Future Generation Computer Systems 93 (2019) 290–303

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Subevents detection through topic modeling in social media posts

Diogo Nolasco [a,*], Jonice Oliveira [b]

[a] Programa de Pós-Graduação em Informática, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil
[b] Departamento de Ciência da Computação, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

HIGHLIGHTS

- A Method for detecting subevents within main complex events is proposed.
- Social sensors are used to detect subevents through social networks.
- Algorithms to represent subevent topics as labels are proposed.
- Proposed method works with many languages.
- Evaluation is made in health and urban events, with results suggesting the method versatility.

ARTICLE INFO

Article history:
Received 31 March 2018
Received in revised form 30 July 2018
Accepted 2 September 2018
Available online 12 September 2018

Keywords:
Topic labeling
Topic modeling
Subevent detection
Social networks
Unsupervised learning

ABSTRACT

Event detection has been a significant topic for a long time, since the onset development of pervasive systems. The ability to gather data from various sensors, in a diverse number of formats, is a challenge due to the continuous growth of data volume. Users of social media act as human sensors, providing data and information in real time about entities and events. Most of the research about event detection – using human or non-human sensors – concentrates only on identifying events. These models assume an event to be a single entity and ignoring that it can be composed of other new events over time. The detection of subevents enriches the understanding of the main event, contextualizing it and creating a powerful knowledge about the scenario. To capture the parts of an event and the information changing over time, we created a scalable and modular topic modeling based algorithm. It identifies subevents and creates labels to represent them more accurately. We evaluate the proposed sub-event detection approach using two large-scale Twitter corpus. The first one is related to Brazil's political protests scenario. The second analyzes the Zika Virus epidemic in the world. Our approach detected several subevents, most of them are related to real subevents. Due to the nature of social networks, with a minimum delay between an event occurrence and its dissemination, these results can open an opportunity for temporal tracking of emergence and outbreak scenarios.

1. Introduction

Social networks have been receiving much attention over the years, and the amount of data generated by them increases each day. Information included in its posts range from politics to sports, from history to breaking news and from private to public interests. Many applications use this data as a source of opinions, sentiments, and information about a given population, culture, and society. One of the many important challenges for these systems is the event and its associated subevents detection. These tasks are useful at analyzing real-time or past events and breaking those into components that can be further assessed.

There are a lot of types of social networks used for different objectives, such as video logs, professional, and interest networks are some examples used for sharing personal experiences, job contacts and common preferences, respectively. We can include the microblogging networks, which are particularly useful for event detection because of their real-time news-propagation nature.

Microblogging is a blogging approach consisting of short texts, photos, videos or audios sent by the users. Some examples include Twitter, Tumblr, Plurk, among others. The main objective is to let users share information in a fast and succinct way, usually leading to more and varied data inputs.

Particular characteristics that make microblogging suitable for event detection are the real-time updates and the presence of a mix of people, companies, and organizations in general. The real-time nature makes it possible to retrieve data from events happening right now. Microblogging users typically update frequently, so we

* Corresponding author.
E-mail addresses: diogo.sousa@ppgi.ufrj.br (D. Nolasco), jonice@dcc.ufrj.br (J. Oliveira).

https://doi.org/10.1016/j.future.2018.09.008
0167-739X/

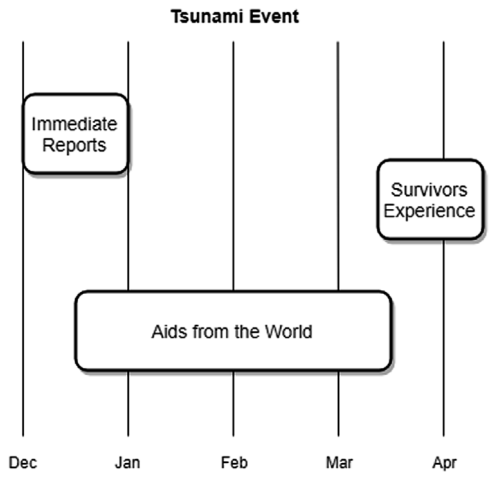

Fig. 1. An example of a complex event with its subevents.

can discover what people are doing or thinking in hours or minutes, differently from blog users who make updates of longer texts, daily.

The amount of updates in microblogging networks results in numerous reports related to events. These can include elections, a soccer game, protests or even emergency events such as a tsunami, accidents, riots, and fires. Events reported on Twitter, for example, demonstrated to have a minimum delay between occurrence and report [1,2].

The miscellaneous information provided for people from different backgrounds, knowledge, locations, companies, and organizations add more levels of details to events and particular subevents, helping on a better contextualization. They can even include different opinions and viewpoints related to the development of some event.

The approach of using humans as sensors, via online social networks, turns every person count as a sensor, a source of data for the environment and context. Users can detect and monitor events and their associated subevents. A subevent is a component of a complex event. An event as a tsunami disaster, for example, can have subevents such as the particular rescue needs, immediate reports, survivor's personal experience, food provision planning, aids from the world and others. Fig. 1 illustrates this event. In this figure, we can see the subevents starting with the initial reports of the main event. While the initial reports travel across the globe, many organizations, such as medical and financial ones, offer support to the victims (some subevents can occur in parallel while others occur sequentially). In the end, while repairing the damage done by this specific tsunami event the survivors began to report their experiences and the final results and statistics. This is a simplified example of a complex event, many other events have a higher amount of subevents; as well as subevents that causes other events. It should be noted however, that those subevents are initially only connected to the main event. An explosion as a result of a tsunami, could cause a blackout subevent while other subevents, like the death of fish, are not related to the explosion, nevertheless both are connected to the main event.

There is considerable research done in event detection using social media data [3,4]. Usually, these event detection algorithms use the variation of the distribution of words to identify an event. There are other ways of identifying events in social media, mainly through tags and classifications (e.g., hashtags and controlled labels). Regardless of the approach, most of them represent an event as a static entity. This representation assumes that an event does not change over time, which is not always true. Moreover, these mechanisms are not applicable to subevents. Microblogging networks rarely include tags for the subevents (as described in the tsunami example), giving preference to capture the main events or subjects. Furthermore, subevent information provided by users through text and tags are scattered across the network. Subevent entities are diverse depending on the location and user preference, and many times they are implicitly described in the text.

Our goal is to address the subevent detection in an unsupervised way so it can find unknown subevents while also producing explanatory labels so a user could identify the main aspects of the subevent without having to know beforehand their details. The main challenge in subevents detection is the similarity of vocabulary since every subevent shares some of the key terms present in the main event. They are also usually in a larger quantity and occurs in small timeframes when comparing to main events. We are not resorting to classification and supervised methods since social networks events are unpredictable. Also, we are concerned about giving an understanding of the event to the user without requiring a field specialist via phrases that can capture the thematic of the social discussions. This way, many events occurring could be deeply investigated at the level of subevents and be used to help in a variety of situations such as emergencies, epidemics, political and cultural movements.

This work focuses on the detection of these components that are present in the reports and conversations but are not detectable without human analysis. We propose a method capable of identifying subevents from main events and subsequently creating labels to represent them, thus facilitating its comprehension by a specialist for example (e.g., an emergency specialist in the tsunami scenario). Textual information from Twitter users are used as input in the experiments because of its data volume and popularity, then we use a topic modeling based algorithm for detection mixed with a topic labeling algorithm for representation of the subevents as concise labels. Two experiments were made in the political and public health scenarios. The first one was made with a series of subevents from political protests that occurred in Brazil in 2013. The second one related to the Zika virus epidemic that occurred between 2015 and 2016.

The main contributions of the method proposed include its viability when used with short and informal texts (A Tweeter post, for example, can have up to 140 characters and usually includes abbreviations, local terms, and specific words). Scalability, because the algorithms can be used with any amount of data with minimal additional processing. Language Independence, as the topic modeling and labeling can be used with any language and do not involve specific language feature processing. These characteristics can be useful for detection of subevents of different sizes with different amount of available data and for detection of global or international subevents that usually includes reports in many languages. It also differentiates itself from the related works by adding the novel labeling methods to display the subevents as entities that could be understood without analyzing the documents or the collection.

This paper is organized as follows: In the next section, we explain and compare related works to give an overview of the main approaches discussed in the literature. In Section 3 we give a brief explanation of the event definitions used and the human as a sensor approach along with a topic modeling background, followed by the proposal in Section 4. In Section 5, we describe the experiments and evaluation performed along with the corresponding discussion. Finally, in Section 6, we conclude with the final considerations.

2. Related work

Many works in the event detection area have been published in the last years, the majority of them focus on main events, how to differentiate an event and noise/false events and news.

The research of [1], for example, proposes a news processing system based on Twitter. They make a classifier to separate news

from irrelevant information and cluster them, although not making labels for the news. A similar work is presented by [5], where the authors track the breaking news through "#breakingnews" hashtags, indexing them to monitor their evolution and group them based on Named Entity Recognition. This method needs more processing and usefulness in large databases in debatable.

The distinction between events and common topics or subjects is addressed by [6], where they use a grouping and classifier technique to distinguish real-world events from topics reported that have no relation to any real-world occurrence.

Other works focus on the detection and monitoring through time as [3] that proposes the construction of wavelet signals based on individual words which could identify the beginning, end, and duration of an event. The same method is used in [7], but with the use of hashtags instead of individual words. For subevents, these methods have an obstacle for application due to the subevents being associated with the main event and thus sharing tags and terms.

The specified event detection has works as [4] that focus on detecting events that generate public discussion in social networks. The discovery is entity based, given a controversial figure (such as a president), their goal is to detect events that generate public discussions about this figure.

Musical concerts event detection is proposed by [8], using a graph model, artists information and event location to detect messages related to the specific event. Local Festivals detection is made by [9] via geotags, which are geographic information included in posts and that can be used to detect agglomeration and movement patterns.

Finally, query expansion is used by [10] to assess the usage of terms over time. In each interval, the usage of terms is computed and compared to the history database to know which terms are at a peak and so more suitable to be related to an event.

