Project Report – CSE 40437/60437
Zikafinder

Diandra Kubo
College of Engineering
University of Notre Dame
South Bend, Indiana (574) 299 – 3699
Email: dkubo@nd.edu

Artur Pimentel
College of Engineering
University of Notre Dame
South Bend, Indiana (574) 292 – 9595
Email: apiment2@nd.edu

***Abstract*—Current methods for reporting disease activity generally take too long to take place. In order to save lives, health authorities should have auxiliary tools that can follow disease spreading in a real-time manner. Given that Twitter has been a useful platform for researchers to track events such as political episodes, commercial trends, and, lately, epidemic spreading, this work focus on the development and deployment of a application, called Zikafinder, that is capable to locate tweets related to two mosquito-related diseases: Zika fever and dengue. At the end, the challenges involved during the development and the ones that will happen in a possible expansion of the work are discussed.**

I. Introduction

Social media has been providing a huge help in the monitoring of many large-scale events. Twitter is one of the most prolific sources for data to keep up with those events, such as important political episodes [1], news exchange [2], commercial trends [3] and epidemic spreading [4].

The latter topic has always been of tremendous importance on health surveillance and epidemiological control, where data collection, processing and analysis are the essential steps that support the decision making process of health departments in order to improve prevention and control efforts [5].

Since last year the Zika virus has been getting more and more attention as it has escalated in South America [6], engendering the need of systems that can accurately monitor its occurrences.

II. Related Work

Similar studies related to detection of epidemic spreading using Tweeter data focus mainly in the correct selection of keywords related to that particular infection in tweets [7] [8] [9] [10]. Differences arise in metrics, set of keywords or the given relevance to those words.

In [7] a set of keywords related to H1N1 influenza, public concern, and vaccine side effects or shortages is adopted in order to collect a number of tweets, displaying them in a map format. Also, that study shows correlations among those different subjects and clusters tweets using Support Vector Regression.

Classes of keywords and other methods are adopted to distinguish tweets that relate to flu but do not report the infection in [8], thus preventing them from corrupting the epidemic tracking.

Linear and multiple regressions to identify flu-related tweets have also been used [9]. Their method starts with a simple hand-chosen set of keywords to generate a first sample of tweets, and then uses this sample to find the 5000 most frequently occurring words, correlating them to ILI rates.

In [10] tweets happening in the surroundings of the main UK urban centers are collected and hand set textual markers receive scores to estimate the Flu rate around cities. They use Porter's Algorithm as in [7] to reduce inflected words and a method such as the one on [9] to generate automatically keywords.

Therefore, it can be seen that recent work using twitter to analyze disease data has been published, but none seem to have used it for the Zika Virus. As this is yet a new topic of discussion, a lot of people in social media are just showing interest or concern about it, and not reporting it as it happens with another diseases transmitted by this same vector, such as dengue.

For that reason, the most common challenges are the selection of relevant tweets to disease surveillance by a correctly choosing keywords and assigning values to them.

III. Problem Statement

Speed is an essential factor when dealing with health issues: an epidemic can spread rapidly if heath authorities do not have means of tracking it down. Because the most popular methods for doing this rely on medical reports, it takes a long time– usually a one to two weeks timespan–for this information to become available [9]. Therefore, we need a tool that can help follow an epidemic in a quicker way, ideally in a real-time manner.

IV. Solution

As a consequence of the need of such a system, we have built our own system architecture for a Twitter data crawler, as Figure 1 shows.

The first step was data acquisition. We collected the data from the Twitter Social Media through the Streaming API [11] by using a definite set of keywords. Table I shows keywords related to disease names and vectors. Table II shows keywords related to the main symptoms of Zika fever and dengue in

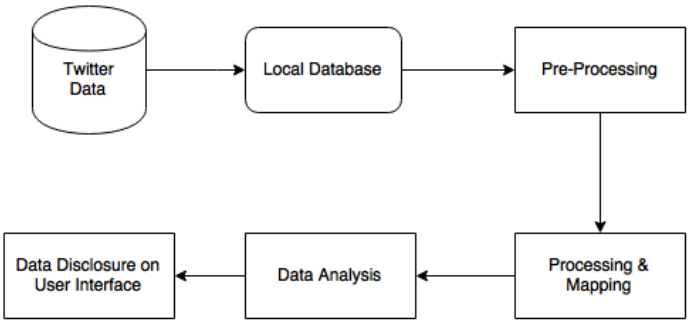

Fig. 1: The system's work flow.

Type of Keyword	Keywords
Disease Name	Zika, dengue, microcephaly, microcefalia
Vectors	mosquitoes, mosquito, Aedes aegypti

TABLE I: Data crawler keywords for disease names and transmitters

different languages in order to cover tweets from most of the Americas.

After those tweets were saved in our local database, we went to the next step, pre-processing. The first problem we encountered was getting tweets with their actual location, because:

- 3.15% of tweets had the bounding box of location on it;
- 59.80% of tweets were wrote by a user with location information

When summing these two categories, we got 62.95% of tweets with location. So the next natural step was to process all the locations available in the profiles of the users who tweeted all the data.

