Twitter Knowledge Inference for Latin American bioSurveillance

Arturo López Pineda, MS
University of Pittsburgh, Department of Biomedical Informatics, Pittsburgh, PA, USA, arl68@pitt.edu

Fernando Suárez Obando, MD, MS
Pontificia Universidad Javieriana, Instituto de Genética Humana, Bogota, DC, Colombia,
fernando.suarez@javeriana.edu.co

Charalampos S. Floudas, MD, PhD, MS
New York University, Center for Health Informatics and Bioinformatics, New York, NY, USA,
Charalampos.Floudas@nyumc.org

ABSTRACT

Disease biosurveillance (DB) in Latin America is primarily done through national level aggregation of reports from sentinel hospitals in each country. In this study we propose a novel approach to DB based on a method for automatic monitoring of content posted on the popular social network Twitter. We used a list of 25 symptoms for 7 syndrome categories and translated them into colloquial terms in Spanish. We then extracted Twitter messages containing those terms and used Natural Language Processing (NLP) to establish the location for each message. Finally we aggregated results by country and created a health map of the Twitter messages of Healthcare Interest (THI). Our results show that the proposed method can be used as a low-cost tool to support the task of biosurveillance for public health in Latin America.

Keywords: biosurveillance, social networks, natural language processing

1. INTRODUCTION

Accurate estimation of country-specific burden of disease is a challenging task because of the organizational complexities it involves. In low-income countries reporting is usually done by paper-based methods, a cumbersome and often unreliable method. However, some excellent examples have been implemented in low-income countries to try to increase access to Electronical Health Records (EHR) and its corresponding biosurveillance systems (Sanjoaquin et al. 2013). In contrast, high-income countries have widespread availability of EHR technologies and can consequently use methods for digital biosurveillance. Such systems have been shown to be able to predict an ongoing outbreak with high degrees of accuracy (Tsui 2003). Countries of the Latin American region have diverse infrastructures and resources. However, the traditional methods of biosurveillance have proven to be inefficient. The measles outbreak in Colombia and Venezuela in 2002, the H1N1 influenza pandemic with earlier cases identified in Mexico and southern USA in 2009, the Haiti cholera outbreak after the earthquake of 2010, are just some examples of the challenges that the region faces.

Timely and accurate biosurveillance tools are greatly needed to assist decision makers in effectively planning the appropriate course of action. Traditional Monitoring of disease has been implemented in different ways: classification of different sources of the EHR, i.e. triage diagnoses (Olszewski 2003), emergency department chief complaints (Chapman et al. 2005), emergency department dictation notes (Tsui et al. 2011). More recently other sources of information that do not rely on the EHR information have been investigated for public health surveillance, like monitoring nation-wide purchases of over-the-counter health products, such as medication and thermometers (Wagner et al. 2004). The explosion of web-based and social media infomartion has further introduced new opportunities of monitoring diseases. Some examples of these include trends on search patterns to detect influenza outbreaks (Ginsberg et al. 2008), detection of influenza outbreaks through analysis of Twitter messages (Culotta 2010), (Signorini et al. 2011). More specifically, Twitter has already proven to be an efficient tool for biosurveillance, as in the case of monitoring the cholera outbreak in Haiti in 2010 (Hirschfeld 2012). In our previous study (Lopez Pineda et al. 2011) we used open source software to show the potential of Twitter for monitoring Influenza in spanish-speaking populations, using a single keyword. For the current study we propose to monitor self-reported content of Twitter messages from Latin America to detect nation-wide prevalence of various syndromes.

2. METHODS

We aggregated a dataset of Twitter messages that were later assigned to a symptom code of the Unified Medical Language System (UMLS). Using a natural language process we filtered out those messages that were adding noise to the aggregation. Finally, we built a classification model for the messages. A brief description of these methods is described ahead in this section.

2.1 DATASET

The syndrome categories that we selected is similar to a previous study (Chapman et al. 2005). The full list of keywords, terms and categories that we used for our study can be seen in Table 1. The set of symptoms used for this study is not an exhaustive set of possible symptoms for the syndromes of interest, but a representative one based on the advice of clinicians. We translated the UMLS codes into colloquial Spanish keywords based on personal experience (two authors are native speakers: Mexico and Colombia). The different translations are menat to capture the richness and regional diversity of the language.

