

Publish or Post: Identification of influences between science and society through intelligent systems

Diogo Nolasco¹ and Jonice Oliveira²

Universidade Federal do Rio de Janeiro, Rio de Janeiro RJ, Brazil

¹diogo.sousa@ppgi.ufrj.br, ²jonice@dcc.ufrj.br

Abstract. Every social community is deeply influenced by scientific discoveries and technology. Research results have impacted our lives directly, such as the cure of diseases and the development of new devices. The interrelationship of the academy and society remains a mystery, despite these influences. How scientific works impact and are recognized by society? Do research works match societal demands? Trying to answer these questions, we create a system that is capable of generating links between scientific and social data. We use the scientific articles as “science sensors” and online social networks as “social sensors”. Topic modeling algorithms enable us to detect and to link main research themes and social events. The proposed system uses heterogeneous sources and can be applied in a variety of scenarios. We evaluate our environment in the Zika domain, using a large-scale Twitter corpus combined with PubMed articles. Our approach detected links among various subevents, suggesting that some degree of the scientific impacts in society can be automatically inferred. Results can open new opportunities for identifying the social consequences and reactions produced by scientific discoveries.

Keywords: Topic Modeling, Social Networks, Topic Labeling, Event Detection.

1 Introduction

Despite the fact that science has a significant impact on society, there is a large gap between scientific communications and general public perception, such as two isolated universes. However, the mutual influence is evident in public discussions and scholar conferences alike. A real case, as a new disease, is a new demand to scientists, whose efforts can produce a treatment. Consequently, all scientific actions and results generate news and discussions in public spaces. The same cycle repeats for discoveries of different magnitudes, such as new devices and systems, and the discovery of new exoplanets or physical particles.

Not all scientific work generate or receive a substantial social influence. A mathematical model or problem solution cannot have direct impact in real physical world. The sight that scientists should work only on social applications is called Baconian model. Otherwise, the Newtonian model states that scientists should conduct research with little concern for practical applications [1]. These two views show different aspects

Copyright held by the author(s)

of the scholar production with different degrees of public interaction depending on the nature and current relevance of a topic.

The relationship between science and society – or the scientific impact in society, and vice versa - is a theme that has rarely been addressed by Computer Science [2]. The main attempts were the use of citation networks, using articles and patents [3]. Recently, the use of altmetrics on the most important scientific digital libraries emphasizes the increasing importance of social influence and knowledge communication [2]. Altmetrics are non-traditional metrics proposed as an alternative or complement to more traditional citation impact metrics, such as h-index or impact factor [4]. It can include citations (on non-scientific publications, e.g. blogs or newspapers), number of views, number of downloads, social media comments and posts, reactions (e.g., likes/dislikes) and bookmarks. The open science and social media increase the range and influence of scientific works. Consequently, the adoption of altmetrics is irreversible.

The social features embodied on altmetric-based systems are still limited, focused on an article or an author. Mostly, only measuring how much a work (or a researcher) is cited in social media, but not the social influence of the research. Influence acts in both directions, and it is worth to consider the interrelationship between them.

This work proposes an integrated system that identifies the latent topics in a scholar dataset and those discussed in social media. From this, we try to identify the relationship among them, connecting social events, new scientific discoveries, and the influences and impacts of communication between science and society. Heretofore, this communication has been mediated only by official news sources. With the adoption of social features by many researchers and scientific venues, we have an increasing of social engagement in the public opinion.

This is a topic-modeling based approach, using the LDA algorithm [5]. Topic Modeling is a suite of algorithms for discovering the main subjects from a large collection of unstructured documents [6]. The same approach is used to process scientific and social data. The connection among scientific and social topics is done through similarity measures in variable time windows. To evaluate a particular event, the time dynamic is very important because the relevance of a science topic can increase/decrease due to variance of its social importance. The method used is capable of representing topics as a set of comprehensible labels via topic labeling which can be used to easily analyze the scenario even for users who are not familiar with the domain.

An experiment was made by using the Zika epidemic occurred between 2015 and 2016, where scientific development and social repercussion were notorious. This evaluation used microblogging posts extracted from Twitter as the social dataset, and PubMed publications about Zika to extract scientific topics.

