Water related health forecast based on social media data

Mini grants data innovation 2015, UN Pulse lab Jakarta

Name of applicant	Bachti Alisjahbana, MD, PhD Chairperson, TB-HIV Research Center, Medical Faculty, Universitas Padjadjaran
Project title	Water related health forecast based on social media data
Partners	Ir. Jurjen Wagemaker, Owner, Floodtags (NL) Dr. Gertjan Geerling, Senior researcher, Radboud University (NL) and Deltares (NL) Project group: Ali hürriyetoglu, researcher, computational linguistics, Radboud University (NL) Konstantin Löser, student researcher, ISIS, Radboud University (NL) Ron Boortman BSc., Floodtags (NL) Nopi Susilawati, Medical faculty, UNPAD Alif Al Birru, Medical faculty, UNPAD

Introduction

Reliable health data is difficult to obtain in Indonesia especially aggregated data on a higher scale level such as provincial or national scales. This hampers adequate policy making and response to outbreaks by health workers on the ground. One of the acute sources adding to the total burden of disease are the water-borne and water related diseases[1]. These often have a seasonal character based upon the weather patterns and also affect whole communities at the same time.

The project at hand started from the notion that floods are already mapped reliably based on data from Twitter by Floodtags (NL). As data from public health is difficult to obtain, applying a social media analysis to acquire these data is an interesting prospect. So, the question is, can we use tweets to map health burden in communities? Furthermore, if we focus at water-related diseases, are floods and water-related health tweets correlated? The potential being that health impacts can be forecasted based on flood tweets.

The project explored the following research questions to be able to build a prototype tool:

- Can we classify and map Indonesian (West-Java) tweets related to water-borne and water-related diseases (focusing on dengue and diarrhea)?
- How do these tweets relate to field data from Puskesmas in the Bandung area?
- Can we find a relation between peaks in tweets about floods and peaks of tweets about diarrhea and dengue?
- Present a prototype health burden mapping and forecasting system based on the found relation between flood and health related tweets.

[1] Diarrhea is often caused by water-borne pathogens and so in that case a water-borne disease. Dengue is a form of water-related disease as the mosquito transferring dengue is dependent on water, not the virus itself.

Method

The method applied consists of several steps, these steps are outlined below.

1. **Collecting Tweets about Floods and Health**
   We used the Twitter API searching for keywords relating to Floods, Dengue and Diarrhea. They are banjir (for floods) and sakit, pusing, pening, nyen mastaka, sakit kepala, lieur, jangar, demam, panas dinoin, meriang, demam berdarah, tipes, tifus, tipus, mimisan, nyamuk, gigitan nyamuk, pendarahan, sembelit, keracunar makanan, disentri, muntah, muntah-, muntah, mules, sakit perut, diare, mencret, eneg, muntaber (for health). The period over which we collected data is October 2014 to July 2015. This yields a raw set of twitter data amounting to around 45 million tweets in the research period.
2. **Clustering, annotating and classifying Tweets**
   In order to get meaningful information from the tweets, they were clustered automatically based on content similarity into 100 bins. From these 100 bins sets of twenty random tweets were selected and were labelled manually into meaningful classes on floods, dengue and health. This was repeated until a satisfactory portion was manually classified to ascertain a reliable classification into tweet-classes about floods and dengue and diarrhea for various tweet sources like a person, news agent or organization. The resulting classifier was applied over the entire database.
3. **Geolocation**
   The tweets in the useful classes were screened on geolocation data, and if available only these tweets were selected for further analysis. Additionally, for tweets without geolocation data but for which the user set a home town, this home town was used as origin of location (assuming there is a good chance that ill people are at home).
4. **Analysis of flood and health relations**
   The last step involved testing against available data from puskesmas and looking for relations between peaks in flood tweets and peaks in health tweets to establish a first prototype rule for predicting health impact based on flood tweets.
5. **Building the prototype**
   The results are inputted in a straightforward rule based prototype “health risk map” available on http://floodtags.com.webhosting109.transurl.nl/dengue-and-diarrhea-forecast.

More information on the classifier we used can be found on:

- Presentation of the interface: http://relevancer.science.ru.nl
- Open source python code: https://github.com/cengelif/Relevancer
- Demo presentation: Ali Hürriyetoğlu, Elif Türkay, Antal van den Bosch, Mustafa Erkan Başar, "Relevancer: Finding and Labeling the Structure in a Tweet Set", ATILA 2015, Antwerp, Belgium. Event Page: http://www.clips.uantwerpen.be/atila15

Results

In the results section we show some of the intermediate results of the steps described in the method section. Just to give an indication of the work undertaken and to be able to show were improvements can be expected in the future.

