile devices, we had to handle with extensive access to the VazaDengue system. Therefore, we designed the Application Server to be a stateless server. Thus, we can replicate the same service on various machines and make load balancing without to worry about sending the same clients for the same servers. Any server can meet the requests of any client at any time.

Besides using a stateless protocol to meet the scalability requirement, we also considered the data storage. We decided to use a primary data storage that makes easy the *sharding*. Thus, we can distribute our data in several servers, in which allows us to meet the performance requirements. Some big companies use free solutions as PostgreSQL (Instagram), Mysql (Facebook), MongoDB (Foursquare). For our system, we decided to choose the PostgreSQL. We chose the PostgreSQL because it offers a spatial extension called PostGIS[5]. PostGIS allows us to meet all the functional requirements related to geoprocessing and also guarantee scalability.

We are using a client/server model to meet the functional requirements. Our server side has a database, a REST service as previous described and a map server as well. Map server component is responsible for integrating our spatial data with the different client maps available on the market (Google Maps, Apple Maps, HERE maps, etc.). Client side features Android applications (Mobile Interface) and web applications (Web Interface).

4 The Tweet Classifier for Dengue

VazaDengue harvests data from both Twitter and Instagram, providing the mosquito-related content to be explored. However, due to the comprehensiveness and diversity observed in previous studies (Section 2.2), we opted by investigating how to automatically filtering and classifying relevant tweets, i.e., content regarding mosquito-borne diseases published by Twitter users. This section describes the steps undertaken to produce the first version of the classifier. This version was coupled into the VazaDengue system and was used to classify tweets during the first two years of the system operation.

As briefly discussed in Section 2.2, machine learning algorithms can use two types of learning: supervised and unsupervised. In the supervised machine learning, the algorithm learns from a training dataset; then it uses the knowledge learned from the training set to classify the input dataset. In the unsupervised machine learning, the algorithm learns itself how to classify the data based on the structures or relationships found in the input dataset. Intuitively, we expect that supervised classification algorithms should be able to provide better accuracy, as well as to give a clear way for selecting actionable content from the most informative classified data. However, the supervised classification suffers from a known limitation regarding the training set. The algorithm requires a training set with the characteristics similar to the characteristic of the content to be classified. Thus, if the content changes, the algorithm needs to be retrained with a new training set. Such requirement may impose a burden if the content is volatile. In the context of our research, a supervised classifier needs to be re-trained at the beginning of each epidemic wave. In this sense, the unsupervised classification may be a more attractive alternative,

[5] http://www.postgis.net/

but it could be challenging to achieve similar accuracy. Thus, considering the advantages and disadvantages of both techniques, we evaluated and compared their contributions to the classification of tweets in the scope of our research. This section is present the results of our earlier research and extends [10].

4.1 Supervised Classification of Twitter Content

We used the Twitter Streaming API to collect two sets of tweets published in Portuguese, harvested over two sub-cycles of the 2015 epidemic cycle: the first and second semester. During these periods, outbreaks of Dengue and Chikungunya were reported in Brazil.

4.1.1 Definition of class labels and ground truth annotations

We used the first of the two sets for training and the standard k-fold based validation. Then, we used the second set for testing only (not training) and further assessment of model accuracy. Tweets can be relatively easily classified according to user sentiment, typically into the three classes: positive, negative, and neutral. However, this classification does not fit our purpose. We are primarily interested in segregating content by its potential relevance to health professionals. Thus, our challenge was to find a set of classes that reflect our purpose and can, at the same time, be represented accurately by a large enough set of manually annotated training instances. Our classification goal was to achieve a finer granularity of tweet relevance than just a binary classification into actionable and noise. After some trials over the initial set of 250 tweets, we found a set of four classes with decreasing relevance. Such relevance was qualitatively measured based on the actionability of the content tweeted. We found that the set presented in Table 1 gave at the same time a good accuracy and granularity.

Table 1 Classification of tweets

Class	Actionability	Content
Mosquito Breeding Sites	High	-Tweets reporting sites that have or probably have the breeding of mosquitoes -Sites that provide conducive environments for mosquito breeding
Sickness	Medium	-Users suspecting or confirming they are sick or aware of somebody who is sick -Users talking about disease symptoms
News	Low (indirect)	-Spreading awareness -Reports on available preventive measures -Information about health campaigns Statistical data about the incidences of the disease
Joke	None	-Combination of jokes or sarcastic comments about Dengue epidemic

Most of the tweets about jokes either make an analogy between Dengue and the users' lives, or they use the words related to Dengue as a pun. A typical pattern is the following:

> *meu [algo como: wpp - WhatsApp, timeline, Facebook, Twitter, etc] está mais parado do que agua com dengue.]*

(My [something like: wpp - WhatsApp, timeline, Facebook, Twitter, etc] is more still than standing water with dengue mosquito.])

In this example, the user was playing with the words when referring to the standing status and inactivity in his Whatsapp account - this is because the breeding sites of the Aedes mosquito are typically found in containers with stagnant water. Many of the jokes in the last epidemic wave were related to Zika, which in Brazilian Portuguese, has been used as a new slang word for failure or any personal problem. It is important to note that the previous work 2 on tracking Aedes-related epidemic waves makes no distinction between Mosquito breeding sites and sickness tweets. News is still indirectly actionable and useful, e.g., to identify the emerging outbreak patterns in specific areas. The detection of jokes requires an understanding of sarcastic tone in short text, which is challenging, as it uses the same general terms as those found in a more informative content.

We extracted from the 2015 epidemic cycle two sets for supervised classification: one from the first semester, having 1,000 instances (tweets), and another from the second semester, having 1,600 instances. These sets were first manually annotated by our group at PUC-Rio, which also included the participation of a medical doctor and an epidemiologist. The first set was used as a training and test set, for the supervised classification using the standard k-fold validation. We use the training set also for comparing the accuracy of different classification models and for selecting the more accurate one. The second set was used for further testing, without training.

The training set of about 1,000 tweets was annotated by three local experts independently, by taking the majority class for each instance, this took 100 hours over three refinement steps used for resolving inconsistencies and ambiguities. The classes are fairly balanced, as can be observed in Table 2.

Table 2 Classification of tweets

Class	Size	Rate
Mosquito sites	257	24%
Sickness	338	31%
News	333	31%
Joke	148	14%

4.1.2 Content pre-processing

Before applying supervised learning algorithms, we need first to establish the set of relevant classes. We called such task as *pre-processing*. We used a technique similar to the one described in [25] to determine a set of filtering keywords for harvesting the tweets. In particular, we started with the unique #dengue hashtag "seed" for an initial collection. After manual inspection of about 250 initial tweets from the first epidemic wave collected (1,000 tweets), our local experts extended the set to include the following most relevant hashtags, approved by all researchers: #Dengue, #suspeita, #Aedes, #Epidemia, #aegypti, #foco, #governo, #cuidado, #febreChikungunya, #morte, #parado, #todoscontradengue, #aedesaegypti.