Main contribution of our system is the possibility of use of heterogeneous data sources (scholar, social and possibly other like technological or media) simultaneously. Another contribution is the possibility of use of different languages. Finally, the assessment made by comparing science and public interaction over time can be used and repeated for different scenarios and events.

This paper is organized as follows: Section 2 describes the proposal in detail. Section 3 shows the application of the system in Zika epidemic scenario as evaluation. Section 4 presents the related works and Section 5 concludes with the final considerations.

2 Proposal

Our proposal is to extract the research and social topics, identifying links among them. These links could be causal (i.e., a new research topic which causes social commotion) or relational (i.e., discussions about an ongoing research that feedback the research itself). The Fig. 1 illustrates an example of these relationships. There are topics appearing overtime on scholar and public domains, both related to a new disease and the links show how social subevents and research areas are dynamically interacting with each other.

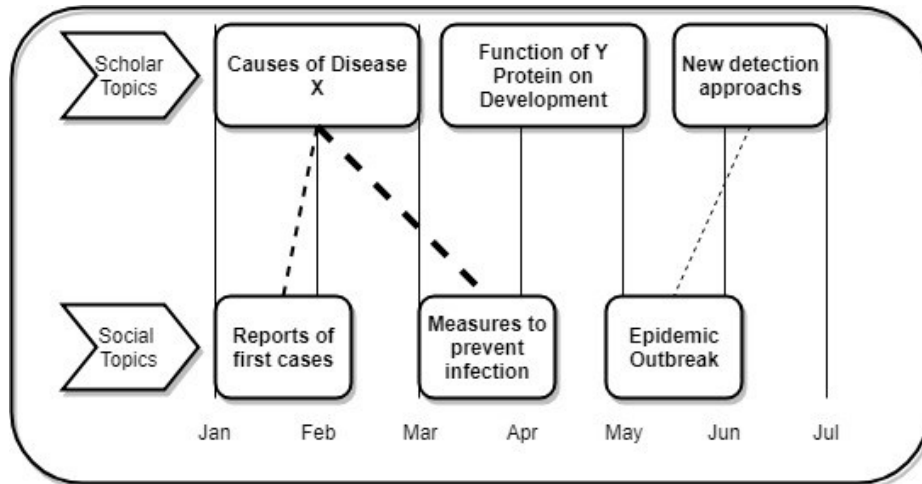


Fig. 1. Example of development of a new disease X and the repercussions on scientific and social domains, the dotted lines shows the links with thickness representing link strength.

Our proposal can be separated into two different tasks: (i) Topic extraction and (ii) Topic Labeling. For the tasks, we apply methods described in [7] as they proved successful with both scholar and social networks data [8]. Each task is described in detail in the next sections.

2.1 Topic Extraction

This task comprises the discovery of subjects in a given collection of documents C . In the academy domain, they will be research fields or topics of interest. In social domain, they will be discussions, opinions, and information dissemination.

We use a topic modeling algorithm to complete this task in this work. Specifically, we will use the Latent Dirichlet Allocation (LDA) [5] algorithm 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 said topic.

The algorithm is usually used in textual data due to its probabilistic nature avoiding the problem of the high number of dimensions included in texts. In traditional clustering

algorithms, each term of the vocabulary is interpreted as a dimension, making data organization a difficult or inaccurate process. The topic models give each word a probability which acts as a measurement of the relevance of a term to the topic, avoiding this problem.

Results show most relevant words on 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 virus, for example, could have relevant words as “vaccine”, “medical” and “treatment” 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 as they appear in most documents.

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 that the user needs to know this previously because it is an input parameter. Social topics cannot be predicted even by specialists, so this feature becomes a problem in this scenario.

To solve this problem, we use a stability analysis approach for topic models presented by [9]. 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 most stable solutions.

For example, a collection of 100 documents could have the K parameter minimum value at 1 topic (all documents belonging to the same topic) and maximum at 100 topics (each document as a different topic). The algorithm will make small samples of the collection to disturb the data and see which K value produces stable results. In the end, the algorithm gives a stability score to each number of topics according to the probable value that is most close to representing the reality of the correct parameter.

