Am. J. Trop. Med. Hyg., 98(1), 2018, pp. 181–191
doi:10.4269/ajtmh.17-0253

Dengue Fever Surveillance in India Using Text Mining in Public Media

Andrea Villanes,[1]* Emily Griffiths,[2] Michael Rappa,[1] and Christopher G. Healey[1]
[1]*Department of Computer Science, North Carolina State University, Raleigh, North Carolina;* [2]*Public Health England, Sheffield, United Kingdom*

Abstract. Despite the improvement in health conditions across the world, communicable diseases remain among the leading mortality causes in many countries. Combating communicable diseases depends on surveillance, preventive measures, outbreak investigation, and the establishment of control mechanisms. Delays in obtaining country-level data of confirmed communicable disease cases, such as dengue fever, are prompting new efforts for short- to medium-term data. News articles highlight dengue infections, and they can reveal how public health messages, expert findings, and uncertainties are communicated to the public. In this article, we analyze dengue news articles in Asian countries, with a focus in India, for each month in 2014. We investigate how the reports cluster together, and uncover how dengue cases, public health messages, and research findings are communicated in the press. Our main contributions are to 1) uncover underlying topics from news articles that discuss dengue in Asian countries in 2014; 2) construct topic evolution graphs through the year; and 3) analyze the life cycle of dengue news articles in India, then relate them to rainfall, monthly reported dengue cases, and the Breteau Index. We show that the five main topics discussed in the newspapers in Asia in 2014 correspond to 1) prevention; 2) reported dengue cases; 3) politics; 4) prevention relative to other diseases; and 5) emergency plans. We identify that rainfall has 0.92 correlation with the reported dengue cases extracted from news articles. Based on our findings, we conclude that the proposed method facilitates the effective discovery of evolutionary dengue themes and patterns.

INTRODUCTION

Communicable diseases remain among the leading mortality causes in many countries, particularly in Asia and Africa.[1] In 2010, of the 52.8 million deaths in the world, 24.9% were due to communicable, maternal, neonatal, and nutritional causes. Moreover, 76% of premature mortality in sub-Saharan Africa in 2010 were due to the same causes.[1]

Combating communicable diseases depends on surveillance, preventive measures, outbreak investigation, and the establishment of control mechanisms.[2] *Public health surveillance* is the process of monitoring trends through data collection, collation, analysis, and dissemination of public health information for evaluation and public health response, to reduce morbidity and mortality.[3–5] *Prevention* is the long-term approach to reduce risk factors of a disease burden,[6] *outbreak investigation* establishes the existence of an outbreak and identifies the source, and *control mechanisms* are meant to cease the spread of a disease to stop its transmission.[7]

Unfortunately, data from surveillance systems are often delayed and reporting is inaccurate, making it difficult to use such data for the detection of outbreaks.[8–12] Moreover, it is estimated that only 35% of communicable disease cases are reported to national health departments.[13–15] Underreporting of these diseases negatively impacts the public health policy makers' abilities to decrease morbidity and mortality.[13,15,16] A recent study by Shepard et al.[17] reported an underestimate of 282 times the number of official reported dengue cases in India for one district under study. Furthermore, national health agencies across the globe publish reports that vary in their timeliness: some agencies report data from the previous week, and some have delays that can range as long as multiple years.[18] In developing countries, existing networks of surveillance systems are not comprehensive for all regions, and relevant communication between countries is often lacking.[9] Moreover, only official reported information is used in disease control programs, determining patterns of disease, and conducting epidemiologic investigations. Given these issues, the underreporting of communicable diseases has a direct impact on the public's health.[13,15]

The main objective of this study is to investigate the creation of a surveillance tool for dengue fever by applying text mining cluster analysis on news articles that discuss dengue. News articles normally contain local and recent information and could be used to overcome delays in official health reports, if actionable information can be extracted from their text. More specifically, text mining cluster analysis enables analysts to 1) differentiate topics being discussed in news sources; and 2) uncover the evolution of dengue topics to aid in monitoring dengue trends, assisting experts to reduce its morbidity and mortality. For this work, we collaborated closely with dengue experts that provided domain knowledge and guidance on data collection, methods, results of our analysis, and interpretation of our results.

In summary, we identified five main topics from Asian newspapers discussing dengue: 1) prevention; 2) reported dengue cases; 3) politics; 4) prevention regarding other diseases; and 5) emergency plans. In addition, we created topic evolution graphs for the topics extracted from news articles in Asia and in India. These evolution graphs help us to identify topic peaks. Finally, when we incorporated and compared the main dengue indicators in our analysis, we found that the "reported dengue cases" topic extracted from news articles matched peaks in dengue cases and associated dengue measures, suggesting that topic clustering may be a good prediction of dengue onset.

Our analysis shows that text mining cluster analysis in news articles can successfully detect dengue trends occurring in a specific geographic region. Because of the lack of current data for dengue cases, and a need for surveillance systems, our methodology may be applicable for detecting trends and taking preventative actions. We also discuss the issue attention cycle defined by Downs in 1972,[19] which can be a limitation to our analysis on news articles. We recommend

* Address correspondence to Andrea Villanes, Institute for Advanced Analytics, North Carolina State University, 901 Main Campus Drive, Suite 230 Raleigh, NC 27606. E-mail: avillan@ncsu.edu

using text mining cluster analysis as a tool for monitoring trends discussing dengue cases, and detecting peaks in reported cases that can affect entire communities.

MATERIALS

In this section, we present information about dengue fever, and relevant public health surveillance systems.

Dengue fever. Dengue is a mosquito-borne viral disease transmitted to humans through infected *Aedes* mosquitoes, a tropical and subtropical species that can be found throughout the world. The principal symptom of dengue is high-grade fever, and can present with any of the following symptoms: facial flushing, skin erythema, body ache, myalgia, arthralgia, and severe headache.[20] Dengue spread rapidly during the twentieth century to infect more than 300 million people in 2010.[21] One in three people live among mosquitoes that transmit the dengue virus, yet there remain major uncertainties over the burden of dengue.[22–26] New, improved methods for assessing this burden are in critical demand.[27]

Public health surveillance systems. Urbanization, population movement, increased global travel, commerce in food and medicinal biologic products, and social and environmental changes are some of the main drivers of the need for global surveillance of communicable diseases. Moreover, in developing countries, it is important to detect communicable disease outbreaks early to reduce the number of deaths, the spread, and the resulting harm. Strong surveillance tools are a necessity for both industrialized and developing countries.[28]

According to Thacker,[11] "public health surveillance is the systematic, ongoing collection, management, analysis, and interpretation of data followed by the dissemination of these data to public health programs to stimulate public health action." Data collected from public health surveillance can be used to detect epidemics; identify health problems in a region; estimate the magnitude of a health problem, including geographic information about the events; uncover changes in health practices; monitor an agent's changes; evaluate control measures; and stimulate research. Public health surveillance is the cornerstone for decision-making, allowing decisions to be made more effectively and in a timely manner.

Several efforts have been conducted by the research community to reduce the gaps in information between surveillance systems sponsored by health ministries, public health institutions, nongovernmental organizations, and multinational agencies. Some of the surveillance systems that mine media sources to detect infectious disease outbreaks include HealthMap,[9,29] BioCaster,[30] The Global Public Health Intelligence Network (GPHIN),[31] MedISy,[32] and EpiSPIDER.[33]

HealthMap (www.healthmap.org) uses Web-based data sources to perform outbreak detection, creates a real-time surveillance system, and updates news on new and ongoing disease outbreaks.[9] The data collected by HealthMap includes several online electronic media sources, for example, news sources from aggregators such as Google News, reports from the World Health Organization (WHO), and the Program for Monitoring Emerging Diseases (ProMED)-mail, which is a globally moderated mailing list that relies on reports sent by volunteers that include first-hand reports, news stories, and additional data related to previously posted reports. HealthMap uses text mining approaches to automatically classify the sources by location and disease, then visualizes the results on a geographic map.[29]

BioCaster is a nongovernmental surveillance system that uses ontology-based text mining to detect and track disease outbreaks. The system has four main steps: 1) topic classification; 2) named entity recognition; 3) disease/location detection; and 4) event recognition.[30]