3. Definitions and background

Since our main objective is the subevent detection using microblogging data (specifically Twitter posts), we will define each post as a document. A collection of posts can be defined as $C = \{d1, d2, d3 \ldots\}$ where d1 is document number one, d2 the number two, and so on. The order of the documents can be retrieved anytime based on timestamp or document indexes and thus, there is no need for the collection to be ordered previously.

In our method and experiments, we will use Twitter posts, specifically, the text contained in the posts. Each text consists of a maximum of 140 characters, which presents a challenge for topic and subject extraction due to the data restriction. Links and tags are considered as "text" and are considered during the processing.

3.1. Event detection

Foremost, we will define an event as:

Definition 1. An event is a significant occurrence limited by time and with an associated location [11].

A subevent can be defined, in the same way, as an event associated with and dependent on a "main event". A subevent is part of an event. An event includes two or more subevents. Thus an event can be imagined as a chain of connected subevents.

Then, a subevent is defined as:

Definition 2. A subevent is an event associated with another event by a composition association.

Event detection can be classified into two types: Specified event detection and unspecified event detection.

Specified event detection is the identification of events which are previously known as planned events, such as a soccer game or a music concert, which have assigned dates and locations.

Unspecified event detection, on the other hand, is the identification of unknown events, which occur quickly and unexpectedly, such as the murder of a political figure or the beginning of an earthquake.

Subevents are often rarely anticipated so the majority of them would be unspecified. This characteristic makes it hard to predict the outcome with traditional trained algorithms or classifiers. Even specified events have unspecified subevents that occur as the main event develops.

This work goal concerning event detection is to detect and differentiate events coming from the same main event. The methods applied can be used in both specified and unspecified subevents, although by nature, the majority of subevents occur without a previous knowledge.

3.2. Humans as sensors

Sensor Networks typically consist of a set of sensors that sends data about environment parameters. The increasing growth in social network data suggests that the largest "sensor network" yet might be human. The use of social data and humans as sensor networks is called the "Human as a Sensor" (HaaS) paradigm [12]. In this approach, individuals are represented by sensors (data sources) who occasionally make observations about the physical world [13]. These observations may be true or false and hence are viewed as binary claims.

The reliability of the system is based on the amount or "correct" observations and it is usually assumed that with millions of users reporting, the amount of "correct" observations surpass by far the "incorrect" ones. However, in the case of human participants, not only is the reliability of sources usually unknown but also the original data provenance may be uncertain.

Humans may report false propaganda based on bias rather than their own experience. Individuals may report observations made by others as their own. Mechanisms like sharing and linking could induce this behavior and the advent of fake news and its reflections on events worldwide has been a subject of debate concerning social media reliability [12].

Despite these concerns, many studies have used social networks as sensor networks with success in tasks like reporting emergency scenarios and disaster recovery [14,15]. Other works displayed good results when analyzing the reliability of microblogging posts on disaster and sports scenarios [2,12].

The focus of this work is the external environment surrounding the human sensor. Social networks can carry some subjective information as posts about the user's feelings or isolated personal opinions. As such we want to detect only discussions and opinions regarding the subevent topic like the sentiment regarding a specific protest or an opinion about a political decision. These examples illustrate the main useful reports the user can send. Every external physical occurrence or state is thus reported by the sensors and as a set they communicate the social environment state. In this case, the physical environment has a unique state, leading to a unique ground truth, according to which these descriptions are either true or false (e.g., either there was a political discourse or not).

Specifically, in the event detection scenario, the user is a sensor of the event that occurs externally. If (s)he sends information about a tsunami occurrence, we consider that (s)he, as a "tsunami sensor" returns a positive value. The posts of this user are considered as sensor readings.

Definition 3. Each user or account in the online social network is considered as a sensor. A sensor can report events via sent posts.

Metaphorically, humans can be active or inactive sensors (if they are writing and sending information to the network or not), and they can be activated by different events. Humans as sensors are noisier than conventional physical sensors, although there is a large number of sensors worldwide.

Definition 4. Each post is considered as a sensor reading. A sensor reading is associated with a time and location.

Each input (post or sensor reading) has its timestamp and its location. In most cases, they are provided for local time zone, and GPS or IP, respectively. Sometimes, location is inaccurate.

The event detection problem can be reduced to spot an event from sensor readings. The task of estimating an object in time and space is a fundamental one in many pervasive computing scenarios [16].

3.3. Topic modeling

The modeling of topics, although it is a statistical method to discover topics in the structure of a corpus, is also seen as a fuzzy or soft "clusterization" [17]. Data clustering or clustering analysis is a multivariate data mining technique that, through numerical methods and only from the information present in the data, aims to automatically group the data of a collection into generally disjointed groups called clusters or groupings. It is considered an unsupervised learning technique that usually involves two basic parameters: N, a number of database items (e.g. documents) and K, the number of groups (e.g. the number of topics).

Unlike the concept of classification (supervised learning technique), "clustering" is a more "primitive" technique where there is no assumption about the groups. In the classification, there are predefined classes and through training with examples of execution, the algorithms "learn" how to allocate the data in each class, hence the name supervised learning. On the other hand, "clustering" does not know the existing classes in advance and does not have examples of how to distribute the data between groups, so it performs unsupervised learning. Unspecified events, for example, are difficult to be tracked or detected by supervised approaches.

The topics extracted by the modeling can then be seen as clusters and the data grouped as the items. In addition, clustering can be divided into two main types: hard clustering and soft clustering [18]. The first is the most usual where each case is associated with one and only one group. The latter, where topic modeling fits, can assign to each case one or more groups with different proportions (which in the case of topic modeling is represented by the probability of each group).

Thus, probabilistic topic modeling is an approach to tackle the problem of grouping and organizing data, mainly textual content, whose main objective is the discovery of topics and the annotation of large collections of documents by thematic classification. These methods quantitatively analyze the words of the original texts to discover the themes present in them. Topical modeling algorithms do not require any prior knowledge of the elements and the topics emerge from the analysis of the original texts [19].

In topic modeling, the Latent Dirichlet Allocation (LDA) model [19] is one of the most popular and served as the basis for the creation of many other probabilistic models. The foundations of the LDA model were based on LSA and pLSI (Probabilistic Latent Semantic Indexing, an evolution of the LSA with the use of probabilistic formulas [20,21].

In the case of topical modeling algorithms, the approach is based on creating a distribution of groups for each term of a textual document and a distribution of groups for each document. Based on these distributions one can group the documents according to the probabilities associated with each group.

The Latent Dirichlet Allocation (LDA) and other topic models are part of the broader field of probabilistic modeling. In this type of modeling, the data are treated as coming from a generative process that contains hidden variables. This process defines a joint probability distribution over observed and hidden random variables, which is used to compute the conditional distribution of the hidden variables given the observed variables. This conditional distribution is also called posterior distribution or simply "posterior". LDA fits into this framework. The variables observed are the words in the documents and the hidden variables are the outline. The computational problem of inferring the outline hidden from a set of documents is the problem of computing the later distribution — the conditional distribution of the hidden variables given the documents.

The generative process in LDA produces text documents and the manipulated data are the words that will form those documents. It is an imaginary process, from which the outline of a collection is obtained by inference from the inversion of that process. Technically, the template assumes that topics are generated before documents. A topic is defined as a probability distribution over a fixed vocabulary. As an example, a topic on genetics will be one that contains genetically related words most likely to occur. In contrast, a topic that relates to any other distinct subject will contain words about genetics with a very low probability of occurrence or zero. All topics contain distributions with probabilities over the entire fixed vocabulary, but these probabilities will only assume higher values in terms related to the topic.

The process that generates the documents in LDA is carried out in two stages. For the generation of each document of the collection, one has that:

1. A distribution on topics is chosen randomly. Example: In a template with only 3 topics, a possible topic distribution for a document A can display topics x, y, and z, respectively.
2. For each word in the document:
    a. A topic is chosen randomly from the distribution obtained in step 1.
    b. A word is chosen randomly from the topic (which is a probability distribution on vocabulary) obtained in 2a.
3. Each document displays topics in different proportions (step 1), each word in each document is obtained from one of the topics (step 2b), which in turn is chosen from the distribution on topics of a particular document (step 2a). This statistical model reflects the intuition that documents exhibit multiple topics, a presupposition that is behind the formulation of the LDA model.
4. The LDA model can also be described more formally by the following notation:
    a. Given the topics $\theta_{1:N}$, where each θ_n is a distribution over vocabulary V.
    b. The proportions of the topics for the dth document are ρ_d, where $\rho_{d,n}$ is the proportion of topic n in document d.
    c. The topic assignments for the dth document are z_d, where $z_{d,i}$ is the assignment of the topic to the ith word in document d.
    d. Finally, the words observed for document d are w_d, where $w_{d,i}$ is the ith word in document d, which is an element of vocabulary V.

With this notation, the generative process in LDA corresponds to the joint distribution of the observed and hidden variables represented by the expression:

$$p\left(\theta_{1:N}, \rho_{1:D}, z_{1:D}, w_{1:D}\right) = \prod_{j=1}^{N} p\left(\theta_d\right) \prod_{d=1}^{D} p\left(\rho_d\right) \left(\prod_{i=1}^{I} p(z_{d,i}|\rho_d) p(w_{d,i}|\theta_{1:N}, z_{d,i}) \right) \quad (1)$$

Although the LDA model is in the active research field in probabilistic topic modeling and allows the automatic segmentation of collections of thousands of documents, which otherwise would not be possible to achieve by human annotation, caution is required in the use and interpretation of the results obtained from this model. The topics and their distribution throughout the documents obtained from LDA modeling and other topic extraction models are not "definitive". Adjusting a topic template to a collection will always produce patterns from the corpus, even though they are not "naturally" present in the collection. Therefore it is important to use it in conjunction with other methods that clearly show the subjects present in the collection, as will be seen later with the labeling of the topics. Consequently, topic templates should be viewed as a useful tool for data mining, where topics provide a summary of the corpus that would be impossible to obtain manually. In any case, the analysis of a topic model can reveal connections between documents and within them that would not be obvious to the naked eye and may still encounter unexpected co-occurrences between terms.