Content pre-processing includes a series of normalisation steps, followed by POS tagging using the tagger from Apache OpenNLP 1.5 series[6], and word lemmatisation using LemPORT [26]. LemPORT is a customised version of Lemmatizer for

[6] http://opennlp.sourceforge.net/models-1.5/

Portuguese language vocabulary. We also normalised the content by replacing 38 kinds of "Twitter lingo" abbreviations, some of which are regional to Brazil by the complete word. For instance, "abs" for "abraço" (hug), "blz" for "beleza" (nice), among others. Emoticons and non-verbal forms of expressions were also normalised. Moreover, we also replaced links, images, numbers, and idiomatic expressions using conventional terms (URL, image, funny,...). We are aware that such language resources are useful to express the sentiment in tweets. However, we found they do no work well as class predictors.

4.1.3 Results

Considering the characteristics of our datasets, we experimented with three classification models: *Support Vector Machines* (SVM), *Naive Bayes*, and *MaxEntropy*.

SVM models, based on quadratic programming, are very popular classification models [27]. An SVM model establishes maximized margins, creating the larger possible distance between the separating hyperplane and the instances on either side of. SVM is well suited to learning tasks in which the number of features is large in comparison with the number of training instances available [27].

Naïve Bayes networks, ensuring a short computational time for training [27], are the most commonly used classifier for text classification [28]. They are simple Bayesian networks composed of directed acyclic graphs with only one parent (unobserved node) and several children (observed nodes). They are easy to use and experiment with and often give effective results. Multinomial Naïve Bayes networks, a version of Naïve Bayes networks, are more suitable for text documents. The only difference is that the Multinomial networks consider the frequency of words and adjust the underlying calculations of probability accordingly while in the Naïve Bayes networks the frequency count does not matter.

Maximum entropy is a general technique for estimating probability distributions from data. The overriding principle of this technique is that when nothing is known, the distribution should be as uniform as possible, that is, to have maximal entropy [27]. Unlike Naïve Bayes and Multinomial Naïve Bayes, Maximum entropy does not incorporate the assumption of feature independence. It is a feature-based model that gives weights to the features which have the maximum likelihood for a class. The higher the weight, the stronger is the indication of the feature for a class.

In our work, the classification performance, measured using standard cross-validation, was similar across different classifier models. We chose Multinomial Naive Bayes as having probabilities associated with each class assignment helped identify the weak assignments, and thus the potential ambiguities in the manual annotations. 10-fold validation reports an overall 84.4% accuracy and 0.83 F-measure.

To further validate these results, we then classified the test-only set (1,600 tweets). This set has a similar class balance to our training set. This set is also independent of the first classification, used for training. The distribution of instances in each class, taken from the ground truth annotations, was not substantially different from that in the training set, except sickness, the more abundant class.

The performance results of the automated classification are reported in Table 3. One can see the results show a good accuracy, especially for sickness and news. Although the precision for sickness was considerably smaller than for the other classes,

it had presented a good recall. Indeed, our main concern is avoiding false negatives, measured through recall. On the other hand, one can see that the precision of sickness was considerably lower. It may be explained by the large range of ways in with someone may tweet about cases of sickness.

Table 3 Performance of Naive Bayes on independent test set

Class	Precision	Recall	F	Accuracy
Mosquito Breeding Site	.79	.74	.76	.74%
Sickness	.63	.85	.72	.85%
News	.79	.85	.83	.86%
Joke	.81	.78	.84	.78%

4.2 Unsupervised Classification of Twitter Content

In this section, we compare our supervised classification approach with Topic Modelling [29], a well-known semantic clustering algorithm that shown useful results in social media content analysis [16, 18, 17]. The supervised classification has the obvious limitation that a re-annotation of a training set is required to react to "concept drift" in the content. It is a real problem in our setting, where online posts reflect the combination of epidemiological and seasonal effects (e.g. epidemics shift from Dengue to Chikungunya and Zika, from year to year). This limitation is discussed further in Section 5. While manually annotating the training set, we also realised that the classification of the individual instance was often ambiguous, making it difficult to draw sharp class boundaries.

Our goal here was to investigate an application of LDA that shows the potential for scalability and flexibility, i.e., by periodically rebuilding the clusters to track a drift in Twitter search keywords. We used the Twitter Streaming API to select a sample dataset composed of 107,376 tweets, harvested in summer 2015 using the standard keyword filtering from the Twitter feed, and containing a total of 17,191 unique words. Raw tweets were pre-processed just like for classification, producing a bag-of-words representation of each tweet. Additionally, as a further curation step, we removed the 20 most frequent words in the dataset, as well as all words that do not recur in at least two tweets. This last step is needed to prevent very common terms from appearing in all topics, which reduces the effect of our cluster quality metrics and cluster intelligibility. [29].

We propose to use the *intra-* and *inter-* cluster similarity as our main evaluation criteria. This is inspired by *silhouettes* [30], and based on the contrast between *tightness* (how similar data are in a cluster) and *separation* (how dissimilar data are across clusters). Specifically, we define the similarity between two clusters C_a, C_b in terms of the cosine TF-IDF similarity of each pair of tweets they contain, i.e., $t_i \in C_a$ and $t_j \in C_b$, as follows:

$$sim(C_a, C_b) = \frac{1}{|C_a|\,|C_b|} \sum_{t_i \in C_a, t_j \in C_b} \frac{\mathbf{v}(t_i) \cdot \mathbf{v}(t_j)}{||\mathbf{v}(t_i)||\;||\mathbf{v}(t_j)||} \tag{1}$$

where $\mathbf{v}(t_i)$ is the TF-IDF vector representation of a tweet. That is, the kth element of the vector, $t_i[k]$, is the TF-IDF score of the kth term. As a reminder, the TF-IDF score of a term quantifies the relative importance of a term within a

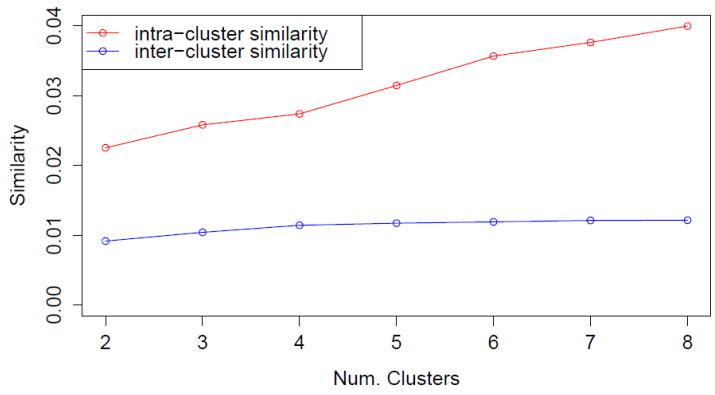

Figure 2 Intra- and Inter-cluster similarities

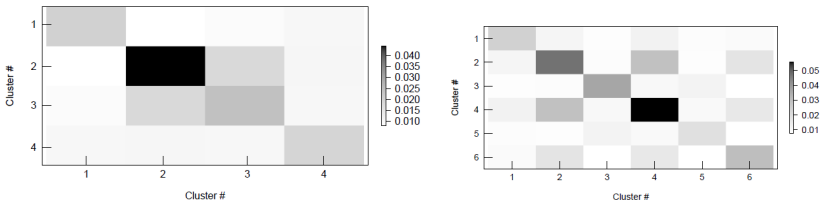

Figure 3 Inter- and intra-similarity for 4 and 6 clusters topic models

corpus of documents [31]. Eq. (1) defines the *inter-cluster similarity* between two clusters $C_a \neq C_b$, while the *intra-cluster similarity* of a cluster C is obtained by setting $C_a = C_b = C$.