GPHIN is a subscription-based Internet system that incorporates news information for global health surveillance. GPHIN relies on two news aggregators: Factiva and Al Bawaba. GPHIN scans, filters, and categorizes information using a taxonomy of keywords and Boolean syntax. The results are then validated by human analysts. In 2005, GPHIN supplied approximately 40% of the WHO's early outbreak warning notifications.[31]

MedISys is an automated early-warning system for food and food-borne hazards. This system monitors daily news articles from more than 2,200 news sites in 50 languages. The articles are automatically categorized into predefined multilingual categories if they satisfy category definitions based on Boolean and proximity operators.[32]

EpiSPIDER is a Web-based visualization surveillance system for infectious disease threats. EpiSPIDER uses ProMED, the WHO, European Surveillance Network RSS feeds, and news syndication sites such as Reuters as the sources for their reports.[34] EpiSPIDER extracts the location for each source, then generates country-level maps for all countries.[33]

A system that analyzes news sources for disease severity trends was created in Pakistan. The system characterizes the severity of dengue outbreaks in Pakistan by using news from six local sources to form input for a Support Vector Machine–based classifier that identifies dengue-related articles. These articles are then used to extract the following features: date, location, number of cases, and number of deaths. A severity index is calculated for each location over a period of time based on a polynomial regression model.[35]

Our tool proposes using text mining cluster analysis to: 1) infer topics being discussed in newspaper articles related to communicable diseases 2) understand the evolution of the topics; 3) detect disease outbreaks, and; 4) understand the information being communicated in news articles related to communicable diseases. The main difference with the other tools described herein, is the use of text mining cluster analysis to extract topics from news articles as our main technique of analysis. The tools described above have used text mining, but as a way to extract and classify features like location and disease based on text. Existing tools are interested in finding a predefined set of topics using keyword matching to perform topic assignment. Critically, and unlike our system, these tools cannot adapt to new topics with manually updating the predefined topic list. Our tool dynamically extracts the appropriate set of topics found in a text corpus, without the need for human intervention.

METHODS

Text mining. Text mining is the process of extracting nontrivial and previously unknown information from large text document collections, converting unstructured text data to a structured matrix form.[36] This is accomplished by converting documents into vectors in some feature space, for example, converting the text in document D into a term vector D_j.

Each entry in D_j corresponds to a specific term t_i, and its value defines the frequency of $t_i \in D_j$. To identify the terms, several representations can be used. The bag-of-words approach is the most common, where each selected word forms a term (or dimension) in the feature space.

Given the size of text documents, feature selection is an important step in text clustering because of high dimensionality and data sparsity. A data collection contains many terms, but only a small number of these normally occur in any individual document. Several sophisticated local and global methods exist for reducing document dimensionality. Local methods remove unimportant or noninformative words, whereas global methods apply a global dimension reduction to transform all documents identically. Popular local methods include: stemming, which reduces words to their stem; stop word removal, which removes noninformative words; and synonym lists, which identify and reduce synonyms to a common word. Global methods include latent semantic analysis, latent Dirichlet allocation (LDA), and nonnegative matrix factorization that characterize documents in terms of *concepts* and sets of terms that represent a more complex idea discussed in a document.

Several techniques are available to identify information in text, such as classification, clustering, and summarization.[37] Our approach begins with clustering, an unsupervised learning technique in which patterns (observations, data items, or feature vectors) are assigned into homogeneous groups called clusters. A clustering task involves the following components: 1) problem representation, including feature extraction and/or feature selection; 2) calculation of similarity between observations; 3) application of a clustering algorithm; 4) clusters labeling; and 5) evaluation.[38,39]

Text mining cluster analysis, which is the combination of text mining and cluster analysis, groups together documents with similar topics or topics with similar meaning.[40] The objective of text mining cluster analysis is to classify documents into groups, or clusters, containing documents that are similar to each other. This allows us to identify the main topics being discussed in the document collection.

Data collection. To collect dengue fever news articles discussing the Asian region, we searched the LexisNexis Academic database, an online academic database that accesses more than 15,000 news, business, and legal sources.

Multiple search criteria (Table 1) were used to create a search query. The words dengue, DEN-1, DEN-2, DEN-3, DEN-4, and break bone were submitted as keywords to locate relevant news articles. The timeline we queried was January 1, 2014 to December 31, 2014 for the continent of Asia. The year 2014 was chosen over more recent years because of the fact that the most recent dengue indicators are only available for 2014, not for later years, and even then only for limited regions.

Each of the search queries (one query per month) produced a resulting HTML file. We used Beautiful Soup, a Python package for HTML parsing, to extract and process the raw HTML. This allowed us to transform the file into a dataset where each row corresponds to a news article for a specific month. For each news article, we extracted the article's full text, the title, the publication date, and the length of the article.

Text parsing. We next created a structured representation of information stored in the text documents, known as a term-document matrix (TDM). To create the TDM, the following preprocessing steps were applied: 1) stemming—find the stem or root form of a term, aggregating different terms with the same root as equivalent (e.g., run, runs, running and ran would all stem to the root run); 2) stop word removal—words that are common in the text but do not contribute to any useful semantic context are removed using a stop list (e.g., a, an, the, which). The words specified in the stop list are excluded from parsing.

TABLE 1
LexisNexis academic search criteria

LexisNexis academic search criteria	Values
Search terms	dengue OR DEN-1 OR DEN-2 OR DEN-3 OR DEN-4 OR break bone
Date	January 1, 2014 through December 31, 2014
Source	*Blank* (any source)
Content type	Newspapers
Language	English
Geographic location	Asia

To transform the document collection into a set of terms relevant to all documents, our stop word list was extended to include all the city names in Asia, "dengue", "dengue fever", and "break bone".

The TDM quantitatively represents information contained in a set of documents, by enumerating the frequency of the terms contained in each document.[39] This converts document D into a term-vector D_j. Each entry $D_{i,j} \in D_j$ corresponds to the number of occurrences (the frequency) of a specific term $t_i \in D_j$. The result of this step is an $m \times n$ TDM corresponding to m unique terms across n documents. Text parsing not only produces a TDM, but also reduces the total number of terms, improving efficiency, and better capturing the content of a document by aggregating terms that are semantically similar.

Term vector weighting. We next construct a weighted TDM by applying term frequency–inverse document frequency (TF-IDF). TF-IDF is an algorithm that gives greater weight to terms that occur more frequently within a document (TF), but infrequently across the document collection (IDF). Intuitively, TF-IDF implies that if a term t_i occurs frequently in a document D_j, it is an important term for characterizing D_j. Moreover, if t_i does not occur in many other documents, it is an important term for distinguishing D_j from other documents. Given an n document collection,

$$\text{TF-IDF}_{i,j} = f_{i,j} * \log(n - n_i) \quad (1)$$

where n_i is the number of documents containing term t_j. This formula varies with the importance of terms based on how frequently the terms occur in individual documents, and how the terms are distributed throughout the document collection. As a final step, each D_j is normalized to remove the influence of document length from the TF-IDF weights because longer documents would have higher TF-IDF scores versus short documents.

The weighted TDM becomes the underlying representation for the collection of documents. Once documents have been converted into a weighted TDM, vectors can be compared

with estimate the similarity between pairs or sets of documents; determine the optimal number of topic clusters and; perform topic clustering.

Calculating the pairwise cosine similarity matrix. A matrix with pairwise document cosine similarities was calculated using the weighted TDM. Here, we use the cosine pairwise similarity measure $s_{u,v}$ to quantify the similarity between a pair of text documents.[38,39] Mathematically, given two documents u and v, the cosine similarity is calculated as follows:

$$\cos\theta = \frac{D_u - D_v}{[D_u] \cdot [D_v]} \quad (2)$$

Because the document vectors are normalized, this reduces Equation (2) to $\cos\theta = D_u - D_v$. Dot product similarity represents the cosine of the angle between two document vectors. As the angle between the document vectors nears zero, the more similar the documents are assumed to be. Given the formula, the cosine of the angle between two identical documents is one, whereas the cosine of the angle between two completely dissimilar documents is zero. The corresponding dissimilarity (distance) measure $d_{u,v}$ is given by $1 - \cos\theta = 1 - \cos(D_u - D_v)$. Dissimilarity $d_{u,v}$ ranges from zero (identical) to one (completely dissimilar). This converts the cosine similarity matrix to a cosine dissimilarity matrix.

