Dengue surveillance based on a computational model of spatio-temporal locality of Twitter

Janaína Gomide[1]

Adriano Veloso[1] Wagner Meira Jr.[1] Fabrício Benevenuto[2]
Virgílio Almeida[1] Fernanda Ferraz[3] Mauro Teixeira[3]

[1]Computer Science Department
UFMG - Brazil

[2]Computer Science Department
UFOP - Brazil

[3]Biochemistry and Immunology Department
UFMG - Brazil

Motivation

- Twitter is a unique social media channel, in the sense that users discuss and talk about the most diverse topics, including their health conditions
- Traditional disease surveillance comprises a set of epidemiological procedures that monitor the spread of a disease and determine how it is spreading
- Social media channels, such as Twitter, offer a continuous source of epidemic information, arming public health agencies with the ability to perform real-time surveillance

Background on dengue

- Dengue is a mosquito-borne infection that causes a severe flu-like illness, and sometimes a potentially lethal complication
- Outbreaks tend to occur every year during the rainy season but there is large variation of the degree of the epidemic in areas with similar rainfall
- Current strategies for prediction of dengue disease are based on surveillance of insects, which provide only a rough estimate of cases
- Once disease outbreaks are detected in a certain area, efforts need to be concentrated to avoid further cases and to optimize treatment and staff - number of cases can reach several hundred thousands
- In Brazil, where there is a functional disease reporting system, detection of important outbreaks may take a few weeks, leading to loss of precious time to tackle the epidemy

Goal

- Our work aims at using the user generated content available on online social media to predict a real-world event. The event we are interested is the dengue outbreaks.
- In this paper, we analyze how dengue epidemics are reflected on Twitter and to what extent that information can be used for surveillance
- We then introduce an active surveillance framework that analyzes how social media reflects epidemics based on a combination of four dimensions: volume, location, time, and public perception

Contributions

In summary, this work has the following contributions

- proposal and application of a four-dimensional framework for assessing the use of social media data in active surveillance
- proposal, implementation and evaluation of an active surveillance system
- instantiation of both the framework and the surveillance system for dengue

Datasets

We employ data collected from two different sources

official dengue reports made available by the Brazilian Health Ministry

- contains the number of dengue cases per city, notified between 2007 and 2010

twitter Twitter messages mentioning the word "dengue"

- from 2006 to July 2009: 27,658 tweets, out of which 90.27% are from 2009
- from December 2010 to April 2011: 465,444 tweets

Methodology

- We propose a methodology to perform active dengue surveillance based on a combination of the four dimensions that are associated with Twitter data
  - Public perception
  - Volume
  - Location
  - Time
- The methodology is divided into the following parts
  - Content analysis
  - Correlation analysis
  - Spatio-temporal analysis
  - Surveillance

Content analysis

- Content analysis is employed in order to
  - provide important clues about the attitude associated with tweets mentioning dengue
  - reduce noisy for surveillance by focusing only on tweets that are related to dengue cases
- Classification techniques may be used to estimate sentiments expressed in tweets

Steps:

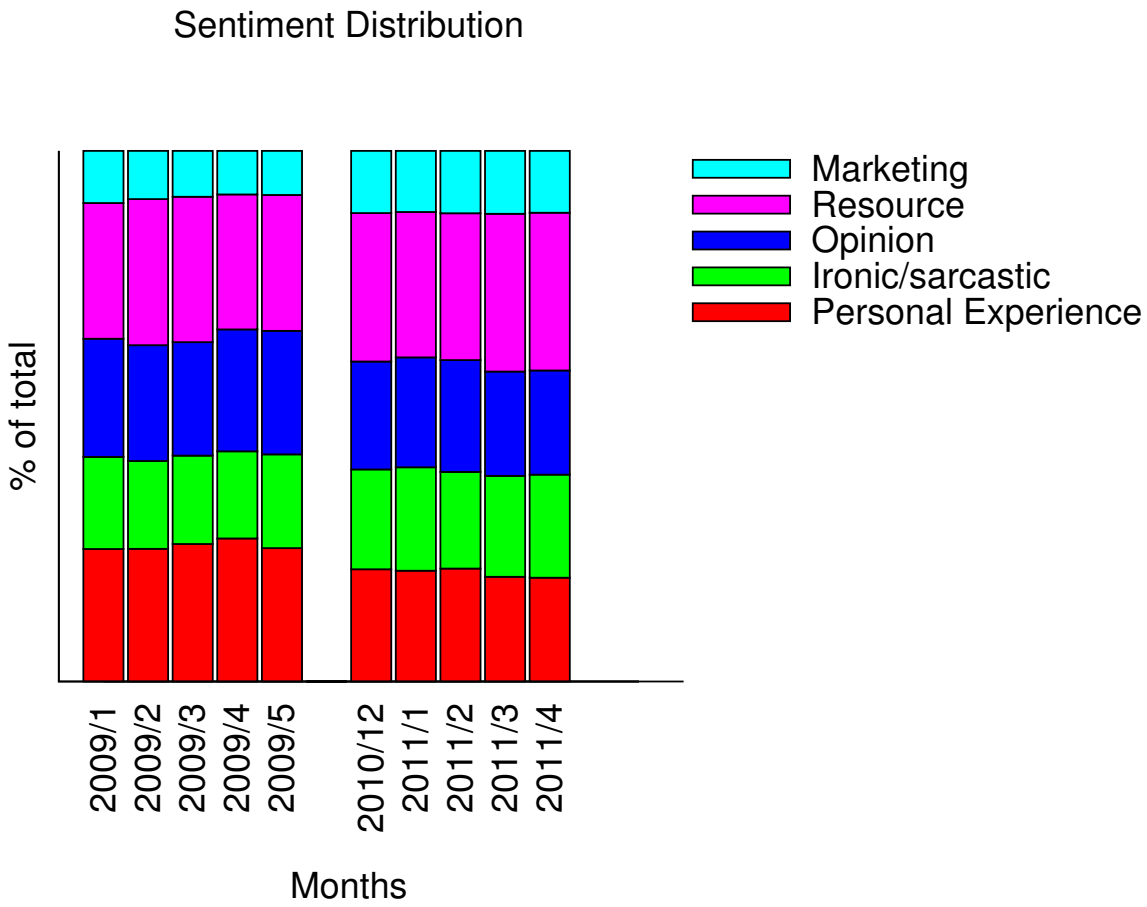
1. Specify the sentiment categories
2. Create a representative training dataset
3. Classify tweets

Content analysis

1. Specify the sentiment categories
   - Personal experience
   - Ironic/sarcastic tweets
   - Opinion
   - Resource
   - Marketing
2. Create a representative training dataset
   - A selective sampling strategy [J. Kivinen and H. Mannila PODS 1994] was carried in order to build a small, but representative training dataset
3. Classify tweets
   - An associative classifier (LAC) [A. Veloso *et. al.* ICDM 2006] produces a sentiment model, which is extracted from the training dataset
   - LAC provides a scoring function which estimates the likelihood of each sentiment being the implicit attitude of tweet