With those previous definitions for LDA, given a vocabulary $V = w_1, w_2, \ldots, w_{|V|}$ consisting in all terms that exist in the collection C, we have the following definitions:

Definition 1. A topic model θ in C is a probability distribution of words such that $\theta = p\left(w_1|\theta\right), p\left(w_2|\theta\right), \ldots, p(w_{|V|}|\theta)$ and $\sum_{w \in V} p\left(w|\theta\right) = 1$. A "tsunami" topic, for example, would assign higher probabilities to the words "disaster", "nature" and "victims" and lower probabilities to uncorrelated words as "Marathon".

We use "events", "subevents", and "topics" as synonyms in the topic modeling scenario. The term "topic" is the jargon in this research area.

By this definition, a topic is a probabilistic distribution where words with high probability are more relevant to the subevent, while low probability words are irrelevant or stop words.

Definition 2. A label l for a topic model θ is a word or sequence of words that expresses the thematic of θ. We will be using words and phrases as labels. By this definition, it is possible to have more than one label for a topic model since synonyms would be valid labels for a topic model θ.

A topic model consists of a set of words with associated probabilities. Consequently, without a label, it is difficult to evaluate the detection because raw words can be misleading.

4. Proposal

The proposed method consists mainly of two big tasks: (i) Subevent Identification and (ii)Subevent Labeling. Given an event collection, (i) will extract subevents as topics from textual data and the (ii) will tag the subevent with representative labels.

In the next sections, we describe all the steps in each process.

4.1. Subevent identification

Given a collection C which contains documents from the online social network about a certain event, this process is responsible for the identification of subevents included and related to the main complex event.

In this work, we use a topic modeling algorithm to achieve this target. We use the Latent Dirichlet Allocation (LDA) [7] in the experiments and tests. LDA is a probabilistic topic modeling algorithm where each topic is represented as a multinomial distribution of words, according to its relevance to the aforementioned topic.

The algorithm is usually used in textual data due to its probabilistic nature making a dimensionality reduction. In traditional clustering algorithms, each term of the vocabulary is interpreted as a dimension, making data organization a difficult or inaccurate process. Consequently, there are latent and visible variables, represented by the subevents and words, respectively. Then, it assigns a probability to each word according to its relation to the topic.

Results show the high relevance of words in a particular topic with higher probabilities and common words with a low probability across all topics. Then, it is possible to identify a topic by its relevant words. A topic about a riot, for example, could have relevant words as "police", "protest" and "commotion" and can be identified by analyzing these relevant words as a set. Words like "a", "they" and "used" are expected to be irrelevant to all topics.

The primary parameter of topic modeling algorithms is the number of topics K. This parameter defines the number of topics to be extracted from the collection. The problem is demanding that the user needs to know this parameter previously because it is an input parameter. As topics here represents subevents, it is not possible to predict the number of subevents included in the collection. Typically, they will be unspecified events which could not be estimated, particularly in big data and streaming scenarios.

To solve this problem, we use a stability analysis approach for topic models presented by [22]. The stability analysis refers to the ability of an algorithm to replicate similar results from data originating from the same source.

This algorithm consists of taking samples from the collection and executing the topic model algorithm with these samples to get the parameter value that provides the most stable solutions.

For example, a collection of 100 documents could have potentially a minimum of 1 topic and a maximum of 100 topics for K. The algorithm will make small samples of the collection to disturb the data and see which K value produce stable results. In the end, the algorithm gives a stability score to each number of topics according to the probable value that is closest to representing the reality of the correct parameter.

In practice, the main input parameter of the algorithm is the range of values that we want to test for K. While in a collection of 100 documents we can theoretically test a range from 1 to 100, in practice is more plausible to evaluate a range between 10 and 60, for example. In real-world applications, we do not expect each tweet to be a new subevent and many tweets reporting a single subevent is much more likely.

After these steps, the output will be subevents represented by multinomial word distributions, which are the output of the LDA and topic models algorithms. The next challenging task is to represent the subevents contents in a comprehensible way to human users with concise and informative labels.

4.2. Subevent labeling

Given a set of topics representing the subevents of the collection, the final task is to create a comprehensible set of labels for each subevent. For this task, we extend the method proposed by the same authors in [23]. This work compares a lot of metrics for topic labeling in formal documents, such as scientific publications and patent reports. We will choose the best techniques as suggested by the authors and apply it to informal and short texts as the microblogging posts.

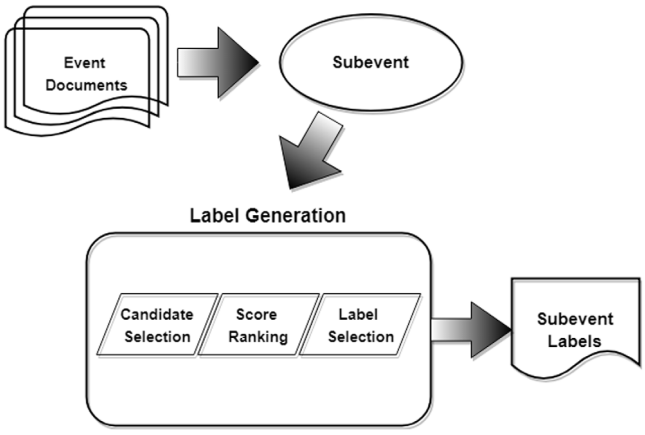

Fig. 2. Labeling Process.
Source:
(Adapted from [23]).

Algorithm 1 Candidate Selection algorithm

```
Input: D, number of top documents from θ, and W, number of top words from θ
Output: L, list of candidate labels

D' ← top D documents extracted from θ
W' ← top W words extracted from θ
for all d ∈ D' do
    P ← extract primitive labels from d
    for all p ∈ P do
        if p contains w |w ∈ W' then
            L ← L + p
return L
```

Fig. 3. Candidate Selection Algorithm.

The Fig. 2. shows the basic process of generating labels for subevents. The major subtasks are *Candidate selection*, *Score Ranking,* and *Label selection*.

4.2.1. Candidate selection

First of all, we need to extract and filter a list of candidate labels L for each subevent. We will use a sample on the collection's documents to filter the most relevant documents according to each subevent. This task is simple in topic modeling because – similar to words – each document can be represented by a distribution of words relative to each topic. Thus, we can eliminate noise from less relevant documents through a sample of documents from a topic.

Each document in the collection has a probability associated with each topic, which shows the document relevance to the given topic. The most relevant documents for a topic θ are those that have the highest associated probability with it. To avoid noise in L and to maintain the scalability of the algorithm in very large datasets, we take a sample of the documents in the collection based on this associated probability. Instead of using the entire collection, we use the top D documents of θ. Using this parameter D, we do not have to apply the algorithm to the entire collection. If necessary, we can increase the collection with more documents and the labels will only change if they belong to D. This characteristic makes this solution scalable to use with data-intensive environments and with frequently evolving sets.

After acquiring the samples, we extract initial labels from them. These primitive labels will be matched with the top W words of the multinomial distribution of θ (the list of words ranked by probability) to generate the candidate labels. The number of words W and the sample of D documents is the input parameters of the algorithm. Fig. 3. shows a formal description of the algorithm.

As a result, this step provides as candidate labels for θ, a list of words and phrases that match or contain some word from W. This helps in filtering common words, such as "with" or "choose" and ensures that words included in generated labels are relevant to the related topic.

The parameter W selects the most relevant words of a topic. Thus, the size of W will influence the number of candidate labels chosen. As the algorithm only parses the label and a subset of documents, the overall complexity of it remains the same of a typical LDA, because D is smaller than C, and L, W are smaller than V.

The sampling reasoning is that we can utilize the output of the topic modeling to catch only the most relevant terms and documents for generating labels as both already contain an associated probability with them.

The extraction of initial labels is done with an algorithm based on the fast keyword extraction algorithm [24], which in turn is based on the fact that labels frequently contain multiple words, but they rarely contain punctuation or stop words. The input of the algorithm is a list of stop words, phrase (punctuation) and word delimiters (spaces). All words or sequence of words among phrase delimiters and stopwords are considered as an initial label.

The algorithm provides a fast way to acquire initial labels. Moreover, it avoids the use of language and domain dependent features. Consequently, it becomes a generalist algorithm capable of extracting keywords in almost any kind of document. An example of the algorithm output is shown in Table 1.

Then, this output is used as a list L of candidate labels for each subevent.

4.2.2. Score ranking

To create scores to these labels, we use a metric proposed by the authors called Modified Label Degree [25], which is capable of balancing the weights of words and phrases through a mix of term frequency and label degree metrics.

Term Frequency (tf) usually gives higher scores to stopwords and non-descriptive terms when used in raw text As we are already filtering common words in the algorithm, tf will tend to give higher scores to words than phrases because words tend to have a higher frequency. It is formally defined as:

$$tf\ (t, d) = f_{t,d} \tag{2}$$

where t is a term, d a document, and $f_{t,d}$ the frequency of the term t in a document d. In this case, the "document" is the list of candidate labels.

The degree (deg) of a word in a collection C is defined as the sum of the frequency of the word in C and the frequency the word appears as a substring in another label. For a phrase, the degree is the sum of the degrees of its words. The label degree (ldeg) is the sum of the frequency of the entire label and the frequency it appears as a substring of another candidate label. Formally:

$$deg\ (w, d) = f_{w,d} + sf_{w,d} \tag{3}$$

$$deg\ (t, d) = \sum_{w \in t} deg(w, d) \tag{4}$$

$$ldeg\ (l, d) = f_{l,d} + sf_{l,d} \tag{5}$$

where w is a word, t a term (which can be a word or a phrase), and l a label (in this scenario a candidate label, but in general it is equivalent to a term). The component $sf_{w,d}$ represents the *substring frequency,* the number of times a word or term appears as a substring of another word or term in the document (In (5) l is the label and d the document). The document here is also the list of candidate labels.