Flood and Health tweet classification results

The figure 1 and 2 below show the results of the classification of all tweets that were gathered in the research dataset, spanning October 2014 to July 2015. More than 45 million tweets were “mined” using dengue and diarrhea related keywords in Bahasa Indonesia speaking territories. Out of 8.441.901 flood tweets, we found 424.109 “unrelated” and out of 36.852.695 health tweets, we found 31.574.557 “unrelated”. The tweets marked as unrelated were excluded from further analysis. In below graphs the results of classification of the remaining 8.017.792 flood tweets and 5.278.138 health tweets. The y-axis shows the amount of tweets, while the x-axis shows the tweet-classes.

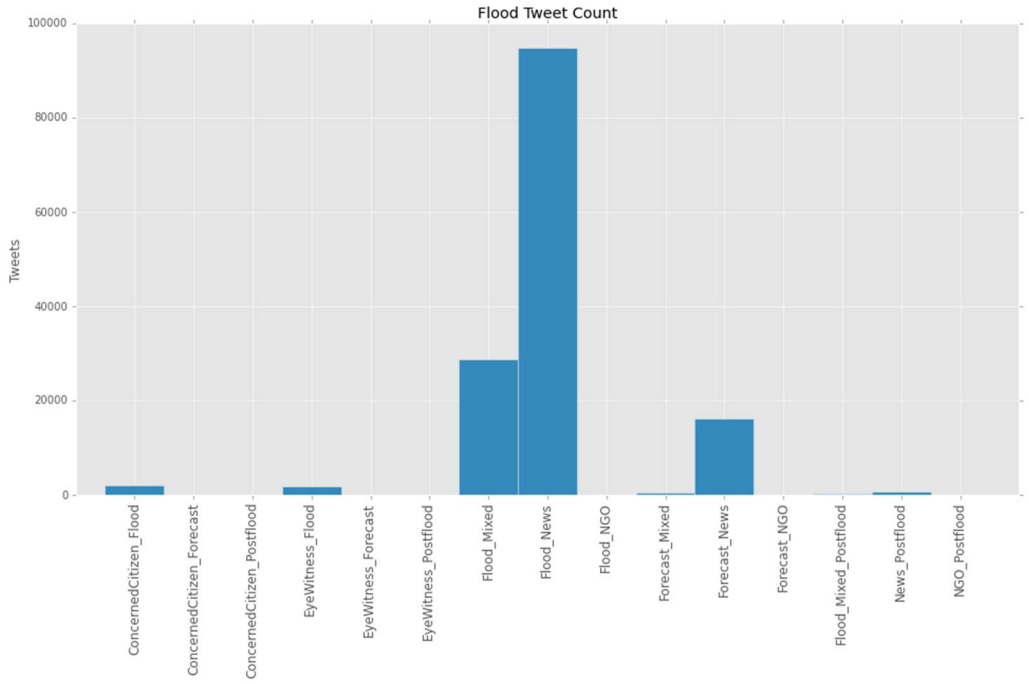

Figure 1. Number of tweets per flood class. The Flood_mixed and Flood_News classes contain the most tweets regarding floods.

