5-19-2016

Developing a Prototype System for Syndromic Surveillance and Visualization Using Social Media Data.

Anup Aryal
axa8556@rit.edu

Developing a Prototype System for Syndromic Surveillance and Visualization Using Social Media Data.

by

Anup Aryal

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Masters of Science

in

Bioinformatics

Thomas H. Gosnell School of Life Sciences, Bioinformatics Program

College of Science

Rochester Institute of Technology

Rochester, NY

May 19th, 2016

Committee Approval:
Dr. Jim Leone, Professor, RIT

Dr. Gary Skuse, Professor, RIT

Dr. Brian Tomaszewski, Assistant Professor, RIT

Acknowledgements

I am immensely grateful to my family especially my mother Ranju Aryal, for her continued support and understanding. This work would not have been possible without continued support from my advisor Dr. Skuse and committee members Dr. Leone and Dr. Tomaszewski.

I am immensely thankful to my friends in Rochester for their support throughout my endeavors, academic or otherwise.

Abstract

Syndromic surveillance of emerging diseases is crucial for timely planning and execution of epidemic response from both local and global authorities. Traditional sources of information employed by surveillance systems are not only slow but also impractical for developing countries. Internet and social media provide a free source of a large amount of data which can be utilized for Syndromic surveillance. We propose developing a prototype system for gathering, storing, filtering and presenting data collected from Twitter (a popular social media platform). Since social media data is inherently noisy we describe ways to preprocess the gathered data and utilize SVM (Support Vector Machine) to identify tweets relating to influenza like symptoms. The filtered data is presented in a web application, which allows the user to explore the underlying data in both spatial and temporal dimensions.

Table of Contents

List of Figures:

List of Tables:

SECTION I

Introduction

Health Surveillance is systematic collection of information regarding public health. This information is needed for planning and evaluating public health practices and policies. More importantly surveillance can help identify early indicators of health emergencies. Health intelligence data is traditionally compiled systematically at regional or national levels by a experts. The primary sources of information are clinical reports, disease reports, laboratory or pathology reports and health registries among others. Most of these reports are complied by health care professionals on a weekly or monthly basis.

Morse describes Health Intelligence as a broad term that applies to any usable information about events that are significant in regard to public health (Morse 2007). Health Intelligence can be classified into Disease surveillance and Syndromic Surveillance.

Disease surveillance refers to gathering information about incidence of a particular disease. Although this is useful to track known diseases, additional systems are required to track unusual patterns that might indicate an emerging disease that might have not been understood yet.

Syndromic Surveillance systems utilize various non-diagnostic systems including but not limited to emergency and ambulance call logs and chief complaints reports from physicians among.

There is a growing need for global Syndromic systems that will help authorities take rapid action in case of an epidemic. Recent outbreaks of H1N1, Swine Flu, Bird Flu and SARS have added urgency to the need to establish a global surveillance system that can serve as an early warning system. (Morse 2007)

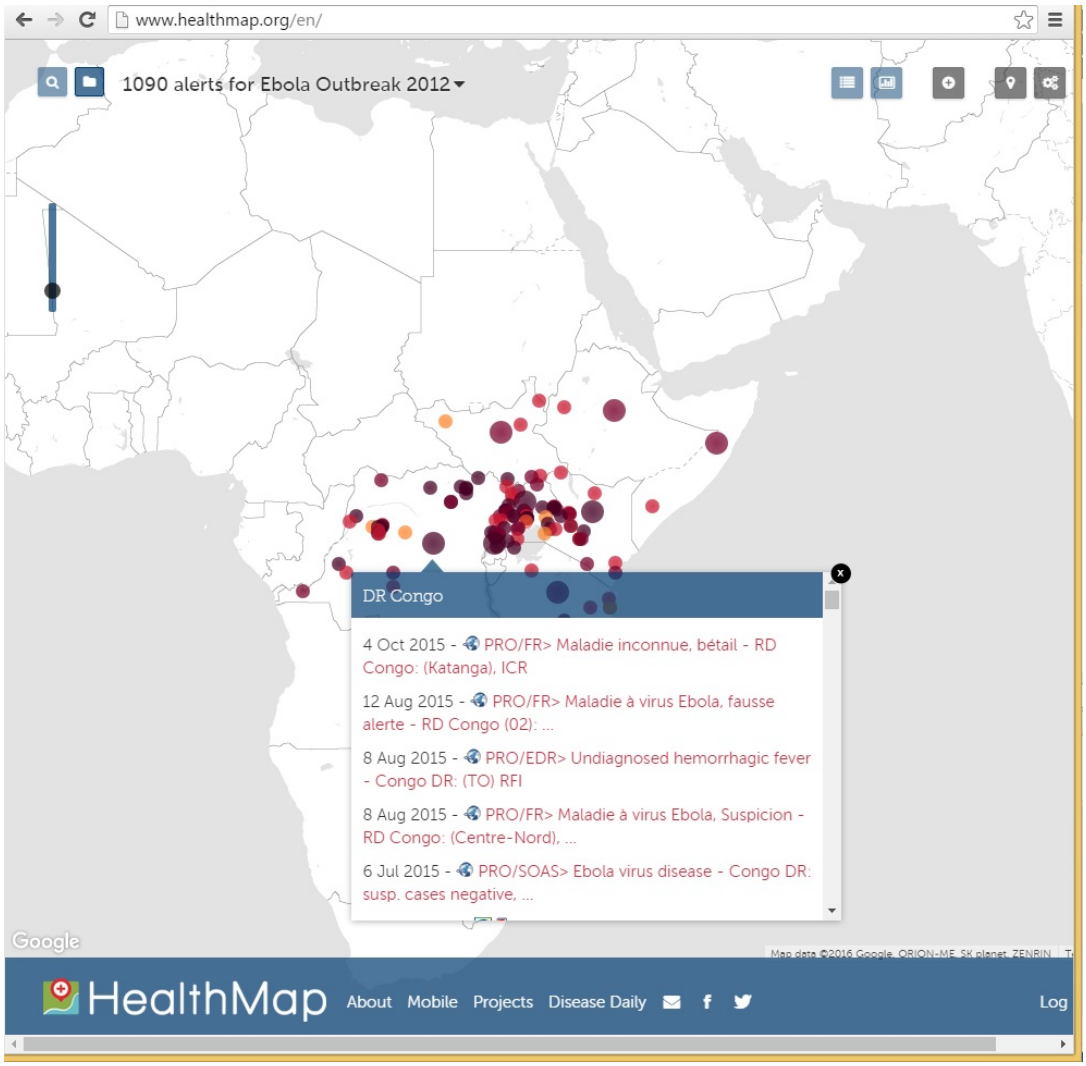
Traditional methods of disease and Syndromic surveillance rely on contributions from experts and are faced with multitude of challenges including lack of resources and established systems and delayed reporting among others. Morse argues that local authorities might have other pressing issues, which causes public health to have an overall lower priority. Further, surveillance might be a secondary priority for local authorities that are focused on responding to existing health problems. Some of these problems pertain more to developing countries; however, a new disease is equally likely to emerge anywhere in the world.