These degree metrics tend to give higher scores to words as term frequencies because is easier for a word to appear as a substring of another label than a phrase of two or three words.

Table 1
Example of output from extraction algorithm.

Original Text	Governor Cabral says that he is going to transfer government helicopters #Cabral #OccupyTheStreets @EpochMagazine
Output	"Governor Cabral", "transfer government helicopters", "says", "he", "going", "#Cabral", "#OccupyTheStreets", "@EpochMagazine"

Algorithm 2 Label Selection algorithm

```
Input: L, list of ranked candidate labels. n, number of labels to be selected.
Output: L', set of selected candidate labels

L' ← n labels from L
for all i ∈ L' do
    for all j ∈ L' do
        if i ⊂ j then
            Remove i from L'
            L' ← L'+ label from L at position n + 1
return L'
```

Fig. 4. Label Selection Algorithm.

The Modified Label Degree (mdeg) then, gives one point for each label that appears as a substring of another candidate label and two points for every occurrence of the entire label. A formal notation would be:

$$mdeg(l, d) = ldeg(l, d) + 2 * tf(l, d) \quad (6)$$

where l is a candidate label and d a document represented by the set of candidate labels for a certain topic.

Comparing, for example, a "political" and "political protests" with a "political protests" label would give "political" a score of one and "political protests" a score of two. Reducing the importance of single words compared to the previous metrics but without ignoring them. A comparative example of metrics results is shown in Table 2. A complete comparison of the modified label degree with other metrics along with the results that show its efficiency in representing topic contents in a number of experiments can be found in [23].

When all labels have received a score and ranked by the modified label degree metric, the last step will be selecting representative labels from the list.

4.2.3. Label selection

The natural selection after having a list of labels ordered by a score function is to choose the top one to represent the subevent content. One question in this step is: "Is one label enough to describe a subevent or it can be doubtful?"

If the answer is "yes", we can take the first label as the most succinct description of the subevent. Otherwise, when more than one label is required, the additional labels need to represent additional aspects of the subevent. In this case, synonyms should not be used, for example.

For example, in a topic about "political protests" in a campaign scenario, a second label "political" adds nothing when compared to "political protests" and could be discarded without losing information. Other labels such as "human rights" or "law 466" could add more insight to the subevent description.

To solve this issue, we are comparing the selected labels eliminating the ones that prove to be a substring of the other. The next one in the ranking replaces it, and the process is repeated as many times as necessary. Fig. 4. shows a formal description of the algorithm. An example of the output of this algorithm is shown in Table 3.

In our experiments, we use a multilabel approach for visualization of the labels. A single label assessment can be done by just considering the first label in all sets. For a complete comparison of representations using a different number of labels, we refer to [23], that suggests the use of sets of labels increase comprehension when compared to single-label approaches.

5. Evaluation

Two experiments were made to evaluate the viability and efficiency of the method proposed for identifying subevents. We conduct the assessments using posts from Twitter databases related to the two events. We use our method to detect latent subevents described in posts and user communication, automatically. A temporal analysis is presented in each scenario to aid in the results interpretation, however, the mutual relationship between subevents is not simply a chain or a sequence of cause/consequence events but instead, a complex relationship of related and independent subevents, making an analysis of such events never simple.

The first experiment uses a database related to political protests that occurred in Brazil [26], the most significant protest in the history of this country and one of the most significant in Latin America.

The second one is related to the scenario of Zika virus epidemic from 2015 to 2016, which contains a variety of subevents like public counter-measures, propagation to various countries, as well as associated diseases and influence on 2016 Olympic Games organization.

5.1. The experiment planning

A quantitative analysis was used comparing subevents encountered by our method with those reported by official sources, such as newspapers and specialized news websites.

A boolean variable called *Relevance* was used. *Relevance* receives the value of 1 if official media notified the subevent, and 0 otherwise. More than one source should report the subevent and any hoax or misleading news receives 0 as relevance value. Formally:

$$Relevance\,(\theta) = 1\ if\ \theta \subset M;\ 0\ otherwise \quad (7)$$

where θ is the subevent represented by the corresponding topic and M is the set of official media news that is available to the public.

We will use α as the number of times the variable (7) takes the value 0 and β the number of times it takes 1. We assume that the experiment was successful if $\beta > \alpha$ is true.

5.2. Datasets

The two databases consist of posts from Twitter in text form extracted through the API provided by the service. Different approaches are used for extraction and treatment depending on characteristics of the scenario.

5.3. Topic extraction and labeling

As said in the proposal, the topic modeling algorithm needs an input parameter K, which will be automatically defined. To choose the best value of the parameter the algorithm requires a range of possible K values. For this range, we used 8 and 20 as the minimum and a maximum number of topics that could be present in the collection respectively.

Table 2
Example of Output of Different Ranking Metrics.

Set of Candidate Labels	Score for Candidate: "Governor Cabral"		
"Governor Cabral", "Governor Cabral Policies", "Politics Votes", "Governor Cabral developments" "#Cabral", "#OccupyTheStreets", "@EpochMagazine"	tf	ldeg	mdeg
	1	3	5

Table 3
Example of Output from Label Selection Algorithm.

Input Label Set	"Governor", "Governor Cabral", "government", "Cabral", "#Cabral", "#OccupyTheStreets"
Output	"Governor Cabral", "government", "#Cabral", "#OccupyTheStreets"

This range was chosen by using the typical cluster estimation for text databases in which the number of clusters for a document collection defined by a document-term matrix (of size m by n, m: number of documents, n: number of terms) can be roughly estimated by the following formula [27]:

$$mn/t \tag{8}$$

where t is the number of non-zero entries in the matrix.

Using this initial estimation, we made empirical tests with the objective of showing a typical range that could be used without affecting the output by merging topics or splitting topics erroneously, the most stable range. After tests with a maximum of 20, 30, 50, 100…1000 topics (the maximum approximate rounded result of (8)), we found that the range used was the smallest stable range when compared to the others tested (there were minimal differences when comparing results). It should be noted however that this range is used here only for exemplification, a real-world application could execute the algorithm with the full range or the range recommended by a specialist for example, and should produce equivalent results.

For the labeling algorithm, we used the top 10 documents and words for D and W parameters in the candidate selection algorithm. These values were recommended by [23] as they yielded the best results in their experiments comparing a range of values for the parameters.

5.4. The experiment execution — political protests database

The main event used in this case was the 2013 Brazil's protests, a complex event consisting of various popular demands. The manifestation began as a protest against public transportation prices and developed in a protest against government corruption and taxes. Such an event could provide many subevents which could help in the analysis of the components and nuances of the main event.

However, to understand the dataset, in this section we have made a description of the main events related to the protests. Brazil is a huge country with practically continental dimensions and is one of the 10 largest economies in the world. Moreover, it has a population of more than 200 million people, spread about in different national regions. Also, more than 102 million of its inhabitants access the Internet, among which 89% use smartphones to access social media [28], to share their opinion about different subjects [29].

In 2013, Brazilians complained about increases in public transport ticket prices, government corruption, among other things. Indeed, since 2012, Brazilians had marched against increases in public transport fares. Despite this, only in July 2013, the major media channels started to report events, because of the increasing of popular adoption to the protests. During the protests, there were excessive police force and repression, resulting in violent incidents. The population used the online social networks to organize the protests, disseminate information (especially about the brutal police action) and spread their opinion.

The extraction of the posts occurred during the protests in the main cities using the "hashtags" (tags provided by the platform): #acordabrasil, #vemprarua, #ForaFifa, #ogiganteacordou, #anonymousbrazil, #MPI, #passelivre, #pec37, #mudabrasil, #ChangeBrazil, #protesto, #foraDilma, #protestorj, #protestabrasil, #primaverabrasileira, #forafeliciano, #ocupa, #copapraquem, #protest, #pec33 e #pec99. Those could be translated as: #wakeupbrazil, #outtothestreets, #OutFifa, #thegiantwokeup, #anonymousbrazil, #MPI, #freepass, #pec37, #changebrazil, #ChangeBrazil, #protest, #outDilma. #protestrj, #protestbrazil, #brazilianspring, #outfeliciano, #occupy, #cupforwho, #protest, #pec33 and #pec99.

The tags were defined by the observation of the posts published throughout the event. The period considered for the event was between 06/01/2013 and 08/01/2013 . In total, 432.975 documents were retrieved. A preprocessing was done eliminating emoticons, links, and accents.

Again, this scenario consists of data from the big protests occurred in Brazil in 2013, which caused major political repercussions and had an expressive amount of popular participation.

As usual to protests and riots, the main event is not uniform. It is composed of a series of incidents and developments that changes the characteristics of the event as time passes and negotiations progress. As modernly organized protests, much of the process is organized in social networks along with continuous publicity of demands, agreements, and movements by both society and government.

Two different moments of the event were analyzed: (1) The first big national protest occurred in 06/17/2013 and (2) the second big protest occurred in 06/20/2013 .

News websites used for subevents validation were "Estadão", "Estado de Minas", "Terra", "EBC", "IG", "Veja", "Fox Sports", "Portal da Câmara dos Deputados" and "G1".

Concerning the two moments considered, first, in 06/17/2013, the Brazilian population organized public demonstrations via protests to advocate against increases in public transportation ticket prices, violence reduction, World Cup costs and poor quality of public services.

To have a broad view of this occurrence, we use three days of posts on online social networks: 16, 17 and 18 of the same month. The reason for this is so that we can observe the protest planning, execution, and repercussion.

The second moment analyzed was the second big national protest that took more than a million protesters to the streets against government corruption, World Cup and PEC37 (a bill that limits or prevent investigation of public agents). As before, we used three days of posts (19, 20 and 21 in this case).

Then, we can divide the main results into two periods as follows:

(1) 06/16/2013 to 06/18/2013 : In this period, 14 subevents were found by the algorithm. Table 4 shows the topic number, labels, corresponding news source and the relevance variable (7) value in the corresponding columns.

(2) 06/19/2013 to 06/21/2013 : In this period, 20 subevents were found. Table 5 shows the related results.

