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, Newcastle upon Tyne, UK
j.carlton@ncl.ac.uk
[2] PUC-Rio, Rio de Janeiro, Brazil
[3] Universidad Federal Fluminense, Niteròi, Brazil

Abstract. Tropical diseases like *Chikungunya* and *Zika* have come to prominence in recent years as the cause of serious health problems. We explore the hypothesis that monitoring and analysis of social media content streams may effectively complement institutional disease prevention efforts. Specifically, we aim to identify selected members of the public who are likely to be sensitive to virus combat initiatives. 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. 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 such as *Chikungunya* and *Zika* viruses are becoming more frequent in subtropical areas around the world [8], and are responsible for thousands of deaths every year [3]. In Brazil, the regional focus of our research, disease prevention programs led by health government authorities have not been particularly effective. It is therefore natural that Brazilians have become heavy users of social channels to share mosquito-related information. This includes complaints about personal health, dissemination of public news, but also, importantly, details about the discovery of mosquito breeding sites in public locations. This presents an opportunity to complement existing disease prevention programs, as real-time social media is potentially a much faster vehicle for information than traditional channels, and we also hope to discover a few users who stand out for the quality and relevance of their contribution to the social media. These users are referred to as *social sensors* [12], as they spontaneously contribute with information on social media channels, which is relevant to a particular topic.

J. Cabot et al. (Eds.): ICWE 2017, LNCS 10360, pp. 437–445, 2017.
DOI: 10.1007/978-3-319-60131-1_30

In this short paper we present our initial investigation into techniques to identify target users who show to be good social sensors, with the aim to engage them in disease prevention programs within their community. Our approach, summarised in the dataflow diagram of Fig. 1, combines content-based automated classification of tweets, aimed at isolating the relevant signal out of generally noisy chatter about Zika (training phase indicated in the left of the figure), followed by a ranking of the users who author such relevant content (online phase, in the right of the figure). Note that *Zika* is a common slang word in Brasilian Portuguese, often used out of context, resulting in a high-recall but particularly noisy harvest from the Tweeter feed. In this "needle in the haystack" problem, the main challenge is to filter out the large proportion of noise and irrelevant news items about Zika (relevant tweets are less than 10% of a typical harvest), as well as the identify the very few target users who consistently tweet relevant content. We present preliminary results on comparing three user ranking metrics, including our own single-topic adaptation of TwitterRank [14], computed from a set of about 200,000 tweets and 180 active users, harvested during 4 months in 2016. Many details are omitted for space reasons. Please refer to our Technical Report [9] for a more complete account.

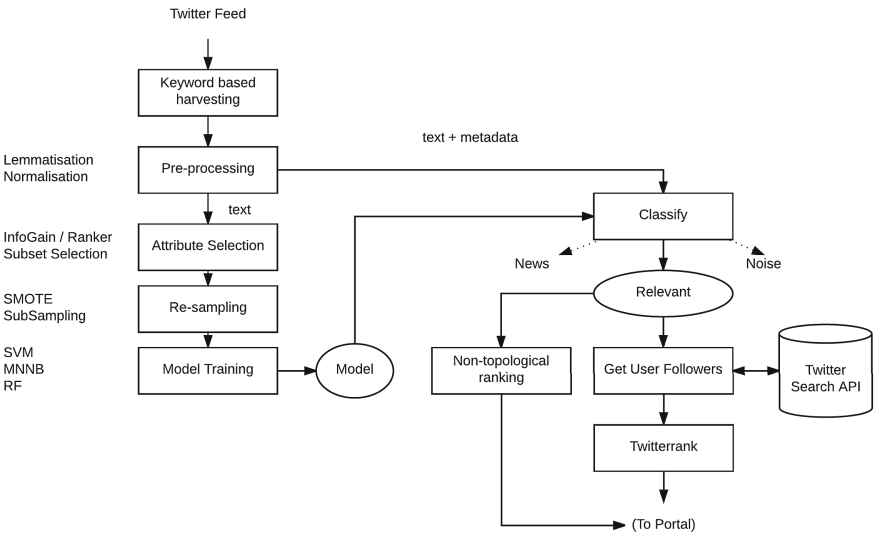

Fig. 1. Dataflow diagram for content classification and user ranking

This work follows on from [10], by re-focusing it on Zika content, extending it to user ranking, and providing a prototype implementation of our streaming architecture. Also, the prototype has been integrated with our VazaZika portal.[1] VazaZika works as an entomological surveillance system in order to combat the

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

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.

Related work. TURank [15] uses link structure analysis on the user-tweet graph to rank Twitter users, including *follow* and *retweet* relationships. While our approach, based on TwitterRank, does not include retweets, unlike TURank we do analyse tweet content.

Another related approach [5] aims to find influential users such as disseminator, expert, etc. However there is an assumption that the follower of someone who is an expert on a topic is also interested in that topic. TwitterRank, in contrast, only considers followers of `Relevant` tweets, who also authored other `Relevant` tweets.

Wei *et al.* [13] 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. We may experiment with incorporating user profile metadata into our analysis in our future work.