A low cost solution to Syndromic surveillance is to utilize variety of Internet resources as primary sources of data. This method is ideal for developing countries where low cost sources such as Internet sources can substitute traditional reporting systems such as established reporting clinics and laboratories. Internet sources of surveillance data are also valuable for developed countries as they can provide near real time data which can be used for planning a response weeks before any red flags are reported by traditional systems.

One of the early attempts to harness the power of information technology and the Internet in Health Intelligence was ProMed. (Morse 1996) Established in 1996, ProMed is a network of public health professionals that share surveillance information via mailing lists. As of 2004, ProMed has 32000 subscribers in 150 countries. ProMed archives are also available via a web portal. Various subcategories

of PubMed mail are also available for subscription as per area of interest and language (Yu and Madhoff 2004).

More recently additional surveillance systems have emerged that aggregate information multiple sources including ProMed, Government reports, WHO (World Health Organization) reports among others (Keller et. al. 2009). HealthMap, Figure 1 (Brownstein et. al. 2008), EpiSpider (Tolentino et. al.) and Global Public Health Intelligence Network (GPIN) (Mykhalovskiy and Weir 2006) are a few prominent aggregate surveillance systems among various others. These methods rely on semi-automated systems where information gathered by automated systems are reviewed by a team of experts and broadcast to the audience. In addition to the variety of sources covered by these systems an even larger portion of information Internet exists as unstructured textual data. Blog posts and social media including tweets have been recently growing in number and are becoming primary sources of news for many. Utilizing this increasing volume of unofficial sources will further cut down reporting delays.



Further, aggregating information from various unofficial sources and drawing scientifically sound conclusions poses additional challenges. HealthMap, EpiSpider and GPIN rely on official publications while ProMed relies on emails from domain experts; both of these types of sources are semi structured and are generally good quality data. Including information from unofficial sources presents new challenges by increasing the volume of available information and amount of noise within the data. The volume of social media data available makes it impractical to employ human experts to verify the inflow

of information. Natural Language Processing (NLP) techniques can be applied to automate extraction of relevant information from unstructured textual data.

Regardless of the challenges, there is convincing evidence that these unofficial secondary sources provide information that closely reflects real world events. For example, search keyword volume has been shown to correlate with real world influenza trends and associated physician visits (Corley et.al. 2010, Eysenbach 2006 and Ginsberg et.al. 2008). Thus, there is a urgent need for efficient automated surveillance systems that include official as well as unofficial sources. Various groups have made significant efforts in this area, a few of which are described below.

Eysenbach utilized the Google ad service to monitor search words in the 2004 and 2005 flu season and pointed out that there is a statistically significant correlation between search keywords and flu data gathered by government agencies such as Center for Disease Control and Prevention (CDC) (Eysenbach 2006). It seems natural that information-seeking behavior which is represented by search term usage frequency is correlated with real world events.

Similarly, correlation between mentions of Influenza in blog posts has been demonstrated to correlate with Influenza data published by the CDC (Corley, et.al. 2010). Corley et.al. used a dataset of a total of a little over 158 million blog posts and news articles and used lexical match frequency of words "flu" and "influenza" including misspellings anywhere in the content of new articles and blog posts and were able to demonstrate correlation between blog posts containing these words and CDC data on influenza for the same season. (Figure 1) It has been pointed out in existing works that one of the weakness of data mining social media data is that only a very small population is actually active on any

given platform or a combination of such platforms. However, the work by Corley et al. shows us that this small sample of active Internet users is a true representative sample of the entire population and can be used to monitor and/or predict real world events. Further, Collier et al. have shown that people's behavior on the micro blogging site Twitter is also correlated with H1N1 incidence data (Collier and Doan 2011). Collier and Doan employed a SVM (Support Vector Machine) model to identify users displaying self-protective behavior, which indicates their exposure to illness.

Brennan et.al. developed a SVM model to predict influenza rate of a city based on travel patterns of the citizens. This was achieved by tracking travel patterns of twitter users and their friends (Brennan et. a. 2013). nEmesis is a similar model developed by Sadilek et. al. to identify food venues that are likely to spread food borne illness by tracking twitter users (Sadilek et.al. 2016). nEmesis relies on tracking tweets and location of users before and after visiting a food venue.

Some other interesting applications of twitter data are: Unsupervised clustering methods to identify popular topics in twitter (Lansley and Longley 2016), identifying drinking patterns in urban and suburban areas (Hossain et. al. 2016), Predicting election outcome (Burnap et.al. 2015). Twitter data has also been utilized in criminology (Williams 2016). Williams demonstrated correlation of tweets describing social disorder to correlate with crime data. Williams' work is based on a model, which assumes twitter users to be individual sensors that report on changes in their environment. We believe that this assumption can be applied to Syndromic surveillance as well. It should be taken into account that a twitter user might be talking about them experiencing symptoms or merely observing others; the sensor model allows us to generalize the observations to a population. While some authors (Paul and Dredze 2011 and Hussein 2016) go further and try to identify if a user is tweeting their own experiences of having a symptom or merely reporting about their friends, most have not made such distinction as it is not relevant in the large scale i.e. while observing trends for a large population.