4.2.1 Results

Figure 2 reports the inter- and intra-cluster similarity scores for each choice of clustering scheme. The absolute similarity numbers are small due to the sparse nature of tweets and the overall little linguistic overlap within clusters. However, we can see that the intra-cluster similarity is more than twice the inter-cluster similarity, indicating a good separation amongst the clusters across all configurations. These results seem to confirm that the LDA approach is sufficiently sensitive for discovering sub-topics of interest within an already focused general topic, defined by a set of keywords.

The plots presented in Figure 3 provide a more detailed indication of the contrast between intra- and inter-cluster similarity at the level of individual clusters. For example, in the 4-clusters case, the average of the diagonal values of the raster plot is the intra-cluster similarity reported in Figure 3, whereas the mean of the off-diagonal values represents the inter-cluster similarity. In these plots, darker boxes indicate a higher (average) similarity. Plots with diagonals darker than the off-diagonal elements are an indication of a high-quality clustering scheme.

An expert inspection carried out by native Brazilian Portuguese speakers, considered both the list of words within each topic and a sample of the tweets from each one. In this case, we found the four topics scheme to be easier amenable to intuitive interpretation. LDA gives the importance of the words as a measure of how well they are represented in the topics. The following is a list of most relevant topics for this scheme:

Topic 1: parado, água, fazer, vacina, até, meu, tão
Topic 2: combate, morte, saúde , confirma, ação , homem, chegar, queda, confirmado, agente
Topic 3: contra, suspeito, saúde , doença, bairro, morrer, combater, cidade, dizer, mutirão
Topic 4: mosquito, epidemia, pegar, foco, casa, hoje, mesmo, estado, igual

Although the initial supervised classification proposed might be improved, we expected that those four core classes should be distinguished in the topic modelling. However, the inspection of the resulting topics suggests that they only partially overlap the a priori supervised classification. Topic 1 is closely related to Jokes. Topic 2 is interpreted as news about increase or decrease of Aedes-borne disease cases as well as individual cases of people who died because of the Aedes-borne diseases, i.e. Dengue, Chikungunya, and Zika. It also contains news about combating the mosquito in certain locations as well. Examples:

> *Rio Preto registra mais de 11 mil casos de dengue e 10 mortes no ano #SP*
> *(Rio Preto reports more than 11 thousand cases of dengue in the year #SP)*

Topic 3 appears to mostly contain news about campaigns or actions to combat or to prevent Aedes-borne diseases, for instance:

> *Prefeitura de Carapicuba realiza nova campanha contra dengue e CHIKUNGUNYA[URL removed]*
> *(Carapicuba City Hall launches new campaign against dengue and CHIKUNGUNYA[URL removed])*

The difference between the news in Topics 2 and 3 is in the type of news. News in Topic 2 is typically about the increase or decrease of Aedes-borne diseases, whereas news in Topic 3 are about campaigns or actions to combat the propagation of the Aedes mosquito. Finally, Topic 4 contains mostly sickness tweets, with some instances of jokes:

> *Será que eu to com dengue ? (I wonder: do I have dengue?)*

Thus, one can see that the unsupervised classifier did not establish a topic covering the most actionable category established in the supervised classification: the mosquito breeding site.

4.3 Supervised vs. Unsupervised Analysis

When we initially looked into the unsupervised classification, our impression was that the most actionable tweets, i.e., those corresponding to the mosquito breeding site, were not easy to spot. In particular, because they do not seem to characterise any of the topics established by the LDA algorithm. To check this intuition, we analysed the content of each topic using our pre-defined four classes as a frame of reference. In this analysis, we used our trained classifier to predict the class labels of all the tweets in the corpus that we used to generate the topics (about 100,000). We then counted the proportion of class labels in each topic, as well as, for each class, the scattering of the class labels across the topics. The results are presented in Table 4 and Table 5, respectively, where the dominant entries for each column are emphasised.

Table 4 Distribution (%) of predicted class labels within each cluster

Class	Topic 1	Topic 2	Topic 3	Topic 4
News	13.9	**72.6**	27.2	**39.4**
Joke	**39.5**	0.1	2.8	4.1
Mosquito Breeding Site	30	4.0	12.3	12.5
Sickness	16.6	23.3	**57.7**	**44.0**
Total	100	100	100	100

Table 5 Scattering (%) of predicted class labels across clusters

Class	Topic 1	Topic 2	Topic 3	Topic 4	Total
News	29.1	28.5	8.9	33.5	100
Joke	**95.0**	0.03	1.05	4.0	100
Mosquito Breeding Site	**79.5**	2.0	5.1	13.4	100
Sickness	**34.8**	9.1	18.8	**37.3**	100

It is worth remembering that these results are based on the predicted class labels. Therefore, they are inherently subject to the classifier's inaccuracy. Furthermore, the predicted class labels were not available to the experts when they inspected the topic content. So they had to perform a new manual classification of a content sample for each topic. Despite the inaccuracies introduced by these elements, Table 4 seems to corroborate the experts' assessment regarding Topics 1 and 2, but less so for Topics 3 and 4. Such differences may be due to the sampling conducted by the experts, which selected the content towards the top of the topic (LDA ranks content by relevance within a topic) and may have come across joke entries which are otherwise scarce in Topic 4. Although the heavy concentration on joke tweets in Topic 1 (see Table 4) seems promising (i.e., the other topics are relatively noise-free), Table 5 shows a problem, namely that Topic 1 is also the topic where the vast majority of Mosquito Breeding Site tweets are found. Thus, although Topic 1 segregates the most informative tweets well, it is also very noisy, as these tweets are relatively scarce within the entire corpus.

Therefore, based on the comparisons performed, we concluded that topic modelling offers less control over the content of topics when compared to a traditional classifier, especially on a naturally noisy media channel. However, although better results from the supervised classifier were expected, we concluded the LDA performance was insufficient to be used in the context of VazaDengue. Thus, we discarded such alternative, leaving all classification of the Twitter content to the supervised classification based on the Multinomial Naïve Bayes.

5 Re-targeting classification for Zika

The classifier presented in Section 4 was launched as part of the VazaDengue system in 2015. Since then, Brazil experienced a surge in Zika and Chikungunya epidemics, which by 2016 had become the primary concern of citizens regarding mosquito-borne diseases, and one of the top public health challenges. The growing evidence of links between Zika with the incidence of newborn children with microcephaly, especially in Brazil, put the disease firmly on the spot.