Finally, [6] aims to discover expert uses on Twitter, assuming that experts will exhibit different Twitter usage patterns than non-experts. 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 Tweets Classification and User Ranking

User ranking requires firstly 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 tweets manually annotated by an expert. This set was collected over three months in 2016. In order to check the consistency of the manual annotations, a representative random sample of the set (margin of error $\pm 5\%$ at the 95% confidence level) was manually re-annotated by a second expert. The agreement between the experts was then assessed by calculating Cohen's Kappa coefficient [7]. We found a substantial agreement (0.70, p-value = 0.000), reaching an almost perfect agreement for the relevant class (0.82, p-value = 0.000).

The need to use the keyword *Zika*, which is a common slang word in Brasilian Portuguese, often used out of context, makes the harvesting and initial filtering difficult and results in a particularly noisy dataset. The process of fine-tuning the data pre-processing and model training pipeline is described in detail in our Report [9]. Here we only report on the final configuration and its performance.

Harvesting content from Twitter (top of Fig. 1) provides content both for manual annotation and model training, as well as for classification and then

Table 1. Classifier accuracy for various choices of N-grams and over- and sub-sampling

	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

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. Filtering keywords were selected in two steps, following an approach similar to that suggested in [11] and using a short list of expert-chosen *seed* keywords, for bootstrapping the process: `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. The top 10 of those were added to the initial seed set: `microcefalia`, `transmitido`, `epidemia`, `transmissao`, `doenca`, `eagypti`, `doencas`, `gestantes`, `infeccao`, `mosquitos`.

2.1 Learning a Relevance Model

Having broadly harvested Zika-related content, the purpose of the tweet classifier is to accurately separate the few interesting tweets, i.e., the `Relevant` class, from a majority of content carrying `News` about Zika, and the general `Noise` class. Based on our prior experience comparing traditional supervised learning (Naive Bayes) with unsupervised topic modelling (LDA [1]), here we focus solely on classification.

The data preparation steps are listed in Fig. 1-left. Firstly, tweet content is reduced to a bag-of-words with N-grams (N = 1,2,3) representation, using POS tagging and lemmatisation and removing common abbreviations, as well as all emoticons and non-verbal forms of expressions. Secondly, we tested two attribute selection approaches, namely Ranking with Information Gain vs Subset Selection, but concluded that neither resulted in improved overall performance. Finally, noting imbalance in the distribution of examples over the classes in the training set: 50.6% `News`, 37.3% `Noise`, 12.1% `Relevant`, we added an extra 600 annotated examples to the `Relevant` class, and applied the SMOTE algorithm [4] to over-sample the `Relevant` class, boosting the examples from 1,214 to 2,428 (12.1% to 24.3%).

We experimented with three popular classification models that have proved effective for short text classification [2]: Support Vector Machines (SVM), Multinomial Naive Bayes (MNNB), and Random Forest (RF). After ruling out SVM due to poor performance (details in [9]), we mapped a space of pre-processing configurations as shown in Table 1. Interestingly, using SMOTE

provides equivalent performance to that obtained by investing extra human annotation effort. Note that the results also show that down-sampling the majority class (`News`) is not as beneficial.

The best overall accuracy figure, 84.1%, is obtained using a Random Forest learner (using an ensemble of 100 trees), with 1,2,3-grams, no attribute selection, and SMOTE-based boosting (weighted average F-measure = 0.84, RMSE = 0.28). This is the classifier we used for the online content relevance detection phase in combination with user ranking, described next.

2.2 User Ranking Metrics

This sparsity of users suggests that TwitterRank may not be effective on this datasets, as it assumes knowledge of the social graph neighbourhood for each candidate user, and requires that meaningful social connections exist within those neighbourhoods. At the same time, note that no ground truth, i.e., explicit knowledge of the top users, is available for evaluation, as our content harvesting was performed purely "in the wild" (the same occurred in the original TwitterRank research [14]).

Our approach is therefore to compare user ranking from TwitterRank with two simple additional, non-topological ranking criteria. For a user u and a set K of keywords, 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 experiment with **Topic Focus** per user, defined as $TF(u) = \frac{R_K(u)}{T_K(u)}$, and with **Overall Focus** per user as $TF(u) = \frac{R_K(u)}{T(u)}$. These count the `Relevant` fraction of u's tweets in the harvest, as an indication of how often user u used the keywords K to express relevant content; and the `Relevant` fraction of u's total tweets in the harvest period, i.e., relative to the user's global interests when posting on Twitter. Note that TwitterRank is agnostic to the set of topics users are interested in. Thus, we modify it to operate on a single topic, and we adapt the original definition of topical differences between two users to work in our context. Please refer to [9] for details.

3 Results

Our experimental dataset consists of a harvest of 278,351 tweets, collected and classified through our online pipeline (Fig. 1) using the keywords presented earlier during a period of 4 months (9–12) in 2016. Using our classifier, we found 15,124 `Relevant` tweets in this set. The distribution of tweets per user is very skewed, with the vast majority of users producing very few `Relevant` tweets during the harvest period: 2 users authored $\geq$20 `Relevant` tweets, 42 between 5 and 20 tweets; 57 : 4 tweets; 209 : 3 tweets, and 12,918 users only one tweet. In practice, the 13,228 candidate users produce 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 2, 3, and 4 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.

