Characterizing dengue spread and severity using Internet media sources

Talal Ahmad[1], Nabeel Abdur Rehman[1], Fahad Pervaiz[1], Shankar Kalyanaraman[2], Maaz Bin Safeer[1], Sunandan Chakraborty[2], Umar Saif[1], Lakshminarayanan Subramanian[2]

[1]Lahore University of Management Sciences, [2]New York University

ABSTRACT

Pakistan witnessed one of its deadliest dengue outbreaks in 2011 resulting in hundreds of deaths throughout the country. Prior to the outbreak, dengue awareness was relatively low and hospitals in the country were not completely prepared to handle the epidemic with limited knowledge about the spread of the disease in each locality. This poster aims to build a system that automatically characterizes the spread and severity of the dengue disease at a fine-grained location granularity based on analyzing news reports from Internet media sources. Based on a detailed analysis of news reports gathered from several leading dailies in Pakistan, we demonstrate the effectiveness of our system to accurately characterize the dengue spread and severity across different locations within Pakistan.

1. INTRODUCTION

Preventing large-scale outbreaks of diseases like dengue, malaria, typhoid constitute an enormous public health challenge, especially in countries with limited infrastructure committed for prevention, spreading awareness and containment of these diseases. In the case of conventional government-managed public health surveillance systems, data collection is handled in a hierarchical manner. This chain-of-command framework has its advantages and disadvantages. While a hierarchical process reduces the occurrence of unnecessary alerts (false positives) that may cause panic, it introduces some bureaucracy in how agencies need to coordinate with each other up and down the order, thereby making it inefficient.

An alternative approach is to use a semi-automated solution, leveraging the ever-expanding ubiquity of the Internet (especially in developed countries). These have some improvements over the manually operated system described above, but still rely on human intervention in the form of experts supervising data collection, analysis and reporting.

Some recent work has focused on taking this further to build completely automated web-based disease surveillance systems [1, 2]. These systems parse news articles from around the world using sources such as Google News and RSS feeds, as well as social media such as Twitter, to filter and classify articles based on the nature of epidemic, location and news sources.

DEV '13, January 11-12, 2013 Bangalore India

At the same time, not much work has been done in the space of automating analysis of longer term trends. This is very crucial from a policy planner's perspective as it gives them a better insight and serves as a backfill for inadequate/incomplete surveillance reports. Furthermore, since the study of disease trends can be time-consuming involving long man-hours at retrieving and analyzing historical records, it makes a fully-automated disease analytics system that can present longer trends even more relevant.

In this poster, we present a system that characterizes severity of a disease automatically. Specifically, we focus on seasonal dengue outbreaks in Pakistan and describe a web-based system, which automatically classifies and characterizes severity of a disease. The main highlights of this system are:

Spatiotemporal model: We capture historical trends for disease spread and severity for the geographical region. While some other systems, notably BioCaster [1] and HealthMap [2], provide maps with an overlay denoting hotspots, and general macroscopic trends, ours is the first system that incorporates time as another parameter and provides finer-grained location-specific trends.

Local media sources: By channeling our insights solely through the lens of local news sources based in Pakistan, we are able to customize and exploit the domain-and location-specific nature of the disease outbreak.

Severity characterization: We propose a severity scale that labels reports from not severe to severe for a given region and time period. To the best of our knowledge, this has not been studied before in the disease surveillance literature.

2. SYSTEM OVERVIEW

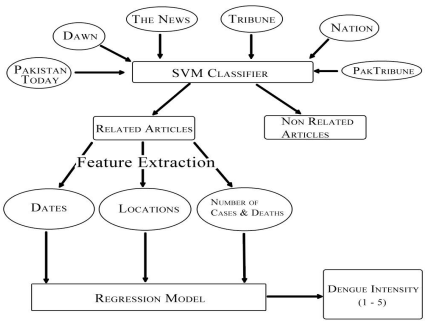

Figure 1: System Architecture

Figure 1 shows the various steps in our system architecture. The architecture consists of four key steps. In the first step, a combination of crawling and search query based extraction techniques to gather a corpus of dengue related articles from different newspapers in Pakistan. In the second step, we use a Support Vector Machine (SVM) based classifier to identify and filter dengue related articles from the larger pool of extracted documents. In the third step, we extract a combination of several dengue related features from these newspaper articles and also tag each article with its corresponding location and the time of creation. Finally, we consider the list of documents corresponding to a locality within a given time period to compute a dengue severity index for the specified location. For this, we use a trained polynomial regression model on our set of extracted features from each article and simplify the score to a dengue severity score in the range 1-5.

3. EVALUATION

We present key observations and results from our evaluation of the proposed dengue severity analysis system. We chose to conduct our evaluation around the peak dengue epidemic outbreak in 2011 in Pakistan to characterize how coverage of the epidemic varied in newspapers as the epidemic peaked and receded.