Determining optimal number of topic clusters. For each month, multiple k-means clustering with increasing k was performed on the month's cosine dissimilarity matrix, to determine the optimal number of clusters k.

The elbow method, which is used to calculate an optimal number of clusters, was applied. This visual technique consists of running k-means for a range of values of k, and for each k, calculating the within-cluster sum of squares (WCSS) variation.[41] The chosen k value for each month represents a point on the WCSS line where the reduction in total WCSS slows:

Input—a TF-IDF weighted TDM W

Output—the optimal number of clusters k

1. Compute pairwise distance (metric = cosine)
2. For i = 1 to 10 do
3. Compute k-means
4. Calculate centroids
5. Compute Euclidean distance between the centroids
6. Calculate WCSS
7. Plot WCSS versus k
8. End for
9. Locate location of a bend (knee) in the plot.
10. Return k

LDA. After determining the optimal number of topic clusters, we chose to use LDA to determine the topics in the document collection. LDA is widely adopted to infer topics from text collections, and is best at learning topics from unstructured text.[42] LDA is an unsupervised topic modeling algorithm, designed to uncover topics (sets of related words) from documents. LDA implements a three-level hierarchical Bayesian model, with each item (document) of a collection represented as a finite mixture over a latent set of topics. LDA assigns a document to a mixture of topics, to characterize the document as a set of associated topic-membership probabilities. This set of topic probabilities is considered an explicit representation of the document.[43]

Cluster identification labeling. After obtaining the topics using LDA for each month, a set of describing terms was used to label each topic. This process allows us to describe a topic, and name our final clusters. With the help of a dengue knowledge expert, a label was manually chosen for every cluster for each of the 12 months in 2014 by using the descriptive terms derived by the LDA algorithm. After calculating the optimal number of topics for each month in 2014 using the elbow method described above, a total of eight different topics were identified in the news articles throughout all of 2014.

Validation. To validate our LDA clustering technique, we performed training and validation in a set of observations. First, we divided all 3,844 observations into two sets: one for training and one for validation. We trained our model on 3,748 observations and validated the result using the remaining 96 observations. For the validation set, we manually labeled each observation with one of the eight topics that were previously identified. We then applied our LDA model to predict topics for the same 97 validation observations. This allowed us to compare the performance of the model versus manual human topic selection.

As a further comparison of LDA versus other topic clustering approaches, we performed clustering of all 3,844 observations into eight topics using k-means, to assess whether LDA provides better classification performance.

Table 2 compares the results from LDA to k-means to assess the performance of LDA versus a simpler clustering algorithm. We wanted to ensure that the added complexity of LDA led to significantly better classification results. Based on the 96 observations in the validation set, LDA outperforms k-means in correctly classifying manually labeled observations. The misclassification rate is 30.2% (29 incorrectly classified documents) for LDA (Clopper-Pearson 95% confidence interval [CI]: [21.25%, 40.43%]), whereas for k-means it is 44.79% (43 incorrectly classified documents) (Clopper–Pearson 95% CI: [34.63%, 55.29%]). Based on this, we chose LDA as our standard classification method. We next discuss the LDA classification results.

RESULTS AND DISCUSSION

Data collection. After our Lexis Nexis search (Table 1), results were processed to remove duplicates and to infer a country for each observation. A total of 3,844 unique news articles were found. Figure 1 contains the number of dengue articles about Asia collected for each month in 2014.

The distribution of dengue news articles by country can be found in Figure 2. Non-Asian countries listed in this graph refer

TABLE 2
Validation comparison LDA and k-means

Labeled topic	Correctly classified LDA	Incorrectly classified LDA	Correctly classified k-means	Incorrectly classified k-means
0	10	2	7	5
1	9	3	8	4
2	9	3	3	9
3	10	2	6	6
4	8	4	8	4
5	5	7	10	2
6	6	6	6	6
7	10	2	5	7

LDA = latent Dirichlet allocation. Total observations in the validation set = 96.

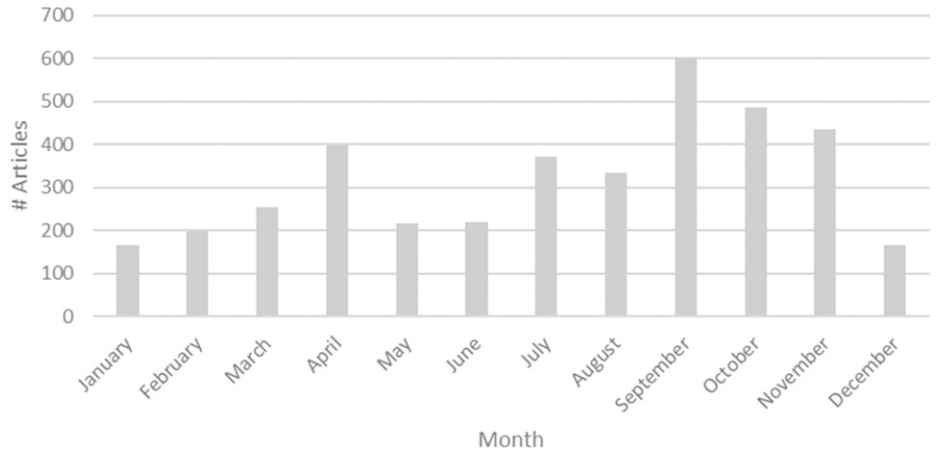

FIGURE 1. Number of articles collected per month.

to news articles originating in non-Asian countries but where a dengue topic from the continent of Asia was discussed. Figure 3 shows the number of individual sources (indexed by LexisNexis) by country. This can be used to explain why some dengue endemic regions show a low number on the number of articles collected. Furthermore, we can see that India has the highest number of sources indexed by LexisNexis, which allows us to focus in India independently.

Figure 4 visualizes the regions and countries where articles originated. In this figure, the darker the color, the more articles collected from that country.

Results. Table 3 details the number of articles discussed for each topic in 2014. The topics were extracted from news articles by running LDA for each of the 12 separate months. The *k* number of topics for each month was chosen using the elbow method. Over all 12 months, the five main topics in articles mentioning dengue in Asia were prevention, reported cases, politics, prevention regarding other diseases, and emergency plan.

Figure 5 visualizes the monthly topic trends found for Asia in 2014. We can see two main peaks—one in April and one in September—in the number of dengue news articles. Based on our topic clustering results, the peak in April is explained because of the increase in news articles discussing prevention (63%), whereas the peak in September is explained because of the increase in news articles that discuss both prevention (49%) and reported dengue cases (39%).

We can also observe that prevention is the only topic discussed throughout most of the year (all months except December), with its primary peaks in April and September. Reported cases have its primary peak in October, with an increasing trend for reported cases from February to May.

Finally, we note that the politics topic starts in May and ends in October, covering the main season of dengue.

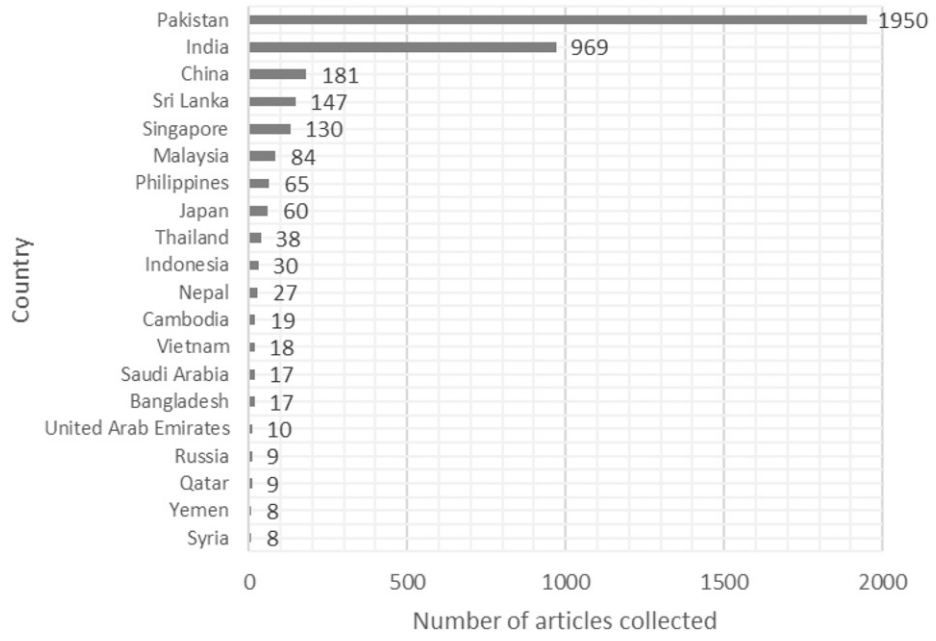

FIGURE 2. Number of articles collected by country.