In our literature review of this topic we have noticed that there are several established systems that rely on official data and also act as a source of health intelligence for the general public. There are also multiple prototype systems and models that can automate health surveillance and utilize unofficial sources. We have observed a gap in these two categories of work. There are not automated systems that collect Syndromic surveillance data from social media and provide a consolidated view. Our goal for this project is to create a prototype system that gathers Syndromic surveillance data from twitter and provide interactive visualizations, which can help users, evaluate the models being used. In the future, our work might be expanded upon by others to provide a platform to evaluate various data mining methods and models being developed worldwide.

In the following sections we describe the software tools used in this project, the methods used for collecting, filtering, classification, storage and visualization of twitter data. Next we present our results and conclusions, which is followed by the discussions.

Twitter

Twitter is a popular micro blogging website that is freely available to the public. It allows users to publicly post short text messages. While there is no limit to how frequently one can post a message each message is limited to 140 characters. Most tweets (the messages posted on Twitter) are either related to what people are doing or are users reaction to real world events. The users themselves choose to be anonymous more often than other social networks and therefore provide less personal information. The number of twitter users is said to have increased by 40% in 2012, reaching 500 million registered users. The tweets posted accessed via Twitters API (Application Programming Interface). This makes twitter the ideal social media for Social Analytics researchers.

Despite having a global user-base, data gathered from twitter has its own limitations. First and foremost only about 1% of users have opted in to publish their precise geographic location (i.e. latitude and longitude) along with their tweets. There has been some work done for predicting users geographic location by mining their tweets. This involves tracking a user over time and developing models that can predict general location of the user given their tweets over time. (Sadilek et.al. 2012, Cho and Leskovec 2011). We collected over 300 Million tweets in little less than a year (March 2013 to Feb 2014), out of which about 73 Million were geo-tagged. Given the volume of geo-tagged tweets available for analysis we did not pursue any methods of establishing location for the rest of the data. As described earlier, twitter users can be considered sensors that broadcast changes in their environment; a large collection of geo tagged tweets can be used to estimate overall patterns for the population.

Another challenge of analyzing social media data including twitter is that the users do not always spell words correctly or use correct parts of speech. Users might also use multiple languages in one sentence or use vocabulary that is unique to their location. Next we describe our attempt at overcoming some of these challenges.

Support Vector Machines (SVM)

In our dataset unrelated tweets outnumber relevant tweets by a large margin. Although we can identify relevant tweets by searching for simple keywords this approach yields tweets that contain the keyword but may or may not be relevant given the context. For example the message: *"it will be back, that thing is like the flu, it always does"* does not relate to flu and yet will be included in a keyword search. A machine learning algorithm can be trained to separate the dataset into two subsets (one pertaining to influenza like illness and the other not pertaining to such illness). The algorithm can be trained with data that contains synonymous keywords and contextual clues but it cannot exclude extraneous and therefore irrelevant tweets. Such an algorithm can better classify any future tweets.

One such machine learning approach is use of Support Vector Machines (SVM) to build a classification model. SVM also known as Support Vector Networks are widely used for various data mining scenarios including filtering tweets. (Chang and Lin 2011, Sadilek 2012, Paul and Dredze 2011, Culotta 2010, Brennan et. al. 2013, Hossain et. al. 2016)

SVM is a supervised learning method in Machine learning used for regression and classification. First introduced by Vapnik in 1982, the current algorithm that can handle non-linearly separable data was first described in 1995 (Cortes et. al. 1995).

Generating a SVM model involves a mathematical approach, which estimates an optimal generalized function i.e. a function that can accurately categorize future data. A data set with n features can be represented as points in an n-dimensional space. The separation function (also called decision surface) can be can be represented as a line separating data points in two dimensional space, a plane in three dimensional space and a hyperplane for higher dimensions.

One key heuristic SVM utilizes is the idea of Support Vectors. Intuitively it can be seen that the performance of the separating hyperplane depends mostly on the distance to the closest data points. These data points are called Support Vectors are a subset of the dataset. SVM maximizes the distance between these hard to separate data points (support vector) and the separating hyperplane. It should be noted that other machine learning approaches such as Linear Regression and Naive Bayes consider all data points while SVM only considers the difficult data points that are close to the decision surface.

Optimal hyperplane, use of kernel functions and use of soft margins are other key aspects of a SVM. The optimal hyperplane allows for the expansion of solution vectors on support vectors. Kernel functions map the input features into higher dimensions and makes the SVM applicable for non-linearly separable data. Also, the kernel function is applied after the optimal hyperplane is calculated (using dot product of the support vectors) in input space. This step SVM from 'curse of dimensionality' i.e. having to perform n^n calculations after projecting input data into n dimensional feature space. Use of soft margins allow for errors in the training data.

Any data mining method can be most useful by using quality data. As we know twitter is full of spam and chatter that does not pertain to our interest (infectious disease surveillance). One effective approach is to use SVM to annotate tweets. As a first step we trained a SVM model to identify if a tweet pertains to flu or not. The SVM model can be trained with a dataset that contains key word mentions of the word 'flu', 'sick', 'influenza' etc. as well as spam and retweets. The resulting SVM model will be able to distinguish tweets that have keyword mentions relating to flu but don't pertain to flu from the ones that do.

Supervised clustering methods have also been used to identify news topics on twitter. (Wold 2015). Some authors have used Regression methods (Culotta 2010 and Brennan et.al. 2013). Culotta compared classification of tweets into categories using SVM and multiple regression methods and concluded that classification outperformed regression in identifying tweets pertaining to influenza.

SECTION II

Objective

The primary goal of this project is to develop an interactive tool that can be used to visualize output of data mining models that utilize social media data. For the prototype we develop a SVM model to filter tweets that pertain to influenza, the output of this model is presented via interactive web based visualization.

The primary objectives of this project are:

1. Develop a framework for data Collection, storage of tweets.
2. Preprocessing the raw data and transformation into formats that can be readily used with web based visualization libraries.
3. Develop a SVM model that can accurately filter tweets relating to influenza.
4. Develop a prototype platform that allows interactive data visualization at various geographic and temporal dimensions.

SECTION III

Materials and Methods

In this section we will briefly describe the software components of this project. The overall flow of the project is described in figure 1 below. The major steps are: Data Collection, Storage, Analysis, and Visualization.