But a huge problem on the location that is available on the users' profile is that in a lot of cases, this information is not correct, with the field being filled with jokes or just other information about the user. To verify these locations, we used the Google Maps API.

Using the Google Maps Geocoding API was another problem on the pre-processing step. This API limits the amount of requests that can be made with the server key they give. So in order to minimize the number of requests made, we first got the unique location descriptions available, so we would process only those ones. And even doing this, we needed 94 keys to be able to process all this locations.

The map with the distribution in the globe of these tweets can be seen on Figure 2, where the green dots are tweets whose location was obtained by the user information, and the red dots are tweets that had a location associated to it.

This process so far creates a database on which we can operate. Considering only the tweets that can be located by the methods cited, we make a second round of searches refining these texts, so that we can start making evaluations of the searches.

Symptom Keyword		
English	**Spanish**	**Portuguese**
fever	fiebre	febre
joint pain	dolor de articulaciones	dor nas juntas
muscle pain	dolor en el cuerpo	dor no corpo
red eyes	ojos rojos	olhos vermelhos

TABLE II: Data crawler keywords for symptoms in the most spoken laguages in the Americas

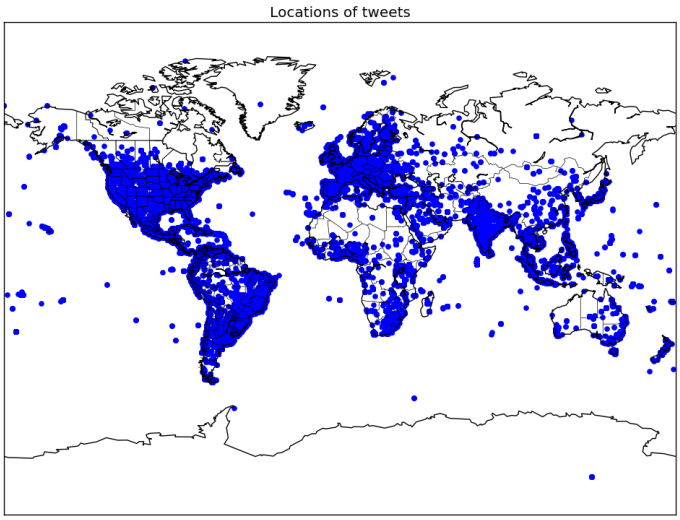

Fig. 2: Tweets distribution

V. Evaluation

The actual processing and Mapping was done after we were able to get the maximum amount of tweets with a location on it.

As we were evaluating which keywords we should search within these tweets that had relevant information for a monitoring system, we were able to perceive initially that a lot of tweets are actually showing concern about the problem, and not reporting its occurrence.

Generally speaking, the concern is mostly towards Brazil, where the cases have been growing exponentially lately. That showed the need for a search with the keywords *Brazil*, *Brasil* (Portuguese) and *Brazilian*. The results can be seen on Figure 3, where these tweets represent 2.38% of the data and we can assume that they are just tweets from people expressing concern about Brazil or sharing news about it. So on all the searches we did after, those tweets are not included, as most of them are just news being shared.

Another combination we tried was searching for the keywords *have fever*, to include people reporting they have fever, as well as *got fever*. Also *tenho febre* and *com febre*, the equivalents in Portuguese. Furthermore *con fiebre* (spanish) and *tengo fiebre*, the counterparts in Spanish. The map with these tweets can be seen on Figure 4.

As fever is the one symptoms that always appears in the dengue and Zika virus manifestations [12], a further look on the report of another symptoms would be more efficient if combined with the report of fever. The first additional

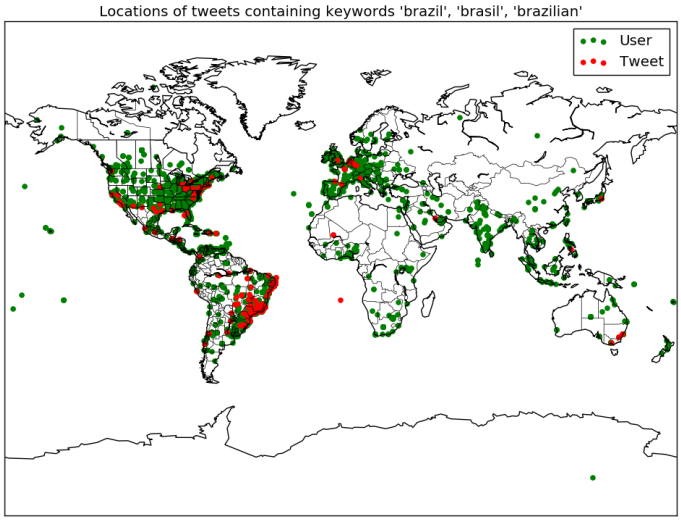

Fig. 3: The combination of 'Brazil', 'Brazilian' and 'Brasil'

