2017

Infodemiology for Syndromic Surveillance of Dengue and Typhoid Fever in the Philippines

Ma. Regina Justina E. Estuar
Ateneo de Manila University

Kennedy E. Espina
Ateneo de Manila University

Custom Citation

Kennedy Espina, Ma. Regina Justina E. Estuar, Infodemiology for Syndromic Surveillance of Dengue and Typhoid Fever in the Philippines, Procedia Computer Science, Volume 121, 2017, Pages 554-561, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2017.11.073.

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 121 (2017) 554–561

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies, CENTERIS / ProjMAN / HCist 2017, 8-10 November 2017, Barcelona, Spain

Infodemiology for Syndromic Surveillance of Dengue and Typhoid Fever in the Philippines

Kennedy Espina[a]*, Ma. Regina Justina E. Estuar[a]

[a]*Ateneo de Manila University, Katipunan Avenue, Quezon City 1108, Philippines*

Abstract

Finding determinants of disease outbreaks before its occurrence is necessary in reducing its impact in populations. The supposed advantage of obtaining information brought by automated systems fall short because of the inability to access real-time data as well as interoperate fragmented systems, leading to longer transfer and processing of data. As such, this study presents the use of real-time latent data from social media, particularly from Twitter, to complement existing disease surveillance efforts. By being able to classify *infodemiological* (health-related) tweets, this study is able to produce a range of possible disease incidences of Dengue and Typhoid Fever within the Western Visayas region in the Philippines. Both diseases showed a strong positive correlation ($R > .70$) between the number of tweets and surveillance data based on official records of the Philippine Health Agency. Regression equations were derived to determine a numerical range of possible disease incidences given certain number of tweets. As an example, the study shows that 10 infodemiological tweets represent the presence of 19-25 Dengue Fever incidences at the provincial level.

Peer-review under responsibility of the scientific committee of the CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies.

Keywords: Social Media; Epidemiology; Infodemiology; Twitter; Disease Outbreak; Visualization; Prediction

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: kespina@ateneo.edu

Peer-review under responsibility of the scientific committee of the CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies.
10.1016/j.procs.2017.11.073

1. Introduction

Online social networking websites in their current form are avenues where people can publicly express emotions, sentiments, reactions, and behavior in real-time[1]. Twitter, a microblogging website, allows data scientists to get user content through the publicly available streaming API (Application Program Interface). This tool provides data that can be used to generate reports for different objectives. In this study, the main objective is to collect, classify, and visualize infodemiological (Health-related) tweets located in the Philippines.

The devolved approach in public health management in the Philippines is a hindrance to obtaining real-time health data at the national level. The country has yet to expand its implementation of the Philippine Health Information Exchange (PHIE) beyond the design to connect fragmented subsystems. This presents a problem since data may come in too late resulting to critical policies not being implemented on time. However, social media data is a space that could be tapped for real-time data gathering, as it may indicate possible outbreaks based on symptoms found in posts. The method in identifying possible public health risks prior an official clinical diagnoses for appropriate response is called Syndromic Surveillance[2,3]. A relatively new term, Infodemiology studies the behavior of health-related information in the Internet. Infodemiology falls under the umbrella term of Syndromic Surveillance, where the source of latent data comes mostly from social media, where people post their current health conditions.

The Philippines is a bilingual country that uses both the Filipino and English language as a means of communicating daily. Moreover, some regions use their own dialects over the national language. This is also seen in the way Filipinos use social networking websites, where they use both languages and the different dialects in their own discretion. In this study, keywords and phrases that indicate health statuses for both languages are taken into consideration when classifying the tweets that were collected. The tweets are then visualized in a spatiotemporal map. Spatially, the map indicates where in the Philippines the concentration of tweets can be found. The temporal aspect helps in visualizing the spread of information over time.

This study does not aim to replace already existing epidemiological and surveillance tools being used by the Philippines' health agency, the Department of Health (DOH). Rather, its objective is to complement what is already being used, and integrate the results of the study to provide a faster and alternative source of data.