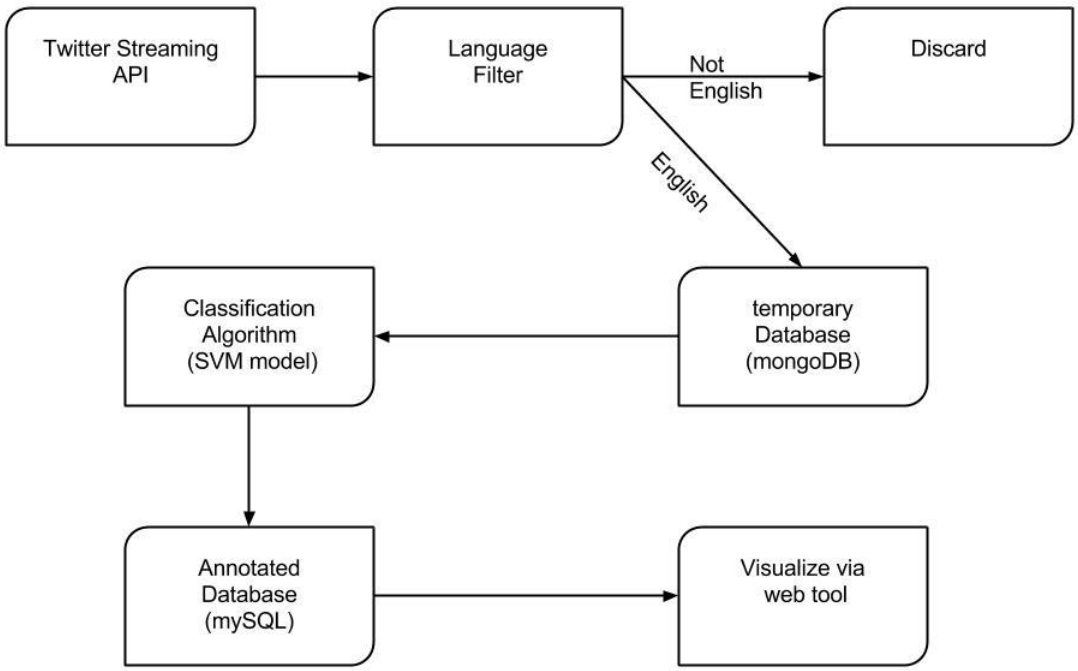


FIGURE 1: DATA FLOW

The data

The data for this project was collected between March 2013 to Feb 2014 using a Java program that utilized Twitter streaming API (Application Programming Interface).

Twitters API allows access to a random sample of tweets that are publicly posted by users. The API also provides an array of metadata for each tweet among which we were most interested in the location information and hash tags. Hash tags are single words preceded by '#' character and are used by tweeter users to tag a message as pertaining to something. We treated these tags as part of the tweet by removing the pound '#' character.

Over 300 million tweets were collected between March 2013 and Jan 2014, which totaled about 8GB of storage.(Table 1) About 6 Million of them were geo-tagged and therefore provided information about the location of the individual posting the tweet.

TABLE 1: VOLUME OF DATA COLLECTED

	Total number of tweets.
Initial raw data (Mar 2013 to Feb 2014)	300 Million
Geotagged tweets among initial data.	6 Million
Tweets relating to influenza that are geotagged.	425,000

The primary data provided by Twitters API is in JSON format. Along with the tweet text the API provides various metadata fields such as timestamp, location (latitude and longitude), username, hashtags. Appendix 1 shows the JSON format used by twitter, it should be noted that twitter has added new attributes over time and that we are mostly interested in the tweet text, timestamp and location.

JSON (Javascript Object Notation)

JSON is a lightweight data exchange format, which is easily readable to both humans and machines. JSON is language and platform independent, which makes it ideal for web, based data. JSON objects can be described as a collection of key value pairs where each value can be a primitive data type or another JSON object. Alternatively, JSON objects can take the form of a series of values similar to an array or a list. Although, these concepts are given different names in different programming languages they are accepted as fundamental data structures and universally supported by most programming languages.

A java program was developed to collect JSON data from twitters API and rapidly store it in a MongoDB database. MongoDB is ideal for this task as it allows to quickly store JSON format data without converting to a structured format required by SQL databases. The data was later processed and the tweets were stored in a relational MySQL database.

MongoDB

MongoDB is one of the most popular noSQL database management systems. The term noSQL refers to non-SQL databases that do not utilize relational algebra and structured tabulation for storage and retrieval of data. MongoDB specializes in storing loosely structured data in JSON format. Since Twitter API provides data in JSON format we can quickly store the incoming data into the database without any preprocessing. Using MongoDB allows us to capture more data without much overhead and have a much stable data gathering system. Using MongoDB also allowed us to start gathering data before the database schema was finalized. We were able to start collecting and exploring data while synthesizing the core objectives of the project. The raw data stored in MongoDB was then filtered and stored MySQL, which was then utilized by the client application for data visualization.

Data Filtering:

Since the available data was large we applied various filtering steps. First of all any tweets that were not in English were discarded. Also, any tweets that did not have latitude and longitude information were removed. Further, a SVM model was applied to identify tweets that are relevant to influenza. The classification model is described below

Classification Algorithm:

In our dataset unrelated tweets outnumbers relevant tweets by a large margin. Although we can identify relevant tweets by searching for simple keywords this approach yields tweets that contain the keyword but may or may not be relevant given the context.

For example the message: *"it will be back, that thing is like the flu, it always does"* does not relate to flu and will be included in a keyword-based search. A machine-learning algorithm can be trained to separate the dataset into two subsets (related and unrelated). The algorithm can be trained with data that contains synonymous keywords and contextual clues. Such algorithm can better classify any future data.

One such machine learning approach is use of Support Vector Machines (SVM) to build a classification model. SVM models are widely used for various data mining scenarios including filtering tweets.

SVM is a supervised classification machine-learning algorithm used for regression and classification. First introduced by Vapnik in 1982, the current algorithm that can handle non-linearly separable data was first described by Cortez et.al. (Cortez et.al. 1995).

Generating a SVM model involves a mathematical approach, which estimates an optimal generalized function i.e. a function that can accurately categorize future data. A data set with n features can be represented as points in an n-dimensional space. The separation function (also called decision surface) can be represented as a line separating data points in two-dimensional space, a plane in three-dimensional space and a hyper plane for higher dimensions.

