Missier P, McClean C, Carlton J, Cedrim D, Silva L, Garcia A, Plastino A, Romanovsky A.

Recruiting from the network: Discovering Twitter users who can help combat Zika epidemics.

***In: 17th International Conference on Web Engineering (ICWE 2017)*. 2017, Rome, Italy: Springer Verlag.**

Recruiting from the network: discovering Twitter users who can help combat Zika epidemics

Paolo Missier[1], Callum McClean[1], Jonathan Carlton[1], Diego Cedrim[2], Leonardo Silva[2], Alessandro Garcia[2], Alexandre Plastino[3], and Alexander Romanovsky[1]

[1] School of Computing Science, Newcastle University, UK
[2] PUC-Rio, Rio de Janeiro, Brasil
[3] Universidad Federal Fluminense, Niteròi, Brasil

Abstract. Tropical diseases like *Chikungunya* and *Zika* have come to prominence in recent years as the cause of serious, long-lasting, population-wide health problems. In large countries like Brasil, traditional disease prevention programs led by health authorities have not been particularly effective. We explore the hypothesis that monitoring and analysis of social media content streams may effectively complement such efforts. Specifically, we aim to identify selected members of the public who are likely to be sensitive to virus combat initiatives that are organised in local communities. Focusing on Twitter and on the topic of Zika, our approach involves (i) training a classifier to select topic-relevant tweets from the Twitter feed, and (ii) discovering the top users who are actively posting relevant content about the topic. We may then recommend these users as the prime candidates for direct engagement within their community. In this short paper we describe our analytical approach and prototype architecture, discuss the challenges of dealing with noisy and sparse signal, and present encouraging preliminary results.

1 Introduction

Mosquito-borne disease epidemics are becoming more frequent and heterogeneous in tropical and subtropical areas around the world. Indeed, we witness the rapid rise to prominence of the *Chikungunya* and *Zika* viruses [9]. These viruses together with the *Dengue* virus are responsible for thousands of deaths every year [4], as well as for long-lasting health problems, especially to children. To make the matter worse, there is a potential relation between Zika virus infection and birth defects [13]. In Brazil, in particular, the regional focus of our research, disease prevention programs led by health government authorities have not been particularly effective. For instance, Brazilian Health System requires that health agents report each Zika case; however, it takes several days to process and publish such information.

Due to the inefficiency of health government programs, no one surprises that the Brazilian population has been so engaged in sharing mosquito-related content on social channels. In fact, the population has shared a variety of types of information, including complaints about personal health, dissemination of public news, but also, importantly, details about the discovery of mosquito breeding sites in public locations. In spite of the volume of mosquito-content, real-time social media is potentially a much faster vehicle for information than traditional channels. Furthermore, together with the shared content, some users stand out for the quality and relevance of their contribution to the social media. These users are namely *social sensors*. The term *social sensors* has been used in similar contexts [14], to denote portions of the online population that spontaneously contribute with information on social media channels, which is relevant to a particular topic.

As social sensors are influential references on social media, this short paper presents an initial investigation into the kind of social sensor signals that can be effectively detected from real-time social media streams. The goal is to rank users who can act as social sensors. Our approach to rank users is based on the classification of relevant tweets. First, we classify tweets automatically based on their content. Classification aims at filtering out relevant tweets from irrelevant ones. Second, we apply an adaptation of the TwitterRank algorithm to rank users who authored the relevant tweets. Additionally, this paper investigates how social sensors can be exploited to complement and support institutional disease combat efforts. Specifically, we investigate the hypothesis that *real time, short content* social media websites such as Twitter, Instagram, etc., when appropriately analysed, are strong allies on the combat and prevention programs. That is, these networks can be exploited to engage the population on health programs by selecting members of online communities (social sensors) to contribute to health vigilance in their *local* communities. Ultimately, we aim to support health authorities, as they need to engage the population to embrace the combat and prevention programs. This support happens when we rank influential users (social sensors) who can engage communities' members.

Our solution to reveal and rank social sensors is integrated to our VazaZika portal [4]. VazaZika works as an entomological surveillance system in order to combat the mosquito that transmits Zika, Chikungunya, and Dengue. The portal and a mobile app allow users to report and visualize occurrences of the mosquito or cases of sick people. VazaZika is integrated to social medias in order to reveal social sensors in such medias. Our solution plays an important role to popularize the surveillance system and the engagement programs provided by the VazaZika portal.

1.1 Overview of the approach

Our approach combines content-based automated classification of tweets, aimed at isolating the sparse relevant signal out of generally noisy chatter about Zika, followed by a ranking of the users who author such relevant content. This is summarised in the dataflow diagrams of Fig. 1.