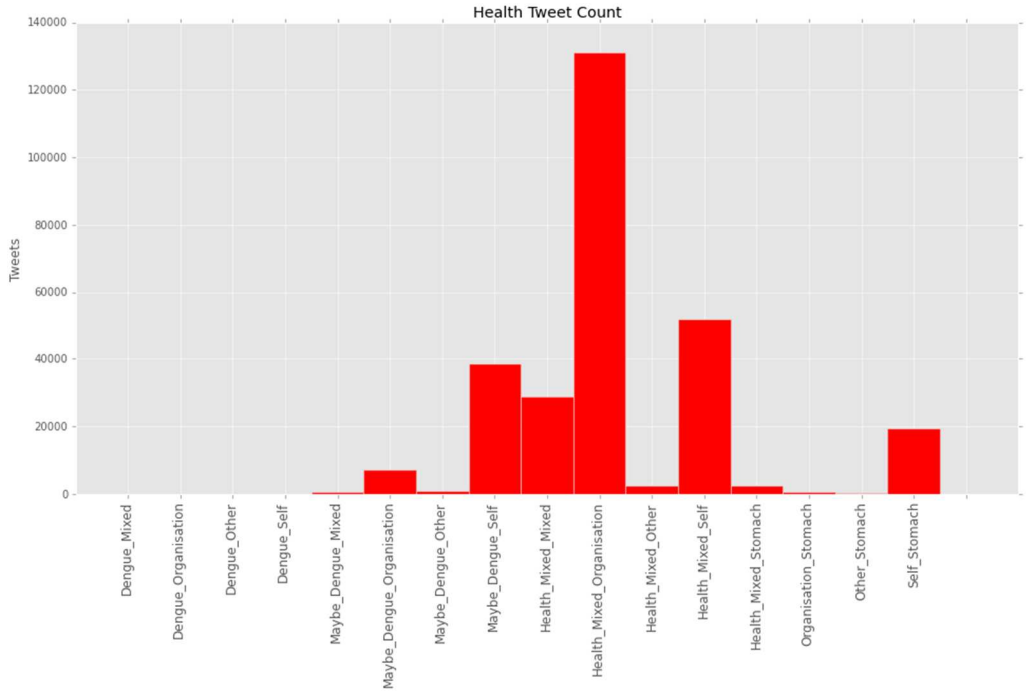

Figure 2. Number of tweets per Health class. In the final result, health classes were grouped to obtain enough spatial resolution.

Figure 1 also shows that three classes are left unused (NGO, Concerned Citizen and Postflood). Apparently, distinguishing these classes on the basis of the used features in clustering is difficult. In below table the validation results for the flood classifiers:

Table 1 and 2: Validation of the flood classifiers

Classes 1: Sample 173 clusters (total 188.000 tweets)	The classifier predicted	While annotators (manually) considered them					
		News	Eye-withness	NGO	Concerned citizen	Mixed	Un-related
News	59448	55083	0	0	0	4365	0
Eyewithness	22265	503	990	0	0	20293	479
NGO	0	0	0	0	0	0	0
Concerned citizen	3497	32	34	0	0	3338	93
Mixed	156642	3509	15384	0	0	130738	7011
Unrelated	7873	219	169	0	0	2595	4890

Classes 2: Sample 188 clusters (total 250.000 tweets)	The classifier predicted	While annotators (manually) considered them				
		Forecast	Flood	Post-flood	Mixed	Unrelated
Forecast	1983	567	630	0	756	30
Flood	64739	472	43369	0	19596	1302
Postflood	1051	67	187	0	784	13
Mixed	165481	930	114692	0	41095	8764
Unrelated	16471	964	5312	0	7831	2364

In below table the validation results for the flood classifiers:

Table 3 and 4: Validation of the health classifiers

Classes 1: Sample 173 clusters (total 700.000 tweets)	The classifier predicted	While annotators (manually) considered them			
		Self	Organi-sation	Mixed	Other
Self	152	115	0	33	4
Mixed	10	3	1	4	2
Organisation	4	1	1	2	0
Other	7	2	0	5	0

Classes 2: Sample 173 clusters (total 700.000 tweets)	The classifier predicted	While annotators (manually) considered them				
		Dengue	Maybe Dengue	Diarrhea	Mixed	Unrelated
Dengue	1	0	0	0	1	0
Maybe dengue	10	0	9	0	0	1
Diarrhea	5	0	0	4	1	0
Mixed	10	0	1	0	3	6
Unrelevant	148	0	0	0	20	128

Geolocation

We used Method 3 to find geo-locations for the tweets. Out of 8.017.792 flood tweets, we could find a location for 623.541 tweets. And from a total of 5.278.138 health tweets, we found a location mark for 279.848 of them. After this step, we finally obtained a flood tweets dataset and a health tweets dataset with the date, location and flood or health (dengue and diarrhea) classified tweets. An example of this data is shown in figure 3 for the Bekasi district on Java Island. The number of flood related tweets is much higher than the number of tweets related to dengue or diarrhea. Therefore, the y-axis of flood tweets is on the left (with highest peak up to 13.000 tweets on a flood printed in blue) and the y-axis of the health tweets is shown on the right (up to 280 tweets for a major peak shown in red). As a future analysis option we also show the rainfall as measured by satellite data.