The evaluation shows some particular aspects of subevents through the labels. Topic 1 from Table 4 refers to a specific protest used in an old dictatorship period where people painted their faces to claim for democracy. Topics 0, 7 and 9 references musical parodies that occurred in the event. Topics 5 and 8 show occurrences of big protests in two of the country's biggest cities. According to the values assumed by the relevance variable (7), we have for this moment $\alpha = 1$ and $\beta = 13$.

From the topics in Table 5, we can cite topic 5 that shows the claims of president renounce. Topic 8 that evidences the violence occurrences during protests. Topics 2, 4 and 7 with different organizations of protests by different groups and topics 7 and 15 that shows occurrences in specific locations in Rio de Janeiro city during the main event. According to the relevance variable (7) we have $\alpha = 2$ and $\beta = 18$.

Within the subevents, we can encounter the expected components such as protests against corruption, specific laws and taxes, but at the same time can find unexpected connections and specific subevents like the ones previously mentioned.

Considering the two periods of the experiment, we a have a total of $\alpha = 3$ and $\beta = 31$. It means that for three times it was not possible to find related data in official media, while in the other 31 topics it was possible to create a connection with some specific news.

It is common, in topic model algorithms, to group irrelevant or miscellaneous minor topics in a single topic destined to contain the data that either did not fit in any topic or were not relevant enough to form a new topic. The labels of topic 13 in Table 4 and topic 10 in Table 5 suggest that this could be the case in these periods.

5.5. The experiment execution — Zika epidemic database

Since 2015, the Zika Virus has been a constant concern due to the unknown effects, its fast contagious and propagation, and the various types of transmission. With an unknown set of symptoms, transmission, and solution, its spreading is more rapid than other epidemic diseases.

In Brazil, Zika Virus (ZIKV) was identified for the first time in 2015. At that time, the Brazilian Northeast was faced with increasing cases of an unidentified disease, characterized by fever, conjunctivitis, rash and joint pain for seven days. When the Federal University of Bahia (UFBA) identified ZIKV, the disease manifestation was quiet, without complications. The disease spread rapidly throughout the country, having been recorded (from January to May 2016) 138,108 probable cases of Zika virus in the country (incidence rate of 67.6 cases/100,000 inhabitants) [30].

Brazilian health authorities detected misshapen babies who born with squashed skulls and neurological problems. On Feb. 1, the World Health Organization formally declared the outbreak of Zika a public health emergency of international concern [31]. Since then, Zika has been spreading worldwide to almost 100 countries.

As per [32], the virus, carried by Aedes aegypti mosquitoes, arose in Tahiti or French Polynesia, where an estimated 66 percent of the population was infected over roughly nine months. Polynesian athletes unwittingly carried the virus to Recife, Brazil, where they participated in the FIFA Confederations Cup soccer games. El Niño, which radically altered rainfall patterns across the Amazon region in 2015, is credited with causing a surge in mosquitoes across the area, fostering the Zika explosion. By this author, nearly all cases identified in the United States and Puerto Rico were related to a traveler who had acquired the virus in another country. Zika was spreading so rapidly in Puerto Rico that United States health officials predicted 25 percent of the population will be infected by August 2016 [32].

Based on these numbers, Zika became a huge international problem, but Brazil has been the most affected with this epidemic. In Brazil, one of the hardest areas hit by the ongoing outbreak, perhaps the most threatening aspect of Zika virus transmission has been its association with microcephaly, other abnormalities of the central nervous system and blindness in fetuses whose mothers were infected during pregnancy [33]. This is a problem with long-term consequences, which demands a restructuring of health systems, public politics, and laws. These children with a wide range of neurological and psychiatric issues require a new approach to their care, from their infancy until their maturity. We will need a new generation of health-care providers [34], including pediatricians and pediatric nurse practitioners. Some women are abandoned by their husbands [35], with lower chances to match a job and the special attention that their children require. As per [32], no tests administered during pregnancy, including sonograms and blood analyses, can definitively assure would-be parents whether their child has been infected, or identify damage to the fetus. This has forced the debate on contraceptive and abortion laws.

In adults, Zika is related to Guillain–Barré syndrome (a rapid-onset muscle weakness caused by the immune system).

Another problem is the vast possibility of contagion. Some researchers suspect the Zika Virus can be also transmitted by: (1) Sexual relationship [36], (2) Urine and Saliva [37], (3) gnat insect [38], (an ordinary mosquito, in Portuguese called 'pernilongo'), (4) blood transfusion [39] and (5) breast milk [39].

With a lot of unconfirmed possibilities, no available vaccines and partial solutions [40], social media has been used as the main alternative to exchange information and create knowledge.

The database for this experiment was made extracting posts from around the world with the #zika "hashtag". In this case, there is little ambiguity with the term (it is hard to confuse Zika with other unrelated relevant subjects). Other terms were not included to avoid introducing noise in the data as the main term is a rather distinctive one already.

The tag was also defined by observing the posts and popularity in different time spans. The time span of the database covers posts between 04/27/2015 and 12/30/2016. A total of 85,601 documents were retrieved. Again, the same preprocessing was done by removing emotes, links and accents from text.

This experiment uses the database related to the repercussion of the Zika epidemic event which started in 2015. Although this is not a new disease, it was believed to be isolated in some parts of Africa and Asia while unheard of in the American continent.

Some critical components of the event include the WHO declaration of the epidemic as a Public Health Emergency of International Concern as evidence grew that Zika can cause congenital disabilities as well as neurological problems. Another one is the discovery that men infected with Zika can transmit the virus to their sexual partners.

As social subevents, we could highlight the travel warnings issued by some countries, making the outbreak expected to reduce tourism significantly. Some States have taken the step of advising their citizens to delay pregnancy until more discoveries about the virus and its impact on fetal development.

International concern was raised regarding the safety of athletes and spectators at the 2016 Olympic Games held in Rio de Janeiro. Also, many subevents of public measures were made to control the epidemic, causing a decrease in the number of cases later in the same year.

These examples show how a complex event can have many subevents as it develops. Epidemic events in general usually behave in cycles of discovery, spread, counter-measures and post-epidemic actions. All these phases are an important source of subevents for analysis.

The outcome of epidemic-related subevents is also important for the population, making it a popular theme in people discussion

Table 4
Results from the first interval.

Topics	Labels	Corresponding News	Relevance
0	'a country', 'very funny', 'could protest' 'had', 'funny not'	http://lulacerda.ig.com.br/	1
1	'face', 'beautiful all', 'everybody', 'it's beautiful', 'protest so beautiful everybody with painted face screaming'	http://www.em.com.br/app/noticia/politica/2013/06/18/interna_politica,407569/onda-de-protestos-e-comparada-as-diretas-ja-e-aos-caras-pintadas.shtml	1
2	'#protestsp #cometothestreet', '#sp17j #medianinja', '#medianinja #freepass', '#sp17j sao', 'sao paulo'	http://g1.globo.com/brasil/noticia/2013/06/protestos-pelo-pais-tem-125-milhao-de-pessoas-um-morto-e-confrontos.html	1
3	'sao paulo', 'sao', 'sao paulo', 'rio de', 'de janeiro', 'rio de janeiro', '65 thousand in sao paulo '	http://g1.globo.com/brasil/noticia/2013/06/protestos-pelo-pais-tem-125-milhao-de-pessoas-um-morto-e-confrontos.html	1
4	'to the street', '#cometothestreets', 'brazil go', 'never seen', 'like never', 'seen before'	http://g1.globo.com/brasil/noticia/2013/06/protestos-pelo-pais-tem-125-milhao-de-pessoas-um-morto-e-confrontos.html	1
5	'human rights', 'needs', 'who needs', 'rights', 'commission of', '#outfeliciano #gaycure immediate renounce of marco feliciano from the presidency of the human rights commission', '#outfeliciano who needs cure', 'human rights commission of the senate approves gay cure', 'of human rights', 'human rights commission'	http://www1.folha.uol.com.br/poder/2013/06/1297075-proposta-sobre-cura-gay-e-aprovada-em-comissao-presidida-por-feliciano.shtml	1
6	'stuck in', 'in the throat', 'throat #cometothestreet', 'was stuck', '#cometothestreet', 'was stuck in the throat #cometothestreet, 'fight #changebrazil', 'brazilian people changed status from eternally lying in splendid crib to', '#changebrazil', 'fight #changebrazil', 'throat #cometothestreet'	http://g1.globo.com/brasil/noticia/2013/06/protestos-pelo-pais-tem-125-milhao-de-pessoas-um-morto-e-confrontos.html	1
7	'day will', 'but this', 'damn one', 'one day', 'this damn', 'its hard to believe for this damn will change one day', '#changebrazil', 'govern against me #willseethatasonofyoursdontrunawayfromafight #changebrazil', 'entire country in a whorehouse because like that we fain more money', 'soon this will change #protestrj'	http://g1.globo.com/brasil/noticia/2013/06/protestos-pelo-pais-tem-125-milhao-de-pessoas-um-morto-e-confrontos.html	1
8	'informations #protestrj'	http://g1.globo.com/brasil/noticia/2013/06/protestos-pelo-pais-tem-125-milhao-de-pessoas-um-morto-e-confrontos.html	1
9	'of the nation', 'future of', 'future of the nation', 'nation of generation', 'sons of' 'future of the nation of coca generation', 'we are bourgeois without religion'	http://g1.globo.com/sao-paulo/noticia/2013/06/protestos-em-sao-paulo-tem-parodia-de-funk-e-puxador-de-hinos.html	1
10	'maracana #protestrj', 'boa vista', 'sight cbn', '#protestrj good', '/6iptrqcbhp maracana', '/6iptrqcbhp maracana #protestrj boa vista cbn psdb', '/6iptrqcbhp maracana #protestrj boa vista cbn', '/cjiixbhimv maracana #protestrj cup psdb dem', 'maracana #protestrj good'	http://g1.globo.com/brasil/noticia/2013/06/protestos-pelo-pais-tem-125-milhao-de-pessoas-um-morto-e-confrontos.html	1
11	'tired of', 'so many', 'brazilian is tired', 'of so much corruption', 'brazilian people', 'brazilian people tired of so much corruption', '#changebrazil', 'it is like a carnival block', '@domrodrigocosta', 'everything'	http://g1.globo.com/brasil/noticia/2013/06/protestos-pelo-pais-tem-125-milhao-de-pessoas-um-morto-e-confrontos.html	1
12	'dont return politics', 'are outdated', '#laws are', '#notaxes ', 'it takes but not fail', 'takes fail', 'justice takes but not fails #laws are outdated #taxes don't return#politics #corrupts bleed', '#politics$steal', 'no return $#politics', 'return$#politics #corrupts'	http://g1.globo.com/brasil/noticia/2013/06/protestos-pelo-pais-tem-125-milhao-de-pessoas-um-morto-e-confrontos.html	1
13	'video @youtube', '@youtube of', 'a video', 'a video @youtube', 'video @youtube of', 'liked a video @youtube from @miticojovems2', 'added as favorite a video @youtube from @ancalado http', 'liked a video @youtube from @edublin http', '@youtube #allrevolutionstartwithaignite #changebrazil #suckdilma #thegiantwokeup', 'liked a video @youtube from @miticojovems2'	–	0

on social networks. Healthcare posts, public advertisement, information about symptoms and prevention are some of the expected conversations to be held in those networks.