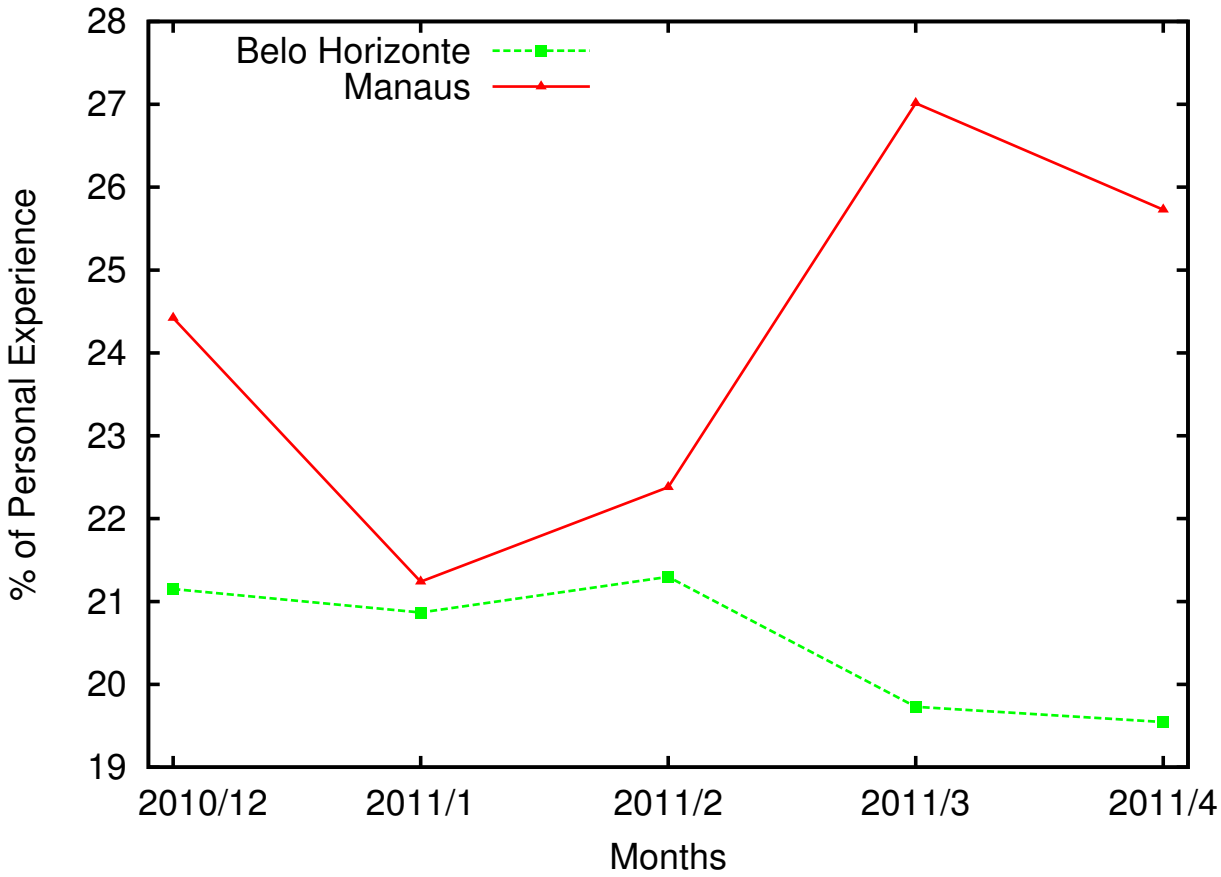
Content analysis

Sentiment distribution for Twitter datasets over time



Content analysis

Verify the discrepancy in the proportion of tweets expressing personal experience, coming from cities that actually showed diferent dengue incidence rates



Correlation analysis

- The correlation analysis enables us to understand whether activity on Twitter indeed reflects dengue incidence in terms of number of tweets refering dengue
- We fit a linear regression model that may approximate dengue incidence rates
- Our models are based on one of the following variables:
  1. The volume of tweets related to dengue, posted by Brazilian users ($\#tweets$)
  2. The ratio of tweets expressing personal experience, posted by Brazilian users ($PTPE$)

Correlation analysis

- The correlation analysis enables us to understand whether activity on Twitter indeed reflects dengue incidence in terms of number of tweets refering dengue
- We fit a linear regression model that may approximate dengue incidence rates
- Our models are based on one of the following variables:
  1. The volume of tweets related to dengue, posted by Brazilian users ($\#tweets$)
  2. The ratio of tweets expressing personal experience, posted by Brazilian users ($PTPE$)
- Linear regression models

$$\#cases_t = \beta_0 + \beta_1 \times \#tweets_t + \beta_2 \times \#tweets_{t-1} + \epsilon$$
$$\#cases_t = \beta_0 + \beta_1 \times PTPE_t + \beta_2 \times PTPE_{t-1} + \epsilon$$