Then, instead of giving K to the algorithm, we can substitute it for a range of possible topic numbers. In theory, a collection of 100 documents can have up to 100 topics, in practice, the number can be adjusted according to the amount and nature of the data (e.g. a range from 10 to 50). In real-world applications, we do not expect each post to represent a new topic as it is much more plausible to expect various posts discussing the same topic.

After execution, the algorithm gives a collection of extracted topics Θ , where each topic is a set of terms with respective probabilities in regard to that topic. From there, the next task is to represent the contents of topics in a comprehensible way to the users, making results and data interpretation possible.

2.2 Topic Labeling

With a set of topics as the output of the topic extraction, the next task is to assign comprehensible labels to them. For this task, we base our algorithms on the methods proposed in [7]. It compares various metrics for labeling topics and tests the results in research areas and event detection using social networks, achieving good results in

both. We choose the best methods suggested by the authors for each type of data and apply it on formal paper documents and informal short microblogging posts.

Fig. 2 illustrates the general process of labeling a topic model that consists of three steps: (i) Candidate Selection, where possible labels are extracted from the results of the topic modeling algorithm, (ii) Score and Ranking, where candidate labels receive a relevance ranking, and (iii) Label Selection, where a set of final labels are assigned for each topic. The next section explains them in further details.

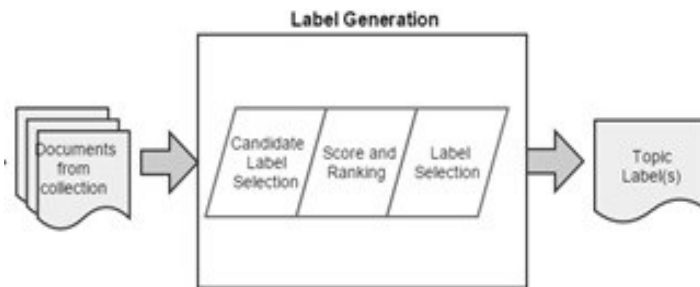


Fig. 2. Labeling process and its steps (Adapted from [8])

Candidate Selection

First of all, we need to extract and filter a list of candidate labels L for each sub-event. 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.

Algorithm 1 Candidate Selection algorithm

```

1: Input: D, number of top documents from  $\theta$  and W, number of top words
   from  $\theta$ 
2: Output: L, list of candidate labels
3:  $D' \leftarrow$  extract the top D documents from  $\theta$ 
4:  $W' \leftarrow$  extract the top W words from  $\theta$ 
5: for d in  $D'$  do
6:   for all primitive label p extracted from d do
7:     if p contains a word from  $W'$  then
8:        $L \leftarrow L + p$ 
return L
  
```

Fig. 3. Candidate Selection Algorithm (based on [7] description)

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.

The extraction of initial labels is done with an algorithm based on the fast keyword extraction algorithm [10], 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 word 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.

Table 1. Output example of candidate selection algorithm

Original Text	A range of quantitative methods is today widely used in research evaluation. Recently, with the increasing popularity of social media, and especially the increasing use of social media in scholarly activities, a new field of research has been introduced, namely, altmetrics, to investigate the use of social media in research evaluation.
Output	“quantitative methods”, “research evaluation”, “popularity”, “social media”, “today widely”, “increasing”, “scholarly activities”, “new field”, “namely altmetrics”, “investigate”, “use”, “research evaluation”

Score and Ranking

With a set of labels extracted from the text the next step is to order than according to relevance so we can select the more important or representative labels for each topic. For this task, we choose to use the metric proposed in [7] called Modified Label Degree,

which uses a mix of term frequency and label degree metrics to rank labels independently of the data used.

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:

$$tf(t, d) = f_{t,d} \quad (1)$$

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:

$$deg(w, d) = f_{w,d} + sf_{w,d} \quad (2)$$

$$deg(t, d) = \sum_{w \in t} deg(w, d), \quad (3)$$

$$ldeg(l, d) = f_{l,d} + sf_{l,d} \quad (4)$$

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. The document here is also the list of candidate labels.