2. Review of Related Literature

2.1. Defining Infodemiology and Syndromic Surveillance

Infodemiology, a portmanteau of "Information" and "Epidemiology", is formally defined as *"the science of distribution and determinants of information in an electronic medium, specifically the Internet, or in a population, with the ultimate aim to inform public health and public policy*[4]*"*. The main goal of epidemiology is to seek the pattern of distribution of a disease to help in gaining the context from which a public health policy may be made[4,5]. Infodemiology, on the other hand, adds a new layer for dealing with epidemiology in such a way that its focus is more on how information about a disease is being talked about online, more than how a disease is actually spreading in the real world as reported by surveillance systems[4]. This means that infodemiology prioritizes the changes on the patterns of gathered data, which in turn may signal a change in the behavior of the people or the users online[4]. This is the key to the epidemiology aspect of infodemiology, since a "pulse" in health-related posts may indicate the prevalence of health concerns within an area.

Syndromic surveillance is a system that is put into place for identifying clusters of illnesses early on before an outbreak happens. This uses population data that are generated before actual diagnoses are confirmed and reported to public health agencies[6]. Doing so provides enough time for public policies to be implemented for the mitigation of possible risks before an actual outbreak happens. Infodemiology then plays a major role in syndromic surveillance by using data that are found in the Internet - a term coined as Infoveillance, the "automated and continuous analysis of unstructured, free text information available on the Internet[7]".

2.2. Classification of Infodemiological Tweets

An initial study was made to determine the feasibility of using a classification algorithm to determine *Infodemiological* tweets in the Filipino language, the national language used in the Philippines. The Naïve-Bayes classification algorithm was used for classifying tweets in our previous study[8]. Using Naïve-Bayes achieved a 79.91% accuracy in identifying health-related tweets. This new research builds on top of the previous study[8] to improve on the classification algorithm and visualize possible disease outbreak incidences in the Philippines. In another study, tweets were used to determine public response and sentiment in healthcare services. A tool was created to collect and classify tweets using the Support Vector Machines (SVM) classification algorithm[9].

2.3. Use of Social Media in Determining Outbreaks

Twitter data was used in tracking different diseases in different parts of the world. In the United States, a study during the A[H1N1] outbreak used Twitter to estimate the disease incidences in real-time[10]. A preliminary study was also made in the Philippines that focuses on the use of Twitter to detect Dengue disease outbreaks in the country, where results showed that there is high correlation in the behavior of Dengue Fever with the number of tweets collected[11]. In the study made on Dengue surveillance in Brazil, four (4) parameters were used to filter social media posts, particularly tweets. These are volume, location, time, and public perception[12]. Added to this, the research proposed five (5) taxonomies [Personal experience, Sarcastic tweets, Opinions, Resource, Marketing] to group the posts according to their sentiments[12].

3. Tweet Collection and Infodemiological Classification

A tool was set up to collect tweets in real-time based on specified keywords. After the collection, a classification model was made to distinguish infodemiological tweets among the collected tweets. The SVM classification algorithm was used to do this task. Word association was done once the infodemiological tweets were selected.

3.1. Tweet Collection

A total of 218 health keywords were used in collecting the tweets in both the English (tagged "en") and Filipino (tagged "tl") language. The keywords include English and Filipino terms gathered from the Philippines' DOH's manual of procedures and related literature. Specific keywords for the diseases being studied - Dengue and Typhoid Fever - were included in the searched keywords. Symptoms such as "fever *(lagnat)*", "cough (*ubo)*", "rashes", "headache *(sakit ng ulo)*", and "stomachache *(sakit ng tiyan),*" and their conjugations were supplied as search parameters into the collector.

A daily average of **360,056 tweets** were collected from August 10, 2016 until September 10, 2016, with 1,931,561 Filipino tweets and 8,150,017 English tweets. Filipino tweets come at an average number of 66,605 tweets per day and English tweets collected come at 281,035 tweets per day.

3.2. SVM Classification

Two separate classification models were produced for English and Filipino tweets to determine whether a tweet is infodemiological or not. For the purpose of building a corpus, retweeted posts or those containing "RT" were not used. The tagged tweets were then randomized, where 75% (Filipino: 1,983 Tweets; English: 2885 Tweets) were used as training data set, and the remaining 25% (Filipino: 662 Tweets; English: 963 Tweets) were used for testing.