In this experiment, were analyzed two perspectives of the event to understand method usability: (1) The subevents occurred internationally as the epidemic spreads to other countries and (2) The local subevents occurred in Brazil, where the disease became more dangerous while hosting at the time the 2016 Olympic Games, thus becoming a strategic place to combat the epidemic.

News websites used for subevents validation were: "Associated Press", "Reuters", "BBC", "Google News", "The Wall Street Journal", "G1", "El Pais". Like the previous experiment, quantitative analysis was used comparing subevents reported by news with those encountered by the method. Also, the same boolean variable called relevance (7) was used that receives a value of 1 if a report of the same subevent could be found in the news and 0 otherwise. We will use α as the number of times the variable takes the value 0 and β the number of times it takes 1, then, for the proposal to be effective $\beta > \alpha$ should be true.

Concerning the two perspectives analyzed, named international and local from now on to simplify results exhibition, both contain posts from 04/27/2015 and 12/30/2016 . The difference was the filtering made to retrieve posts from all locations (in the first case) and local posts from Brazil (second case).

Table 5
Results from second interval.

Topics	Labels	Corresponding News	Relevance
0	'human rights', 'no prejudice', 'elect a', 'in the commission', 'of rights', 'hours to elect a president without prejudice in the human rights commission in the senate', '#outfeliciano', 'president without prejudice'	http://g1.globo.com/pi/piaui/noticia/2013/06/opcao-sexual-nao-e-doenca-diz-manifestante-no-pi-sobre-cura-gay.html	1
1	'globo network', 'gaucho affiliated', 'media gaucho', 'media gaucho affiliated, 'globo has', 'group media gaucho affiliated with globo network has ideology revealed', 'media gaucho group affiliated with globo network', ' #sp18j brazil fortaleza medical act national forces #cometothestreets'	http://g1.globo.com/ceara/noticia/2013/06/estacionamento-de-hospital-e-invadido-em-protesto-em-fortaleza.html	1
2	'/6iptrqcbhp #cometothestreets', 'dangerous relationships', '#cometothestreets #protest', '#protest ifcs', '/dlhnvvw9io niteroi #cometothestreets globo network psdb pps mpl',	http://ultimosegundo.ig.com.br/brasil/2013-06-21/1-milhao-de-pessoas-vai-as-ruas-e-vandalismo-se-espalha-pelo-pais.html	1
3	'of jacobina', 'jacobina too', 'students of', 'go too', '#cometothestreets students', '#cometothestreets students from jacobina go too', 'streets today', '#cometothestreets', '#17h know more ', '#thegiantwokeup #politicalreformnow'	–	0
4	'#changebrazil', '#cometothestreets what', 'photo', 'guy #cometothestreets', 'what you said', 'what #cometothestreets dude', '#cometothestreets what', '#changebrazil #thegiantwokeup', '/xufvuxt8sz #changebrazil #thegiantwokeup'	http://g1.globo.com/ceara/noticia/2013/06/estacionamento-de-hospital-e-invadido-em-protesto-em-fortaleza.html	1
5	'#outdilma', '#shutupdilma #outdilma', 'of dilma', 'dilma was', 'was so', 'of dilma', 'a joke because if she goes out it is worse to brazil'	http://g1.globo.com/ceara/noticia/2013/06/estacionamento-de-hospital-e-invadido-em-protesto-em-fortaleza.html	1
6	'pride of', 'today i feel', 'truth for', 'feel pride', '#protestsbrazil #cometothestreets today i feel truly pride of being brazilian'	http://g1.globo.com/ceara/noticia/2013/06/estacionamento-de-hospital-e-invadido-em-protesto-em-fortaleza.html	1
7	'#cometothestreets', 'to the manifestation', 'being at', 'in protest', 'americas go', 'going to stay', 'americas av. is going to be in protest'	http://g1.globo.com/ceara/noticia/2013/06/estacionamento-de-hospital-e-invadido-em-protesto-em-fortaleza.html	1
8	'#protest', 406, '#protestrj manifestants', 'threw firecrackers', 'manifestants threw', 'firecrackers against', '#gocorinthians rt @oglobonewspaper #protestrj	http://g1.globo.com/ceara/noticia/2013/06/estacionamento-de-hospital-e-invadido-em-protesto-em-fortaleza.html	1
9	'cup of', 500, 'of confederations', 'confederations cup', 'for the cup', 125, 'tickets to', 'tickets to the confederations cup #cometothestreets',	https://www.foxsports.com.pe/news/107182-fifa-nega-cancelamento-da-copa-das-confederacoes	1
10	'a video', '@youtube of', 'video @youtube', 'of a', 'a video @youtube', 'a video @youtube', 'liked a vídeo from @youtube of @kaka_craft ',	–	0
11	'of until', 'daily of', 48, 'cup', '#cometothestreets', 'daily of until', '#cupforwho'	http://g1.globo.com/ceara/noticia/2013/06/estacionamento-de-hospital-e-invadido-em-protesto-em-fortaleza.html	1
12	'rio de', 'de janeiro', 'a Thousand people', 'rio de janeiro', '71 cities', 'de janeiro', 'more 71 cities aw #cometothestreets #protestbr'	http://g1.globo.com/ceara/noticia/2013/06/estacionamento-de-hospital-e-invadido-em-protesto-em-fortaleza.html	1
13	'following the', 'following day', 'to the protests', 'following to the protests', 'day following the', 'streets ', 'of manifestations', '/x6tq6xyaus #thegiantwokeup',	http://g1.globo.com/ceara/noticia/2013/06/estacionamento-de-hospital-e-invadido-em-protesto-em-fortaleza.html	1
14	'your son', 'do not', 'a son', 'do not run away', 'your son don't', 'your son do not run away from the fight', 'fight', '#cometothestreets', '#thegiantwokeup',	http://g1.globo.com/ceara/noticia/2013/06/estacionamento-de-hospital-e-invadido-em-protesto-em-fortaleza.html	1
15	'#protestrj', 'following in', 'belford roxo', '@radargreencoast @drylawrj', 'purple following', '#westshopp by #cetrio without manifestations here #protestrj	http://g1.globo.com/rio-de-janeiro/noticia/2013/06/caxias-e-belford-roxo-rj-anunciam-reducao-nas-tarifas-de-onibus.html	1
16	'petition now to protect the investigation powers', 'public ministery', '#pec37', '#pec37 #wakeupbrazil', 'no'	http://veja.abril.com.br/politica/presidente-da-camara-adia-votacao-da-pec-37/	1
17	'to the street', 'come to', '#cometothestreets', 'come to the street', 'street', 'rt @estadao', '#thegiantwokeup #cometothestreets', 'come to the street come #thegiantwokeup'	http://g1.globo.com/ceara/noticia/2013/06/estacionamento-de-hospital-e-invadido-em-protesto-em-fortaleza.html	1
18	'in sao', 'sao paulo', 'in sao paulo','by bus', 'reduction of', 'alckmin is going to announce bus tax reduction in sao paulo #geledes #freepass	http://g1.globo.com/ceara/noticia/2013/06/estacionamento-de-hospital-e-invadido-em-protesto-em-fortaleza.html	1
19	'#changebrazil', 'health', 'education', 'needs to increase', 'no', 'no need', 'no need to increase'	http://g1.globo.com/ceara/noticia/2013/06/estacionamento-de-hospital-e-invadido-em-protesto-em-fortaleza.html	1

Internationally, in the timeline covered by the database, occurs the initial Zika virus infection in Brazil issued by the Pan American Health Organization in May 2015. After that, the virus becomes endemic and spreads to several countries in Central and South America. In January 2016, the Center for Disease Control issued a travel warning for pregnant women traveling to regions where Zika virus was spreading. In February 2016, the World Health Organization (WHO) declared the virus and its suspected link to

congenital disabilities an international public health emergency. In the same month, Dallas County Health and Human Services report its first case of sexual transmission by the virus in the United States. Finally, in November 2016, WHO announced the end of the Zika epidemic.

Locally, after the initial case in May 2015, in November of the same year the first evidence of connections between Zika and microcephaly were detected. At the same time, many instances of Guillain–Barré syndrome were reported suggesting another association. In January 2016, local authorities presented a plan to try to prevent the spread of the Zika virus during the 2016 Summer Olympics in Rio. In the same month, international partnerships were made and local studies were published containing data about how to combat the epidemic and the mosquito vector. Finally, in February 2016 a national campaign started with the use of military forces to inform the population and to help in combating the mosquito.

The algorithm results are divided as follow:

(1) International: The algorithm applied in the whole database resulted in 14 subevents. Table 6 shows the topic number, labels, corresponding news source and the relevance variable (7) value in the corresponding columns.