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

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:

$$mdeg(l, d) = ldeg(l, d) + 2 * tf(l, d) \quad (5)$$

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

Label Selection

This is the final step and given the labels already ranked by the metrics, the process is as simple as selecting the one with the higher score value. The only problems arise when using a multiple label approach, as a set of labels can have many term intersections.

Multiple labels can help the user interpretation of the topic by presenting multiple layers of significance. A “Virus” label could be paired with a “vaccine” and “treatment” labels, emphasizing that the topic is about disease treatment instead of infections causes

or transmission topics. The same “Virus” label paired with labels like “virus results” and “virus model” would not add as much meaning layers to the topic as different labels could potentially show.

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.

3 Evaluation

An experiment was made using the Zika epidemic as base scenario, aiming to evaluate the efficiency of the proposed method for detecting research and social topics. Moreover, we would like to identify the relationship among different topics, especially among science and social topics.

The evaluation was made using two datasets: i) a Twitter dataset with posts related to Zika, ii) and a PubMed corpus with scientific articles about Zika. The scenario covered by these datasets are relative to the Zika epidemic from 2015 to 2016, which contains a variety of topics such as reports, propagation to various countries, associated diseases and influence on 2016 Olympic Games organization.

A quantitative analysis was made comparing topics detected by our proposal with those reported by official sources. The comparison is made using two sets of “golden standards”. For social topics, we used: i) A timeline of the Zika epidemic communications report [11], and ii) news reported by media for social topics. For science topics, we use the mapping of research for Zika virus response of the World Health Organization [12] and the research agenda published by the same organization [13].

A boolean variable called Relevance (Equation 6) was used for the comparisons. Relevance receives the value of 1 if official venues notified the topic or event, and 0 otherwise. Formally:

$$Relevance(\theta) = 1 \text{ if } \theta \in M ; 0 \text{ otherwise} \quad (6)$$

Where θ is the topic and M is the set of official sources that were used as a representation of available public topics content.

We will use α as the number of times the variable (6) 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.

3.1 Datasets and Scenarios

Although not new, the appearance of Zika Virus (ZIKV) cases on uncommon locations, led to an expressive outbreak started in 2015. With an unknown set of symptoms, transmission, and solution, its spreading was faster than other epidemic diseases.

In Brazil, ZIKV was identified in 2015 for the first time. At that time, the Brazilian Northeast was faced with increasing cases of an unidentified disease, characterized by fever, conjunctivitis, rash and joint pain until seven days. 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) [14].

On Feb. 1, the World Health Organization formally declared the outbreak of Zika a public health emergency of international concern [15]. Since then, Zika has been spreading worldwide, with cases in almost 100 countries.

The database for this experiment was made extracting posts from around the world with the #zika “hashtag” and articles from PubMed database containing the keyword Zika. The term is popular in both domains and has little ambiguity, so the addition of other terms could introduce more noise to the data. Both datasets cover documents created from May. 2015 to Dec. 2016. A total of 85.601 tweets and 1.769 articles were collected. A preprocessing was made in these data by removing emotes, links and accentuation from text.

3.2 Planning and Execution

The topic modeling algorithm needs an input parameter K , which will be automatically defined. To choose the best value of the parameter the proposal algorithm requires a range of possible K values. For this range, we used 4 and 20 as the minimum and a maximum number of topics that could be present in the collection respectively.

For the labeling algorithm, we used the top 10 documents and words for D and W parameters in the candidate selection algorithm.

Some critical issues are the WHO declaration of the epidemic as a Public Health Emergency of International Concern, and the evidence that Zika can cause congenital disabilities and neurological problems. Another one is the discovery that men infected with Zika can transmit the virus to their sexual partners. Also, there was the international concern regarding the safety of athletes and spectators at the 2016 Olympic Games, to be held in Rio de Janeiro (Brazil).

In this experiment, we analyzed two different periods of scientific and social discussion: 1) From May. 2015 to Feb. 2016, covering the start of the epidemic and first counter-measures and 2) From Mar. 2016 to Dec. 2016, covering the worst moment (with the increasing and highest number of cases) and the subsequent decline.