[4] Available at http://vazadengue.inf.puc-rio.br/

Initially, in the *offline phase* (left in the figure), a classifier is trained on a collection of manually annotated tweets. The classifier aims at segregating the *target tweets* that are indicative of user interest in aspects of the Zika problem, as opposed to news feeds, e.g. those originating from news agencies, as well as background noise. The main challenges in achieving good classification performance is the high levels of noise found in the filtered harvest. This is mainly due to the idiosyncratic use of critical keywords, such as *Zika* itself, which in Brasil happens to be used as a common slang word completely out of the context of discourse about the virus or the disease. In the wild, the target tweets are less than 10% of a typical harvest.

In the *online phase* (right side of the figure), tweets are continuously harvested from the raw twitter feed, using a set of filtering keywords that we have chosen to provide high recall relative to the set of target tweets. We denote as *candidate users* the authors of all tweets that are classed as `Relevant`. These are ranked using a variation of the TwitterRank algorithm [17], which we modify to operate on a single topic. For this, the users connections in the Twitter social graph are retrieved (specifically, the set of users' followers) and used to rank the candidate users according to their relative relevance. Ideally, this approach provides a set of top-k target users, which is continuously updated as the live feed is tracked over time.

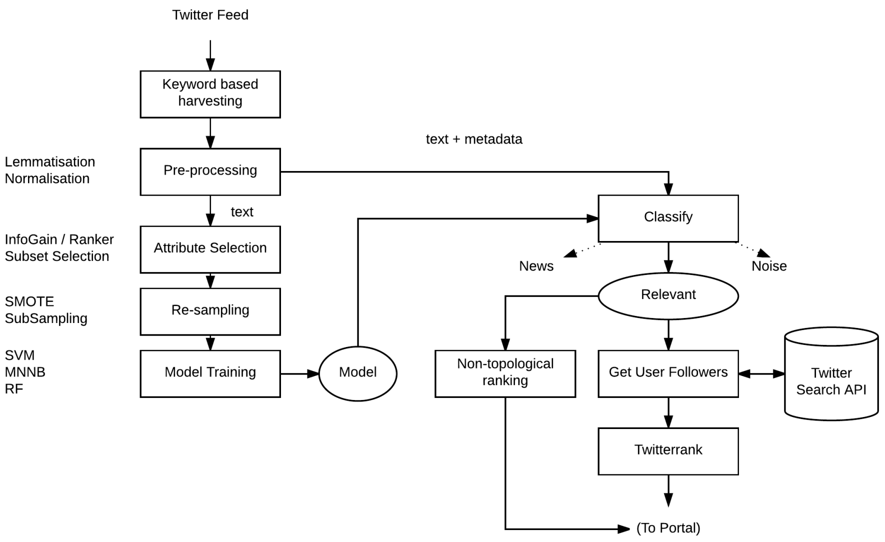

Fig. 1: Dataflow diagram for content classification and user ranking

1.2 Challenges and Contributions

In this short paper we present our technical approach, including our experimental selection of a suitable classification approach, our single-topic adaptation to TwitterRank, and some early results. While classification accuracy is acceptably good (84.1%, F-measure = .84), we find that acquiring a sufficient number of relevant tweets per user requires a long harvest time. On a 3-months batch in 2016, we have identified about 13,000 relevant tweets, with most users only contributing one single relevant tweet. This sparsity of users suggests that at this time scale it is difficult to apply any ranking criterion, and that TwitterRank is both ineffective and inefficient. Indeed, TwitterRank assumes knowledge of the social graph neighbourhood for each candidate user, and requires that meaningful social connections exist within those neighbourhoods. Thus, it is an inefficient approach because it requires retrieving all followers for a large number of single-contribution candidate users; and it is ineffective because the vast majority of these followers will not be candidate users themselves, which means they will not contribute to the ranking of other users. In reaction to these observations, in Sec. 4 we propose alternative, and simpler, ranking criteria that do not rely upon the topological properties of the social graph around the users, and compare those with the TwitterRank top-k users. Note that, as in the original TwitterRank research [17], no ground truth, i.e., explicit knowledge of these top users, is available for evaluation, as our content harvesting was performed purely “in the wild”. Thus, our discussion on the results is necessarily based on a comparison of relative merits of these metrics.

1.3 Related work

Similar to our work, [18] propose a method to rank Twitter users using a variation of graph analysis called TURank. The authors perform link structure analysis on the user-tweet graph they introduce consisting of tweets and users as nodes, and follow and retweet relationships as edges. While we do not consider the retweet relationships that users form between each other, as our ranking phase does not allow for it, we do evaluate the tweets that users are posting.