Correlation analysis

- The correlation analysis enables us to understand whether activity on Twitter indeed reflects dengue incidence in terms of number of tweets refering dengue
- We fit a linear regression model that may approximate dengue incidence rates
- Our models are based on one of the following variables:
  1. The volume of tweets related to dengue, posted by Brazilian users ($\#tweets$)
  2. The ratio of tweets expressing personal experience, posted by Brazilian users ($PTPE$)
- Linear regression models

$$\#cases_t = \beta_0 + \beta_1 \times \#tweets_t + \beta_2 \times \#tweets_{t-1} + \epsilon \qquad R^2 = 0.7829$$

$$\#cases_t = \beta_0 + \beta_1 \times PTPE_t + \beta_2 \times PTPE_{t-1} + \epsilon \qquad R^2 = 0.9578$$

Spatio-temporal analysis

- We are interested in discovering groups of cities that are near each other and have similar dengue incidence rates at a given point of time
- This enables government agencies to concentrate efforts on critical locations in the right time
- ST-DBSCAN is a density-based clustering algorithm and is used for clustering spatial-temporal data

Steps:

1. Calculate the incidence rate associated with each city
2. Determine the input parameters for ST-DBSCAN and run ST-DBSCAN

Spatio-temporal analysis

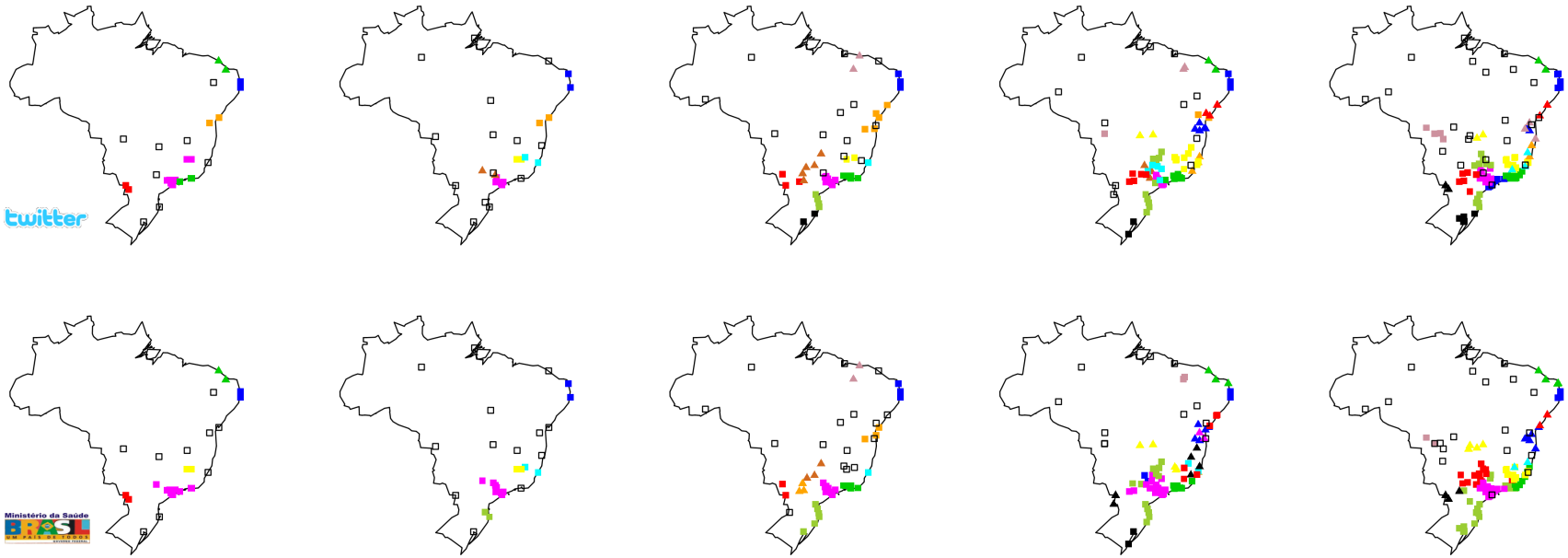
- To calculate the incidence associated with each city rate we used
  - the volume of tweets
  - the PTPE value

Spatio-temporal analysis

- To calculate the incidence associated with each city rate we used
  - the volume of tweets $\quad$ *Rand Index* $= 0.8506$
  - the PTPE value $\quad$ *Rand Index* $= 0.8914$