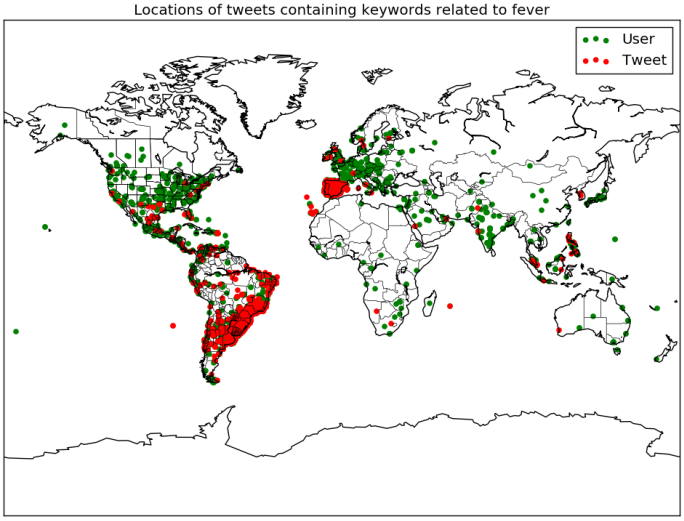

Fig. 4: Searching for fever reports

symptom we looked for was joint pain, another main symptom [13].

To be inclusive of English, Portuguese and Spanish tweets, we did a very broad search looking for the following keywords: *joint pain*, *dor nas juntas* (portuguese), *dolor en las articulaciones*, *dolor de articulaciones* and *dolor articular* (spanish).The map containing that search can be seen on 5.

What appears to be only one tweet are actually two from a very close area. On table III the text from these tweets can be seen, and they show that these tweets are actually reports from the disease we are looking for, what shows this combination of keywords can be used for monitoring the disease on future data, even though they correspond to a very tiny amount.

Next we looked for tweets containing the keywords *red eye* (English), *olho vermelho*, *olhos vermelhos* (Portuguese), *ojo rojo*, *ojos rojos* (Spanish), as it is another symptom linked to this diseases' [14]. The map can be seen of Figure 6. These

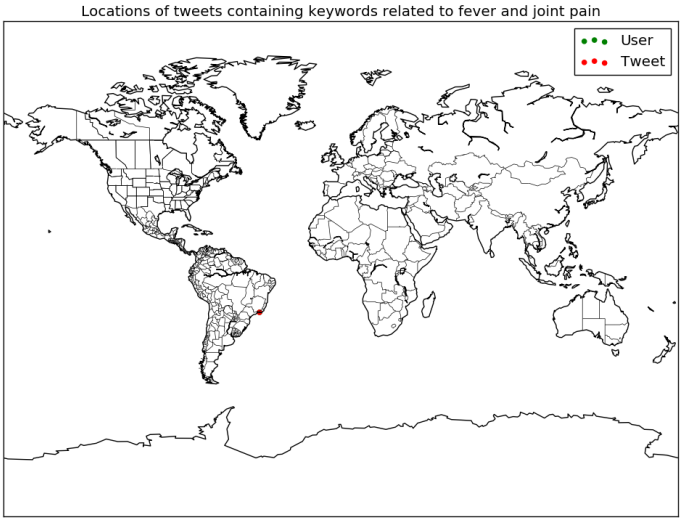

Fig. 5: The combination of 'joint pain' and 'fever'

Original Tweet	Translation
continuo com febre e com dor nas juntas	"I still have fever and joint pain"
vocês lembram que dia desses eu disse que provavelmente havia sido picado por um aedes aegypti? pois é, acordei com febre e dor nas juntas.	"Do you remember that one of these days I said that I probably got bitten by an aedes aegypti? Well, just woke up with fever and joint pain"

TABLE III: Tweets containing joint pain and fever report

tweets represented only 1.07% of the data, and in combination with fever yielded no tweets at all.

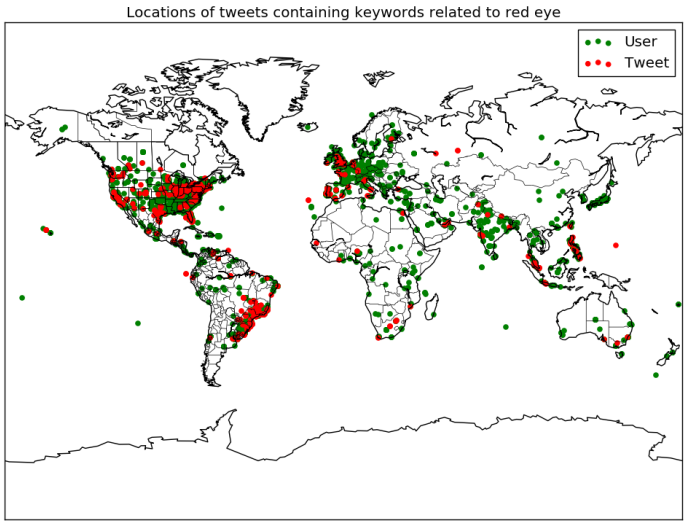

Fig. 6: The search of 'red eyes'

An additional symptom combined with fever was body pain. The search included the following keywords: *muscle pain*, *muscle ache*, *body pain* (English), *dor no corpo*, *corpo dolorido* (Portuguese), *dolor en el cuerpo*, *cuerpo dolorido* (Spanish). The resulting map can be seen on Figure 7.