In [6] they propose a clustering algorithm to partition influential users into five categories; fan, disseminator, expert, celebrity, and others. Their work relates to ours as the clustering of users into the five influence role categories can be related to our topic-specific communities, however, their communities are formed of users that are experts, for example, on multiple topics. The authors introduce a limitation in their work; they pick the top 10 users within a topic and crawl each of their followers, making an assumption that those followers also have an interest in that specific topic. Our approach tests the validity of the followers of a relevant user by only considering followers that are relevant themselves.

In our previous work [10], we used topic modelling similar to that shown in [15], however, we focused on pre-defined classes of interest specifically related to Zika epidemics. They use community detection (Louvain modularity) and the encoding of random walks to detect community structures within the topics they previously defined. However, in our experiments, we found that the data, tweets and users, are too sparse to form communities, thus other approaches are required, i.e. ranking.

An alternative approach to finding authoritative users and ranking them is presented in [16] by Wei *et al.* They use a combination of Twitter lists (a grouping of followers per a criterion), the follower graph and the users profile information to produce a global authority score for each user in their data set. In this paper, we do not use the lists nor the user profile but it demonstrates that Twitter is a great resource, with many attributes, that can be utilised to produce results similar to ours.

A heuristic-based approach for automated identification of expertise on Twitter is presented in [8] and is based on the premise that experts will use Twitter differently to that of non-experts. They find that experts tend to receive information from many friends, filter and distil it, and that experts tend to be old, in relation to the length of time passed since the creation of their accounts, Twitter users. Our work differs as we aim at seeking out users who stand out not because of their expertise but because of their demonstrated interest in engaging with a specific topic.

2 Twitter relevance model training

We mentioned in the introduction that user ranking requires first of all the capability to identify with high precision the few tweets that are relevant to the Zika topic, amongst a large amount of Twitter noise. For this, we tuned a harvester on a set of relevant keywords, and then trained a supervised classifier on an initial set of about 10,000 manually annotated tweets, collected over multiple time windows, between Sept and Dec. 2016. As mentioned earlier, however, the need to use the keyword *Zika* makes the harvesting and initial filtering difficult and results in a particularly noisy dataset. Fine-tuning of data pre-processing and model training was therefore required. We discuss these issues in the rest of this section.

2.1 Selecting Twitter feed harvesting keywords

The first task, Twitter harvesting (top of Fig.1) provides content both for manual annotation and model training, as well as for classification and then user ranking. High recall is important in the initial filtering, as the relevant tweets we seek to isolate are no more than about 10% of the feed. At the same time filtering out clearly irrelevant content is required for reducing the noise prior to classification. The choice of keywords to harvest from the live Tweeter feed[5] is therefore critical to striking this balance.

Filtering keywords were selected in two steps, following an approach similar to that suggested in [12]. Firstly, a short list of *seed* keywords was *bootstrapped*

[5] For this we used the Twitter stream API through the Twitter4j library.

from sample tweets content using manual, expert inspection, and borrowing from our earlier work [10]. These are the top 8 keywords: `dengue`, `combateadengue`, `focodengue`, `todoscontradengue`, `aedeseagypti`, `zika`, `chikungunya`, `virus`.

Those keywords were then used to harvest an initial corpus of tweets, whose terms were then ranked according to their TF-IDF score relative to the corpus. The top 10 of those were added to the initial seed set, after removing common stopwords and those words that experts deemed to be out of context. The resulting additional terms, listed here below, were used together with the seed terms above, as filtering keywords on the Tweet stream API, both for harvesting the training set, and for ongoing harvesting for continuous user ranking: `microcefalia`, `transmitido`, `epidemia`, `transmissao`, `doenca`, `eagypti`, `doencas`, `gestantes`, `infeccao`, `mosquitos`.

2.2 Learning a relevance model

We aimed to learn a classifier that effectively provides an operational definition of *relevance* of tweets in the context of our topic. In our previous work on detecting Dengue-related tweets [10] we used four target classes with the same purpose: `Mosquito-focus`, `Sickness`, `News`, and `Joke`, representing (i) content that is strictly relevant to the topic, (ii) content that describes symptoms by affected people, (iii) content from news agencies or that echoes news from agencies, and (iv) content from people who make mostly sarcastic or humorous remarks about Dengue, respectively. In that setting, both `Mosquito-focus` and `Sickness` tweets would be considered relevant.