We extracted on May 2, 2013 a set of 100 records for each keyword. The messages time of creation spans from May 1, 2013 to May 3, 2013. For each message we recorded the username, message content, timestamp of the message, and self reported location of the user from the user account.

Table 1: List of keyword terms used for search

Syndrome Category	UMLS Code	UMLS description	Colloquial Spanish search terms
Respiratory	C0021400	Influenza	*gripa, gripe*
	C1740837	Acute nasal congestion	*congestión nasal, mocos, rinorrea*
	C0010200	Coughing	*tosiendo, tos, carraspera*
Botulinic	C1858502	Ocular movement abnormalities	*tic en el ojo*
	C1848464	Ocular muscle abnormalities	*me duele el ojo*
	C0011168	Deglutition disorders	*se me cerró la garganta*
	C1527347	Difficulty speaking	*se me fue la voz, afónico*
Gastrointestinal	C0027497	Nausea	*nausea*
	C0042963	Vomiting	*me vomito*
	C0000737	Abdominal pain	*dolor de estómago, dolor de barriga*
Neurologic	C0018681	Headache	*tengo dolor de cabeza, cefalea*
	C0149931	Migraine disorders	*migraña*
Constitutional	C0015967	Fever	*tengo temperatura, fiebre*
	C0085593	Chills	*escalofrio*
	C0231218	Malaise	*malestar*
Hemorrhagic	C0019080	Hemorrhage	*hemorragia, sangrado, sangrar*
Allergic	C1971712	Rash	*sarpullido, roncha*
	C0349790	Exacerbation of asthma	*ataque de asma, ahogo, falta de aire*
	C0423153	Lacrimation	*me llora el ojo*
	C0022281	Itching of eye	*me pica el ojo*
	C0033774	Pruritus	*tengo comezón*
	C0235267	Redness of eye	*ojos rojos*
	C2220053	Bilateral puffy eyelids	*ojos hinchados*
	C1260880	Rhinorrhea	*moqueando*
	C0235564	Sneezing excessive	*estornudo*

2.2 Natural Language Processing

A Natural Language Processing (NLP) was used to parse the search query results of the Twitter website (http://search.twitter.com). The NLP parsed the full website to select only the content of interest, and stored it into a tab separated values file.

In Twitter, there is the option for automatically displaying geolocation information for each message, however, in our study, most users do not have this option activated. Thus, the NLP retrieved the self-reported location from each user's account. Finally, we cleaned this location description and assigned it to the most probable country given a predefined list of countries, cities and states.

Another option in Twitter is the use of Re-Tweets (RT), which is primarily generated by a user different from the one who is posting the content. Re-Tweets were eliminated, as well as duplicate messages from different users.

3. Results

The total number of Twitter messages of Healthcare Interest (THI) aggregated by country and syndrome category is shown in Table 2. This table also includes the USA and Spain, because we constrained our Twitter search for language and not for region. Therefore, Latin American countries that do not speak Spanish, i.e. Brazil, are presented with smaller counts for each syndrome.

Table 2: THIs by country and syndrome

	Respiratory	Botulinic	Gastrointestinal	Neurologic	Constitutional	Hemorrhagic	Allergic	**Total**
Argentina	34	54	21	12	17	15	119	**272**
Mexico	31	36	18	21	31	19	87	**243**
Venezuela	52	12	18	13	26	14	51	**186**
Colombia	60	10	26	32	22	6	22	**178**
Brazil	24	2	0	0	0	35	9	**70**
Ecuador	7	3	4	17	6	4	7	**48**
Guatemala	8	3	18	6	4	2	7	**48**
Dom. Rep.	3	24	3	3	3	3	5	**44**
Chile	4	2	7	4	4	1	8	**30**
Peru	4	6	3	1	4	3	8	**29**
Puerto Rico	7	2	5	4	2	2	3	**25**
Panama	5	1	3	4	3	0	4	**20**
El Salvador	0	3	1	4	1	1	2	**12**
Paraguay	1	2	0	0	4	0	3	**10**
Uruguay	1	1	0	0	0	1	6	**9**
Costa Rica	2	0	2	0	1	0	1	**6**
Nicaragua	2	0	1	0	0	1	0	**4**
Honduras	0	0	0	2	0	0	1	**3**
Bolivia	0	0	0	0	0	1	0	**1**
USA	7	6	3	5	6	4	20	**51**
Spain	29	24	23	7	17	12	85	**197**
Total	**281**	**191**	**156**	**135**	**151**	**124**	**448**	**1486**

For comparison, we normalized the THIs per country population and also the by estimated Twitter users in each country (http://gs.statcounter.com). The adjusted indices allow for better interpretation of the results and comparison across countries. Figure 1 shows a radial graph of THIs.