The SVM algorithm was implemented using the RTextTools package for R. The tagged Filipino and English tweets were used to train the SVM algorithm for classification, respectively. Additional Filipino stopwords were compiled and used for the cleaning of the tweets in the Filipino language. Table 1 shows the results with an overall accuracy of 90.09% in classifying the infodemiological and non-infodemiological tweets correctly. The model has a high positive recall of 77.42% and negative recall of 95.17%.

Table 1. SVM Classification Result

	Number of Infodemiological Tweets	Number of Non-Infodemiological Tweets	Total Predicted Tweets
Predicted Infodemiological	360 Precision: 86.53%	56	416
Predicted Non-Infodemiological	105	1104 Precision: 91.32%	1209
Total Tweets	465	1160	1625

3.3. Word Association

Word association is used to show how possible syndromes – groups of related symptoms – show up on the way Filipinos tweet online. A list of symptoms for both Dengue and Typhoid Fever were compiled to see the commonly used words associated to them when people tweet. Some words that were used can be seen in Table 2.

Table 2. Sample words used for association

Keyword List		
Fever (English)	Pain (English)	Rash (English)
Lagnat (Filipino)	Ubo (Filipino)	Sipon (Filipino)

Two networks of the associated words were produced using ORA-PRO, a network analysis software, for Filipino and English tweets, seen in Figures 1 and 2 respectively. Figure 1a shows the connection between "*suka"* (translated: "*vomit")* and *"nahihilo"* (translated: *"feeling dizzy"*), which is an indication of a sickness, particularly of Typhoid Fever. Figure 1c shows that there are also connections among three (3) symptoms, namely *"ubo"* (translated: *"cough"*), "*sipon"* (translated: *"colds"*), and "*lagnat*" (translated: *"fever"*). These are the symptoms for Dengue Fever, which illustrates a syndrome that could be tracked from Filipino tweets.

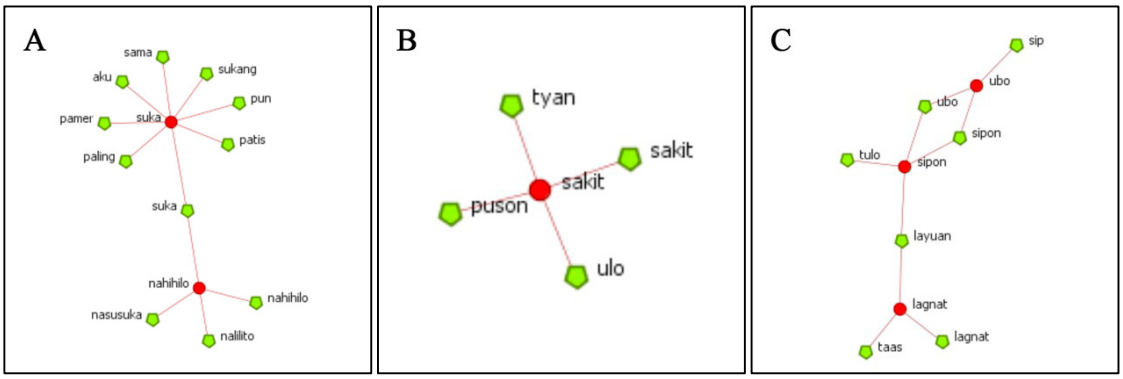

Fig. 1. (A) Network of *"suka"* and *"nahihilo"*; (B) Network of *"sakit"*; (C) Network of *"ubo"*, *"sipon"*, and *"lagnat"*

Unlike the Filipino tweets, the English tweets seen in Figure 2 do not form any significant network of symptoms based on word associations. It could be mildly inferred that Filipinos more often tweet the collection of their symptoms (syndromes) in the Filipino language, rather than in the English language. This could be attributed to the fact that the Filipino language is used more conversationally in the country, which is then also evident in the way Filipinos post in social media.