For Zika-related content, we focused initially on two "relevance" classes, namely information `provider` and `receiver`, with a view to engage two groups of users: those who are shown to volunteer information about possible infestation locations, the `providers`, as well as those who may need assistance because they talk about their experience being infected, i.e., the `receivers`. However, the more noisy nature of Zika content relative to the Dengue content, along with the scarcity of instances in each of the two relevant classes, contributed to poor accuracy in our early experiments, suggesting that merging the two classes might be beneficial. In this work, we therefore only use three classes: `Relevant`, `News`, and `Noise`.

In the work just cited, we contrasted a traditional supervised learning approach (a Naive Bayes model) with unsupervised topic modelling, using variations of the LDA algorithm [2], which has proven popular in recent research [7, 11]. We concluded that LDA under-performs when topics are pre-selected and topic modelling is expected to discover "sub-topics", and that a relatively small annotation effort (2,000 tweets at the time) was sufficient (.83 F-measure across the classes).

Having noted earlier that high recall *Zika* Twitter harvests are going to be more noisy than the more specific *Dengue* tweets, in this work we have focused solely on supervised classification, using 10,000 labelled examples. We experimented with a number of supervised classification models as well as multiple

	1-grams	**1+2-grams**	**1+2+3-grams**
SVM	73.96 (0.68)	73.97 (0.70)	74.01 (0.70)
MNNB	81.21 (0.81)	81.74 (0.82)	81.81 (0.82)
RF	81.1 (0.80)	80.65 (0.80)	79.97 (0.79)

Table 1: Baseline classifier accuracy

data preparation steps, illustrated in Fig. 1, left side, all implemented using the Weka toolkit.

The final configuration, described below, is the result of exploration over a space of available alternatives and parameter settings at each step of the data preparation pipeline. This includes (i) representing tweets using bag-of-words and a choice of N-grams (N=1,2,3); (ii) attribute selection using Ranking with Information Gain vs Subset Selection; and (iii) whether to rebalance class distribution in the training set, i.e. using class over- and sub-sampling (note that Attribute Selection methods, namely Ranking with Information Gain and Subset Selection, did not improve performance and are therefore not discusssed further).

For the initial *text normalisation* we used POS tagging and lemmatisation[6], also removing common regional "twitter lingo" abbreviations, as well as all emoticons and non-verbal forms of expressions. While those are crucial to understanding the *sentiment* expressed in a tweet, we found that they are not good class predictors. Links, images, numbers, and idiomatic expressions were also replaced by conventional terms (*url*, *image*, *funny*,...).[7]

In searching for a suitable combination, we then heuristically reduced the space of possible configurations by first establishing a baseline classifier performance, for three popular classification models that have proved effective for short text classification [3]: Support Vector Machines (SVM), Multinomial Naive Bayes (MNNB), and Random Forest (RF). This includes a choice of N-grams, but no attribute selection and no class re-sampling. Table 1 reports the overall accuracy and F-measure (in parenthesis) for each of these classifiers. Based on these early results, We ruled out SVM, which performed substantially more poorly, and focused solely on MNNB and RF.

Class rebalancing. As mentioned above, one of the main classification challenges is the relative scarcity of `Relevant` tweets in the Twitter feed for user ranking. This imbalance in the minority class is naturally reflected in the class proportions observed in the training set: 50.6% `News`, 37.3% `Noise`, 12.1% `Relevant`, and may reduce accuracy. To address this issue, we experimented with two complementary approaches. Firstly, we added an extra 600 annotated examples to the `Relevant` class. Secondly, we applied statistical over-sampling to the `Relevant` class, using

[6] We used the tagger from Apache OpenNLP 1.5 series (`http://opennlp.sourceforge.net/models-1.5/`), and the LemPORT Lemmatizer customised for Portuguese language vocabulary.

[7] Note that these steps are the same as described in [10]).

	RF			MNNB		
	1-grams	1+2-grams	1+2+3-grams	1-grams	1+2-grams	1+2+3-grams
SMOTE over-sampling	83.5	83.1	**84.1**	81.2	80.9	81.2
Sub-sampling (Spread)	75.8	76.3	76.1	77.5	78.9	79.95
Over- and sub-sampling	82.5	82.7	83.6	80.6	80.0	80.95
+600 Relevant samples	80.8	80.5	80.4	80.5	81.0	81.2

Table 2: Classifier accuracy for various choices of N-grams and over- and sub-sampling

the SMOTE algorithm [5] to boost the examples from 1,214 to 2,428 (12.1% to 24.3%).

The results, reported in Table 2 for various combinations of N-grams and MNNB vs RF, show that there is no real advantage in investing extra human annotation effort, as boosting using SMOTE provides equivalent performance. Note that the results also show that down-sampling the majority class (News) is not as beneficial.

