DATA INNOVATION MINI GRANTS

EARLY WARNING FOR WATERBORNE DISEASES

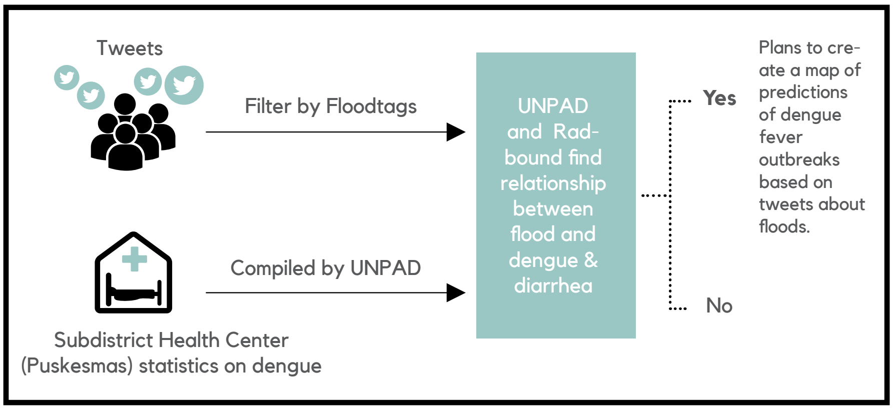
BACKGROUND

Reliable health data can be difficult to obtain, especially aggregated data on a provincial or national scale. This hampers evidence-based policy making and the rapid response to outbreaks by health workers on-the-ground. Many diseases are related to water and are seasonal in character. Some data is available both on health as well as on water resources, but these data are rarely used and seldom combined. The combination of these data can provide public health insights at a greater scale than is currently possible.

IDEA

Analysis of the data from social media and government statistics on floods can help predict waterborne disease outbreaks. The prototype aims to:

1. Explore relationships between floods and public health concerns, such as dengue fever and diarrhea:
   a. Using twitter data on floods.
   b. Using twitter data on dengue fever and diarrhea.
   c. Using puskesmas (community health center) data on dengue fever and diarrhea.
2. Monitor floods to predict dengue fever and diarrhea outbreaks
   a. Using twitter data on floods.
   b. Using established relationships between floods and dengue fever and diarrhea.


Tweets
Filter by Floodtags
Subdistrict Health Center (Puskesmas) statistics on dengue
Compiled by UNPAD
UNPAD and Radbound find relationship between flood and dengue & diarrhea
Yes
No
Plans to create a map of predictions of dengue fever outbreaks based on tweets about floods.

IMPLEMENTATION

TB-HIV UNPAD, Radboud University and Floodtags worked with three datasets:

- **TWEETS ABOUT FLOODS:** Using the twitter API to collect tweets that contain the word “banjir”. Volume of database: 7.9 million tweets.
- **TWEETS ABOUT DENGUE & DIARRHEA:** Using the twitter API to collect tweets, that contain keywords that indicate various diseases. Volume of database: 35 million tweets.
- **DATA FROM THE PUSKESMAS:** Data from Bandung District.

Using above-mentioned datasets, the partners undertook the following steps:

1. Filtering datasets 1 and 2

STEP 1: Create clusters for a subset of the database, on the basis of a large number of features of the tweets.

STEP 2: Determine what we want to know from the tweets using classes

STEP 3: Annotation of the central clusters.

STEP 4: The annotated clusters are separated from the subset of the database.

STEP 5: With the remaining tweets in the subset of the database, repeat clustering and annotation three times.

STEP 6: On the basis of all the classified clusters, create an algorithm that can classify new incoming tweets into classes.

STEP 7: The classifier is applied over the entire database and unrelated tweets are filtered out. It is now possible to see over the entire database, which classes are represented and to what extent.

2. Geocoding of the tweets by looking for location information in the body text and in the user profile. The location data is matched with the Open Street Map database to obtain coordinates.

3. Analysis and comparison of the three datasets. The datasets of tweets are now available as time series with class and geolocation in CSV (and other formats). These must now be compared to each other and to the dataset of the Subdistrict Health Center(Puskesmas).

NEXT STEPS

This type of data gathering using a specific type of classification will provide convenient access to predictions on disease outbreaks through social media. By understanding the correlations, local government will be able to react faster in the case of disease outbreaks.

Australian Aid
INNOVATION FOR DEVELOPMENT
UNDP Empowered lives. Resilient nations.
PULSE LAB JAKARTA