One key heuristic SVM utilizes is the idea of Support Vectors. Intuitively it can be seen that the performance of the separating hyper plane depends mostly on the distance to the closest data points. These data points are called Support Vectors are a subset of the dataset. SVM maximizes the distance between these hard to separate data points (support vector) and the separating hyper plane. It should be

noted that other machine learning approaches such as Linear Regression and Naive Bayes consider all data points while SVM only considers the difficult data points that are close to the decision surface.

Optimal hyper plane, use of kernel functions and use of soft margins are other key aspects of a SVM. The optimal hyper plane allows for the expansion of solution vectors on support vectors. Kernel functions map the input features into higher dimensions and makes the SVM applicable for non-linearly separable data. Also, the kernel function is applied after the optimal hyper plane is calculated (using dot product of the support vectors) in input space in order to avoid having to perform n^n calculations after projecting input data into n dimensional feature space. Use of soft margins allow for errors in the training data.

Data Preprocessing:

Before any data-mining algorithm can be applied to the data collected multiple preprocessing steps were applied in order to improve the quality of the classification model.

As we know twitter is full of spam and chatter that does not pertain to our interests. Tweets also contain text in various languages and links to webpages. Using SVM model on textual data involves tokenizing the input text. The text data is first split into individual words and a word vector is used to represent the text. In this approach the text is represented as a vector in an n dimensional space, where n is the number of words in the vocabulary used.

As part of cleaning the data we removed any http links, or username mentions which start with @. Further, we removed any non-alphanumeric character. All punctuations were also removed.

Cleaning the text data by removing punctuations, URL links and username mentions simplifies the SVM model by limiting the number of ways a given words might appear in the vocabulary. In most uses, stemming is also used to convert each word to a root word e.g. reading is converted to read. However, removing stop words did not improve our algorithm; hence stop words were not removed. Further, word bigrams were used instead of individual words as it improved the accuracy of the model by 4%.

In addition, we noticed that users were misspelling words by repeating a single character e.g. "soooooo" instead of "so". Any consecutive repeating characters were replaced by itself if repeating more than twice. The text was also converted to all lower case to further simplify the word vectors generated for the model.

LibShortText

We implemented SVM by using libShortText (Yu et. al.) a SVM library for Python specifically designed for short text classification. LibShortText is a modification of the widely used LibSVM library. The advantage of using this particular library is not having to tune the algorithm for the optimum kernel function and penalty factor as the tool comes pre tuned for short text classification.

Short texts messages have unique features that distinguish short text analysis from analysis of other data. Yu et.al. have demonstrated that since short text data have more features than the length of the data use of linear kernel is ideal. (Yu et.al 2010)

We trained the SVM model using data from Yu et.al. (Yu et.al. 2013). The dataset consisted of annotated tweets from 2009 and 2012. Each tweet was annotated using Amazons Mechanical Turk service, an online service for work such as creating a training dataset that can only be done manually. Several annotators categorized each tweet and the consensus was stored. Input form annotators who disagreed with other annotators more often were excluded. The resulting dataset contains 2366 tweets from 2009 and 4760 tweets from 2012 totaling to 7126 status ids. Since twitters does not allow for sharing collected data, only status ids for the tweets were included in the dataset. A simple python script was used to fetch the text for these ids from twitters servers. However, a lot of the these tweets were no longer available . We were able to successfully collect a little over 2000 tweets. We used a thousand of them for training the SVM model and the rest for testing.

The SVM model was testing using the built in analyzer tool that was included in libShortText. Table 1 shows a confusion matrix for the data. Our model performed accurately 97.9% while tested against 1000 tweets in the test dataset; this observation is consistent with the accuracy reported by Yu et.al. However, the model performed poorly (accuracy = 60%) when tested against our data. Upon further investigation, we noticed that the dataset had more mentions of swine flu than seen in our dataset.

In order to develop a data model that is accurate for our data we manually annotated 700 tweets. We identified any tweet as being related to flu if the text was about the user or someone else experiencing

symptoms or seeking care. Train and test dataset were complied by randomly splitting the dataset which included 401 and 299 instances each. The SVM model performed much better in our dataset with accuracy = 81.60%. We applied this classification model to our entire dataset.

TABLE 2 : CONFUSION MATRIX FOR TESTING SVM MODEL (N=299) ACCURACY = (216 + 28)/299 = 81.60%

		Predicted Classes	
		Related	Unrelated
Actual Classes	Related	216	21
	Unrelated	34	28

MySQL Database

Although MongoDB has various advantages MySQL is still ideal for use with web-based applications. The raw data stored in MongoDB was filtered and stored in MySQL, which allowed for indexing and rapid access. Figure n describes the tables in the MySQL database.