The Table also reports the best overall accuracy figure across all configurations, namely 84.1%, obtained from a Random Forest learner (using an ensemble of 100 trees), with 1,2,3-grams, no attribute selection, and SMOTE-based boosting. More in detail, the performance measures of this configuration is: weighted average F-measure=0.84 across the three classes is, with F-measure=0.83 for the Relevant class, and RMSE=0.28. This is the classifier we used for the online content relevance detection phase in combination with user ranking, described next.

3 User ranking

In the next phase of our study, we collect all users that have authored at least one Relevant tweet and experiment with three ranking criteria to select the top-k users. While we hope these may be ideally suited for engagement by the health authorities on Zika combat campaigns, we have no ground truth about the effective attitude of these users, as our study is conducted entirely *in the wild.* Thus, we are going to present our results in the form of a comparative analysis across the three types of rankings. Specfically, we compare our own variation of the TwitterRank algorithm [17], which is based on the social media graph, with non-topological metrics that simply count the fraction of relevant tweets per user within the harvest set and within the whole twitter stream. Firstly, we describe these metrics.

3.1 TwitterRank

In [17] a method of assigning a topic-specific rank to the users of Twitter is proposed, called TwitterRank (TR). The approach is an extension of PageRank, however, TR differs as it measures importance by taking both the topical similarity between users and the underlying social network structure into account. They propose the formula below to calculate the topic-specific rank for a user.

$$\overrightarrow{TR_t} = \gamma P_t \times \overrightarrow{TR_t} + (1 - \gamma)E_t$$

$\overrightarrow{TR_t}$ is the TR score associated with a user for topic t. P_t is the transition probability of a random surfer moving from follower to friend. E_t is the teleportation vector of the random surfer in topic t, i.e. how many times a user's tweets have been assigned to topic t. γ is a variable that controls the probability of teleportation. The lower γ is the higher the probability that the random surfer will teleport to users according to E_t and vice versa [17].

3.2 Adaptation of TwitterRank

While TR can fit with our work, we found that TR does not contextually translate perfectly and needed adapting slightly in order to work with our data sets. The authors, in [17], use a multi-column matrix to store the rank of a user within their data set, with each column representing a topic and each row a user. We limit this matrix to a single topic, as we're interested in discovering highly ranked users within a topic-relevant virtual community. Furthermore, they propose a topical difference between two users, which isn't applicable in our context so we introduce a new metric; the normalised occurrences for a user: υ_t.

The transition probability is calculate as shown below; this determines the likelihood that a random user will start at follower s_i and then move to s_j.

$$PT_t(i,j) = \frac{|\tau_j|}{\sum\limits_{a:s_i follows s_a} |\tau_a|} \times sim_t(i,j)$$

This formula, fundamentally, remains the same for us, however, we redefine components of it. To start, τ_j, the number of tweets published by s_j, is changed to the number of tweets published by s_j within the topic (rather than overall). τ_a becomes the number of tweets published by all of s_i's friends, that are within the topic, rather than the sum of tweets published by all of s_i's friends across all topics. Finally, $sim_t(i,j)$ calculates the similarity between s_i and s_j within topic t. We changed this to find the absolute different υ_t for users i and j, rather than the absolute topical difference between users i and j. The change is shown below respectively.

$$sim_t(i,j) = 1 - |DT'_{it} - DT'_{jt}|$$

$$sim_t(i,j) = 1 - |\upsilon_{it} - \upsilon_{jt}|$$

The final modification that we made was to the teleportation vector for the random user, E_t. Originally this described the t-th column of the topical difference matrix and is the column-normalised form of the matrix DT such that $||DT''_{.t}||_1 = 1$. This isn't a metric that we use, however, it forms part of the overall TR calculation. Therefore, we instead use the normalised occurrences for a user i in topic t.

3.3 Application of TwitterRank

We create a Java-based application in order to implement the adaption of TR presented previously. To start, the followers for each user within the data set are collected using a crawler previously developed that queries the Twitter public REST API. Once the followers are collected, we only consider followers of a user if all of those or a subset of followers are also in the data set. This approach iteratively builds topic-specific communities starting with one user and then expanding outwards, potentially linking communities together. We decided to do this as there would be a lot of noise introduced in calculating the TR score as the vast majority of those in the social network would not be relevant to our goal; if all followers are considered, and it reduces a computation overhead discussed in [1]. Finally, as per the original TR approach, we set the teleportation vector as $\gamma = 0.85$.