Table 2 shows results extracted from social media in the two periods and Table 3 shows the research topics extracted at the same periods. The column “Relevance” takes the value 1 if the topic was able to be found in the standard sets, and 0 otherwise. Topics which could not be found, do not have corresponding news or research areas in the standard sets.

For the first period, the proposal found 9 social topics and 5 scientific topics. The second period had 10 social topics and 8 scientific topics.

The results show varied topics in the two periods. At the first period of social topics, topic 1 is related to a discussion about a rumor of a possible relationship between a company (Monsanto) and microcephaly. Topic 2 is about the chance of epidemic affecting the Olympic Games preparation to be held in Brazil.

Event-related topics are found in topics 3, 4, and 7 which are related to the WHO declaration of the epidemic as a Public Health Emergency of International Concern, the news about travel warnings for pregnant women, and the cases exemplifying how the

virus spread through various countries. Topics 5 and 8 are related to discussions about cases, in other languages (Spanish and Portuguese specifically), on the most affected countries.

The last topic could not be found in the sources. It could be reminiscent of a merge of minor topics or just a topic that aggregates irrelevant terms, something common with topic modeling.

Table 2. Social Topics for Zika Epidemic

Period 1 – From May, 2015 to Feb. 2016		
Topics	Labels	Relevance
1	zika virus” — doctors expose monsanto linked pesticide, birth defect microcephaly, birth defect	1
2	zika virus #zikainrio #zika #zika @rio2016_en, cancelling rio olympics due, skipping #2016olympics due	1
3	world health organization director general declares #zika virus outbreak, world health organization declares spread, intl health regulations emergency committee	1
4	miami beach #zika #zika #advisory #miami #beach area	1
5	prevenir el #zika #zika pandemia ubicada como peligro mundial hoy @hijoslakebuena, si estas embarazada redobla el cuidado contra el mosquito del dengue, #zika el virus zika es causado por la picadura de	1
6	zika vaccine candidates #zika #zika #cdc #nih #niaid #vaccines \$gsk \$sny, zika vaccine candidates #zika #zika #cdc #nih #niaid #vaccines \$sny \$gsk, zika \$nlk #zika #zika-virus #vaccines #pharma #nih #cdc \$sny \$gsk \$mrk	1
7	zika virus spreads #zika #automotive #india, zika virus spreads, risk low	1
8	caso de, primeiro caso, zika virus	1
9	Protect, use, know measures, for	0
Period 2 – From Mar. 2016 to Dec. 2016		
Topics	Labels	Relevance
1	neutralizing human antibodies prevent #zika virus #zika replication, human protein ifitm3 blocks #zika virus replication, human fetal neural stem cells	1
2	2016, transmission, #cdc, sexual, cdcgov	1
3	fight #zika #doyourjob @housegop @senategop #zika-virus, fight #zika virus ravaging fl, fighting #zika virus fails	1
4	#nc governor pat mcrory, dilemma, #miamibeach	1
5	#cuba reports 1st #zika travel case, #breaking beijing reports 3rd case, chp confirms #zika virus case	1
6	asian zika virus mutated negatively & zika virus mutated negatively	1
7	mosquito repellent zika virus protection, 99 free ship	1
8	Sports, game, 2016, july	0
9	#zika virus, cientistas #vooz, #vooz #zika virus, solucoes baseadas em #dados para fazer frente ao #zika virus	1
10	suspensa por, de janeiro, ser suspensa	0

At the second period of the social topics, we have late concerns of the public opinion. Topic 5, for example, is composed of posts reporting the spread of the virus to other

countries not initially affected. Topics 3 and 7 are related to preventing mechanisms for contamination and dissemination of the disease. Topic 9 concentrates Brazilian discussions about the theme and topic 1 has the main information about the disease. Finally, topic 4 shows population concerns about the Matthew Hurricane that hit Central and North America in the period.