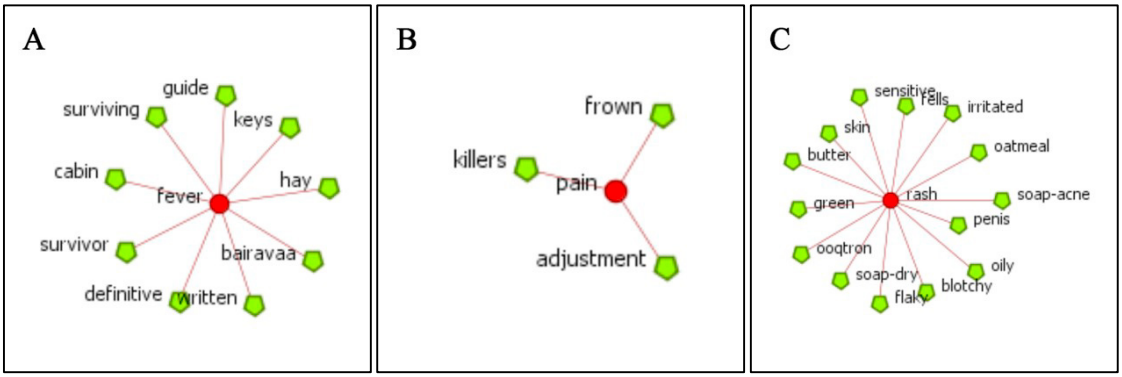

Fig. 2. (A) Network of *"fever"*; (B) Network of *"pain"*; (C) Network of *"rash"*

4. Infodemiological Visualization

Geocoding was used in order to get the lat-long data for tweets without embedded geolocations. After geocoding, two overlays were created for the visualization of the tweets, namely 1) the choropleth map and 2) location points. The choropleth map was used to display the density of tweets in certain locations, and the location point map was used to see what tweets can be found on a certain point of the map.

4.1. Tweet Geocoding

A script was made using the RDSTK package for R to programmatically geocode the tweets that were collected. Tweets that were not successfully given a lat-long data were disregarded.

Table 3. Sample Output File of Geocoding

Tweet	Tweet Username	User-Specified Location	Lon	Lat	Is_Geocoded
Sakit ng ulo koooo	BABEbibobu28	Caloocan, Philippines	120.97	14.65	TRUE
Ang saket ng ulo kooooo	Azzhmrda	Antipolo, PH	121.26	14.65	FALSE

Table 3 shows a sample output of the geocoding scripts. The table contains the columns Tweet, Tweet Username, User-Specified Location, Lon, Lat, and Is_Geocoded. The Is_Geocoded column signifies whether the tweet used the original embedded tweet location (tagged "FALSE") or the geocoded location (tagged "TRUE").

4.2. Choropleth Map

A script using the GISTools package for R was used to count the number of tweets within polygons of the Philippine shapefile. The script goes through each row of the geocoded csv files containing location details of the tweets. The script bounds each tweet within a polygon, and cumulatively counts the number of tweets that falls inside it. Figure 3 shows a sample visualization (tweets from August 10, 2016) with polygons at the municipal/city level. The choropleth map helps in seeing the concentration of the tweets within the country, where a darker shade means there are more incidents in an area. As seen in Figure 3, there are concentration of tweets in the Metro Manila Area, an urban area where people have easy access to Internet connection compared to other parts of the country. When a user clicks an

area, details such as the name of the province and municipality, and the number of tweets found inside a polygon are shown.

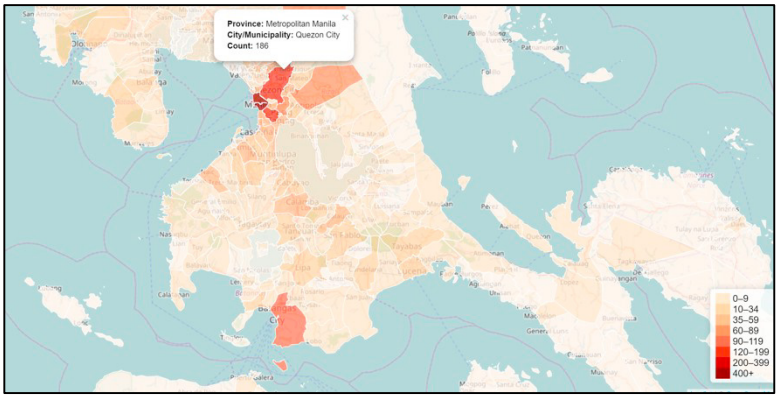

Fig. 3. Clicked Choropleth Map Showing the Tweet Details in Quezon City, Metro Manila, Philippines