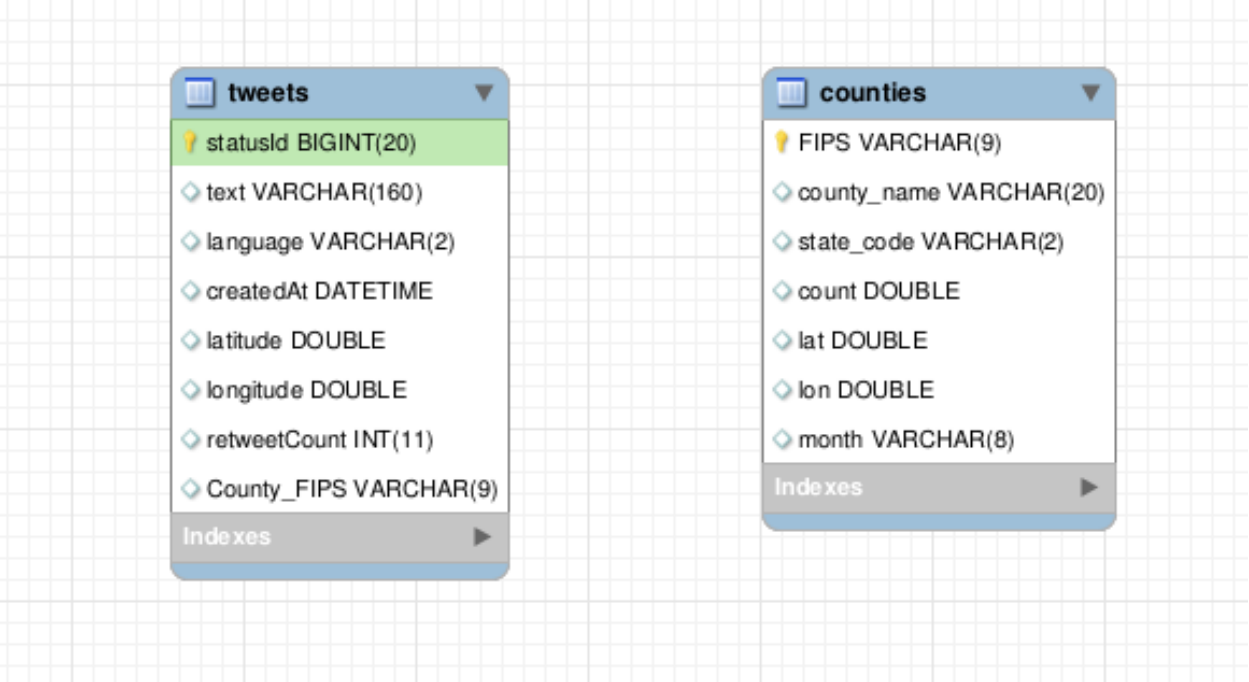


FIGURE 2: MYSQL DATABASE SCHEMA

Tweets table:

Selected attributes from the tweets JSON object were parsed and stored in tweets table. Indexes were added for 'createdAt', latitude and longitude for easy retrieval of data. This table was used to store tweets that were classified as pertaining to influenza by the SVM model described above. We utilized Federal communications commissions (FCC) Census block conversion API to identify which county each tweet belonged to. The API translates (reverse geocode) latitude longitude data into state and county. The county is identified by its unique five-digit identifier (County FIPS), we update our dataset with the county information, which allowed to create aggregation table described below.

Counties Table:

In order to rapidly retrieve data the tweets were aggregated per county. A simple python script was designed to calculate sum total of tweets per county per month. The normalized count of tweets for a county was based on the maximum and minimum number of tweets for any county.

Client Application

The web client application was built using HTML and JavaScript. The primary function of the web tool is to visually present the summarized data. The client application features were designed to have an intuitive and easy to use interface to explore the data and aid in hypothesis generation. Figure n shows the layout of the web client application. Below we explore each aspect in detail.

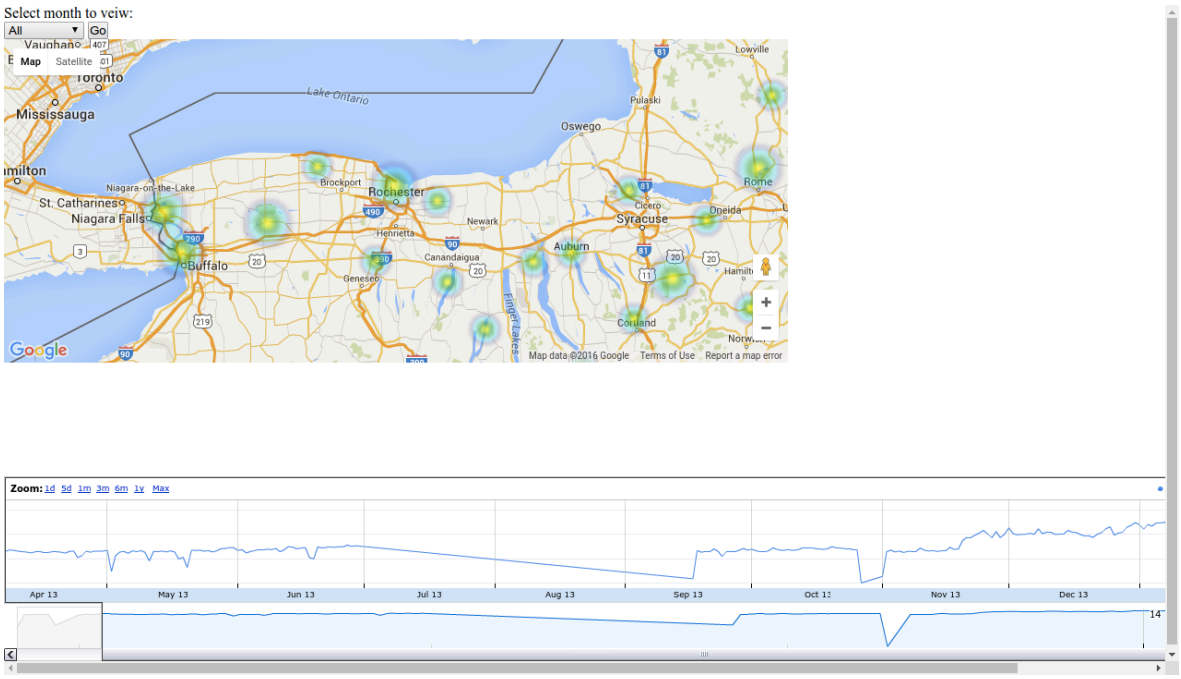


FIGURE 3: WEB BASED INTERACTIVE VISULIZATION OF TWEETS RELATING TO INFLUENZA

Visual aspects

Maps

Maps are ideal visualization tools that allow communicating spatial features of data. Google's map API service provide a very maps service for web tools that require mapping information. The API also allows for the addition of information and the visualization of data points. We experimented with various methods of displaying data points on an interactive map. Displaying each tweet with location feature on a map as a single icon (marker) on the map can lead to visual overload while displaying a large number of markers. Due to the size of our dataset this was not a good option.

We implemented clustering [Gary Little, Marker Clusterer Plus] an open source library that allows for clustering data points based on zoom level.

While viewing larger areas the markers are clustered into a color-coded icon that includes green (few markers), yellow (moderate markers) and red (lots of markers). Using clustered icons we can communicate more information on the map without overloading the visual display with too many visual elements. However, the drawbacks of this approach were that it was only applicable to visualize a moderate number of data points using javascript. Since the clustering was performed on the client side the application was still very slow to display each cluster.