Spatio-temporal analysis

- To calculate the incidence associated with each city rate we used
  - the volume of tweets — *Rand Index* = 0.8506
  - the PTPE value — *Rand Index* = 0.8914

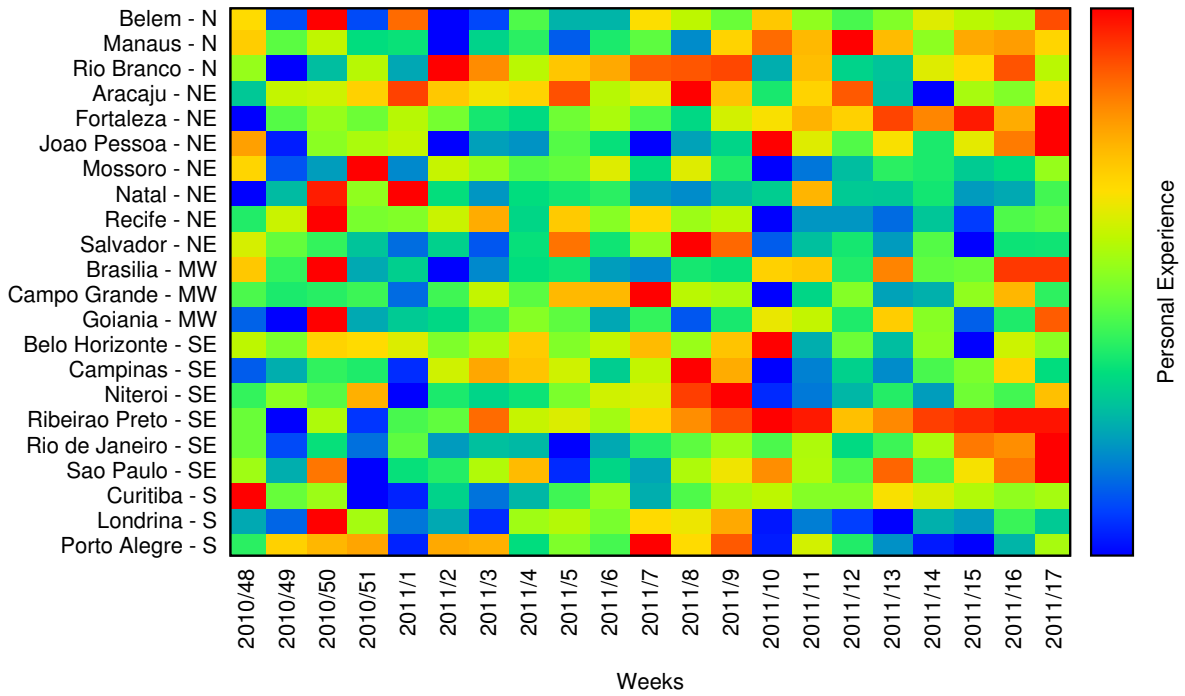


Surveillance

- We aim to analyze the proportion of tweets expressing personal experience in a weekly basis
- The intuition is that, an abrupt increase on PTPE may indicate outbreaks in the corresponding cities
- The visualization of surveillance system based on heat maps is able to capture variations on PTPE

Surveillance

- Brazil is divided into 26 states, which are grouped into five regions: N, NE, MW, SE and S
- About 68% of dengue cases notified concentrated in 7 states which are all represented below



Conclusions

We show the potential of Twitter data for the sake of dengue surveillance
We proposed a methodology based on four dimensions: volume, location, time and content

- We speculate how users refer to dengue in Twitter with sentiment analysis and use the result to focus only on tweets that express personal experience about dengue.
- We constructed a highly correlated linear regression model for predicting the number of dengue cases using the proportion of tweets expressing personal experience
- We showed that Twitter can be used to predict, spatially and temporally, dengue epidemics by means of clustering

Thank you!

Janaína Gomide
janaina@dcc.ufmg.br

observatóriodadengue

http://www.observatorio.inweb.org.br/dengue/