4.3. Location Point Map

The second visualization created complements the choropleth map by plotting the tweets' location points. As seen in Figure 4, users may click a pin and see details about that particular tweet including the username of the person who tweeted it. Since there are thousands of tweets being put on the map, the *MarkerCluster* plugin for Leaflet.js was used to cluster nearby points, which gives the number of tweets that were clustered together.

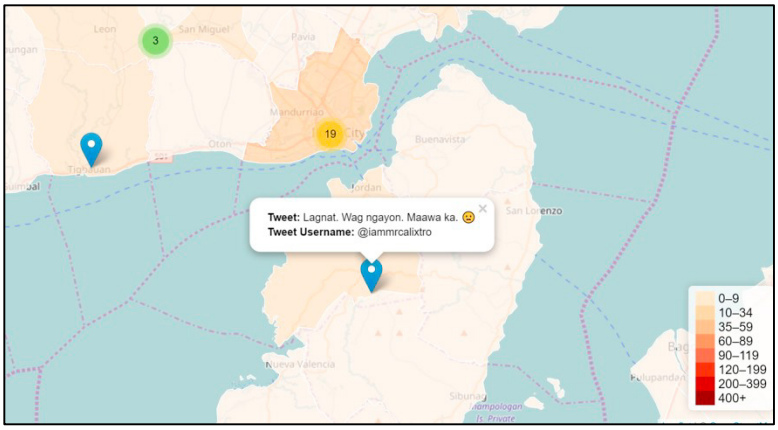

Fig. 4. Clicked Location Pin Showing the Tweet and the Username

5. Regression Modeling

The collected Twitter data were used to correlate and create the regression model to generate the predicted values of disease incidences based on tweets using the Dengue and Typhoid Fever surveillance data coming from the Philippines' (DOH). For this particular study, the data used came from the Western Visayas Regional office of the DOH.

5.1. Dengue Fever

Seen in Table 4 is the Pearson's correlation coefficient of the Dengue surveillance data and Twitter infodemiological data at the provincial and municipal/city level. Both levels present a positive and statistically significant relationship with the tweets.

Table 4. Correlation Result for Tweets and Dengue Data

	Provincial Level	Municipal/City Level
R	0.8451608	0.6772134
p-value	< 2.2e-16	<2.2e-16

The number of infodemiological tweets and the recorded Dengue incidences were used to create a predictive model using linear regression, where the surveillance data was used as the dependent variable. Using the slope and the intercept that was generated, the number of possible cases may be projected through the use of the collected infodemiological tweets. For example, at the municipal level, if there are ten (10) tweets identified in an area, the linear regression model forecasts that there are 22 possible disease incidences using the equation. Added to this, using a 95% confidence interval, the linear regression model can also give a lower bound and upper bound incidence count. Table 5 shows different test data and results using the linear regression model for Dengue Fever at the provincial level.

Table 5. Regression Result for Tweets and Dengue Data

Tweet Count	Fit Result	Lower Bound	Upper Bound
1	1.60277	-2.284252	5.489792
10	22.1407	19.10462	25.17678
50	113.4204	105.3649	121.4759

5.2. Typhoid Fever

The same methodology used in Dengue Fever was used to get the results for Typhoid Fever. Table 6 shows the result of the correlation test of the tweets with the Typhoid Fever data at the provincial and municipal level. At the provincial level, a statistically moderate positive relationship between the two data sets was seen. On the other hand, at the municipal/city level, the relationship becomes significantly weaker as indicated by the lower correlation coefficient.