In turn, this caused a change in the Twitter patterns on which we had trained our classifier 4. In particular, the online chattering about Zika turned out to be much noisier than expected, not least because "Zika" is pronounced in Portuguese like "zica", a slang word that has historically been used in multiple unrelated contexts in different regions of Brazil, generally referring to "something bad". And, with the surge of Zika epidemics, many other meanings had emerged.

This, along with the "concept drift" shown by the online posts led to a progressive degradation in the actual accuracy of the classifier, compared to its theoretical validation (Sec. 4), triggering re-training. Learning from our earlier experience, we took this as an opportunity to revisit the learning strategy through the entire model-building pipeline, implementing a number of enhancements from harvesting to class selection, to manual labelling and training. In the rest of this section, we report on this new classifier, which powers the current version of the VazaDengue system.

5.1 Keyword selection for high recall

Firstly, a new set of seed keywords for harvesting were selected manually to align to the new Zika lingo:

dengue, *combate-a-dengue*, *foco-dengue*, *todos-contra-dengue*, *aedes-eagypti*, *zika*, *chikungunya*, *virus*.

Those were used to harvest an initial corpus of tweets and then refined using a TF-IDF ranking of the terms found in the harvest (after removing common stopwords and those words that experts deemed to be out of context). This gave us a rich set of keywords for high-recall harvest:

microcefalia, *transmitido*, *epidemia*, *transmissao*, *doenca*, *eagypti*, *doencas*, *gestantes*, *infeccao*, *mosquito*.

5.2 Class labels

Secondly, we adopted the view that the classifier would serve as a preliminary "noise reduction" step as part of a more complex analytics processing, with the ultimate aim to identify influential Twitter users who either post relevant content or follow/retweet relevant news items. Thus, we simplified our initial 4-class model of Sec.4 to only include `Relevant`, `News`, and `Noise` classes, leading to a more balanced class representation in the training examples, as well as a simpler manual labelling and automated classification task. These class labels are described in Table 6.

5.3 Manual labelling

Next, we addressed the issue of training set size (initially, only 1,000 examples) as well as of ambiguous class labelling by experts, who were effectively called upon to

Table 6 Classification of tweets

Class	Actionability	Content
Relevant	Medium-High	-Tweets reporting mosquito breeding sites -Sites that provide conducive environments to mosquito breeding -Users suspecting or confirming they are sick or they are aware of somebody else who is sick -Users talking about disease symptoms
News	Low (indirect)	-Spreading awareness -Reports on available preventive measures -Information about health campaigns -Statistical data about the incidence of the disease
Noise	None	-News citing a mosquito-borne disease but without providing useful content -Use of filter terms such as "Zika" out of context -Combination of jokes or sarcastic comments about the mosquito and diseases

give an operational definition-by-example of "content relevance" in our Zika setting. For this, we adopted a consensus approach where 15,000 tweets were independently labelled by two experts, with tied tweets submitted blindly to a third annotator (in the rare instances where all three classes get a vote at this point, a final independent tie-breaker was called upon).

5.4 Training set selection

A rich set of keywords gives high recall, but it may also make it challenging for the classifier to achieve good precision. We therefore simulated more limited harvest by repeatedly selecting subsets of keywords and filtering the training set for examples containing only those keywords.

This exploration revealed that best model performance in this setting is achieved using the follwoing small set of keywords:

mosquito, *dengue*, *zika*, *aedesaegypti*, and *aegypti*.

Filtering the full 15,000 instances training set using these keywords, we are left with the following class distribution on the training set:

Relevant: 1,906 from 2,258
News: 5,180 from 5,720
Noise: 6,218 from 7,022
Total: 13,304 from 15,000

5.5 Feature selection and meta-features

From this training set, we extracted bag-of-words features as indicated in Sec.4, to which we added 1,2, and 3-grams (with minimum term frequency in the training set of 3), resulting in 11,446 n-grams features. Importantly, we also added a number of *meta-features*, which we have previously shown to improve model performance in Twitter content classification for sentiment analysis [32]. These 28 meta-features are listed in Table 7

5.6 Class rebalance, feature selection, and results

As noted, one persistent problem in this modelling exercise is the class imbalance that results from the scarcity of *Relevant* content (1,906 from 13,304, 14%). To

Table 7 List of meta-features used in the Zika content classifier

Features	Emoticon Features	Punctuation Features
1. hasRetweet	11. hasQuestionMark	20. hasEmoticon
2. hasHashtag	12. hasExclamationMark	21. hasPositiveEmoticon
3. hasUsername	13. numberOfQuestionMark	22. hasNegativeEmoticon
4. hasURL	14. numberOfExclamationMark	23. isLastTokenPositiveEmoticon
5. hasRepeatedLetters	15. lastTokenContainsQuestionMark	24. isLastTokenNegativeEmoticon
6. numberOfCapitalizedWords	16. lastTokenContainsExclamationMark	25. numberOfPositiveEmoticons
7. numberOfWordsWithAllCaps	17. numberOfSequencesOfQuestionMark	26. numberOfNegativeEmoticons
8. numberOfWords	18. numberOfSequencesOfExclamation Mark	27. numberOfExtremelyPositive Emoticons
9. numberOfCapitalLetters	19. numberOfSequencesOfQuestionAnd ExclamationMarks	28. numberOfExtremelyNegative Emoticons
10. numberOfRepeatedLetters		

address this, we applied a standard SMOTE filter to double the size of the minority class (1,816 synthetic examples), resulting in a more balanced class distribution: Relevant: 25%, News: 34%, Noise: 41%.

We then selected the top 1,500 features using an Information Gain approach (the InfoGain filter with Ranker from the Weka suite). Interestingly, 22 out of the 28 meta-features listed in Table 7 appear in the top-500 features, which shows that these can be as significant in this context as they are in sentiment analysis.

Given this training set, we compared popular model builders (Multinomial Naive Bayes, Random Forest, SVM) achieving our top accuracy of 86.13% (F-measure 0.862) across all classes, using Random Forest with standard k-fold cross validation. Accuracy figures per class are *Relevant*: 93% (F: 0.856), *News*: 89% (F: 0.891), *Noise*: 80.8% (F: 0.84).

Our analysis shows that using Random Forest we have achieved an accuracy/F-measure that are similar to that found in the first classifier for its training set (84.4%, 0.83). We believe this to be a successful result considering the new challenges imposed by the 2016 epidemic cycle. In 2016, the incidence of Zika had significantly grown in Brazil, making the disease a popular topic among Brazilian Twitter users, especially during the 2016 Olympic Games. Indeed, more jokes about Zika were reported than about other diseases. Moreover, we found the use of the term "Zika" was also extended by Brazilians to denote things that different from the disease, such as referring to experts ("He is *Zika* in playing soccer"); characterizing good, reliable people ("That lady is *Zika*"); referring to lovers ("I met my *Zika* yesterday"). Moreover, several tweets referring to the news citing Zika but irrelevant for our purpose were published in the period. Most of them addressed the concern of particular sportsmen and celebrities on getting Zika during the 2016 Olympic Games in Rio de Janeiro, Brazil.