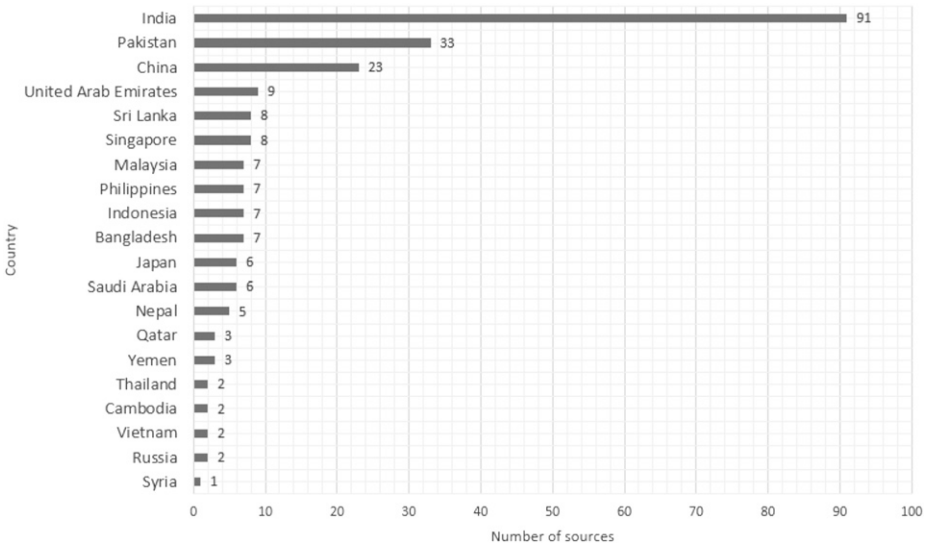

FIGURE 3. Number of individual sources by country.

Prediction. To further explore country-specific details, and to compare our trends with other dengue indicators, we analyzed the country of India separately. Figure 6 shows the topics found in India. Two main trends for 2014 appear: prevention and reported cases. Prevention has two primary peaks in April and in September, whereas reported cases have its primary peaks in July and November. The purpose of comparing our reported cases trend with existing dengue main indicators is to assess how well our reported cases trend can help us predict dengue outcomes.

Historically, researchers have investigated several factors to try to predict dengue cases, such as socioeconomic status and human settlement patterns; migration; temperature and precipitation fluctuations; and Breteau Index levels, which report the number of water containers where dengue is present per 100 houses inspected, documenting the breeding potential of the dengue vector *Aedes aegypti* and *Aedes albopictus*.[44] Herein, we compare the reported dengue cases extracted from news articles with rainfall and the Breteau Index. We then calculate the correlation between these trends to asses our work.

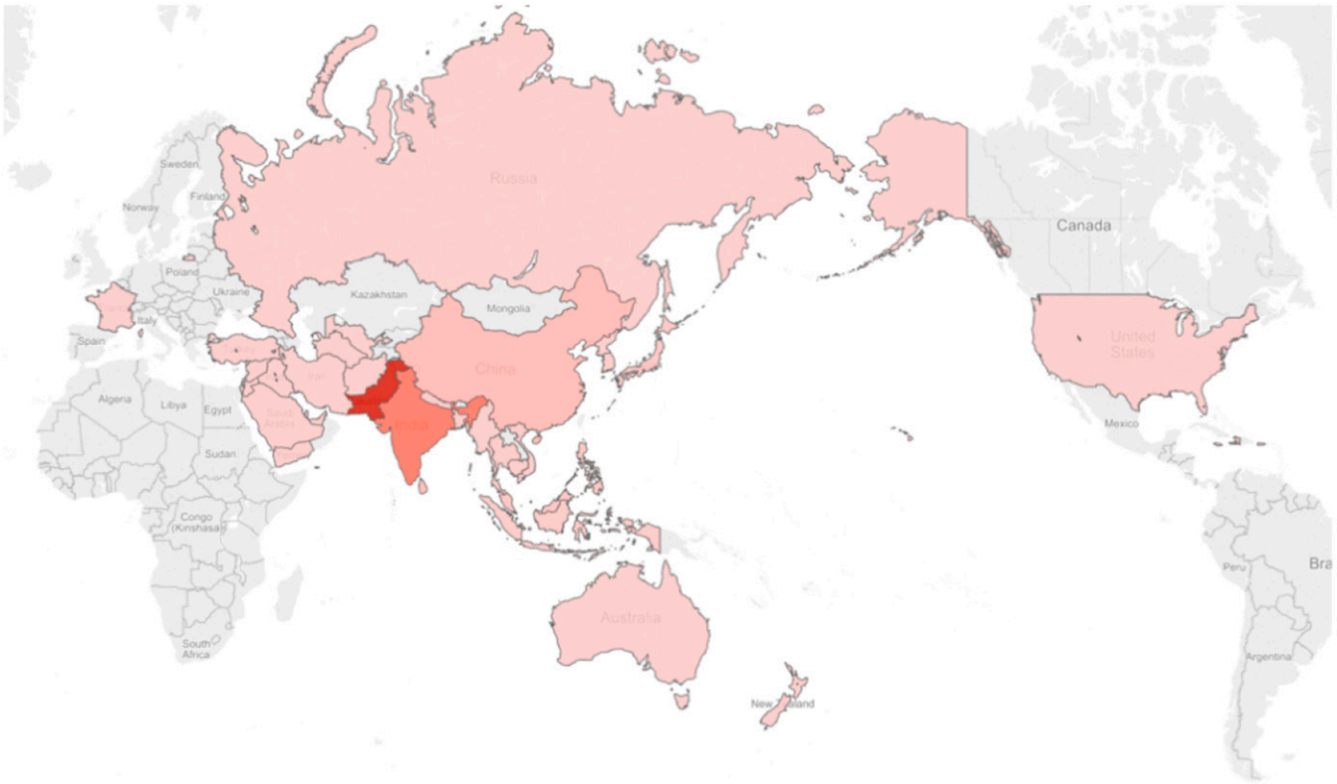

FIGURE 4. Number of articles collected per country. This figure appears in color at www.ajtmh.org.

TABLE 3
Overall number of articles per topic

Topic label	Number of articles
Prevention	1,546
Reported cases	1,460
Politics	407
Prevention–other diseases	270
Emergency plan	68
Research	42
Vaccines	29
Miscellaneous	22
Total	3,844

In Figure 6, we see that the primary peak for average monthly rainfall in India occurs in July, which is the same month where the trend for reported cases starts. The monsoon period for India occurs from June through September, possibly when the dengue vector is more active because conditions favor stagnant water. The June through September period has an average Breteau Index of 8.25, increasing to 21.33 in October, and declining to 0 in February.[45] Although the Breteau Index can be a strong predictor of dengue outbreaks, it is often not available because it requires a time-consuming and costly individual household inspection.

In Figure 7, the primary peak for the number of news articles on reported cases in India occurs in November, 2 months after the monsoon season declines, and 1 month after the Breteau Index peaks. However, as mentioned before, the Breteau Index is labor intensive and cannot be easily collected in each geographic location.

Finally, the appearance of reported cases aligns with the increase in the precipitation index. Transmission of dengue increases during the monsoon season,[46–49] as confirmed in our results. Stagnating water after rainfall favors breeding of the mosquito vector, resulting in an increased incidence of dengue.