We then implemented a heat map layer that displays a heat map image as an additional layer on the map. The viewable map is broken into multiple rectangles and a heat map image is created by querying the database for data points that fall within the tile. We implemented a PHP library created by Oliver G (G, Oliver, blog.gmapify.fr). The heat map tile server is able to generate a heat map image based on latitude and longitude data. We randomly sampled the database for data points that fall within a tile (a heat map image), this approach yielded a map that showed warm regions near larger cities and cold regions in less populated areas.

Finally we implemented a normalized heat map by county. This was achieved by grouping the raw data by counties. We utilized Federal Communications Commission's Census Block Conversions API, to identify the state and county based on latitude and longitude information.

We were then able to calculate the total number of flu related tweets for a county and normalize the value across all counties. The final heat map displays a larger warm region for a county with higher normalized count.

Timeline

Timelines allow visualizing data over time. Patterns on a timeline might indicate changing trends of keyword usage in response to some real world event. A timeline that shows frequency of tweets related to flu over time was created using Google charts API. The timeline allows to zoom in and out to view specific date range.

Month Dropdown

A dropdown menu was added to allow filtering the data being displayed on the map and the timeline by one click. This allows the visualization of changing patterns over time.

SECTION IV

Results and Conclusion

Recently there has been considerable amount of work done in developing syndromic surveillance systems that collect data from social media. Leveraging freely available social media data such as Twitter has two major advantages over traditional surveillance systems: low cost and speed.

We collected a large volume of tweets which were filtered based on language and presence of location information. Further this data was classified as being related to flu or not using a SVM model. Figure n shows the total number of tweets relating to influenza as predicted by the SVM model. There is a clear increase in tweet volume starting late November. The sections in the figure with sharp drops indicate missing data: 2013 May 26 through 31st, 2013 June 29th through August 17th, 2013 October 27th through November 1st and 2014 Jan 13th through 20th. We noticed a sharp increase in the total tweet volume on Charistmas Day (2013 Dec 25). This was marked by a sharp increase in false positive classification by our model possibly due to the increased volume of people active on twitter. Figure n shows the overall distribution of tweets and volume between October and November.

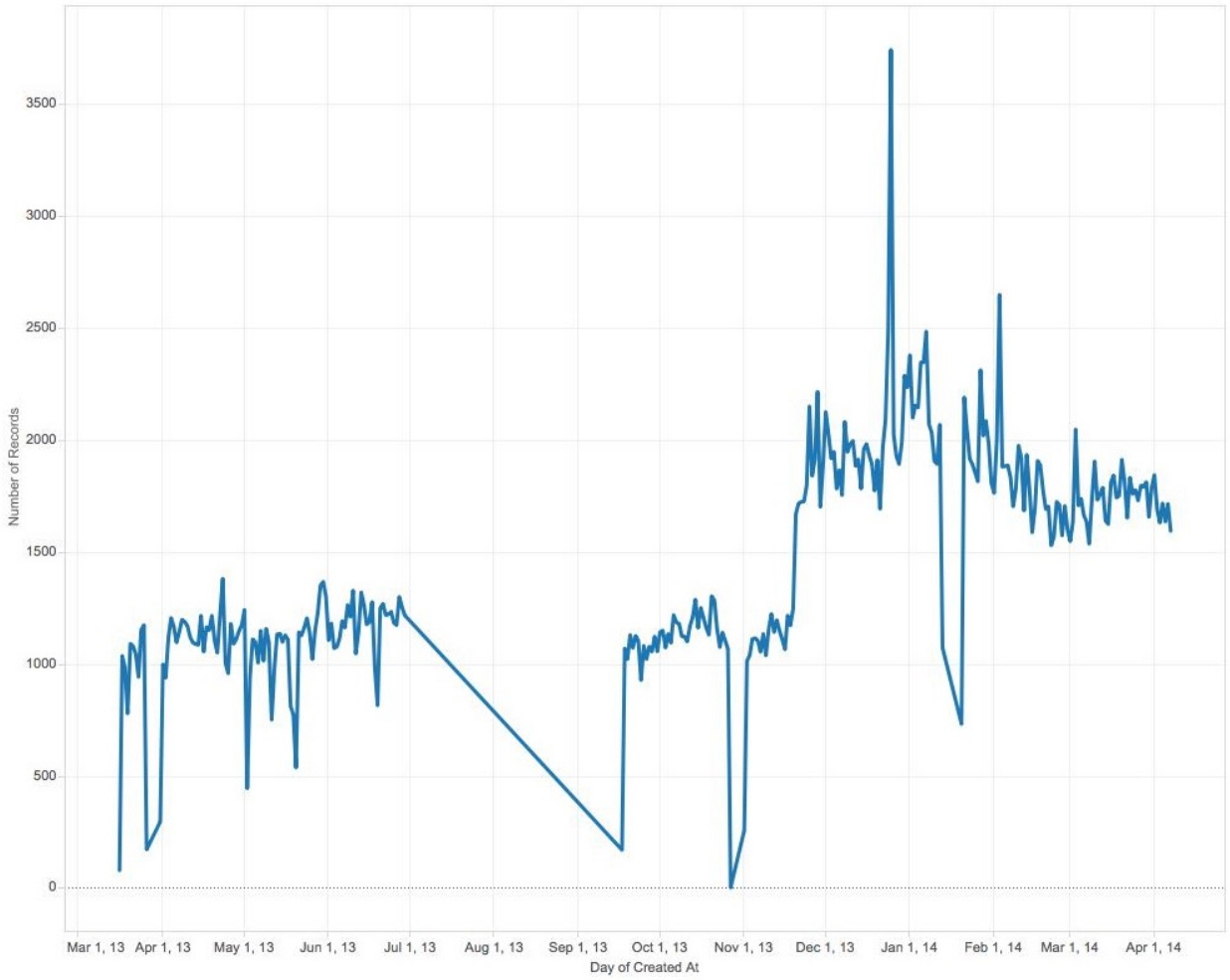


FIGURE 4: NUMBER OF TWEETS RELATED TO INFLUENZA OVER TIME.