Table 6. Correlation Result for Tweets and Typhoid Fever Data

	Provincial Level	Municipal/City Level
R	0.7175732	0.2211671
p-value	< 2.2e-16	<2.2e-16

Using the slope and the intercept generated through the regression modeling, Table 7 shows test cases for the prediction of Typhoid Fever disease incidences using tweets at the provincial level.

Table 7. Regression Result for Tweets and Typhoid Fever Data

Tweet Count	Fit Result	Lower Bound	Upper Bound
20	0.7789518	0.6865051	0.8713985
50	1.865208	1.632204	2.098213

100	3.675635	3.207489	4.143782

6. Conclusion

The use of social media for different agenda has grown continually over the years, and this study presents another case that uses Infodemiology for Syndromic Surveillance in the context of the Philippines. Three (3) major phases/steps were created to arrive at the final result of being able to predict possible disease incidences within certain locations in the Philippines using infodemiological data from Twitter. By streaming data, the study is able to show that infodemiological tweets could provide an additional layer in spatio-temporal syndromic surveillance. One possible benefit of which is to create timely public policies before an actual outbreak happens.

This study was able to show that using Support Vector Machines (SVM) for the classification of Infodemiological tweets was effective in identifying health-related tweets, as shown in its accuracy of 90.09% and a positive recall value of 77.42%. The classified infodemiological tweets were then geocoded and visualized on a spatio-temporal map on a web interface. Lastly, a regression equation was derived from the Twitter infodemiological data and official recorded disease surveillance data. The regression model can be used to determine possible disease incidences. As a recommendation for future studies, it is suggested to include data from other regions in the country for further validation of results, as there are differences in both disease incidences and Twitter usage in different parts of the country.

Acknowledgements

We would like to acknowledge the following for their significant contribution to this study: Philippine Council for Health Research and Development (PCHRD), Department of Health (DOH), Department of Science and Technology (DOST), the Ateneo Java Wireless Competency Center (AJWCC), Ateneo Social Computing Science Lab (ASCS Lab) and the Ateneo de Manila University.

References

1. Chikersal P, Poria S, Cambria E, et al. Modelling public sentiment in twitter: using linguistic patterns to enhance supervised learning. In International Conference on Intelligent Text Processing and Computational Linguistics, pages 49-65. Springer, 2015.
2. Nascimento T, Dos Santos M, Danciu T, et al. Real-time sharing and expression of migraine headache suffering on twitter: a cross-sectional infodemiology study. Journal of medical Internet research, 16(4), 2014.
3. Lall R, Abdelnabi J, Ngai S, et al.. Advancing the use of emergency department syndromic surveillance data, new york city, 2012-2016. Public Health Reports, 132(1 suppl):23S-30S, 2017
4. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. J Med Internet Res, 11(1):e11, 2009.
5. Nascimento T, Dos Santos M, Danciu T, et al. Real-time sharing and expression of migraine headache suffering on twitter: a cross-sectional infodemiology study. Journal of medical Internet research, 16(4):e96, 2014.
6. Henning K. What is syndromic surveillance? Morbidity and Mortality Weekly Report, pages 7–11, 2004.
7. Eysenbach G. Infodemiology and infoveillance tracking online health infor- mation and cyberbehavior for public health. Am J Prev Med, 40(5 Suppl 2):S154–8, May 2011.
8. Espina K, Estuar MRJ, et al. Towards an infodemiological algorithm for classification of filipino health tweets. Procedia Computer Science, 100:686–692, 2016.
9. Ali A, Magdy W, and Vogel S. A tool for monitoring and analyzing healthcare tweets. In HSD workshop, SIGIR 2013, 2013.
10. Signorini A, Segre AM, and Polgreen P. The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. PloS one, 6(5):e19467, 2011.
11. Coberly J, Fink C, Elbert Y, Yoon I, Velasco MJ, Tomayo A, et al. Tweeting fever: can twitter be used to monitor the incidence of dengue-like illness in the philippines? Johns Hopkins APL Tech Dig, 32(4):714– 25, 2014.
12. Gomide J, Veloso A, et al. Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In Proceedings of the 3rd International Web Science Conference, page 3. ACM, 2011.