To the best of our knowledge, official monthly reported positive cases for India in 2014 are not publically available. Because of this, a monthly estimate for the five regions north, south, east, west, and central was calculated for 2014 based on monthly hospital data $\{h_{\text{north,Jan}}, h_{\text{north,Feb}}, \ldots, h_{\text{north,Dec}}\}$, $\ldots, \{h_{\text{central,Jan}}, h_{\text{central,Feb}}, \ldots, h_{\text{central,Dec}}\}$ found in the literature for each region in India,[50–54] and the annual 2014 annual number of dengue cases reported by the Government of India $\{G_{\text{north}}, G_{\text{south}}, \ldots, G_{\text{central}}\}$ for each state. We extrapolated hospital statistics to the state level based on the official annual positive cases reported by the Indian Government as follows for each of the five regions. For example, we used hospital statistics to calculate a January estimate for the north region as

$$H_{\text{north}} = \sum_{i=\text{Jan}}^{\text{Dec}} h_{\text{north},i}$$
$$h_{\text{north,Jan\_pct}} = \frac{h_{\text{north,Jan}}}{H_{\text{north}}}$$
$$G_{\text{north,Jan}} = G_{\text{north}} \times h_{\text{north,Jan\_pct}} \quad (3)$$

Identical calculations were used to estimate cases for the remaining months and four additional regions. Because our methodology does not rely on the number of absolute positive dengue cases, but rather on the trend of positive cases, a monthly estimate for India in 2014 suffices our purpose. In Figure 8, we observe that the estimated number of positive dengue cases in India has its primary peak in October, whereas the news articles reporting dengue cases in 2014 peak in November. We also compared our estimated results with historic data found in the literature from 2006 through 2008 for Chennai.[49] We observed that the pattern for our estimates and the reported dengue case numbers in Chennai were very similar, as observed in Figure 9. The lack of data of positive reported cases in India for 2014 reinforces our

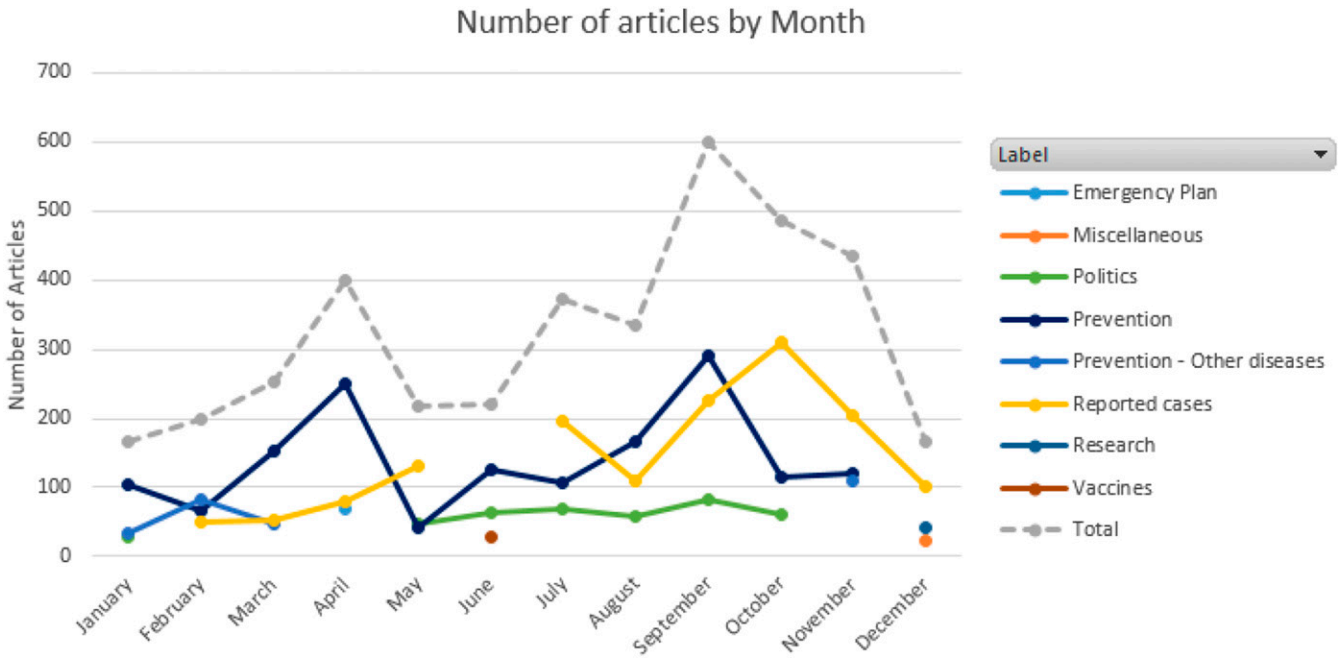

FIGURE 5. Dengue trends for 2014 in Asia. This figure appears in color at www.ajtmh.org.

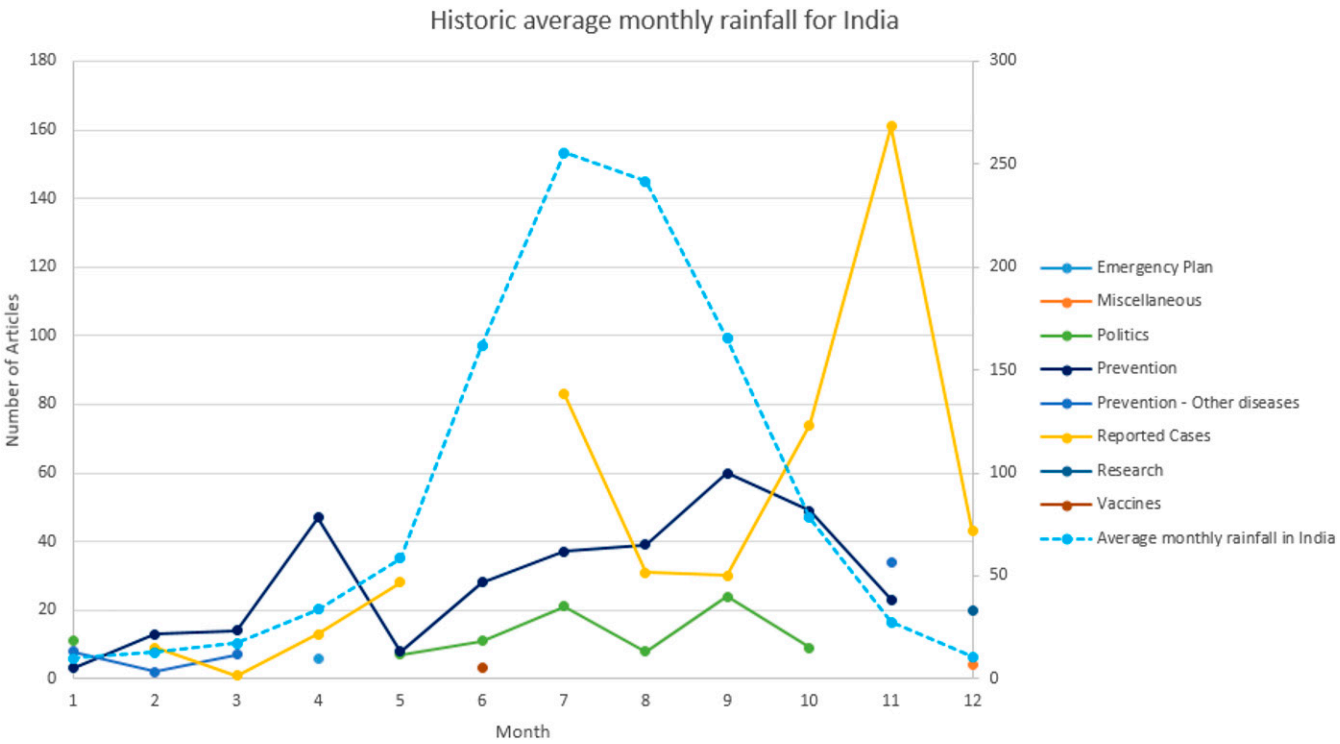

FIGURE 6. Historic average monthly rainfall for India. This figure appears in color at www.ajtmh.org.

premise that official data are not readily available. Moreover, Singh et al. reported 216 positive dengue cases in one hospital in 2014 for Lucknow in the state of Uttar Pradesh. However, the Government of India reported a total of only 200 positive dengue cases for the entire state of Uttar Pradesh. This report which supports the belief that dengue data are underreported[17] and new tools to provide surveillance are needed.