Publisher	Article count
Dawn	1552
Nation	592
Pakistan Today	1014
PakTribune	104
TheNews	434
Tribune	589

Table 1: Source wise breakdown of articles

Table 1 gives a source-wise breakdown of our document set. In all, we retrieved more than 4,200 articles for the period of August 2010 to April 2012 at an average of about 210 articles per month. Note that these are only articles mentioning the word "dengue." This indicates a fairly high level of coverage and attention devoted to the epidemic as it spread through the country.

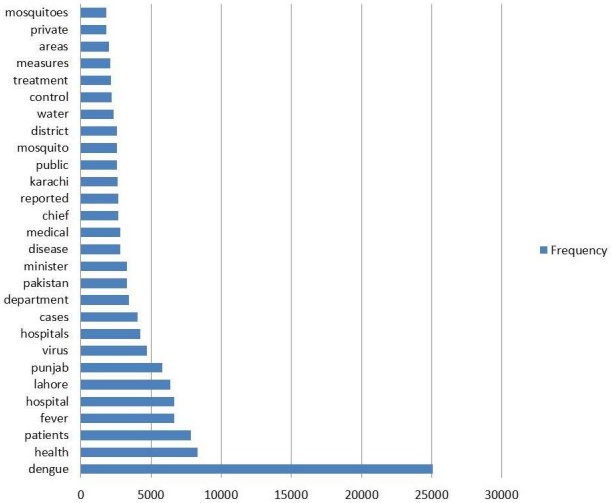

Figure 2: Most frequently occurring words in dataset

In contrast with content published on Twitter, there is more structure and order in news articles, and this renders it easily amenable for text processing. We exploit this and consider simple unigrams as potential features in our system and sort them in descending order of TF-IDF scores. Figure 2 captures this, and speaks to the high correlation and co-occurrence of words from the public health terminology.

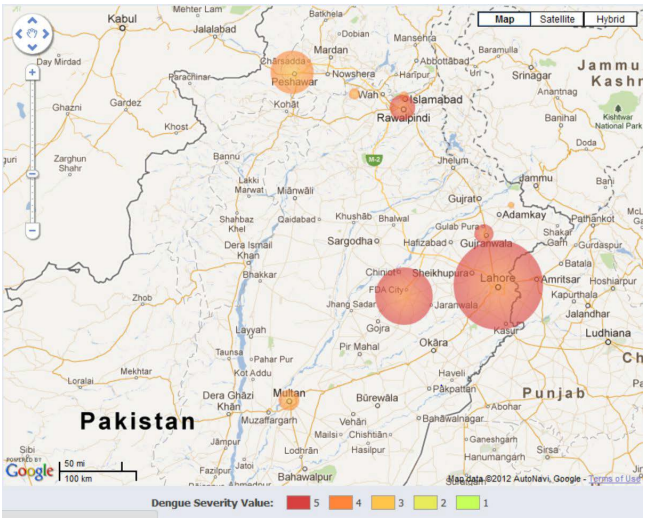

Figure 3: Snapshot 11/11/2011 to 11/18/2011

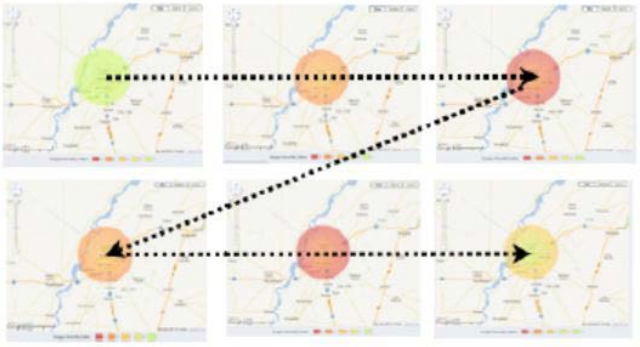

Figure 4: Tracking Dengue severity in Multan 8/2011 to 10/2011

Our spatiotemporal model is visualized in Figures 3 and 4. Of particular note is the fact that our model was able to precisely capture the peak of the dengue outbreak during the month of September 2011. Since our model can extract locations and generate scores for each location, we are able to do a deep-dive and analyze regions at the town level, as shown in Figure 11 which captures the dengue outbreak severity trend from August 11, 2011 to November 23, 2011.

4. REFERENCES

[1] Collier N, Doan S, Kawazoe A, Goodwin RM, Conway M, and et al., *Biocaster: Detecting public health rumors with a web-based text mining system*, Bioinformatics 24 (2008), 2940U″ 2941.

[2] Freifeld CC, Mandl KD, Reis BY, and Brownstein JS, *Healthmap: global infectious disease monitoring through automated classification and visualization of internet media reports*, J Am Med Inform Assoc 15 (2008), no. 2, 150–157