The last symptom we looked for in combination with fever

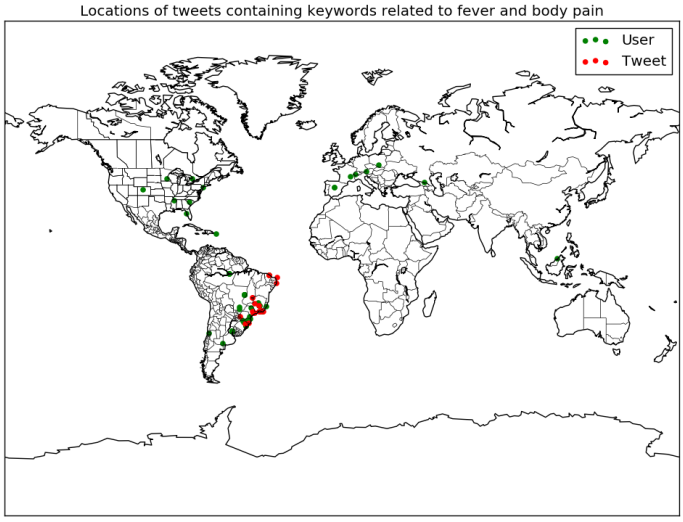

Fig. 7: The search of fever and body pain

was rash, as it is cited as one of the main symptoms as well [15]. As a rash is also associated with an itch, the combination of keywords was: *rash*, *itch*, *mancha*, *coceira* (Portuguese), *erupciones*, *prurito* (Spanish). This combination yielded also a good result as it can be seen on Figure 8, with the texts available at table IV.

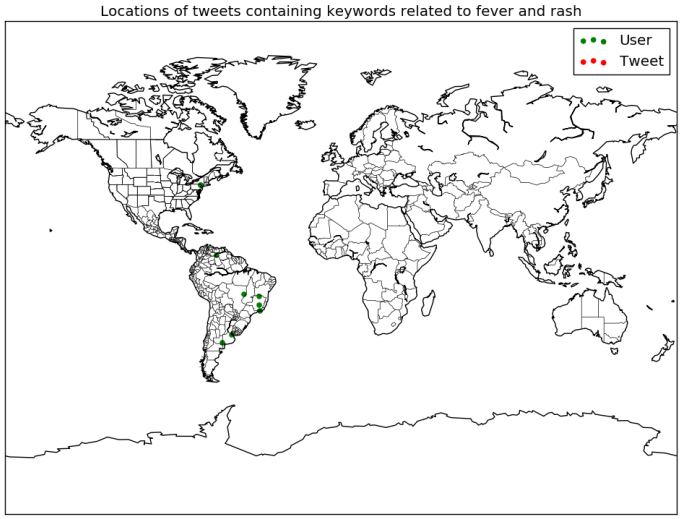

Fig. 8: The search of fever and rash

Given that dengue is a better known disease with a feasible diagnostic, we also looked for the reports of it, as the same vector (Aedes aegypti) that transmits dengue also transmits the zika virus [16], [17]. The keywords we looked for were *have dengue* (English), *tenho dengue*, *estou com dengue*, *to com dengue* (Portuguese), *estoy con dengue*, *tengo dengue* (Spanish). The resulting map can be seen on Figure 9.

Still in the search of the vector that transmits the disease, we also looked for reports of people getting bitten by mosquitoes. A way of doing that was looking for the keywords *got bit* (English), *mordid*, *picad* (verb radicals in Portuguese), *picadura* (Spanish) and *mosquito* (same word for the insect in the three languages). The map with the results can be seen on Figure 10.

Original Tweet	Translation
quem estou com febre, dor de garganta e mancha no corpo? pqp	“ Guess who has fever, throat pain and rash all over the body? Oh god ”
@raashirashi no no, tengo fiebre y no puedo dormir!	“ No no, I have fever and cannot sleep! ”
tengo fiebre y la cara manchada con tintura no me canso de triunfar nunca	“ I have fever and rash on my face, I do not get tired of winning ”
@badwaterwitch nao sei se é isso, mas to com febre e dor no corpo	“ I do not know if it is that, but I do have fever and body pain”
rt @alexandrashb: mañana en la mañana: - mama me siento demasiado mal, creo que tengo fiebre - no seas guevon, no quieres ir a clases	“ Morning after morning - mom, I feel too bad, I think I have fever - I’m not being lazy, I do not want to go to class’
espero que essas manchas sejam so uma alergia, pq acabo de saber que to com febre tmb	“ I hope these rashes are just an alergy, because I just got word also have a fever ”
@tranquilist omg really? do you have fever or do they itch?	Same
@tranquilist but mine dont itch and i dont have fever omg	Same
@apgorito não estou de folga, fiquei a semana toda de cama com febre, e hoje apareceu manchas muita coceira, então não dá pra dormir.	“ Im not on a break, but have been sick all week with a fever, and today some rashes appeared and a lot of itching, so I cannot sleep. ”