The results presented show a correlation between monthly trends of reported cases, rainfall, and the Breteau Index. Correlation is a metric to measure the connection between two variables, ranging from –1 and 1. A correlation of –1

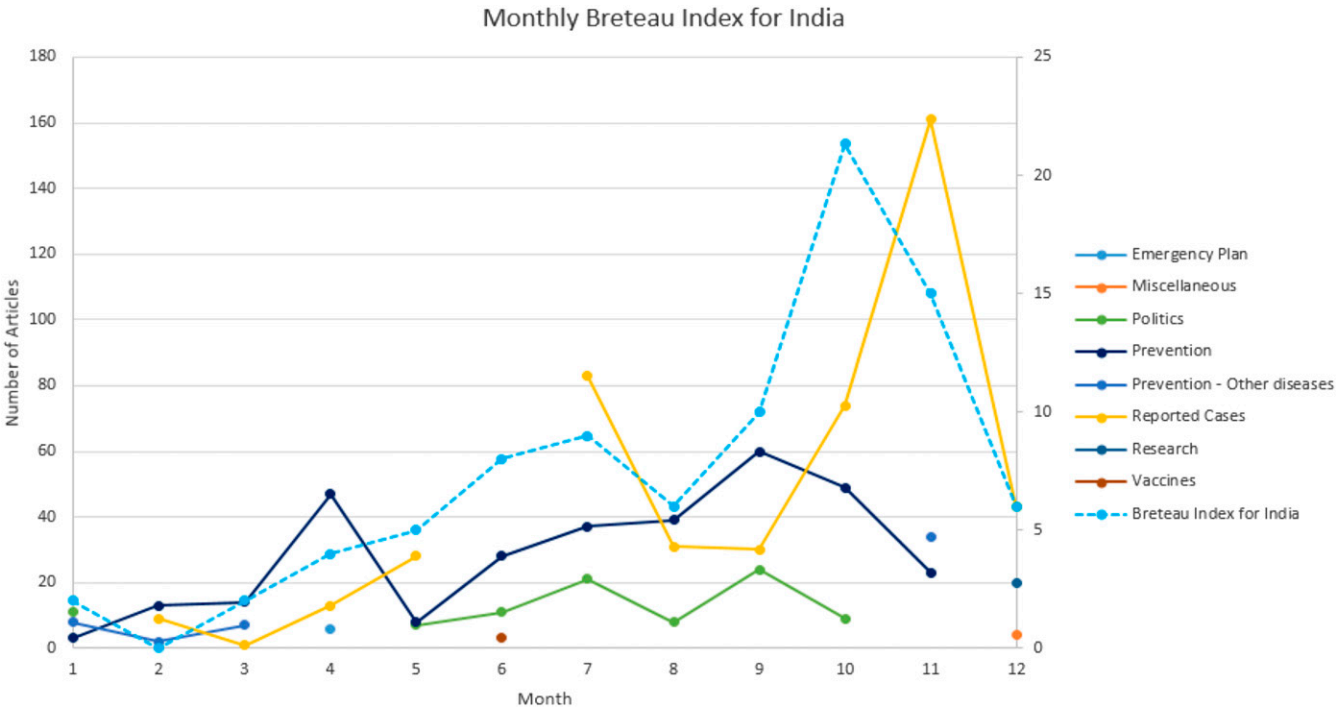

FIGURE 7. Breteau Index for India. This figure appears in color at www.ajtmh.org.

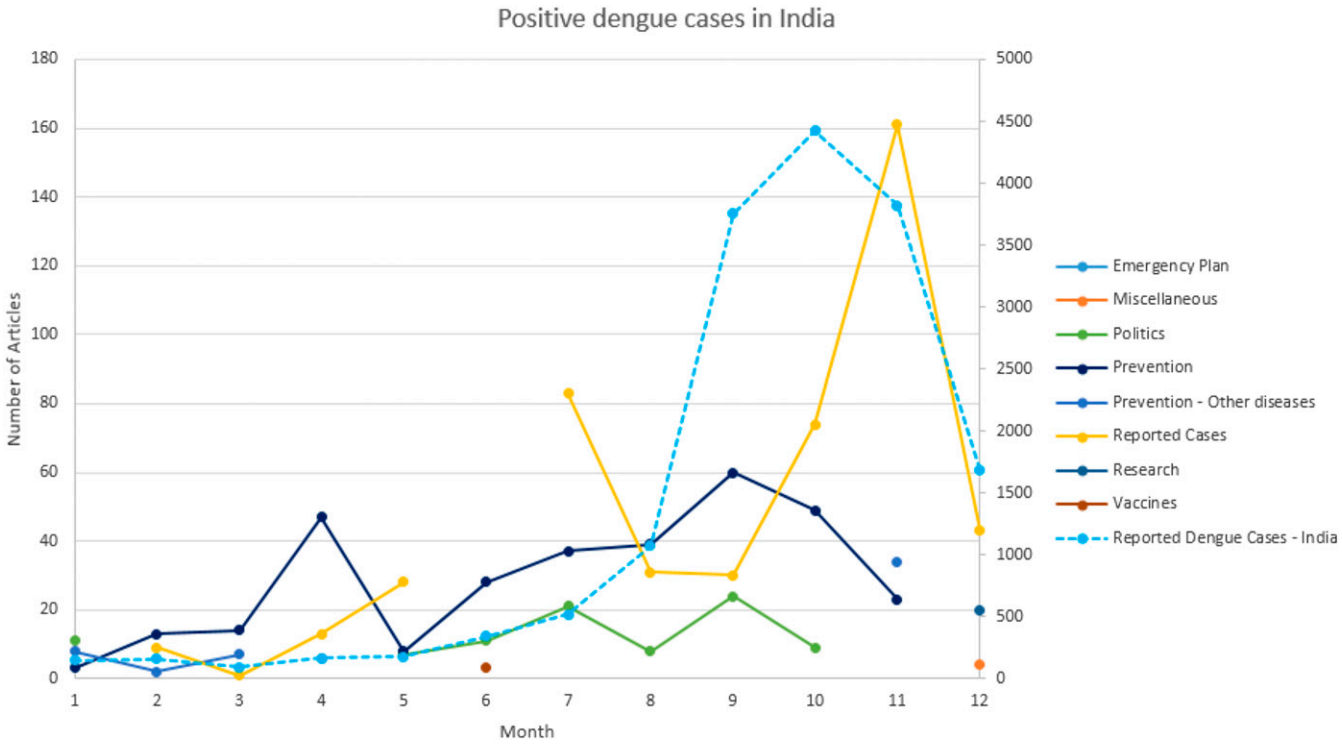

FIGURE 8. Estimated monthly positive dengue cases in India in 2014. This figure appears in color at www.ajtmh.org.

indicates a perfect negative relationship between the two variables (an increase in one produces a corresponding decrease in the other), and a value of 1 indicates a perfect positive relationship. Our results, as indicated in Table 4, show a strong positive correlation (0.92) between the "reported dengue cases" topic extracted from news articles, and the lag of 3 months of average monthly rainfall. In addition, we can see a correlation of 0.69 in relationship with officially reported dengue cases in India with a lag of 1 month. For our prediction purposes, we can infer, given the strong correlation to the "reported dengue cases" topic, that our extracted topics can be used in lieu of dengue trends, such as rainfall and/or official