Regarding TwitterRank, we note firstly that the small absolute figures are not indicative, as the original paper [14] does not provide any reference figures

Table 2. Top 10 TwitterRank candidate users

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 3. Top 10 topic focus candidate users

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

Table 4. Top 10 overall 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

at all. However we note a significant spread (150%) between the top and bottom ranks in the top-10 list. As noted earlier, however, the significance of this ranking is questionable, because our candidate users have very few connections amongst each other. Despite this, looking at the social connections amongst some of our candidate users, as in Fig. 2 (Appendix A) reveals few but interesting connections, and indeed even a few friends (shown with the double arrow). Note also that all of our top-10 TwitterRank users appear 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.

Comparing this ranking with the other two metrics, we note (Table 3) 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 Table 2 shows that all top 10 TwitterRank users are top-30 Topic Focus users, suggesting that high TwitterRankx may correlate well with high Topic Focus. Interestingly, in Table 4 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`.

4 Conclusions

We have begun exploring the hypothesis that social media analytics can be used to identify individuals who are actively contributing to social discourse on the 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 multiple 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, we found that the very popular TwitterRank algorithm [14] is not very effective. 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 soon the top-k users computed using our metrics will be shown and continually updated.

A Social Graph Fragment

See Fig. 2.

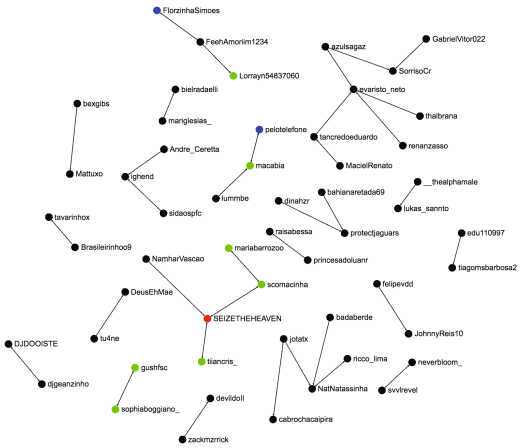

Fig. 2. Fragment of followers and friends graph for candidate users in our experimental dataset. Green nodes are in the top 10 TwitterRank. Blue nodes are in top 10 TwitterRank *and* top 10 Overall Focus. Red nodes are in top 10 TwitterRank *and* top 10 Topic Focus.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
2. 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, pp. 769–777 (2016)
3. CDC: Centers for Disease Control and Prevention (2015). http://www.cdc.gov/dengue/. Accessed 15 Dec 2015

4. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
5. Chen, C., Gao, D., Li, W., Hou, Y.: Inferring topic-dependent influence roles of twitter users. In: Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2014, NY, USA, pp. 1203–1206. ACM, New York (2014)
6. Horne, B.D., Nevo, D., Freitas, J., Ji, H., Adali, S.: Expertise in social networks: how do experts differ from other users? In: Proceedings of the Tenth International AAAI Conference on Web and Social Media, vol. 10, pp. 583–586 (2016)
7. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**, 159–174 (1977)
8. Miles, T., Hirschler, B.: Zika virus set to spread across americas, spurring vaccine hunt, January 2016
9. 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. Research report, School of Computing Science, Newcastle University (2017). https://arxiv.org/abs/1703.03928
10. Missier, P., Romanovsky, A., Miu, T., Pal, A., Daniilakis, M., Garcia, A., Cedrim, D., Silva Sousa, L.: Tracking dengue epidemics using twitter content classification and topic modelling. In: Casteleyn, S., Dolog, P., Pautasso, C. (eds.) ICWE 2016. LNCS, vol. 9881, pp. 80–92. Springer, Cham (2016). doi:10.1007/978-3-319-46963-8_7
11. Nagarajan, M., Gomadam, K., Sheth, A.P., Ranabahu, A., Mutharaju, R., Jadhav, A.: Spatio-temporal-thematic analysis of citizen sensor data: challenges and experiences. In: Vossen, G., Long, D.D.E., Yu, J.X. (eds.) WISE 2009. LNCS, vol. 5802, pp. 539–553. Springer, Heidelberg (2009). doi:10.1007/978-3-642-04409-0_52
12. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of WWW 2010, p. 851 (2010)
13. Wei, W., Cong, G., Miao, C., Zhu, F., Li, G.: Learning to find topic experts in twitter via different relations. IEEE Trans. Knowl. Data Eng. **28**(7), 1764–1778 (2016)
14. 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)
15. Yamaguchi, Y., Takahashi, T., Amagasa, T., Kitagawa, H.: TURank: twitter user ranking based on user-tweet graph analysis. In: Chen, L., Triantafillou, P., Suel, T. (eds.) WISE 2010. LNCS, vol. 6488, pp. 240–253. Springer, Heidelberg (2010). doi:10.1007/978-3-642-17616-6_22