TABLE IV: Tweets containing rash and fever report

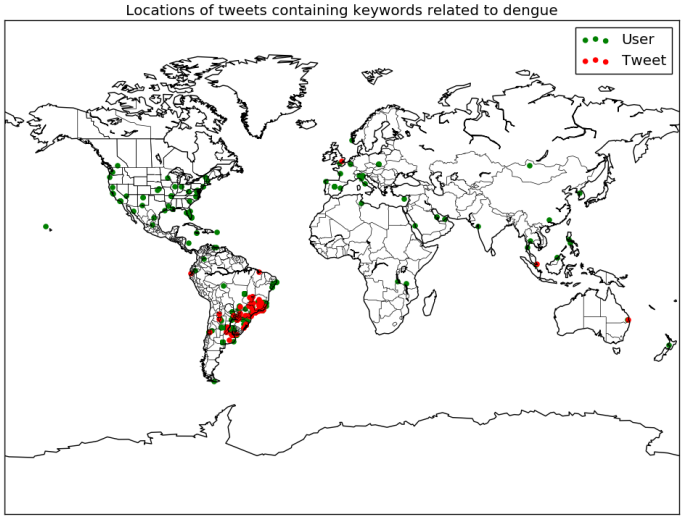

Fig. 9: The search of dengue

After looking for the reports of the disease, we also took a look at some important keywords that give a scale of the amount of discussion around the problem.

On Figure 11 we can see all the tweets containing *microcephaly* and *microcefalia* (the latter for Portuguese and Spanish)

We can see all the tweets that contain *Aedes aegypti*, the vector, on Figure 12.