The scientific topics of the first period suggest concerns about the initial stages of the epidemic, as can be identified by topics 1 and 5. They have labels linked to infection vectors, how it is transmitted, and possible threats. Topic 2 refers to the time when the researchers already considered the disease an outbreak. Topics 3 and 4 refer to studies about the first cases in Brazil (Particularly Bahia, a state where local researchers identified the Zika virus for the first time in the region) and the relationship between Zika virus (ZIKV) and Dengue virus (DENV), another virus that shares the same transmission vector (*Aedes aegypti* mosquito).

At the second period of the scientific topics, we have more advanced research topics showing solidified knowledge about new forms of transmission such as saliva (topic 7), human semen (topic 4). With a better understanding of the epidemic, we also have studies focused on prevention (topic 1) and how to diagnose the disease. Topic 2 is also an important topic, which shows the developed research discovering the link between the occurrence of microcephaly and Zika virus infection in pregnant women. Topic 6 suggests a research between Zika and neurological problems caused by it like the Guillain-Barré syndrome.

In general, we have topics that cover the first aspects of the disease at the first period, a possible link can be made among social topics 8, 7, and 5 with science topics 3 and 2. All they are related to the beginning of the outbreak. Social topics seem to give more importance to topics 2 and 3, events that refer to 2016 Olympic Games and WHO epidemic declaration, respectively. The mention in social topics about pregnant women affected by the virus are strongly related to studies that appear in the second period of science topics (topics 1 and 2).

In the second period, we could link social topic 2 with science topic 4, as both are related to the discovery of new forms of transmission. The science topics investigating prevention and counter-measures (topics 1 and 5) influenced some social topics (topics 7 and 3), which communicate how to fight the virus and prevent infection. Both social topics did not appear in the previous period. Such the first period, social topics seem to emphasize specific events and authorities, and have a large number of topics. Science topics have produced more formal labels, which can be a consequence of the scientific language.

Table 3. Science Topics for Zika epidemic

Period 1 – From May, 2015 to Feb. 2016		
Topics	Labels	Relevance
1	ZIKV, virus, infection	1
2	Zika, emerging doorstep, outbreak	1
3	Brazil, Bahia, Americas	1
4	Following dengue, dengue spread, zika	1
5	Zika virus infection, co-infection, new threat	1
Period 2 – From Mar. 2016 to Dec. 2016		
Topics	Labels	Relevance
1	Zika virus prevention, travellers concern, emerging infectious diseases	1
2	Congenital fetal malformations, pregnant women, congenital microcephaly	1
3	Zika virus infection, emergency department, ZIKV IgM	1
4	Dengue Virus, human semen, pregnant	1
5	Counter zika virus, diagnostic challenge, detecting	1
6	Neurologic inhibition, inflammatory, imported arbovirus	1
7	Mosquito-borne arboviruses, African, saliva	1
8	Survey measures, results, attitudes and practices	0

Considering the total of the social topics, we have the values $\alpha = 3$ and $\beta = 16$. The total of the science topics produced $\alpha = 1$ and $\beta = 12$.

Finally, considering both topics and the periods, results show that $\alpha \ll \beta$, only for 4 times it was not possible to identify a corresponding topic in our reference documents (in a total of 32 topics). Many topics - from social and science side - were identified as related. Social topics have more variety, quantity, and have also links to events and organizations. Science topics show the development indicated by public claims at different periods. This opens new opportunities for systems that need to detect the more advantageous research paths to meet social demands or forecast the impact of new technologies when they reach the public.

4 Related Works

Despite the increasing importance of online social networks in scientific communication, their use is not completely explored. Most existing works in Computer Science - especially in the area of Text Mining - do not consider the social aspect when modeling topics or building author networks [16–18]. However, two research fields consider it: altmetrics and citation networks.

In the altmetric scenario, we can mention [19] and [20] which address many different types of alternative metrics and compare them. They explore Twitter, Facebook, Mendeley and other collaborative mechanisms, to show how the impact of an article is compared to citation impact. They found high correlation among social and scholarly citations, highlighting Twitter and Mendeley as the most effective channels of social impact. Another interesting point is the frequency of social citation, higher than scholarly citations. These works proposed specific metrics to analyze the impact of an article. In our approach, we focused on a broad view of research and social areas (and their mutual impact).