6 Content Evaluation

VazaDengue was designed for supporting health professionals and citizens, in general, to be aware of relevant information about the mosquito-borne diseases. We expect to contribute to health professionals identifying tempestively the geographic distribution of current epidemic waves and the upcoming of new ones. In the particular case of community health agents, it is expected that information provided by the users and relevant content filtered from Twitter would be used as input for supporting the conduction of immediate prevention and combating activities. In both cases, the precise classification of content mined in large-scale from Twitter plays a key role.

This section presents two distinct studies conducted to evaluate the potential contributions of the tweets mined and classified by VazaDengue to prevent and combat mosquito-borne diseases. Section 6.1 presents a survey conducted with Brazilian community health agents aiming at evaluating the relevance of tweets. Section 6.2 compares official reports of 2015-2016 epidemic waves in Brazil with the distribution of tweets and news mined by VazaDengue during these waves.

6.1 Community Health Agent's Opinion

Community health agents work performing continuous preventing and combating activities such as identifying and eliminating mosquito breeding sites, disseminating preventing information and applying insect killer solutions in houses. Part of their action is grounded on citizen calls to the health department of the city hall. Such calls include complaints about the incidence of mosquito breeding sites and the incidence of mosquito-borne diseases. Thus, we expected that relevant tweets filtered by VazaDengue may help them to perform their professional activities. Therefore, we conducted a survey aiming at characterizing the perceived *relevance* of tweets by these professionals.

The survey questionnaire is composed of two main parts. In the first part, the subjects are asked to provide information about their location and professional background. They are also asked about their experience in identifying relevant content in social networks. In the second part, subjects are asked to evaluate the perceived relevance of 20 real tweets for supporting prevention and control activities. We established this limited number of tweets to prevent subjects from giving up the survey [33].

To make the scope of the evaluations more comprehensive, we opted by distributing different sets of tweets among the subjects. We established four sets of 20 different tweets each, resulting in 80 tweets to be evaluated. These tweets were randomly sampled from the 590 tweets filtered from the 2016 epidemic cycle and annotated as relevant by the classifier and by the researchers that performed the manual annotation (Section V.D). Therefore, four different versions of the survey questionnaire with different sets of tweets were applied, named as Q1, Q2, Q3, and Q4. We applied a four-level Likert to ask the subjects about the perceived relevance of the tweets: *totally irrelevant, partially irrelevant, partially relevant, and totally relevant.*

Survey research requires the identification of samples aligned with the research objectives [34]. We found on the social network Facebook a potential source of population composed of several discussion groups of Brazilian community health agents. All the groups used in the study are classified as *closed*, i.e., groups in which Facebook users should be previously accepted by an administrator to become new members. Some of these groups were composed of more than 50,000 community health agents located in different Brazilian cities. After one week of subscription, we were accepted into five groups. Based on the size of the groups, we distributed different versions of the survey questionnaire. We also shared the questionnaires with community health agents from the researchers' personal networks.

6.1.1 Results and Analysis

After three days of recruitment, we sent reminders in each group. After one week, 21 professionals had answered the survey, totalizing 420 tweet evaluations. These professionals are active community health agents from 18 different cities located in 12

different Brazilian states. All survey participants had reported previous professional experience on preventing and control of mosquito-borne diseases. Moreover, 20 subjects had declared previous experience on identifying relevant content regarding these activities in social networks. Table 8 presents a summary of the respondents' characteristics, grouped by the questionnaire answered by each one. The subjects' characterization suggests a heterogeneous sample of experienced community health agents that fit the subject's profile desired in our study.

Table 8 summarizes the subjects' answers by questionnaire, also presenting the results of the agreement test (Cohen's Kappa test) applied between the respondents of each questionnaire. Cohen's Kappa test [35] measures inter-rater agreement for categorical items. It is generally thought to be a more robust measure than calculating the simple percent agreement since it takes into account the possibility of the agreement occurring by chance. The value of Cohen Kappa may range from -1 to 1 (perfect agreement).

Table 8 Summary of the survey results by questionnaire

Quest.	#Resp.	Average Exp.	Totally Irrelevant	Partially Irrelevant	Partially Relevant	Totally Relevant	Cohen's Kappa	p-value
A	6	10.2	35.71%	20.24%	29.76%	14.29%	.1247	.0002
B	4	5	28.57%	16.07%	26.79%	28.57%	-.1278	.9899
C	7	8.42	24.49%	17.35%	31.63&	26.53%	-.0039	.5555
D	4	5.25	23.21%	14.29%	19.64%	42.86%	.0282	.3026

One can see it was found agreement only among the six subjects that had answered questionnaire Q1. However, the agreement level obtained in Q1 is very low. In the other questionnaires, the insufficient p-values do not allow us to draw any conclusion. Although the small sample sizes would influence the results of the Kappa test, we observed a frequent divergence of opinion in all questionnaires. Such divergences could be influenced by the different perceptions of the reported by the health agents about which content they may consider relevant in social networks to support preventing and control of mosquito-borne diseases. For instance, a considerable number of subjects reported before evaluating the tweets that publishing news and prevention guidelines in the social network could be useful. However, such content is typically annotated as *news* by the classifier. As previously discussed in section IV.B, news has been considered a secondary source of information, once it is not directly actionable. Therefore, the news category was not included in the survey questionnaire. On the other hand, few subjects reported the use of social networks to identify directly actionable issues, such as the identification of mosquito breeding sites. These findings indicate the potential innovation offered by VazaDengue.

Once the subjects are experienced professionals, we applied a different criterion from that used in the researchers' manual annotation (majoritarian opinion 5) to depict the final classification of each tweet. However, the criteria applied in both cases are similar in terms of the minimum absolute number of positive evaluations: if two health agents agree that certain tweet is relevant, this tweet will be annotated as relevant. However, In the case of the professionals' evaluation, such rule works apart from the number of health agents that had classified the tweet as partially or completely irrelevant.

After applying this criterion, we found that 60 from the 80 evaluated tweets are relevant, resulting in an overall precision of 75%. This result evidence that health

agents tend to identify relevance in the tweets annotated as well in the context of the VazaDengue system. In other words, our definitions of "relevant" are suitable to the context of preventing and control of mosquito-borne diseases. After analyzing the results by tweet classified as relevant, we found the following patterns:

- Tweets reporting users' cases or suspects of mosquito-borne diseases.
- Tweets reporting cases or suspects of the mosquito-borne diseases in other individuals, mainly parents and other Twitter users.
- Tweets reporting the incidence of suspected mosquitoes in the user location or very close to them.