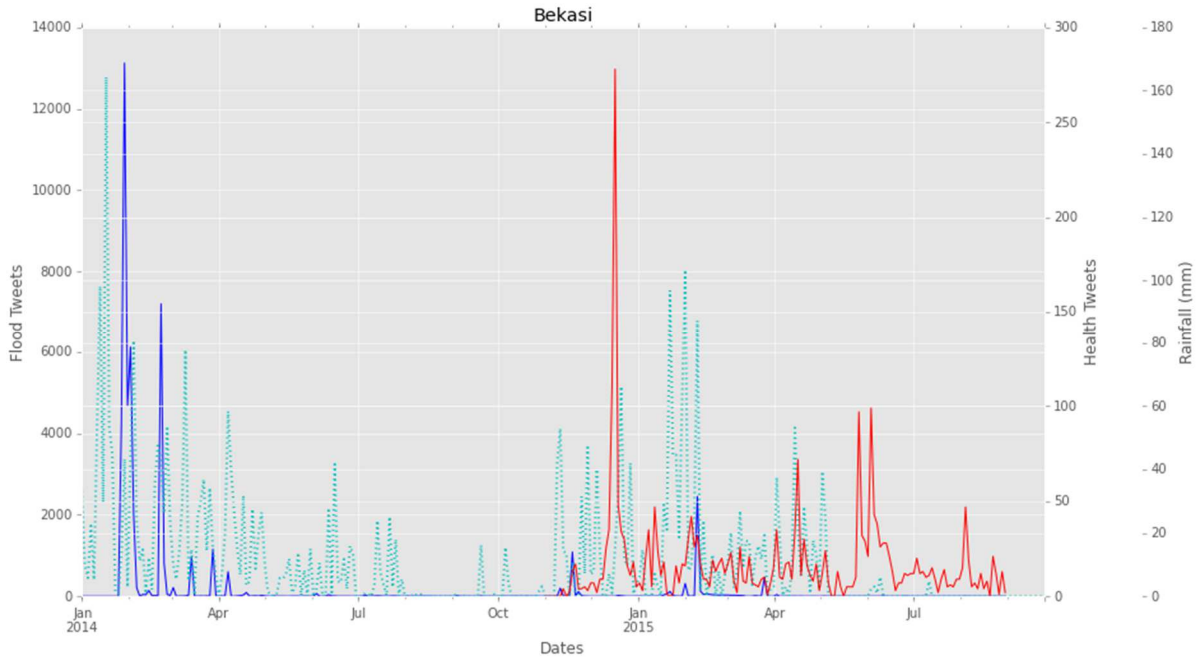

Figure 3. An example from the Bekasi district of the number of flood tweets (blue; left y-axis), number of health tweets (red; right y-axis) and satellite derived rainfall data (dotted line; second right y-axis).

Correlation analysis between floods and health related tweets

We expected a disease (dengue or diarrhea) to emerge some time after a flood. So, in our prototype analysis, we tested for the best possible match between floods and health peaks. To determine the time at which a health effect emerges after a flood, we shifted the flood tweets in different steps of 2 days and tested the "match" (i.e. cross-correlation) between flood twitter peaks and health twitter peaks at every step.

Figure 4, clearly shows and increase of the strength of the match until day 6, at day 8 the match decreases, indicating that the highest response in health tweets occurs in general 6 to 7 days after a flood.

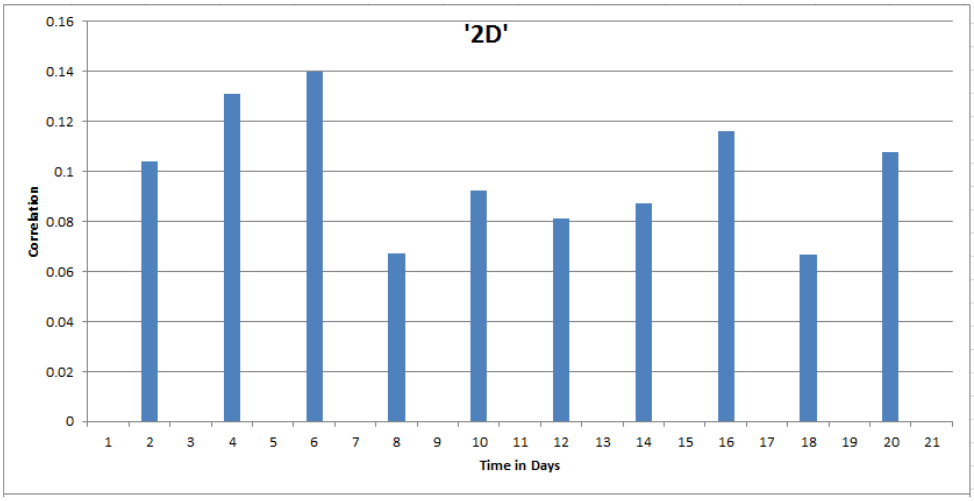

Figure 4. Strength of the "match" between flood twitter peaks and health twitter peaks for the whole dataset of the West-Java province. The strength of the match decreases after 6-7 days.