Puerto Rico, Venezuela and Argentina have the highest number of THIs per population. Guatemala, but also Argentina and Puerto Rico have the highest number of THIs per Twitter user. Argentina and Venezuela are in the top ranking of countries with more THIs overall. In Argentina, the top syndrome is Allergic, while in Venezuela has almost a tie between Allergic and Respiratory. Meanwhile, Guatemala has a high proportion of Gastrointestinal THIs.

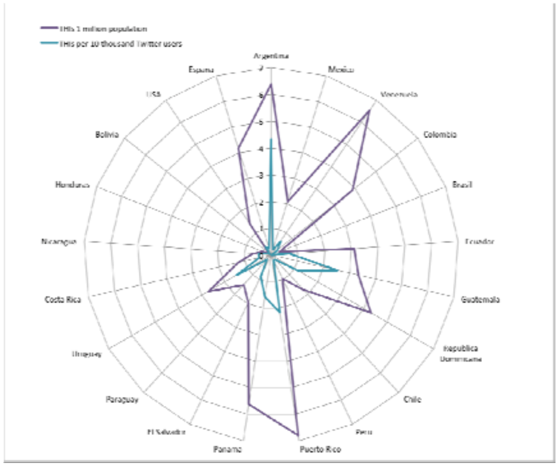

Figure 1: Radial graph of Tweets of Healthcare Interest

The Pan American Health Organization (PAHO 2013) reports an increased activity of Influenza-Like-Illness during the same time of our study for Colombia and Argentina, and very low activity for North America, Central America and other countries in South America (Venezuela is not listed in the report). It is possible that our graphs reflect the same trend. Figure 2 shows a heat map of THIs per country adjusted per population for three of the symptom categories: Respiratory, Gastrointestinal and Allergic. It is notable that Respiratory syndrome has a higher prevalence in Venezuela and Colombia and the surrounding countries tends to fade out. For the Allergic Syndrome, Argentina and Uruguay has a higher prevalence. Meanwhile for Gastrointestinal, it is Guatemala and Puerto Rico. A similar trend is observed in the other syndromes with neighboring countries having similar values, indicating possible spread of contagious vectors and similarities of environmental influences.

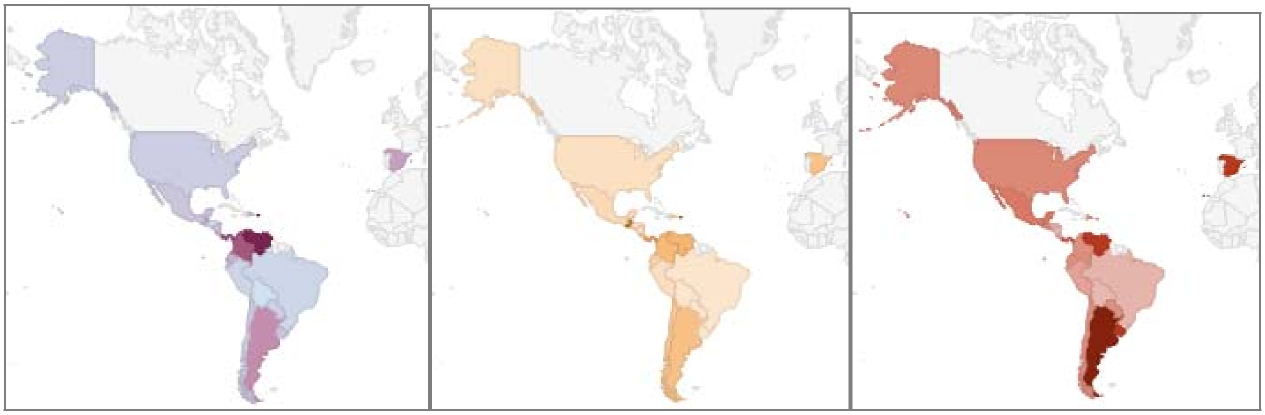

Figure 2: Heat maps of THIs per million population for Respiratory (left), Gastrointestinal (center) and Allergic (right) syndromes