3.4 Non-topological metrics

For a user u and a set K of keyword, let T_K denote the entire harvest, $T_K(u)$ the number of tweets in T_K that are attributed to u, $R_K(u)$ the number of `Relevant` tweets in $T_K(u)$, and $T(u)$ the *total* number of tweets posted by u during the harvest period.

We define the **Topic Focus** per user as $TF(u) = \frac{R_K(u)}{T_K(u)}$. This is the fraction of u's tweets in the harvest, which are `Relevant`, an indication of how often user u used the keywords K to express relevant content;

We define the **Overall Focus** per user as $TF(u) = \frac{R_K(u)}{T(u)}$. This is the fraction of u's total tweets in the harvest period, which are `Relevant`. We take this as an indication of the focus of the user on the topic, considering the user's global interests when posting on Twitter.

4 Results

4.1 Experimental dataset

Given a keyword-based harvest from the Twitter feed, we refer to the set of users who have posted at least one `Relevant` tweet as the *candidate* users. The *target* users are the top candidate users according to some ranking criteria. In this Section we report our preliminary findings on characterising candidate users and ranking them to discover target users.

Our experimental dataset consists of a harvest of 278,351 tweets, collected and classified through our online pipeline (Fig.1) using the keywords presented in Sec. 2.1 during a period of 4 months (9-12) in 2016. Using our classifier, we found 15,124 `Relevant` tweets in this set.

Firstly, we note that the vast majority of those users only produced one single or very few `Relevant` tweets during the harvest period, as shown in Tab.3. This means that there are very many candidate users (13,228 in our batch), each

Relevant Tweets	Users count
≥ 20	2
(10,19)	1
(5,9)	41
4	57
3	209
2	1058
1	11860

Table 3: Distribution of Relevant tweets per candidate user

producing a very weak signal both in terms of generated content and in terms of their social connections to other candidate users.

To deal with this long tail and to strike a balance between strength of content signal and numerosity of candidate users, we only considered users who posted at least 3 Relevant tweets. Out of these 310 users, however, we had to exclude a further 139 whose followers could not be obtained due to privacy settings, leaving 171 candidate users for ranking. The results presented below concern these users.

Tables 4, 5, and 6 show the top 10 users ranked according to each of our three criteria (TwitterRank, Topic Focus, and Overall Focus), respectively. For each of these users, each table also shows the values for the other two metrics, and the position of that user when ranked according to those metrics.

Screenname	**Twitterrank (x100)**	**Relevant count**	**Overall focus (x100)**	**OF Rank**	**Topic focus**	**TF Rank**
FlorzinhaSimoes	0.84	20	14.28	3	71.428	15
Lorrayn54837060	0.64	3	0.1708	142	75	14
pelotelefone	0.41	7	6.1947	7	87.5	7
SEIZETHEHEAVEN	0.39	7	0.3693	65	100	1
macabia	0.39	3	0.44	55	100	5
gushfsc	0.37	6	0.30	85	60	18
tiiancris	0.37	3	0.19	128	50	24
scomacinha	0.35	3	0.13	164	33.33	28
sophiaboggiano	0.35	3	0.14	160	75	14
mariabarrozoo	0.34	3	0.11	169	60	19

Table 4: Top 10 TwitterRank candidate users

Regarding TwitterRank, we note firstly that the small absolute figures are not indicative, as the original paper [17] does not provide any reference figures at all. However we note a significant spread (150%) between the top and bottom ranks in the top-10 list. The significance of this ranking, however, is questionable. TwitterRank only yields interesting rank values when, for each user u, at least some of its followers are also candidate users. When this is not the case the approach is not very effective, because u's followers' TwitterRank is a default

Screenname	Topic Focus	Relevant count	All tweets count	Overall focus (x100)	OF Rank	TR (x100)	TR position
SEIZETHEHEAVEN	100	7	1895	0.3693	65	0.39	XX
LairaMaia	100	6	799	0.7509	35	0.07	XX
llGueto	100	6	1427	0.4204	58	0.07	XX
Giovannacoosta	100	5	960	0.5208	45	0.06	XX
pakito_lucas	100	5	2149	0.2326	111	0.06	XX
Lorranna_Castro	100	5	1573	0.3178	84	0.06	XX
laricrvlh	100	5	951	0.5257	43	0.06	XX
mauriciooasn	100	4	495	0.8080	33	0.04	XX
masoqmath_	100	4	2412	0.1658	145	0.04	XX
isaah13_ferreir	100	4	272	1.4705	19	0.04	XX