(2) Local: In this case, 9 subevents were found. Table 7 shows the results.

Results show a substantial group of subevents in both cases. In the international perspective (Table 6), topic 0 contains general virus information targeted to the public knowledge. Topic 9 shows the first infection cases in many countries as the virus spreads, leading to WHO declaring it an international public health emergency as shown by topic 5. Topics 4 and 2 show the effects of the epidemic on the Olympic Games preparations and main government measures to combat the outbreak respectively. Topic 11 gives reports of vaccine development as a counter-measure that raised social network discussions.

A particularity of the method is in topic 8 that is a fusion of the topics 0 and 5 in the Spanish language. As the algorithm can be used with any language, it can group same topics of different locales.

Other subevents that generated social relevance across the network were not among the main expected subevents but rather a specific one like topics 3 and 6. The first is about a report of a connection between microcephaly and a pesticides company (Monsanto), denying Zika connection to the disease. The latter, regarding the outcome of Hurricane Matthew that hit North and Central American countries, which could worsen the epidemic situation.

Finally, topic 1 shows the information about the sexual transmission of the virus and topic 7 the spread of news alerting pregnant women to avoid or delay travels to active Zika zones. All data in the form of subevents could be related to a real subevent so the international perspective found $\alpha = 0$ and $\beta = 14$ in this case.

The local data from Table 7 shows some of the same important events as well as other specific ones related to certain locations. Topics 5 and 6 replicates the international subevents of the concerns for Olympic Games preparations and WHO declaration of a public health emergency. Topic 0 shows the repercussion of the first official case detected. Topics 1 and 2 the measures to combat the virus using transgenic mosquitoes and bacteria, as well as information regarding public agents visiting various cities to fight the virus vector respectively.

More specific subevents can be found in topics 7 and 8. The first is related to findings suggesting virus transmission by the saliva. The latter has a mix of information but also people sharing the best apps to help prevention and virus fight. Considering the total subevents that could also be found in the news, we have the values $\alpha = 1$ and $\beta = 8$ for the boolean variables.

Comparing with the main subevents of the epidemic event, most of them were able to be found by the algorithm. Exceptions were the WHO declaring the end of the Zika epidemic at the international level and the Guillain–Barré syndrome reported cases at the local level.

The end of an epidemic is not a specific event in time, but instead, it is the decreasing number of cases. Thus, the announcement could not attract social attention or people were much more concerned with the rise of the virus, overshadowing the topic. It could have been the reason for the Guillain–Barré syndrome receiving less attention than microcephaly cases, for example.

Considering the two perspectives addressed by the experiment, we a have a total of $\alpha = 1$ and $\beta = 23$. In comparison with the first experiment ("Political protests"), although similar, this one had better results. A reason could be that the first experiment did not have a clear hashtag for the main event, forcing the introduction of noise in the data when extracting posts by tags.

In general, using the variable values of the two experiments, results show that $\alpha << \beta$, only 4 times it was not possible to retrieve corresponding news in 57 topics. It also suggests that the applicability of automatic subevent detection in microblogging achieves good results even with limited data in each document, its informal nature and without processing the entire collection. The social relevance of subevents can also be addressed by the results when the main event is somewhat independent of people (like the second experiment, a virus epidemic). It can open new applications in big data and streaming environments as the samples used are not updated as frequently as the collection and in pervasive social systems, as the method suggests relations between subevents occurrences and society reports and concerns.

6. Conclusions

Social networks provide abundant data that can be used to assess from personal views to population sentiment. Microblogging networks, in particular, can be used as a sensor network where the humans are the sensors and the events happening in real-time can be tracked through user updates.

One way of using those networks is in the event detection, as users report and discuss real-time events and are frequently updating the status of the occurrences. As components of complex events, subevents usually do not receive the same attention by algorithms despite the fundamental importance of understanding the event development and assessing action courses.

This paper proposes a method that can automatically detect subevents given an event and create a representation of this subevent as a comprehensible label. We use topic modeling algorithms for event mining from raw text and topic labeling methods to assign representative labels to them.

We evaluate the relevance of the subevents extracted by comparing them with existing news in the media and results showed a good relevance score across all the subevents examined. It also indicates that even with the limiting factors of language and size it is still possible to process this kind of data and achieve good results without human interference. A temporal analysis is presented along with the results to facilitate comprehension because the events are complex in nature and so are the analysis made on them.

The main contributions of this work are:

- A method which combines various features to make possible the subevent extraction from the main events.
- A fast and scalable method suitable for use with a large amount of data, since the processing is done only in relevant documents and samples.
- Modularity, as the components of the method, can be changed depending on the goal.

Table 6
International Topics for Zika Epidemic.

Topics	Labels	Corresponding News	Relevance
0	neutralizing human antibodies prevent #zika virus #zikv replication, human protein ifitm3 blocks #zika virus replication, human fetal neural stem cells	https://www.sciencedaily.com/releases/2017/05/170504083012.htm via Google News	1
1	2016, transmission, #cdc, sexual, cdcgov	http://www.bbc.com/news/health-35757541	1
2	fight #zika #doyourjob @housegop @senategop #zikavirus, fight #zika virus ravaging fl, fighting #zika virus fails	https://www.wsj.com/articles/u-s-spending-bill-frees-up-1-1-billion-to-fight-zika-1475162490?mod=searchresults&page=1&pos=7	1
3	zika virus" — doctors expose monsanto linked pesticide, birth defect microcephaly, birth defect	https://www.sciencealert.com/argentinian-report-says-monsanto-linked-pesticide-is-to-blame-for-microcephaly-outbreak-not-zika via Google News	1
4	zika virus #zikainrio #zikavirus @rio2016_en, cancelling rio olympics due, skipping #2016olympics due	http://www.bbc.com/portuguese/noticias/2016/02/160128_zika_olimpiada_jp	1
5	world health organization director general declares #zika virus outbreak, world health organization declares spread, intl health regulations emergency committee	http://www.who.int/mediacentre/news/statements/2016/1st-emergency-committee-zika/en/ via Google News	1
6	#nc governor pat mccrory, dilemma, #miamibeach	https://www.reuters.com/article/us-usa-matthew-idUSKCN12807E	1
7	miami beach #zikavirus #zikazone #advisory #miamibeach, caution pregnant women advised, #miami #beach area	https://www.nytimes.com/2016/08/20/science/5-zika-cases-were-transmitted-in-miami-beach-florida-governor-says.html via Google News	1
8	prevent the #zika #zikavirus pandemic located as a global danger today @hijosdlakebuena, if you are pregnant redouble the care against the dengue mosquito, #zikavirus the zika virus is caused by the bite of	http://www.elmundo.es/salud/2016/02/01/56af91c946163f8f328b45d7.html via Google News	1
9	#cuba reports 1st #zika travel case, #breaking beijing reports 3rd case, chp confirms #zika virus case	http://www.ejinsight.com/20170427-chp-confirms-first-imported-case-of-zika-virus-this-year/ via Google News	1
10	asian zika virus mutated negatively &amp, zika virus mutated negatively &amp, zika virus mutated negatively	https://www.nbcnews.com/storyline/zika-virus-outbreak/zika-virus-mutation-may-explain-spread-study-n556741 via Google News	1
11	zika vaccine candidates #zika #zikavirus #cdc #nih #niaid #vaccines $gsk $sny, zika vaccine candidates #zika #zikavirus #cdc #nih #niaid #vaccines $sny $gsk, zika $nlnk #zika #zikavirus #vaccines #pharma #nih #cdc $sny $gsk $mrk	https://www.reuters.com/article/us-health-zika-who-idUSKCN0V61JB	1
12	mosquito repellent zika virus protection, 99 free ship	http://www.miamiherald.com/living/health-fitness/article93159537.html via Google News	1
13	zika virus spreads #zikavirus #automotive #india, zika virus spreads, risk low	https://www.cnn.com/2016/01/28/health/zika-virus-global-response/index.html via Google News	1

Table 7
Local Topics for Zika Epidemic.

Topics	Labels	Corresponding News	Relevance
0	case of, first case, zika virus	http://g1.globo.com/rio-de-janeiro/noticia/2015/06/rj-registra-oficialmente-o-primeiro-caso-de-zika.html	1
1	zika virus, population reduction, mosquitoes genetically modified by a British company	http://g1.globo.com/bemestar/noticia/2016/02/oms-defende-teste-com-mosquitos-transgenico-e-bacteria-contra-o-zika.html	1
2	agents against, virus are, agents, #zika virus are working in the neighborhoods	http://g1.globo.com/pa/santarem-regiao/noticia/2015/11/agentes-intensificam-combate-zika-dengue-e-chikungunya-em-santarem.html	1
3	#zika virus, #vooz scientists, #vooz #zikavirus, #database-based solutions for #zika virus	https://veja.abril.com.br/saude/casos-de-zika-no-brasil-ja-chegam-a-165-932-em-2016/	1
4	to protect, to know, to be	–	0
5	suspended for, from Rio, be suspended	http://www.brazilianvoice.com/bv_noticias/zika-especialista-diz-que-olimpiadas-no-rio-devem-ser-suspensas.html via Google News	1
6	zika virus, about zika, per outbreak, #oms	http://g1.globo.com/bemestar/noticia/2016/02/zika-e-emergencia-de-saude-publica-internacional-declara-oms.html	1
7	can be transmitted by, #zika # virus can be transmitted by #breast #milk, #zika virus be found actively in saliva	http://www.bbc.com/portuguese/noticias/2016/02/160205_zika_saliva_jp_fd	1
8	indicates the best, best apps, google indicates the best apps for, zika virus #dengue #zikavirus	http://idgnow.com.br/mobilidade/2016/02/11/google-indica-os-melhores-apps-para-combater-a-dengue-e-o-zika-virus/ via Google News	1

- Multilanguage for extraction and labeling since we only depend on a stop word list and these can be generated for most languages, even automatically

Even achieving good results, there is a gap for improving subevent association and representation. Some research questions that arise are: "A subevent can disassociate from its main event

as time passes?" or "How to represent a subevent more closely to how a human analyst would do?"