It is important to note that the location of other individuals than the tweet' user is not available. For instance, one may tweet about the sickness of a friend but he lives in another city. However, the health agents still had classified such content as relevant. The incidence of users tweeting about mosquito breeding was scarce in the whole population of 5,000 tweets used in the analysis. However, we believe that reporting mosquito breeding sites is one of the main contributions that a user could report for preventing and control of mosquito-borne diseases. Health agents may use these reports to take immediate actions on verifying and eliminating these sites. The survey results may also help us to understand possible types of tweets which community health agents tend to do not consider as relevant. By analyzing each one of the 20 tweets no classified as relevant, we found the following anti-patterns:

- Tweets reporting users' vaccination and possible side effects of the vaccines
- Tweets reporting past contraction of mosquito-borne diseases
- Tweets reporting hypothetical consequences of the disease in a long term
- Tweets excessively using bad words and jokes, even when reporting potentially relevant content.

The findings of the presented survey indicate the definitions of the categories used to annotate the tweets 6 is suitable to the context but they could be also refined. Such conclusion touches divergences observed among the annotations performed by the researchers. For instance, researchers stayed divided on annotating as relevant or noise those tweets about vaccination. However, the opinion of the health agents indicated that such type of content should not be considered as relevant.

6.2 Evaluating the Concentration of Tweets

As explained in Section 3, the VazaDengue system has a component, namely Data Crawler, that filters and harvests the content from social media as Twitter. In the case of Twitter content, tweets are classified according to categories, and after the classification, they are provided to the users as reports. The users can use these classified tweets to have a notion of what the other users are talking about the mosquito and its diseases. That is, users can explore the classified tweets and the categories to have an understanding of what mosquito-related content that other users are talking without having to search for it on Twitter. On the other hand, the users may want to use the classified tweets for other purposes. For instance, Figure 4 presents the distribution mosquito-related tweets across Brazil. The users can use this distribution to have a notion of the concentration of mosquito-related tweets in a particular area. This information can be useful for users that want to avoid risk areas or for users that want to monitor their area.

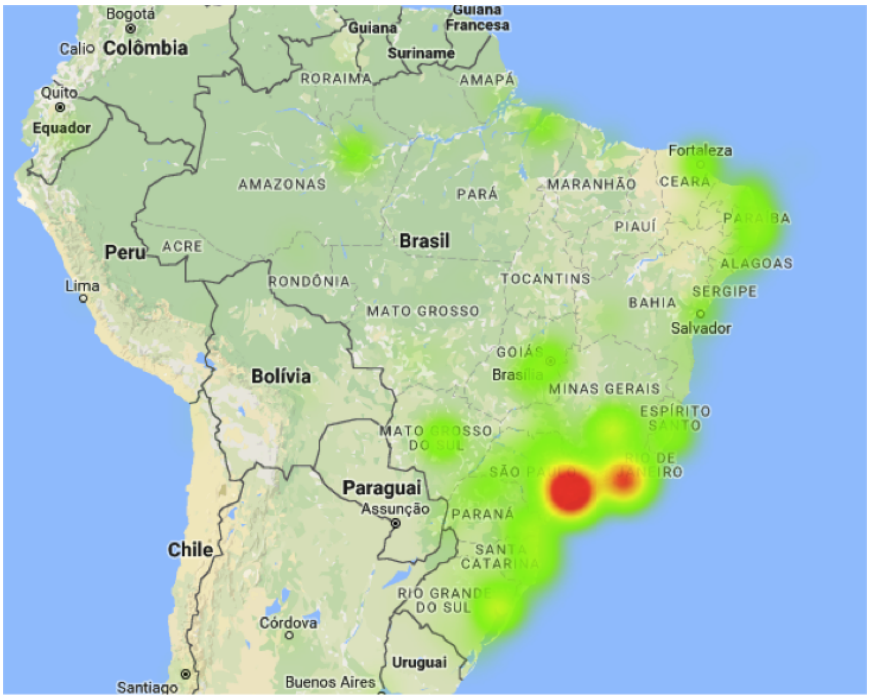

Figure 4 Concentration of mosquito-related tweets

Before allowing users to explore the concentration of tweets for monitoring purposes, we need to investigate if the areas with high concentration of mosquito-related tweets are the same areas reported as areas with high incidence of dengue cases. If there is an intersection between these areas, then the users can use the concentration of tweets to monitor areas with the incidence of dengue cases. In order to conduct this investigation, we compared official reports of 2015-2016 epidemic waves in Brazil with the distribution of tweets harvested by VazaDengue during these waves. As our investigation relies on official reports to perform the comparison, we used the epidemiological report that the Brazilian Health Department releases every single year. Among the information available in these reports, we are interested in big cities with a high incidence of dengue cases. We focus on big cities due to the probability of containing geolocated tweets. The literature reports that less than 1% of tweets are geolocated [36, 37]. Therefore, the more people a city has, the more likely it is that people will tweet with a location. Hence, we narrowed down the scope to big cities in order to be able to compare the cities with the highest incidence of dengue cases and cities with the concentration of tweets. Although we have reduced the scope to big cities, that does not guarantee that these cities contain mosquito-related tweets; what justifies our investigation.

We used the Data Crawler to harvested the tweets that contain location coordinates. After identifying these tweets, we used the Google Maps Geocoding API[7] to determine the cities from the coordinates. Then, we organized the tweets according to the five regions of Brazil: North, Northeast, Central-West, South, and Southeast. Finally, we performed the comparison. Table 9 presents the big cities reported with

[7] https://developers.google.com/maps/documentation/geocoding/intro

Table 9 Incidence of tweets in Big Cities with the highest prevalence of dengue cases

Year	Region	City	Incidence of Tweets	Population Size
2015	Northeast	Recife (Pernambuco)	17	A
		Fortaleza (Ceará)	17	A
	Central-West	Goiânia (Goiás)	18	A
		Aparecida de Goiânia (Goiás)	1	B
	Southeast	Sorocaba (São Paulo)	2	B
		Campinas (São Paulo)	8	A
		Uberlândia (Minas Gerais)	5	B
		São José dos Campos (São Paulo)	3	B
		Guarulhos (São Paulo)	5	A
		Contagem (Minas Gerais)	3	B
2016	North	Porto Velho (Rondônia)	7	B
	Northeast	Fortaleza (Ceará)	48	A
	Central-West	Goiânia (Goiás)	34	A
		Aparecida de Goiânia (Goiás)	7	B
		Cuiabá (Mato Grosso)	5	B
	South	Londrina (Paraná)	12	B
	Southeast	Belo Horizonte (Minas Gerais)	147	A
		Ribeirão Preto (São Paulo)	29	B
		Campinas (São Paulo)	30	A
		Guarulhos (São Paulo)	26	A

A = more than 1 million people
B = between 500 and 999 thousands people

the highest incidence of dengue cases in 2015 and 2016 according to the five regions. First column presents the reported year. Second column presents the region to which the city belongs. Third column presents the city and its state between parentheses. Fourth column contains the number of mosquito-related tweets that was tweeted from the city. Finally, sixth column presents the city's population.