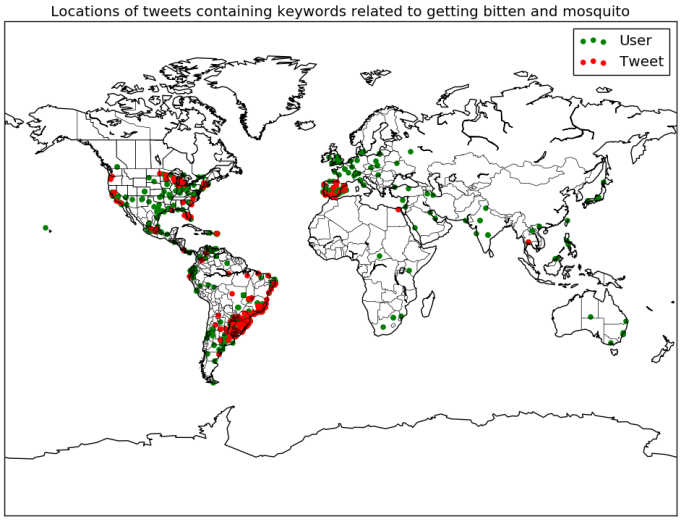

Fig. 10: The search of mosquito and getting bitten

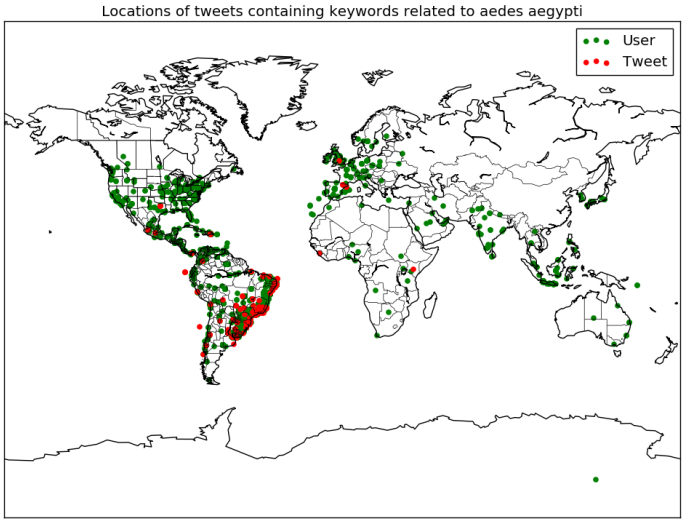

Fig. 12: The search of aedes aegypti

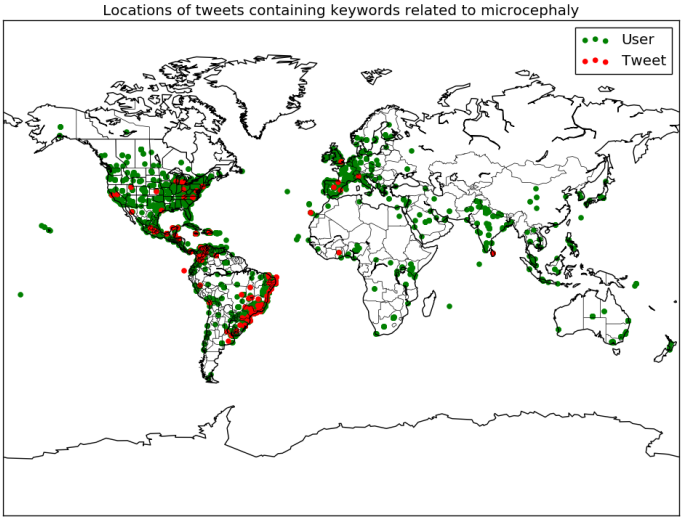

Fig. 11: The search of microcephaly

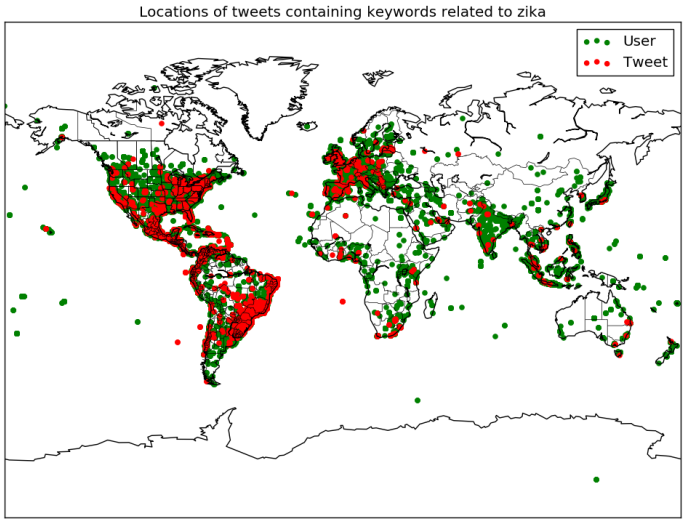

Fig. 13: The search of zika

And finally, we can see on Figure 13 all tweets containing the keyword *Zika*.

The data disclosure and User interface chosen to be developed was through a website with the main maps. The website is already online, and can be checked at *http://www3.nd.edu/~ dkubo/* with a bigger resolution for the images.

VI. Discussion

The searches done so far show there is a great potential in this project to locate and follow the spread of mosquito-related diseases such as dengue and Zika fevers. Although it still lacks refinement, we can draw some basic conclusions from the results shown at the previous section.

To begin, it can be stated from the maps in Figures 5-11 that the second round of searches–the one that happens inside our own database–can be a little more selective.

This happens because we looking for more complex forms of semantics, using key sentences like *got bitten* and so forth, and combinations of such with other keywords, as can be seen in Figure 10.

Furthermore, Figure 3 shows a huge concentration of tweets with words related to Brazil in non-tropical developed countries (USA and European countries), which we considered to be almost certainly related to concernment tweets. Our reasoning is that tweets with the radical *Brazil* are possibly related to discussions about the recent spread in that country and/or pointers to news articles about dengue and Zika virus. Such a behavior of tweet location is similar to the one in Figures 11 and 13, and indicate kinds of words that are possibly useful to cut out non-infection tweets.

This is specially interessant to make an early classification of tweets and not use them in the Gmaps API, which would give us a more eficient search and also mitigate false positives.

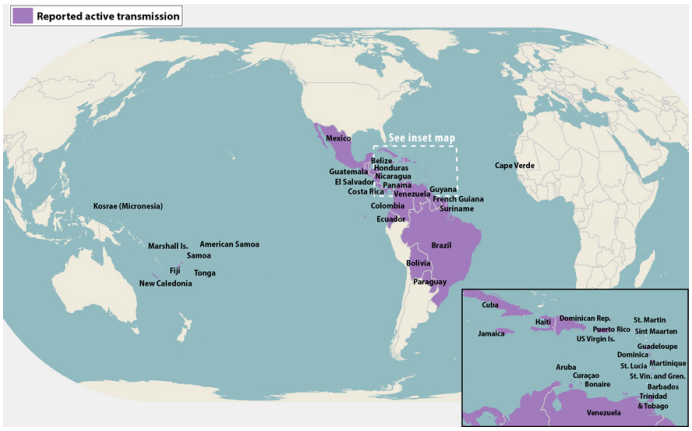

Fig. 14: CDC map of countries with reported activity of Zika virus by 04/26/2016 [23]

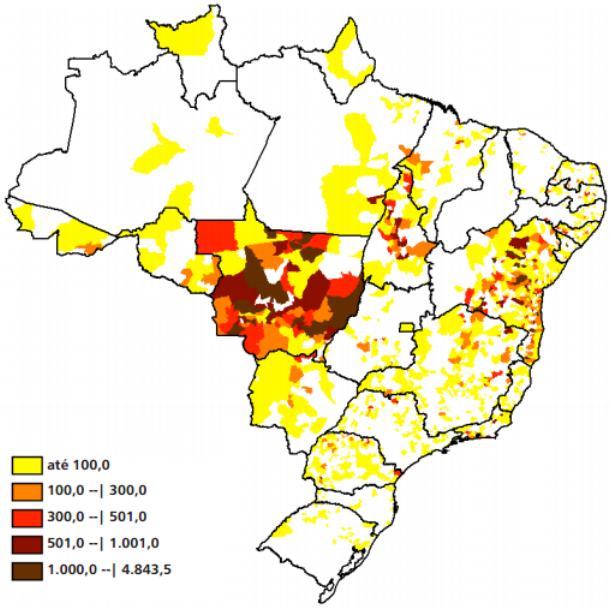

Fig. 15: Official report from Health Ministry of Brazil of Zika activity by municipalities until 04/07/2016. Each color indicates the incidence rate (/100,000 hab.) [25]

It can be seen also that our current tool has little sparsity. This is possible because our initial database of tweets is very large, since it conveys texts from 03/24/2016 to 04/25/2016, or roughly one month of collected data. This also opens the oppotunity in the future of making a weekly analysis of these tweets, so we can keep up the evolution of the disease with better time resolution and make more observations about this tool.