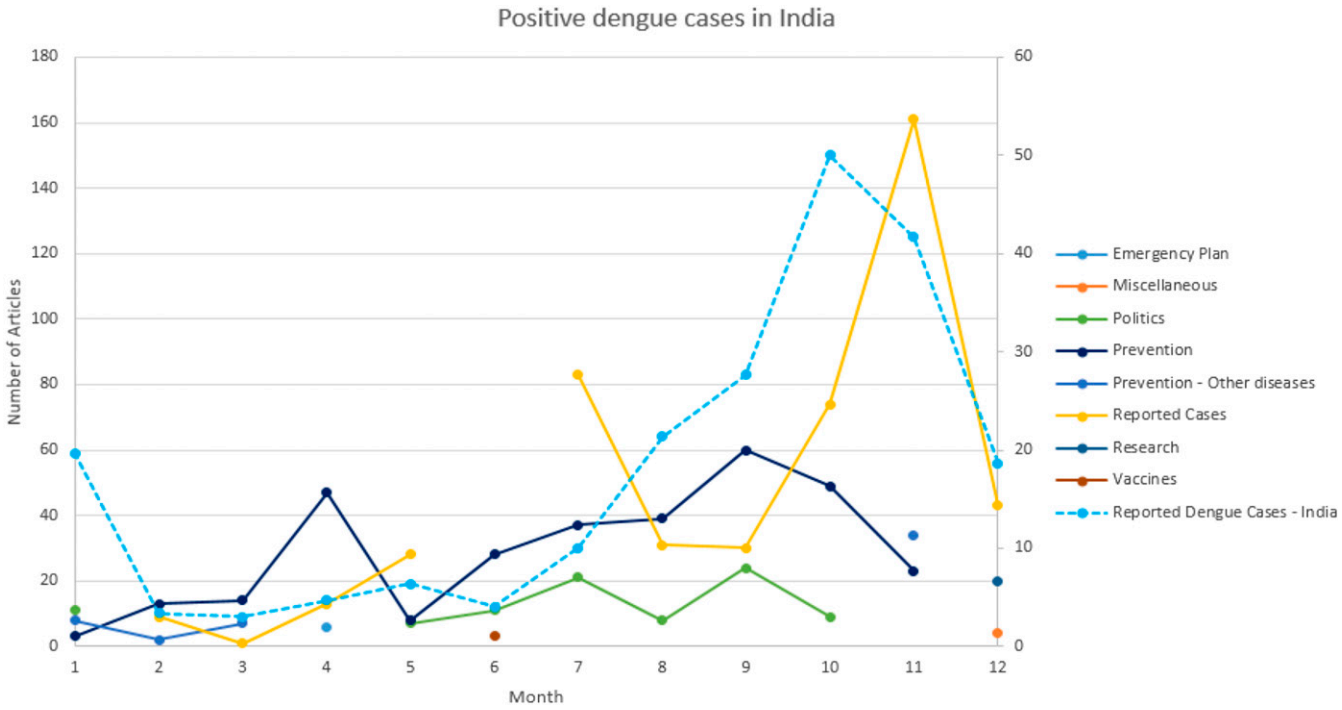

FIGURE 9. Positive dengue cases in India based on historical data from 2006 through 2008. This figure appears in color at www.ajtmh.org.

TABLE 4
Spearman correlation coefficients

	Breteau index in India (Lag 1 month)	Average monthly rainfall for India (Lag 3 months)	Official reported dengue cases for India (Lag 1 month)
Reported dengue cases in news articles	0.27	0.92	0.69

reported cases, supporting the potential usefulness of the proposed methodology in public health surveillance of communicable diseases.

Topic evolution. In all the dengue trend figures we have presented for Asia and India, we see an abrupt decline in news articles in a given country reporting dengue cases after a peak has occurred. This can be explained by the "issue-attention cycle" identified by Peretz in 1972. He describes how an issue abruptly leaps into prominence (alarmed discovery), remains in strong attention for a short time, and then slowly fades from the center of attention (unresolved most of the time).[19]

This cycle can help us to understand what happens in the trends we see in the dengue news articles: 1) a preproblem stage when dengue prevention takes place, and when the trend for reported cases starts to appear; 2) an alarmed discovery when the number of reported cases peaks; 3) a realization of the cost of significant progress when we start to see the trend of dengue research appearing in the news; 4) a gradual decline in intense public interest after the main peak of reported cases occur; and 5) a postproblem stage in the months of December, January, and February.

CONCLUSIONS

In this article, we introduce the use of text mining topic clustering to infer topics from news articles discussing dengue; construct topic evolution graphs; analyze the life cycle of dengue news articles in India, and; relate them to rainfall, monthly reported dengue cases, and the Breteau Index. Our work provides the following novel contributions versus existing approaches: 1) topics extracted from news articles offer not only information on dengue trends in a specific geographic area but also information about other topics extracted from news articles, such as prevention, politics, prevention relative to other diseases, and emergency plans; 2) the evolution of topics throughout the year can be used by dengue experts, health care officials, public health policy makers, communicators, and journalists to obtain insight on relationships to a specific communicable disease; 3) although the rainfall and Breteau Index can be used to detect patterns for dengue, this information may not be promptly available, or may not be collected in a specific region. Our proposed methodology can help close the gap and provide reliable information in a specific region; and 4) although interpretation of the clusters may require human input, our analysis can be automated to reduce the delay in receiving official data, and improve the availability of data needed to decrease morbidity and mortality of communicable diseases.

Our future work includes creating a surveillance system for communicable diseases that combines text mining cluster analysis and sentiment analysis. To perform sentiment analysis, we will create a domain-specific sentiment dictionary for communicable diseases. Next, we will study ways to automatically identify sentiment transitions in a given text, to infer the sentiment in the entire document collection. We will investigate ways to visualize the topics and their associated sentiment estimates to identify relationships between topics. Finally, a Web-based tool will be built to facilitate the surveillance of communicable diseases in different regions of the world.

Received March 28, 2017. Accepted for publication September 10, 2017.

Published online October 23, 2017.

Authors' addresses: Andrea Villanes, Michael Rappa, and Christopher G. Healey, Institute for Advanced Analytics, North Carolina State University, Raleigh, NC, E-mails: avillan@ncsu.edu, mrappa@ncsu.edu, and healey@ncsu.edu. Emily Griffiths, Public Health England, Sheffield, United Kingdom, E-mail: emilycgri@gmail.com.

REFERENCES

1. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, Abraham J, Adair T, Aggarwal R, Ahn SY, 2013. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the global burden of disease study 2010. *Lancet 380:* 2095–2128.
2. World Health Organization, 2006. *Setting Priorities in Communicable Disease Surveillance.* Available at: http://www.who.int/en/. Accessed October 10, 2017.
3. Baker MG, Fidler DP, 2006. Global public health surveillance under new international health regulations. *Emerg Infect Dis 12:* 1058–1065.
4. Langmuir AD, 1963. The surveillance of communicable diseases of national importance. *N Engl J Med 268:* 182–192.
5. Calain P, 2007. Exploring the international arena of global public health surveillance. *Health Policy Plan 22:* 2–12.
6. Beaglehole R, Bonita R, 2001. Challenges public health in the global context-prevention and surveillance. *Scand J Public Health 29:* 81–83.
7. World Health Organization, Regional Office for the Western Pacific, 2004. *Practical Guidelines for Infection Control in Health Care Facilities.* Manila: WHO Regional Office for the Western Pacific, 52.
8. Farrington CP, Andrews NJ, Beale AD, Catchpole MA, 1996. A statistical algorithm for the early detection of outbreaks of infectious disease. *J R Stat Soc Ser A Stat Soc 159:* 547–563.
9. Brownstein JS, Freifeld CC, Reis BY, Mandl KD, 2008. Surveillance sans frontieres: internet-based emerging infectious disease intelligence and the HealthMap project. *PLoS Med 5:* e151.
10. Liu Y, 2004. China's public health-care system: facing the challenges. *Bull World Health Organ 82:* 532–538.
11. Thacker SB, Qualters JR, Lee LM, Centers for Disease Control and Prevention, 2012. Public health surveillance in the United States: evolution and challenges. *MMWR Suppl 61:* 3–9.
12. Davies SE, 2012. The challenge to know and control: disease outbreak surveillance and alerts in China and India. *Glob Public Health 7:* 695–716.
13. Konowitz PM, Petrossian GA, Rose DN, 1984. The underreporting of disease and physicians' knowledge of reporting requirements. *Public Health Rep 99:* 31–35.
14. McKenzie JF, Pinger RR, 2013. *An Introduction to Community Health Brief Edition.* United States of America: Jones & Bartlett Publishers, 43–45.
15. Beatty ME, Stone A, Fitzsimons DW, Hanna JN, Lam SK, Vong S, Guzman MG, Mendez-Galvan JF, Halstead SB, Letson GW, 2010. Best practices in dengue surveillance: a report from the