Table 5: Top 10 Topic Focus candidate users

Screenname	Relevant Count	Keyword count	All tweets count	Overall focus (Rel/All)	Topic Focus	TF rank	TR	TR position
leilaquintsepe	4	4	19	21	100	=4	0.04	70
DCGRodrigues	3	3	18	16.6	100	=5	0.03	169
FlorzinhaSimoes	20	28	140	14.2	71.4	15	0.8	1
RobelioValle	3	4	31	9.6	75	=14	0.03	156
iaedayana	3	3	37	8.1	100	=5	0.03	125
iPedersoly	4	5	51	7.8	80	=10	0.04	81
pelotelefone	7	8	113	6.1	87.5	=7	0.4	3
tacianebielinki	6	10	136	4.4	60	=18	0.07	32
isaldcunha	3	4	98	3	75	=15	0.03	147
onelastovada	7	9	285	2.4	77.7	11	0.1	24

Table 6: Top 10 Overall Focus candidate users

value, which does not influence the TwitterRank of u at all. In our dataset, we find that our candidate users have very few connections amongst each other. This becomes clear when looking at the social connections amongst some of our candidate users, as in Fig. 2. The graph shows very promising results, as even in our small residual candidate set we discover interesting connected components, and indeed even a few friends (shown with the double arrow). Note also that all of our top-10 TwitterRank users appears in some connected component of the graph, which is natural as it is their connectivity that contributes to their TwitterRank. On the other hand, the number of followers of any user who actually influence the user's rank is very small.

We therefore compared this with the other two metrics. Tab. 5 shows that for each of the top-10 Topic Focus users, *all* of their tweets in the harvest ($T_K(u)$), however few (¡10), are `Relevant`. Furthermore, the `TF Rank` column in Tab.4 shows that all top 10 TwitterRank users are top-30 Topic Focus users, suggesting that high TwitterRankx may correlate well with high Topic Focus.

We also note that the top-10 TwitterRank user `SeizeTheHeaven` is also in the top-10 Topic Focus[8]

[8] User `macabia` is also in the top-10, but not shown as evidently the list of users with Topic Focus = 100 is longer than 10.

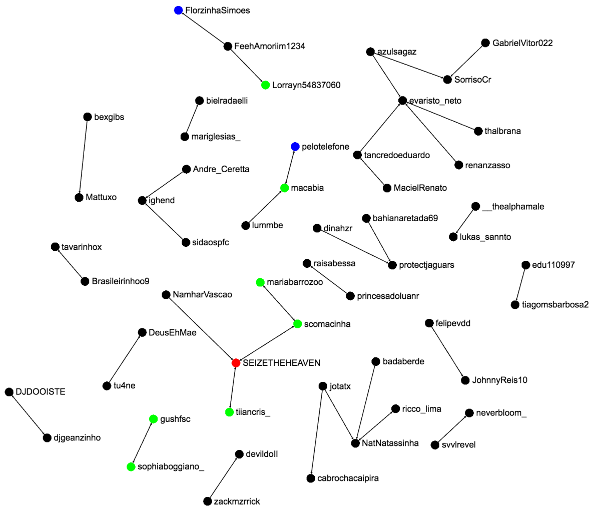

Fig. 2: Fragment of followers and friends graph for candidate users in our experimental dataset. Green nodes are in the top 10 TwitterRank. Blue nodes are in top 10 TwitterRank *and* top 10 Overall Focus. Red nodes are in top 10 TwitterRank *and* top 10 Topic Focus.

Interestingly, if we turn to Tab. 6 we see that the top-10 Overall Focus users also have a high Topic Focus, and rank within the top-20. Again in this list we find users that rank high in other lists: `FlorzinhaSimoes` and `pelotelefone`.

5 Conclusions

The research hypothesis we have explored in this paper is that social media analytics can be used to identify individuals who are actively contributing to social discourse on rthe specific topic of the Zika virus and its consequences, and are thus likely to be sensitive to health promotion campaigns. We tested this hypothesis by focusing on Twitter content related to the Zika virus and its effect on people. We trained a classifier to separate the very sparse interesting signal from large amounts of noise in the feed, and then applied various ranking criteria to the set of candidate users who authored such interesting content.

Given the sparsity of the contributors and their limited connections within the social graph, it is not surprising to find that the very popular TwitterRank algorithm [17] is not particularly effective in this instance. Despite facing a "needle in the haystack" problem, however, we report promising results which indicate that non topology-based metrics that count relevant tweets by user appear to be equally effective, and that a few interesting connections indeed exist in the graph amongst the top ranked users. We are currently experimenting with larger datasets which we continually harvest from the live twitter feed.