Table 9 shows that all the big cities with the highest incidence of dengue cases contain mosquito-related tweets. We highlight in these results the increasing of tweets in 2016. It was an increasing of 436.71% compared with 2015. This 4-times increasing of tweets was due to the Zika virus wave. Since Zika virus was not an issue in Brazil before that, its wave drastically impacted the social media as Twitter, increasing the number of mosquito-related tweets. This increase also reflects the number of cities with the incidence of dengue cases around the country. Before 2015, dengue cases concentrated in three regions (Northeast, Central-West, and Southeast). However, as we can see in the table, cities from other regions also become areas with high incidence of dengue cases in 2016.

The results presented in the Table 9 indicate that the users can use the concentration of tweets to monitor big cities. In fact, we found that the distribution of the tweets by month is similar to the distribution of the dengue cases. This result can be observed in Figure 5, which presents the comparison between the incidence of Dengue cases and incidence of mosquito-related tweets from March, 2105 to August, 2016. As the VazaDengue started to harvest tweets in April, 2105, we considered the tweets from March, 2105 to plot on the map. We considered the tweets until August, 2106 because the 2016 report contained monthly information until August, 2016. From there on, the information is presented in the report quarterly.

We can notice in Table 9 that although all the cities contain mosquito-related tweets, some of them contains only a few tweets (Aparecida de Goiânia city has only one tweet). Thus, the same incident of tweets and dengue cases may be related to a mere chance. This low number of tweets can be explained by the well-known limitation regarding the use of tweets: the location. Unfortunately, it is reported in the literature that less than 1% of tweets are geolocated [36, 37]. Due to this

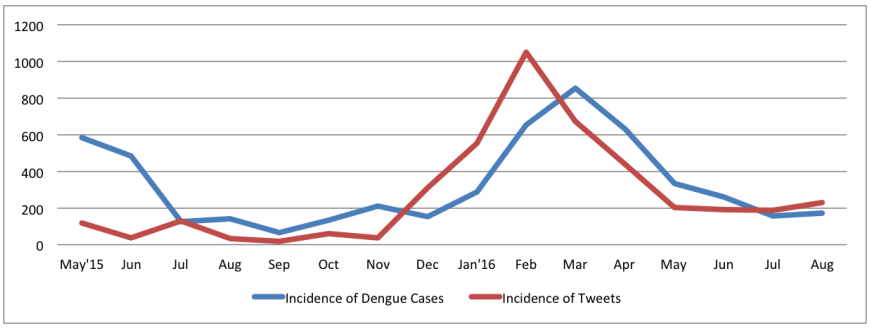

Figure 5 Distribution of mosquito-related tweets and dengue cases by month

limitation, we decided to focus on big cities since they are most likely of containing geolocated tweets. Indeed, most of the cities with more than 1 million people contains several tweets (cities with the A value in the Table 9). Even after our decision of narrowing down the scope, we still found few cases of geolocated tweets. This limitation indicates that the VazaDengue system needs to implement strategies to infer the tweets' location as described in [38].

As previously mentioned, our evaluation is based on big cities with a high incidence of dengue cases reported by the Brazilian Health Department. However, these reports have some limitations that difficult its usage. For instance, the reports are released at the end of each year. Thus, the citizens need to wait for these reports, what it would not be useful if they want a real-time monitoring. In addition, these reports only cover the cities with the highest incidence of dengue cases. Thus, there are cities with a high incidence of dengue cases, but not higher enough to be part of the report. Hence, the citizens will not be able to find these cities in the reports. On the other hand, there are several of cities with mosquito-related tweets. If these cities have users tweeting about dengue, maybe these cities should also be the focus of health agencies. For instance, these cities can be the target for preventing campaigns.

7 Conclusion and Future Work

Mosquito-borne diseases represent concrete risks to the health of citizens from several countries, including Brazil. Brazil is one of the most populous countries of the world and also one of the countries historically with the highest incidence of Dengue in the last decades. In the last years, the outbreaks of Zika and Chikungunya have been also observed in the country. Preventing and controlling mosquito-borne diseases depends on combining efforts of authorities and citizens. However, the state of practice has been showed that traditional approaches for promoting the prevention and control of mosquito-borne diseases do not suffice the promotion of an effective engagement of Brazilian communities more exposed to such diseases.

Brazilians are one of the biggest users of social networks, frequently reporting their daily and sharing relevant news. Considering this scenario, we developed VazaDengue, an innovative platform that offers for it users a platform for preventing and combating mosquito-borne diseases. The main goal of VazaDengue is to strengthen the entomological surveillance of the mosquito that transmits Dengue,

Zika, and Chikungunya by providing geolocated reports, represented through dynamic maps. These reports may be directly included by the users or harvested from social networks such as Twitter and Instagram. In the case of the tweets, VazaDengue also automatically classifies the harvested content, distinguishing those useful from noise. Thus, reaching high accuracy in the classification is a big challenge, even more because each epidemic cycle has particular characteristics that would reflect on the content posted on Twitter.

This paper presented the VazaDengue system, describing its architecture and the evolution of its classifier since its first version, launched in 2015. It also presented two studies of the content generated by the system. In the first study, Brazilian community health agents were surveyed regarding their perceived relevance of the harvested tweets on performing their professional activities. In the second study, it was evaluated the geographical concentration of potentially relevant tweets from the 2016 epidemic cycle in comparison with official reports. The results from both studies indicate that the tweet classifying approach offered by VazaDengue have the potential for supporting different prevention and controlling activities of mosquito-borne diseases. However, opportunities for improvement were also identified.

The browser and Android version of VazaDengue is available since 2015. NNNN downloads of the Android application were made since then and NNN,NNN potentially useful tweets were mined and classified. We plan to launch the IOS version of the VazaDengue application until the end of this year (2017). Next research steps include evolving the platform for stimulating its regular use, also promoting the report of direct contributions in the system. In this sense, we are investigating gamification technologies that would encourage local users to contribute to their communities. We are also negotiating with Brazilian health programs the dissemination of the technology in the context of educational activities. Another research step includes extending the classification of Instagram content, including an intelligent evaluation of pictures associated with potentially relevant posts.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This work is supported by the MRC/UK Newton fund project entitled A Software Infrastructure for Promoting Efficient Entomological Monitoring of Dengue Fever. We are grateful to Alexandre Plastino from UFF (Brazil) for sharing with us the results of his work on tweet classification. We would like to thank Leonardo Frajhof (UNIRIO), Oswardo Cruz (Fiocruz, Brazil), Soeli Fiori (PUC-Rio, Brazil) and Wagner Meira (UFMG, Brazil) for their contribution. We are grateful to the following students of Newcastle University (UK) who contributed to the earlier work on the the topics of the paper: Atinda Pal, Michael Daniilakis, Callum McClean and Jonathan Carlton. Also, we would like to thank PPSUS (Programa Pesquisa para o SUS, Brazil) and FAPERJ (Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro, Brazil) for their support.

Author details
[1]Department of Informatics, PUC-Rio, Rio de Janeiro, Brazil. [2]School of Computing Science, Newcastle University, Newcastle, United Kingdom.