The use of temporal information is limited to the evaluation and analysis of the results and could be improved by using the methods on real-time streams of text, detecting particularities of a complex event while it is occurring.

Future works could include the use of the links and external information to make better representations, such as news and user's social relevance. Another future work is the identification of lifespan and life cycle of subevents. Finally, an online topic modeling algorithm could be used to detect real-time subevents and to automatically create a history of complex events developments.

Acknowledgements

We would like to thank CAPES, CNPq and FAPERJ for the support.

References

[1] J. Sankaranarayanan, H. Samet, B.E. Teitler, M.D. Lieberman, J. Sperling, TwitterStand: News in Tweets, in: Proc. 17th ACM SIGSPATIAL Int. Conf. Adv. Geogr. Inf. Syst. - GIS '09, 2009, p. 42.
[2] S. Zhao, L. Zhong, J. Wickramasuriya, V. Vasudevan, Human as real-time sensors of social and physical events: A case study of twitter and sports games, no. june 2011, 2011, p. 9.
[3] J. Weng, B. Lee, Event detection in Twitter, in: Fifth Int. AAAI Conf. Weblogs Soc. Media, no. 98, 2011, pp. 401–408.
[4] A. Popescu, M. Pennacchiotti, Detecting controversial events from twitter, in: Proc. 19th ACM Int. Conf. Inf. Knowl. Manag. - CIKM '10, 2010, P. 1873.
[5] S. Phuvipadawat, T. Murata, Breaking news detection and tracking in Twitter, in: 2010 IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol., 2010, p. 120–123.
[6] H. Becker, M. Naaman, L. Gravano, Beyond trending topics: Real-world event identification on Twitter, p. CUS-012-11, 2011.
[7] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (4–5) (2003) 993–1022.
[8] E. Benson, A. Haghighi, R. Barzilay, Event discovery in social media feeds, Artificial Intelligence 3 (2–3) (2011) 389–398.
[9] R. Lee, K. Sumiya, Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection, in: Proc. 2nd ACM SIGSPATIAL Int. Work. Locat. Based Soc. Networks, 2010, pp. 1–10.
[10] D. Metzler, C. Cai, E. Hovy, Structured event retrieval over microblog archives, in: Proc. 2012 Conf. North …, 2012, pp. 646–655.
[11] A. Boettcher, D. Lee, EventRadar: A real-time local event detection scheme using Twitter stream, in: 2012 IEEE Int. Conf. Green Comput. Commun., Nov. 2012, pp. 358–367.
[12] D. Wang, M. Amin, S. Li, et al., Using humans as sensors: an estimation-theoretic perspective, in: IPSN 2014 - Proceedings of the 13th International Symposium on Information Processing in Sensor Networks, 2011, pp. 35–46.
[13] A. Wolisz, M.T. ACM Digital Library, S. Association for Computing Machinery. Special Interest Group on Embedded Systems, T. Abdelzaher, L. Kaplan, S. Gu, C. Pan, H. Liu, C.C. Aggarwal, R. Ganti, X. Wang, B. Mohapatra, H. Le, Proceedings of the 13th international symposium on Information processing in sensor networks, IEEE Press, 2014.
[14] B. Takahashi, E.C. Tandoc, C. Carmichael, Communicating on twitter during a disaster: An analysis of tweets during Typhoon Haiyan in the Philippines, Comput. Human Behav. 50 (2015) (2015) 392–398.
[15] A.L. Hughes, Twitter Adoption and Use in Mass Convergence and Emergency Events, no. May, 2009.
[16] V. Fox, J. Hightower, L. Liao, D. Schulz, Bayesian filtering for location estimation, IEEE Pervasive (2003).
[17] J. de Oliveira, W. Pedryczs, Advances in fuzzy clustering and its applications, 2007.
[18] P. Arabie, L. Hubert, An overview of combinatorial data, Clust. Classif. (1996).
[19] D.M. Blei, Probabilistic topic models, Commun. ACM 55 (4) (2012) 77.
[20] M. Steyvers, T. Griffiths, Probabilistic topic models, Handb. latent Semant. Anal. (2007).
[21] D. Blei, J. Lafferty, Topic models, Classif. Clust. Appl. (2009).
[22] D. Greene, D. O'Callaghan, P. Cunningham, How many topics? Stability analysis for topic models, Mach. Learn. Knowl. Discov. Databases (2014).
[23] D. Nolasco, J. Oliveira, Detecting knowledge innovation through automatic topic labeling on scholar data, in: Proceedings of the Annual Hawaii International Conference on System Sciences, 2016–March, 2016, pp. 358–367.
[24] M.W.J.K. Berry, Text Mining Applications and Theory, John Wiley & Sons, West Sussex, UK, 2010.
[25] D. Nolasco, Automatic Research Areas Identification in S & T, in: UFRJ, 2016.
[26] B. Lauand, Contextualization of Social Media Information for Use in Emergency Situations, nUiversidade Federal Do Rio De Janeiro, 2016.
[27] F. Can, E.A. Ozkarahan, Concepts and effectiveness of the cover-coefficient-based clustering methodology for text databases, ACM Trans. Database Syst. 15 (4) (1990) 483–517.
[28] P. Brasil, Research reveals that more than 100 millions of Brazilians have internet access 2016. [Online]. Available: http://www.brasil.gov.br/ciencia-e-tecnologia/2016/09/pesquisa-revela-que-mais-de-100-milhoes-de-brasileiros-acessam-a-internet. (Accessed 01 March 2018).
[29] A. Monroy-Hernandez, E. Spiro, How is the Brazilian Uprising Using Twitter? 2013. [Online]. Available: http://blogs.harvard.edu/andresmh/2013/07/how-is-the-brazilian-uprising-using-twitter/ (Accessed 04 February 2018).
[30] P. da Saúde, Epidemic Situation / Zika Data, 2016. [Online]. Available: http://u.saude.gov.br/index.php/situacao-epidemiologica-dados-zika (Accessed 01 March 2018).
[31] W.H. Organization, WHO Director-General summarizes the outcome of the Emergency Committee regarding clusters of microcephaly and Guillain-Barré syndrome, 2016. [Online]. Available: http://www.who.int/mediacentre/news/statements/2016/emergency-committee-zika-microcephaly/en. (Accessed 01 March 2018).
[32] D. McNeil, Zika: The Emerging Epidemic, WW Norton & Company, 2016.
[33] BBC, Concern Zika causes baby eye problems, 2016. [Online]. Available: http://www.bbc.com/news/health-36366946?SThisFB. (Accessed 01 March 2018).
[34] B. Berkrot, Zika's impact on children will require new medical approach: expert, 2016. [Online]. Available: http://www.reuters.com/article/us-health-zika-children-idUSKCN0Z61S6.
[35] F. Resk, Mens abandon mothers of babies with microcephaly in PE 2016. [Online]. Available: http://saude.estadao.com.br/noticias/geral,homens-abandonam-maes-de-bebes-com-microcefalia-em-pe,10000014877. (Accessed 01 March 2018).
[36] D.G. McNeil, Sex May Spread Zika Virus More Often Than Researchers Suspected, 2016. [Online]. Available: https://mobile.nytimes.com/2016/07/05/health/zika-virus-sex-spread.html?smid=tw-nytimes&smtyp=cur&referer= (Accessed 01 March 2018).
[37] H. Coelho, Fiocruz detects the zika virus with infection potential by saliva and urine 2016. [Online]. Available: http://g1.globo.com/bemestar/noticia/2016/02/fiocruz-detecta-virus-zika-com-potencial-de-infeccao-em-saliva-e-urina.html. (Accessed 01 March 2018).
[38] M. Felix, Gnat can transmit zika virus, a new study reveals 2017. [Online]. Available: https://veja.abril.com.br/saude/pernilongo-pode-transmitir-o-virus-zika-revela-novo-estudo/. (Accessed 01 March 2018).
[39] M. Pagan, Zika virus could be transmitted by 4 other ways besides mosquitos, 2016. [Online]. Available: http://www.vix.com/pt/bdm/saude/zika-virus-pode-ser-transmitido-de-mais-4-formas-alem-de-picada-de-mosquito-conheca. (Accessed 01 March 2018).
[40] A. Dollinger, Devouring 1000 Mosquitoes an Hour, Bats Are Now Welcome Guests as Zika Fears Rise, 2016. [Online]. Available: http://www.nytimes.com/2016/07/05/nyregion/devouring-1000-mosquitoes-an-hour-bats-are-now-welcome-guests-as-zika-fears-rise.html?smid=tw-share. (Accessed 01 March 2018).

Diogo Nolasco is a researcher in Data Mining and Big Data areas. Master in Informatics from Federal University of Rio de Janeiro (UFRJ), he actively collaborates with Prof. Jonice Oliveira research group. Acts in the Big Scholar Data and Social Networks research areas, working in the temporal detection, representation, and correlation between research fields and topics focusing on technological innovation, epidemic scenarios and urban problems. His research has been applied in healthcare and smart cities fields of study as well, and he currently works with automatic detection and tracking in those fields, finding correlations between events and topics of social relevance along with dynamic analysis over time. His research interests include databases, non-supervised learning methods, big data, social IoT, and data mining.

Jonice Oliveira is a professor of computer science, social networks analysis, data mining and data science at the Federal University of Rio de Janeiro (UFRJ). She coordinates the Postgraduate Program in Informatics (PPGI) at the same university. Her research interests include Big Data and Social Network Analysis areas, trying to solving real-world problems, with a special interest in crowd dynamics at large scale events, urban problems and how the analysis of social networks can help in promoting wellness in neglected populations. She has published over 280 journal and conference papers, and during her academic life, she was honored with some national and international awards (IBM Ph.D. Fellowship Award, PQ2-CNPq, Young Scientist of Our State-FAPERJ, etc.). Currently, she is participating in projects (as P.I. or as a member) related to the use of social media and social IoT to better understand cities and their citizens, especially to deal with urban problems as emergence, epidemics, and traffic. More information: http://www.joniceoliveira.net/.