We have developed a public-facing portal where Relevant tweets that are also geo-located are placed on a map of Brasil, and the top-k users computed using our metrics are shown and continually updated.

References

1. Bartoletti, M., Lande, S., Massa, A.: Faderank: An Incremental Algorithm for Ranking Twitter Users. In: Web Information Systems Engineering – WISE 2016, vol. 10042, pp. 55–69 (2016), http://link.springer.com/10.1007/978-3-319-48743-4{_}5
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. J. Mach. Learn. Res. 3, 993–1022 (mar 2003), http://dl.acm.org/citation.cfm?id=944919.944937
3. Carvalho, J., Plastino, A.: An Assessment Study of Features and Meta-Level Features in Twitter Sentiment Analysis. In: {ECAI} 2016 - 22nd European Conference on Artificial Intelligence, 29 August-2 September 2016, The Hague, The Netherlands - Including Prestigious Applications of Artificial Intelligence {(PAIS} 2016). pp. 769–777 (2016), http://dx.doi.org/10.3233/978-1-61499-672-9-769
4. CDC: Centers for Disease Control and Prevention. http://www.cdc.gov/dengue/ (2015), [Online; accessed 15-december-2015]
5. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research 16, 321–357 (2002)
6. Chen, C., Gao, D., Li, W., Hou, Y.: Inferring Topic-Dependent Influence Roles of Twitter Users. Proceedings of the 37th international ACM . . . (2014)
7. Dela Rosa, K., Shah, R., Lin, B., Gershman, A., Frederking, R.: Topical Clustering of Tweets. SIGIR 3rd Workshop on Social Web Search and Mining (2011)
8. Horne, B.D., Nevo, D., Freitas, J., Ji, H., Adalı, S.: Expertise in Social Networks : How Do Experts Differ From Other Users ? Proceedings of the Tenth International AAAI Conference on Web and Social Media 10, 583–586 (2016)
9. Miles, T., Hirschler, B.: Zika virus set to spread across americas, spurring vaccine hunt (Jan 2016), http://www.reuters.com/article/us-health-zika-idUSKCN0V30U6
10. Missier, P., Romanovsky, A., Miu, T., Pal, A., Daniilakis, M., Garcia, A., Cedrim, D., Sousa, L.: Tracking Dengue Epidemics using Twitter Content Classification and Topic Modelling. In: Procs. SoWeMine workshop, co-located with ICWE 2016. Lugano, Switzerland (2016)
11. Morstatter, F., Pfeffer, J., Liu, H., Carley, K.: Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. Proceedings of ICWSM pp. 400–408 (2013), http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/viewPDFInterstitial/6071/6379
12. Nagarajan, M., Gomadam, K., Sheth, A.P., Ranabahu, A., Mutharaju, R., Jadhav, A.: Spatio-temporal-thematic analysis of citizen sensor data: Challenges and experiences. LNCS 5802 LNCS, 539–553 (2009)
13. Rasmussen, S.A., Jamieson, D.J., Honein, M.A., Petersen, L.R.: Zika virus and birth defects — reviewing the evidence for causality. New England Journal of Medicine 374(20), 1981–1987 (2016), http://dx.doi.org/10.1056/NEJMsr1604338, pMID: 27074377

14. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: Procs. WWW '10. p. 851 (2010), http://portal.acm.org/citation.cfm?doid=1772690.1772777$\delimiter"026E30F$npapers3://publication/doi/10.1145/1772690.1772777
15. Surian, D., Nguyen, D.Q., Kennedy, G., Johnson, M., Coiera, E., Dunn, A.G.: Characterizing Twitter Discussions About HPV Vaccines Using Topic Modeling and Community Detection. Journal of Medical Internet research 18(8), e232 (aug 2016), http://www.ncbi.nlm.nih.gov/pubmed/27573910http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5020315
16. Wei, W., Cong, G., Miao, C., Zhu, F., Li, G.: Learning to Find Topic Experts in Twitter via Different Relations. IEEE Transactions on Knowledge and Data Engineering 28(7), 1764–1778 (jul 2016), http://ieeexplore.ieee.org/document/7426825/
17. Weng, J., Lim, E.P., Jiang, J., He, Q.: Twitterrank: finding topic-sensitive influential twitterers. In: Proceedings of the third ACM international conference on Web search and data mining. pp. 261–270. ACM (2010)
18. Yamaguchi, Y., Takahashi, T., Amagasa, T., Kitagawa, H.: TURank: Twitter User Ranking Based on User-Tweet Graph Analysis. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 6488 LNCS, pp. 240–253 (2010), http://link.springer.com/10.1007/978-3-642-17616-6{_}22