A generalization of altmetrics can be found in [4], which uses the alternative metrics to rank journals, similarly as impact-factor metrics measure them. Although a journal is broader and aggregator than an article, it is insufficient to map scientific areas.

Related to citation networks, we have [21] working with social and traditional citations to create (citation and co-citation) networks. They also study author's profiles in networks to evaluate academic engagement in social relations. Similar to altmetrics studies, they found that social interactions impact higher than traditional citations.

Finally, some works compare the use of different networks such as Mendeley [22], Scopus [23], Twitter [24] and found statistics about user participation, comparing them. Most of these works identifies the same findings, which are: the increasing use of alternative means to communicate scientific results, and a higher impact when compared to common citation analyses.

5 Conclusions

This article describes a method to automatically extract topics related to scientific and social perspectives. The same algorithm is used to identify topics from both sides, identifying thematic links in common. We use topic modeling algorithms to extract topics from texts, and topic labeling methods to assign representative labels to them. Labeled topics are essential to understand the results, and to show relationships and differences among topics discussed under social and scientific perspectives.

We conduct an experiment using the Zika epidemic scenario, an important event and highly mentioned in social networks and scientific literature. We evaluate the possible links connecting research and social topics. Afterwards, we compare our results with topics created for domain specialists and official news media.

Results pointed a moderate relationship among social and scientific topics, despite of a common sense that there are two separate and incommunicable universes. We recognize some aspects of this relationship, such topics that are specific to a perspective (social or scientific) and the evolution of the links.

The main contributions of this work are:

- The application of a topic modeling method to extract topics from heterogeneous data composed of science articles and microblogging posts.
- Application of a labeling system for different data (formal and informal texts), making heterogeneous data comparable at a higher abstraction level.
- Assessment of relation of scientific and social topics in a real-life event.

The topic association suggested by this work can be used in new scenarios, envisioning the identification of impact and influence of media news or technology reports. Another possible research topic is the development of altmetrics that operate at the topic level, indicating the impact of research areas in society and how the research development changes public perception.

References

1. G. Sonnert, G. Holton, *Ivory bridges: connecting science and society*, MIT Press Cambridge, MA, 2002. <https://mitpress.mit.edu/books/ivory-bridges> (accessed June 6, 2018).
2. K. Holmberg, The meaning of altmetrics, (2014). <http://docs.lib.purdue.edu/iatul/2014/altmetrics/1/> (accessed June 6, 2018).
3. A. Verbeek, K. Debackere, M. Luwel, Linking science to technology: Using bibliographic references in patents to build linkage schemes, *Scientometrics*. 54 (2002) 399–420. <http://link.springer.com/article/10.1023/A:1016034516731> (accessed May 24, 2014).
4. T. V. Loach, T.S. Evans, Ranking Journals Using Altmetrics, *ISSI 2015*, 15th Int. Soc. Sci. Inf. Conf. (2015) 6. doi:dx.doi.org/10.6084/m9.figshare.1461693.
5. D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet Allocation, *J. Mach. Learn. Res.* 3 (2003) 993–1022. doi:10.1162/jmlr.2003.3.4-5.993.
6. D. Blei, L. Carin, D. Dunson, Probabilistic topic models, *IEEE Signal Process. Mag.* 27 (2010) 55–65. doi:10.1109/MSP.2010.938079.
7. D. Nolasco, J. Oliveira, Detecting knowledge innovation through automatic topic labeling on scholar data, in: *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, 2016: pp. 358–367. doi:10.1109/HICSS.2016.51.
8. D. Nolasco, J. Oliveira, Intelligent Subevent Detection Based on Social Network Data, in: *2017 IEEE 15th Intl Conf Dependable, Auton. Secur. Comput. 15th Intl Conf Pervasive Intell. Comput. 3rd Intl Conf Big Data Intell. Comput. Cyber Sci. Technol. Congr.*, IEEE, 2017: pp. 820–827. doi:10.1109/DASC-PICOM-DataCom-CyberSciTec.2017.139.
9. D. Greene, D. O’Callaghan, P. Cunningham, How Many Topics? Stability Analysis for Topic Models, *Mach. Learn. Knowl. Discov. Databases*. (2014).
10. M.W.J.K. Berry, *Text Mining Applications and Theory*, John Wiley & Sons, West Sussex, UK, 2010. doi:10.1002/9780470689646.
11. R. Fundação Oswaldo Cruz., I.S. Araujo, A mídia em meio às ‘emergências’ do vírus Zika: questões para o campo da comunicação e saúde, Fundação Oswaldo Cruz, 2007. <https://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/1088> (accessed June 4, 2018).
12. WHO | Mapping social science research for Zika virus response, WHO. (2016). <http://www.who.int/risk-communication/zika-virus/rcce-activities/en/> (accessed June 4, 2018).
13. WHO | WHO Zika virus research agenda, WHO. (2017). <http://origin.who.int/reproductivehealth/zika/zika-virus-research-agenda/en/> (accessed June 4, 2018).
14. P. da Saúde, Situação Epidemiológica / Dados Zika, (2016). <http://u.saude.gov.br/index.php/situacao-epidemiologica-dados-zika> (accessed March 1, 2018).
15. W.H. Organization, WHO Director-General summarizes the outcome of the Emergency Committee regarding clusters of microcephaly and Guillain-Barré syndrome, (2016).