References

1. Organization, W.H.: Dengue fact sheet. http://www.who.int/mediacentre/factsheets/fs117/en
2. Organization, W.H.: Zika virus fact sheet. http://www.who.int/mediacentre/factsheets/zika/en
3. Organization, W.H.: Chikungunya virus fact sheet. http://www.who.int/mediacentre/factsheets/fs327/en/
4. Dengue, R.: RadarDengue. Accessed in 01/11/2017. Published at *Android Apps on Google Play*. https://goo.gl/xhooZu
5. UFRN: Observatório do Aedes Aegypti. Accessed in 01/11/2017. http://observatoriodadengue.telessaude.ufrn.br/

6. Vieweg, S., Hughes, A.L., Starbird, K., Palen, L.: Microblogging during two natural hazards events: What twitter may contribute to situational awareness. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '10, pp. 1079–1088. ACM, New York, NY, USA (2010). doi:10.1145/1753326.1753486. http://doi.acm.org/10.1145/1753326.1753486
7. Gerber, M.S.: Predicting crime using twitter and kernel density estimation. Decision Support Systems **61**, 115–125 (2014)
8. Chen, X., Cho, Y., Jang, S.Y.: Crime prediction using twitter sentiment and weather. In: 2015 Systems and Information Engineering Design Symposium, pp. 63–68 (2015)
9. Group, O.R.: VazaDengue. http://www.vazadengue.inf.puc-rio.br/
10. Missier, P., Romanovsky, A., Miu, T., Pal, A., Daniilakis, M., Garcia, A., Cedrim, D., da Silva Sousa, L.: In: Casteleyn, S., Dolog, P., Pautasso, C. (eds.) Tracking Dengue Epidemics Using Twitter Content Classification and Topic Modelling, pp. 80–92. Springer, Cham (2016). doi:10.1007/978-3-319-46963-8$_7$. http://dx.doi.org/10.1007/978-3-319-46963-8$_7$
11. Missier, P., McClean, C., Carlton, J., Cedrim, D., Silva, L., Garcia, A., Plastino, A., Romanovsky, A.: Recruiting from the network: discovering twitter users who can help combat zika epidemics. arXiv preprint arXiv:1703.03928 (2017)
12. UNA-SUS: UNA-SUS Dengue. Accessed in 01/11/2017. Published at *Android Apps on Google Play*. https://play.google.com/store/apps/details?id=com.all4mobile.unasus.dengue
13. Brasil, D.: Dengue Brasil App. Accessed in 01/11/2017. http://www.dengue.org.br/app/
14. Codeco, C., Cruz, O., Riback, T.I., Degener, C.M., Gomes, M.F., Villela, D., Bastos, L., Camargo, S., Saraceni, V., Lemos, M.C.F., Coelho, F.C.: Infodengue: a nowcasting system for the surveillance of dengue fever transmission. bioRxiv (2016). doi:10.1101/046193. http://www.biorxiv.org/content/early/2016/03/29/046193.full.pdf
15. Twitter: Twitter Usage. Accessed in May 2017. https://about.twitter.com/company
16. Morstatter, F., Pfeffer, J., Liu, H., Carley, K.M.: Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. arXiv preprint arXiv:1306.5204 (2013)
17. Weng, J., Lim, E.-P., Jiang, J., He, Q.: Twitterrank: finding topic-sensitive influential twitterers. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, pp. 261–270 (2010). ACM
18. Rosa, K.D., Shah, R., Lin, B., Gershman, A., Frederking, R.: Topical clustering of tweets. Proceedings of the ACM SIGIR: SWSM (2011)
19. Instagram: Instagram API. https://www.instagram.com/developer/
20. Lampos, V., Cristianini, N.: Tracking the flu pandemic by monitoring the social web. In: 2010 2nd International Workshop on Cognitive Information Processing, pp. 411–416 (2010). doi:10.1109/CIP.2010.5604088
21. Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., Liu, B.: Predicting flu trends using twitter data. In: 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 702–707 (2011). doi:10.1109/INFCOMW.2011.5928903
22. Gomide, J., Veloso, A., Meira, W. Jr., Almeida, V., Benevenuto, F., Ferraz, F., 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, pp. 3–138. ACM, New York, NY, USA (2011). doi:10.1145/2527031.2527049. http://doi.acm.org/10.1145/2527031.2527049
23. Zhu, J., Xiong, F., Piao, D., Liu, Y., Zhang, Y.: Statistically modeling the effectiveness of disaster information in social media. In: 2011 IEEE Global Humanitarian Technology Conference, pp. 431–436 (2011). doi:10.1109/GHTC.2011.48
24. (2017). http://www.paho.org/bra/index.php?option=com_content&view=article&id=1895&Itemid=777
25. Nagarajan, M., Gomadam, K., Sheth, A.P., Ranabahu, A., Mutharaju, R., Jadhav, A.: Spatio-temporal-thematic analysis of citizen sensor data: Challenges and experiences. In: International Conference on Web Information Systems Engineering, pp. 539–553 (2009). Springer
26. Rodrigues, R., Gonçalo Oliveira, H., Gomes, P.: Lemport: a high-accuracy cross-platform lemmatizer for portuguese. In: OASIcs-OpenAccess Series in Informatics, vol. 38 (2014). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik
27. Kotsiantis, S.B., Zaharakis, I., Pintelas, P.: Supervised machine learning: A review of classification techniques (2007)
28. McCallum, A., Nigam, K., *et al.*: A comparison of event models for naive bayes text classification. In: AAAI-98 Workshop on Learning for Text Categorization, vol. 752, pp. 41–48 (1998). Citeseer
29. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
30. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987)
31. Aggarwal, C.C., Zhai, C.: A survey of text clustering algorithms. In: Mining Text Data, pp. 77–128. Springer, ??? (2012)
32. Carvalho, J., Plastino, A.: An assessment study of features and meta-level features in twitter sentiment analysis. In: ECAI, pp. 769–777 (2016)
33. Torchiano, M., Fernández, D.M., Travassos, G.H., de Mello, R.M.: Lessons learnt in conducting survey research. In: Proceedings of the 5th International Workshop on Conducting Empirical Studies in Industry, pp. 33–39 (2017). IEEE Press
34. Linaker, J., Sulaman, S.M., Höst, M., de Mello, R.M.: Guidelines for conducting surveys in software engineering v. 1.1 (2015)
35. Cohen, J.: Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological bulletin **70**(4), 213 (1968)
36. Graham, M., Hale, S.A., Gaffney, D.: Where in the world are you? geolocation and language identification in twitter. The Professional Geographer **66**(4), 568–578 (2014)
37. Zheng, X., Han, J., Sun, A.: A survey of location prediction on twitter. arXiv preprint arXiv:1705.03172 (2017)
38. Ajao, O., Hong, J., Liu, W.: A survey of location inference techniques on twitter. J. Inf. Sci. **41**(6), 855–864

(2015). doi:10.1177/0165551515602847