Comparing the maps with fewer tweets, Figures 6-10 with the ground truth data for Zika virus and dengue in Figures 14, 15, and 16, it can be seen that there is some overlapping, which indicates that this path is good, but still have problems of precision.

However, Table IV shows a that we successfully collected

Fig. 16: WHO report for countries with dengue activity by 04/29/2016 [24]

meaningful tweets, ones that contain two or more of the main symptoms of dengue and Zika fevers. This indicates that the following challenge is to eliminate as many false positives as possible to make a better mapping of the epidemic.

Although we already have a reasonable amount of data to run tests and help on the design of our system, we are still collecting more for the future. We also intend to start running new searches using more keywords that represent the symptoms of the disease, and possibly also search for the report of cases of Chikungunya, which is another disease that people tend to report on social media, that has also been growing and that share the same vector, the Aedes Aegypti, with the Zika Virus [18].

That way, we hope to broaden the reach of the tweet collection, since it is reasonable to suppose that people oftentimes will not look for a doctor to diagnose a possible Zika infection and thereafter tweet that they actually have Zika fever. It may happen because the disease has often mild symptoms that can last less than a week [19]. One group of people, however, is potentially good to follow: pregnant women.

Given the correlation between the fever and microcephaly [20], we believe that pregnant women and her families are more likely to report suspicions about Zika virus. This indicate that using keywords related to pregnancy can be a possibility. Maybe these people can report cases of Zika they have heard or seen at their neighbourhoods, what would be valuable as a source of infection mapping.

Obviously, this expansion of keywords also increases the challenge of filtering relevant tweets. With this new sprectre of keywords, other types of filters, such as a sentiment analysis tool, could be useful for this study. By using that, we could filter out positive tweets, which we believe that will hardly represent a text of someone who is telling followers about disease suspicions. Nevertheless, this idea is still open to evaluation.

Another idea comes from a recent study [21] that has been successful in modeling disease spreading from social interactions on twitter. A similar model for relating Aedes presence and fever outcome prediction in an specific area could be built. It can work by using geo-tagged tweets about the mosquito activity to estabilish a region where the chances

of getting Zika is bigger. Other tweets that come from tha same region and report suspicious symptoms could have bigger chance of reporting Zika.

This would also open up this project to categorize different tweets according to their contents in classes such as *infection*, *suspicion*, and *potential disease activity*. In that way we can stabilish a better functionality for the Zikafinder.

A point to be made is that user location, although widely used, is just one way of guessing the location of Zika ocurrence when there is no geo-tagging for the tweet. When using that, we can happen to locate a report in a different place of where it should be, because a user location can only be, for example, the hometown of the user, while the user is actually in a entirely different place.

To bypass this, we can use the content of that user tweets in general to reveal the right place of ocurrence. If that user's profile contains tweets like "It's sunny here in [place]", they can be used to give a better guess of the user's location.

Other limitation is the possibility of a report to be potentialy not located by the user's position. This can happen in tweets with text like: "My Mother has been feverish last two days". The user location can not be the place where the fever is happening, but it is still the best guess when the tweet provides no other information.

Even though all of those features are implemented, there is a limitation of location resolution, and we may not be more precise in locating a infection than the city it happened. Health authorities in city level are the ones more affected by this.

There is also the issue of separating public awareness tweets from actual disease reporting. So far we consider every tweet collected as an actual zika or dengue finding, which is clearly false. Most tweets come from news report and fear from the disease. And although many of these tweets relate to new official data about the Zika fever, there are those that are simply reporting a new public effort in fighting the epidemic or how to deal with Aedes aegypti.

Those kind of tweets could receive a lower score in a hipotetical model for classifying tweet reporting. Surely works such as [8] can give us some insights for this. According to this paper certain combinations of subject-verb-object sequences are more frequent in infection tweets than in concerned awareness ones. Developing a score to those kind of semantic characteristics can be helpful. Another characterstic of awareness tweets is the use of numbers, generally from recent cases that are shown in the news.

There is a question, however, on how all those heuristics would be implemented. Is it necessary to have a machine learning model to read those tweets and classify them according to its contents? Or is it better to synthetisize those ideas in a simpler manner? We could develop a system of scores given all the elements in a tweet to decide wheter it reports Zika or not, and a separate system to locate tweets. Possible ideas lie in the use of models like in [8] for the first task, and the system for locating local words for the second task, such as the one in [22]. It can be useful to take a look in the rest of the tweets of a user to gather more information related to the user location.

The steps taken so far showed that there are and will be many challenges to tackle in future work with this project. We need to braoden our tweet collection, separate relevant tweets from others, locate Zika, dengue, and Chikungunya cases given the information cotained in the tweet, and get more ground truth data to evaluate those models.

VII. Conclusion

The making of a tool to follow an epidemic automatically is something with reasonable applications today, where the velocity of disease reporting can be crucial to save as many lives as possible.

With this project, we show that by searching for specific keywords and its combinations we can collect meaningful reports and suspicion of disease activity. Tweets with more complex sentences and symptom description generally give neater, less polluted results.

One challenge is to reduce noise as much as possible at the same time that we do not lose important information in other kinds of Twitter text. Another one is to develop better ways to locate an infection.