- <http://www.who.int/mediacentre/news/statements/2016/emergency-committee-zika-microcephaly/en> (accessed March 1, 2018).
16. M. Steyvers, M. Steyvers, P. Smyth, P. Smyth, M. Rosen-Zvi, M. Rosen-Zvi, T. Griffiths, T. Griffiths, Probabilistic author-topic models for information discovery, *Proc. Tenth ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* (2004) 315. doi:10.1.1.10.2031.
 17. A. Perianes-Rodríguez, C. Olmeda-Gómez, F. Moya-Anegn, Detecting, identifying and visualizing research groups in co-authorship networks, *Scientometrics*. 82 (2010) 307–319. doi:10.1007/s11192-009-0040-z.
 18. N. Shibata, Y. Kajikawa, Y. Takeda, K. Matsushima, Detecting emerging research fronts based on topological measures in citation networks of scientific publications, *Technovation*. 28 (2008) 758–775. doi:10.1016/j.technovation.2008.03.009.
 19. Fs.-U. Hassan, U.A. Gillani, Altmetrics of “altmetrics” using Google Scholar, Twitter, Mendeley, Facebook, Google-plus, CiteULike, Blogs and Wik, *J. Chem. Inf. Model.* 53 (2013) 1689–1699. doi:10.1017/CBO9781107415324.004.
 20. R. Costas, Z. Zahedi, P. Wouters, Do “altmetrics” correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective, *J. Assoc. Inf. Sci. Technol.* 66 (2015) 2003–2019. doi:10.1002/asi.23309.
 21. A. Martin-Martin, E. Orduna-Malea, J.M. Ayllon, E.D. Lopez-Cozar, The counting house: measuring those who count. Presence of Bibliometrics, Scientometrics, Informetrics, Webometrics and Altmetrics in the Google Scholar Citations, ResearcherID, ResearchGate, Mendeley & Twitter, 2016. doi:<http://doi.org/10.13140/RG.2.1.4814.4402>.
 22. E. Mohammadi, M. Thelwall, S. Haustein, V. Larivière, Who reads research articles? An altmetrics analysis of Mendeley user categories, *J. Assoc. Inf. Sci. Technol.* 66 (2015) 1832–1846. doi:10.1002/asi.23286.
 23. N. Maflahi, M.T.-J. of the A. for, undefined 2016, When are readership counts as useful as citation counts? Scopus versus Mendeley for LIS journals, *Wiley Online Libr.* (n.d.). <http://onlinelibrary.wiley.com/doi/10.1002/asi.23369/full> (accessed June 4, 2018).
 24. S. Haustein, V. Larivière, M. Thelwall, ... D.A.-I.-I., undefined 2014, Tweets vs. Mendeley readers: How do these two social media metrics differ?, *Degruyter.Com.* (n.d.). <https://www.degruyter.com/view/j/itit.2014.56.issue-5/itit-2014-1048/itit-2014-1048.xml> (accessed June 4, 2018).