Asia-Pacific and Americas dengue prevention boards. *PLoS Negl Trop Dis 4:* e890.
16. Suaya JA, Shepard DS, Beatty ME, 2007. *Dengue: Burden of Disease and Costs of Illness. Scientific Working Group: Report on Dengue (Vol. TDR/SWG/08).* Geneva, Switzerland: WHO.
17. Shepard DS, Halasa YA, Tyagi BK, Adhish SV, Nandan D, Karthiga KS, Chellaswamy V, Gaba M, Arora NK, INCLEN Study Group, 2014. Economic and disease burden of dengue illness in India. *Am J Trop Med Hyg 91:* 1235–1242.
18. Woodall J, 1997. Official versus unofficial outbreak reporting through the internet. *Int J Med Inform 47:* 31–34.
19. Downs A, 1996. *Up and Down with Ecology: The "Issue-Attention Cycle".* The Politics of American Economic Policy Making.
20. World Health Organization, Special Programme for Research, Training in Tropical Diseases, World Health Organization, Department of Control of Neglected Tropical Diseases, World Health Organization, Epidemic, Pandemic Alert, 2009. *Dengue: Guidelines for Diagnosis, Treatment, Prevention and Control.* Geneva, Switzerland: World Health Organization.
21. Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, Drake JM, Brownstein JS, Hoen AG, Sankoh O, 2013. The global distribution and burden of dengue. *Nature 496:* 504–507.
22. Moreira LA, Iturbe-Ormaetxe I, Jeffery JA, Lu G, Pyke AT, Hedges LM, Rocha BC, Hall-Mendelin S, Day A, Riegler M, 2009. A wolbachia symbiont in Aedes aegypti limits infection with dengue, chikungunya, and *Plasmodium*. *Cell 139:* 1268–1278.
23. Reyes M, Mercado JC, Standish K, Matute JC, Ortega O, Moraga B, Avils W, Henn MR, Balmaseda A, Kuan G, 2010. Index cluster study of dengue virus infection in Nicaragua. *Am J Trop Med Hyg 83:* 683–689.
24. Montoya M, Gresh L, Mercado JC, Williams KL, Vargas MJ, Gutierrez G, Kuan G, Gordon A, Balmaseda A, Harris E, 2013. Symptomatic versus inapparent outcome in repeat dengue virus infections is influenced by the time interval between infections and study year. *PLoS Negl Trop Dis 7:* e2357.
25. Olkowski S, Forshey BM, Morrison AC, Rocha C, Vilcarromero S, Halsey ES, Kochel TJ, Scott TW, Stoddard ST, 2013. Reduced risk of disease during post-secondary dengue virus infections. *J Infect Dis 208:* jit273.
26. Tissera H, Amarasinghe A, De Silva AD, Kariyawasam P, Corbett KS, Katzelnick L, Tam C, Letson GW, Margolis HS, De Silva AM, 2014. Burden of dengue infection and disease in a pediatric cohort in urban Sri Lanka. *Am J Trop Med Hyg 91:* 132–137.
27. Shepard DS, Undurraga EA, Halasa YA, 2013. Economic and disease burden of dengue in southeast Asia. *PLoS Negl Trop Dis 7:* e2055.
28. Heymann DL, Rodier GR, 1998. Global surveillance of communicable diseases. *Emerg Infect Dis 4:* 362.
29. Freifeld CC, Mandl KD, Reis BY, Brownstein JS, 2008. HealthMap: global infectious disease monitoring through automated classification and visualization of internet media reports. *J Am Med Inform Assoc 15:* 150–157.
30. Collier N, Doan S, Kawazoe A, Goodwin RM, Conway M, Tateno Y, Ngo Q, Dien D, Kawtrakul A, Takeuchi K, 2008. BioCaster: detecting public health rumors with a web-based text mining system. *Bioinformatics 24:* 2940–2941.
31. Mykhalovskiy E, Weir L, 2006. The global public health intelligence network and early warning outbreak detection: a Canadian contribution to global public health. *Can J Public Health 97:* 42–44.
32. Rortais A, Belyaeva J, Gemo M, Van der Goot E, Linge JP, 2010. MedlSys: an early-warning system for the detection of (re-) emerging food-and feed-borne hazards. *Food Res Int 43:* 1553–1556.
33. Tolentino H, Kamadjeu R, Matters M, Pollack M, Madoff L, 2007. Scanning the emerging infectious diseases horizon-visualizing ProMED emails using EpiSPIDER. *Advances in Disease Surveillance 2:* 169.
34. Madoff LC, Woodall JP, 2005. The internet and the global monitoring of emerging diseases: lessons from the first 10 years of ProMED-mail. *Arch Med Res 36:* 724–730.
35 Ahmad T, Rehman NA, Pervaiz F, Kalyanaraman S, Safeer MB, Chakraborty S, Saif U, Subramanian L, 2013. Characterizing Dengue Spread and Severity Using Internet Media Sources. *Symposium on Computing for Development 3*. 18.
36. Hotho A, Nrnberger A, Paa G, 2005. A brief survey of text mining. *GLDV Journal for Computational Linguistics and Language Technology 20:* 19–62.
37. Khan A, Baharudin B, Lee LH, Khan K, 2010. A review of machine learning algorithms for text-documents classification. *J Adv Inf Technol 1:* 4–20.
38. Jain AK, Murty MN, Flynn PJ, 1999. Data clustering: a review. *ACM Comput Surv 31:* 264–323 (CSUR).
39. Feldman R, Sanger J, 2007. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* United States of America: Cambridge University Press, 291.
40 Larsen B, Aone C, 1999. Fast and Effective Text Mining Using Linear-Time Document Clustering. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 5*. 16–22.
41. Kodinariya TM, Makwana PR, 2013. Review on determining number of cluster in K-means clustering. *Int J 1:* 90–95.
42 Stevens K, Kegelmeyer P, Andrzejewski D, Buttler D, 2012. Exploring Topic Coherence Over Many Models and Many Topics. *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning 5*. 952–961.
43. Blei DM, Ng AY, Jordan MI, 2003. Latent Dirichlet allocation. *J Mach Learn Res 3:* 993–1022.
44. Sanchez L, Vanlerberghe V, Alfonzo L, Marquetti Md C, Guzman MG, Bisset J, Van Der Stuyft P, 2006. *Aedes aegypti* larval indices and risk for dengue epidemics. *Emerg Infect Dis 12:* 800–806.
45. Vikram K, Nagpal BN, Pande V, Srivastava A, Gupta SK, Anushrita VP, Singh H, Saxena R, Tuli NR, Yadav NK, 2015. Comparison of *Ae. aegypti* breeding in localities of different socio-economic groups of Delhi, India. *International Journal of Mosquito Research 18:* 20.
46. Reiter P, 2001. Climate change and mosquito-borne disease. *Environ Health Perspect 109:* 141.
47. Hii YL, Rocklv J, Ng N, Tang CS, Pang FY, Sauerborn R, 2009. Climate variability and increase in intensity and magnitude of dengue incidence in Singapore. *Glob Health Action 11:* 2.
48. Barrera R, Amador M, MacKay AJ, 2011. Population dynamics of *Aedes aegypti* and dengue as influenced by weather and human behavior in San Juan, Puerto Rico. *PLoS Negl Trop Dis 5:* e1378.
49. Gunasekaran P, Kaveri K, Mohana S, Arunagiri K, Babu BS, Priya PP, Kiruba R, Kumar VS, Sheriff AK, 2011. Dengue disease status in Chennai (2006–2008): a retrospective analysis. *Indian J Med Res 133:* 322.
50. Singh J, Dinkar A, Atam V, Himanshu D, Gupta KK, Usman K, Misra R, 2017. Awareness and outcome of changing trends in clinical profile of dengue fever: a retrospective analysis of dengue epidemic from January to December 2014 at a tertiary care hospital. *J Assoc Physicians India 65:* 42.
51. Deshkar ST, Raut SS, Khadse RK, 2017. Dengue infection in central India: a 5 years study at a tertiary care hospital. *International Journal of Research in Medical Sciences 5:* 2483–2489.
52. Poddar S, Sengupta P, Chandra G, Hati AK, 2016. Effects of the weather on dengue infections in Kolkata, India. *J Mosquito Res 6:* 1–5.
53. Oza JR, Patel UV, Gajera KD, 2016. Clinico-epidemiological profile of dengue fever cases admitte at tertiary care hospital, Rajkot, Gujarat, India. *Int J Com Med Pub Hlth 3:* 2667–2671.
54. National Health Mission, 2017. *Puducherry State Health Mission.* Available at: http://www.nhmpuducherry.org.in/. Accessed October 10, 2017.