There is a lot that can be done with those initial results, and future work reserves many features to be added to the Zikafinder.

References

[1] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing, "A latent variable model for geographic lexical variation," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010, pp. 1277–1287.

[2] K. Lerman and R. Ghosh, "Information contagion: An empirical study of the spread of news on digg and twitter social networks." *ICWSM*, vol. 10, pp. 90–97, 2010.

[3] B. Hollerit, M. Kröll, and M. Strohmaier, "Towards linking buyers and sellers: detecting commercial intent on twitter," in *Proceedings of the 22nd international conference on World Wide Web companion*. International World Wide Web Conferences Steering Committee, 2013, pp. 629–632.

[4] E. Aramaki, S. Maskawa, and M. Morita, "Twitter catches the flu: detecting influenza epidemics using twitter," in *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2011, pp. 1568–1576.

[5] R. Bonita, R. Beaglehole, and T. Kjellström, *Basic epidemiology*. World Health Organization, 2006.

[6] A. S. Fauci and D. M. Morens, "Zika virus in the americasyet another arbovirus threat," *New England Journal of Medicine*, 2016.

[7] A. Signorini, A. M. Segre, and P. M. Polgreen, "The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic," *PloS one*, vol. 6, no. 5, p. e19467, 2011.

[8] A. Lamb, M. J. Paul, and M. Dredze, "Separating fact from fear: Tracking flu infections on twitter." in *HLT-NAACL*, 2013, pp. 789–795.

[9] A. Culotta, "Towards detecting influenza epidemics by analyzing twitter messages," in *Proceedings of the first workshop on social media analytics*. ACM, 2010, pp. 115–122.

[10] V. Lampos and N. Cristianini, "Tracking the flu pandemic by monitoring the social web," in *Cognitive Information Processing (CIP), 2010 2nd International Workshop on*. IEEE, 2010, pp. 411–416.

[11] Twitter, "Streaming api," https://dev.twitter.com/streaming/overview, 2016.

[12] M. R. Duffy, T.-H. Chen, W. T. Hancock, A. M. Powers, J. L. Kool, R. S. Lanciotti, M. Pretrick, M. Marfel, S. Holzbauer, C. Dubray *et al.*, "Zika virus outbreak on yap island, federated states of micronesia," *New England Journal of Medicine*, vol. 360, no. 24, pp. 2536–2543, 2009.

[13] C. V. Ventura, M. Maia, B. V. Ventura, V. V. D. Linden, E. B. Araújo, R. C. Ramos, M. A. W. Rocha, M. D. C. Carvalho, R. Belfort Jr, and L. O. Ventura, "Ophthalmological findings in infants with microcephaly and presumable intra-uterus zika virus infection," *Arquivos brasileiros de oftalmologia*, vol. 79, no. 1, pp. 1–3, 2016.

[14] J. C. Kwong, J. D. Druce, and K. Leder, "Zika virus infection acquired during brief travel to indonesia," *The American journal of tropical medicine and hygiene*, vol. 89, no. 3, pp. 516–517, 2013.

[15] J. Olson, T. Ksiazek *et al.*, "Zika virus, a cause of fever in central java, indonesia," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 75, no. 3, pp. 389–393, 1981.

[16] S. Christophers *et al.*, "Aedes aegypti (l.) the yellow fever mosquito: its life history, bionomics and structure." *Rickard.*, 1960.

[17] N. Marchette, R. Garcia, A. Rudnick *et al.*, "Isolation of zika virus from aedes aegypti mosquitoes in malaysia." *American Journal of Tropical Medicine and Hygiene*, vol. 18, no. 3, pp. 411–415, 1969.

[18] C. for Disease Control, P. (CDC *et al.*, "Chikungunya fever diagnosed among international travelers–united states, 2005-2006." *MMWR. Morbidity and mortality weekly report*, vol. 55, no. 38, p. 1040, 2006.

[19] V. Sikka, V. K. Chattu, R. K. Popli, S. C. Galwankar, D. Kelkar, S. G. Sawicki, S. P. Stawicki, T. J. Papadimos *et al.*, "The emergence of zika virus as a global health security threat: A review and a consensus statement of the indusem joint working group (jwg)," *Journal of Global Infectious Diseases*, vol. 8, no. 1, p. 3, 2016.

[20] J. Mlakar, M. Korva, N. Tul, M. Popović, M. Poljšak-Prijatelj, J. Mraz, M. Kolenc, K. Resman Rus, T. Vesnaver Vipotnik, V. Fabjan Vodušek *et al.*, "Zika virus associated with microcephaly," *New England Journal of Medicine*, vol. 374, no. 10, pp. 951–958, 2016.

[21] A. Sadilek, H. A. Kautz, and V. Silenzio, "Modeling spread of disease from social interactions." in *ICWSM*, 2012.

[22] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 759–768.

[23] "Cdc map of zika virus," http://www.cdc.gov/zika/, accessed: 2016-02-01.

[24] "World health organization map of risk diseases," http://apps.who.int/ithmap/, accessed: 2016-02-01.

[25] S. de Vigilância em Saúde Ministério da Saúde, "Boletim epidemiolgico volume 47," http://goo.gl/mjhqIB, 2016, accessed: 2016-02-01.