4. Limitations and Future Work

A limitation of our study stems from the limited list of symptoms and symptom-related keywords, but a longer list would perhaps introduce more noise. Information about the location of the tweets also poses a significant hurdle, because self-reported location might not be accurate or up-to-date. The short period of extraction results in a limited set of messages and therefore reduced ability to identify trends. The choice of date and time for the extraction might introduce bias related to temporal patterns of syndromes and the possible propensity of users to post content at night, or during week days, etc. The list length and period of extraction will both be increased in the second post-exploratory phase of our study. Our results are highly dependant on the accuracy of the NLP. We eliminated many THIs because of non-realistic locations, i.e. “wonderland”, “in the sky”. Some other locations were not clearly identified with a single country, i.e. “Merida” can be found at least in Mexico, Venezuela and Spain. The most probable country due to google maps search was selected.

In future work we would like to include other languages, particularly portuguese to better reflect the trends in Brazil. We would also like to refine our NLP, to include temporality, person of interest and intention of messages. Finally, the use of a machine-learning process might be beneficial to trigger alerts for syndrome trends per country.

5. CONCLUSION

We have shown that Twitter can be a useful tool to monitor the geographic and temporal distribution of symptoms related to different syndromic categories. Twitter messages of Healthcare Interest (THI) are indeed ocurring in Latin American region, and it is posible to indicate subregions where a given syndrome is more prevalent. Such a tool could function as a low-cost and reliable platform facilitating the work of public health authorities in each country. Furthermore, it can help trans-border biosurveillance for international organizations such as the World Health Organization (WHO), as it is independent of the reporting framework and capabilities of each country. With the continuing increase in the availability of the Internet and mobile technologies it can be expected that the benefits of such a system for tracking outbreaks across Latin America will be increasing as well.

REFERENCES

Chapman, W.W., Dowling, J.N. & Wagner, M.M., 2005. Classification of emergency department chief complaints into 7 syndromes: a retrospective analysis of 527,228 patients. Annals of emergency medicine, 46(5), pp.445–455.

Culotta, A., 2010. Detecting influenza outbreaks by analyzing Twitter messages.

Ginsberg, J. et al., 2008. Detecting influenza epidemics using search engine query data. Nature, 457(7232), pp.1012–1014.

Hirschfeld, D., 2012. Twitter data accurately tracked Haiti cholera outbreak. Nature.

Lopez Pineda, A. et al., 2011. Monitoring Twitter content related to influenza-like-illness in Spanish-speaking populations. Emerging Health Threats Journal, 2011(4), p.11185.

Olszewski, R.T., 2003. Bayesian classification of triage diagnoses for the early detection of epidemics. In Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference.

PAHO, 2013. Regional Update EW 16, 2013. Influenza and other respiratory viruses. pp.1–9.

Sanjoaquin, M.A. et al., 2013. Surveillance Programme of IN-patients and Epidemiology (SPINE): Implementation of an Electronic Data Collection Tool within a Large Hospital in Malawi. PLOS Medicine, 10(3), pp.e1001400–e1001400.

Signorini, A., Segre, A.M. & Polgreen, P.M., 2011. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. PLoS ONE, 6(5), pp.e19467–e19467.

Tsui, F.-C., 2003. Technical Description of RODS: A Real-time Public Health Surveillance System. Journal of the American Medical Informatics Association, 10(5), pp.399–408.

Tsui, F.-C. et al., 2011. Probabilistic Case Detection for Disease Surveillance Using Data in Electronic MedicalRecords. Online Journal of Public Health Informatics, pp.1–17.

Wagner, M.M. et al., 2004. National Retail Data Monitor for public health surveillance. Morbidity and Mortality Weekly Report, 53 Suppl, pp.40–42.

Authorization and Disclaimer

Authors authorize LACCEI to publish the paper in the conference proceedings. Neither LACCEI nor the editors are responsible either for the content or for the implications of what is expressed in the paper.

ALP is supported by The International S&T Fulbright Award and CONACyT-Mexico.