The map of the West-Java province shows the spatial variability of the strength of the match between floods and health after a 7 day shift, see figure 5. Stronger colors indicate that the correlation is higher after a shift of one week for certain locations. For example, Kota Cirebron has a correlation factor of 0.6. The variability of the correlations is interesting and depends on various factors that can be researched in the future, such as: there is a variation in the available tweets per district and this influences the ability to correlate peaks; different regions might have different relations between water and health like coastal regions versus upland regions.

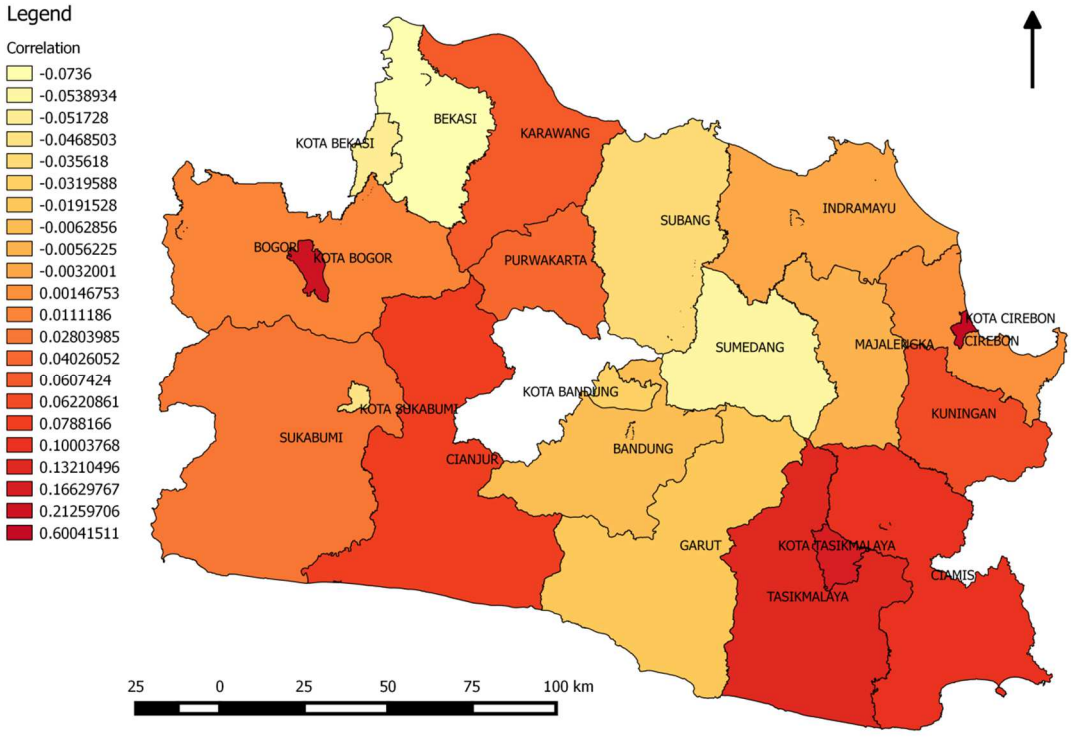

Figure 5. Strength of the "match" or cross-correlation for districts in West Java province for a 7 day time lag between flood and health incidences.

Prototype

Based on our analysis shown above, we developed a prototype live forecast map, based on the relation we found (time lag) between a flood twitter peak occurrence and a health peak twitter occurrence.

Our rules for the prototype are: A flood peak is defined as a peak when the tweet count is 2 times higher than the standard deviation of the previous week. This will give a health warning for that district for the next 7 days.

Figure 6: A screenshot of the health incidence map, taken from: http://floodtags.com.webhosting109.transurl.nl/dengue-and-diarrhea-forecast.

Conclusions and outlook

The main findings are:

- We were able to classify and map dengue and diarrhea related tweets.
- We were able to classify and map flood tweets.
- The flood and health tweet peaks for West-Java showed a relation that was strongest between 6-7 days after a flood event.
- Based on the classification and relation we were able to deliver a prototype health forecast.
- A large amount of tweets end up in the “mixed” class. Method 2 can be improved importantly, by improving the clustering techniques and annotation of classes for floods as well as for health tweets.
- For a large amount of tweets, we were not able to find geolocations. Also, we could not do a sound validation of the found geolocations. Method 3 can be improved, by improving the geocoding techniques and validating its results.

It is our intention to further research the possible relations between water and health in a 4-year PhD project both at UNPAD and the Radboud University in collaboration with Deltares (NL), the Dutch Royal Meteorological Institute and Floodtags. At the time of writing early possibilities for this are explored. The aim is to refine the prototype and to connect to potential users of this information, including the UN Pulse initiative.