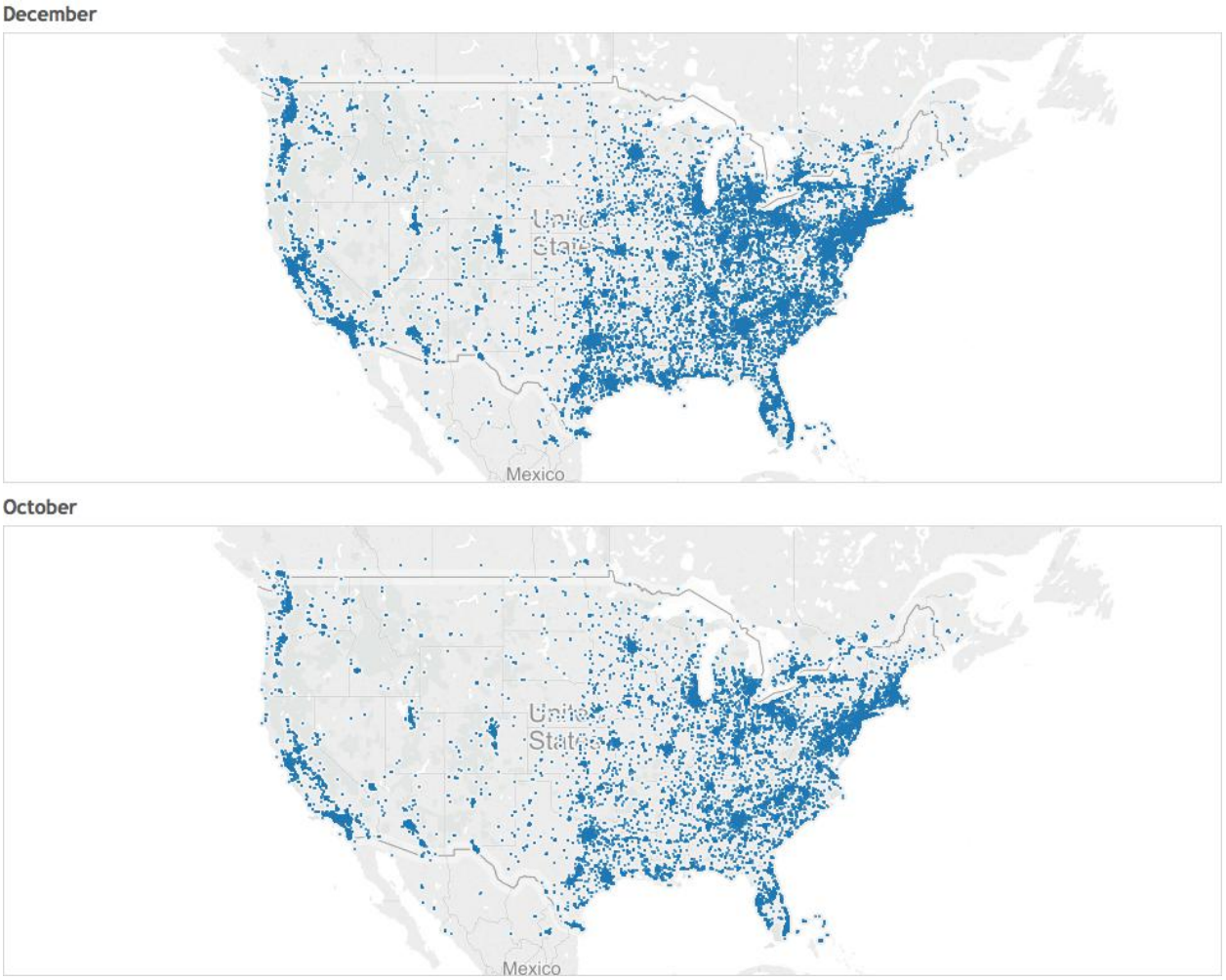


FIGURE 5: GEOGRAPHIC DISTRIBUTION OF TWEETS FOR OCTOBER AND DECEMBER 2013

The SVM model performed at 81.6% accuracy. The resulting data was then stored in MySQL and a summary table was created for easy retrieval of total counts at county level. The county level data was visualized in an interactive map and timeline using HTML and JavaScript. A time filter was added to the visualization to allow for filtering specific date range of data to be visualized. The web tool developed serves as a spatial as well as temporal representation of the tweets identified by our classification model.

SECTION V

Discussion

For this study we focused on training the SVM model to identify tweets related to influenza. With sufficient training and testing data the model can be trained to identify ailments, treatment and/or medications.

Our SVM tagging model was 81.6% accurate. The confusion matrix (table n) shows that the model is biased towards false positives. Based on the confusion matrix we might be including more tweets that don't relate to influenza than excluding actual mentions on influenza.

The SVM model was trained using manually classified dataset of 1000 tweets. The model did not improve by including training datasets from other authors.

Although Twitter has a global user base and tweets are created most popular languages every day for this study we limited our data collection to tweets in English only. Moreover we filtered tweets originating from continental United States by filtering using a rough rectangular boundary.

It was clear from our data that only 1% of all the tweets collected were geo-tagged. Although the web tool allows interactive visualization of the data, in the absence of geo-tags only a small fraction of the data could be displayed on a map. Recently there have been considerable amount of work published about determining user location from the tweets. However, the best methods surveyed were either not accurate enough (50%) (Cheng et. al 2010) or provided better spatial resolution than city level (Mahmud et.al. 2012).

Twitter allows users to post a message originally posted by someone else, these repeated tweets are known as retweets. These retweets can be considered as repeated data points and removed from the analysis. We adopted a different approach and included them in our analysis assuming that increase in repeated tweets about a topic indicates increased awareness among the public. In the future the visualizations generated might be compared with or without including retweets. Also, twitter data can be analyzed entirely based on retweets to measure spread of ideas over time.

Regarding future research in this field, authors might want to consider using a spam filter to avoid spam messages on twitter. Others might also expand on this work by training a classification model for a border Syndromic surveillance that cover other symptoms of common contagious diseases.

SECTION VI

References

1. Brennan, Sean Padraig, Adam Sadilek, and Henry A. Kautz. "Towards Understanding Global Spread of Disease from Everyday Interpersonal Interactions." In IJCAI. 2013.
2. Brownstein, John S., Clark C. Freifeld, Ben Y. Reis, and Kenneth D. Mandl. "Surveillance Sans Frontieres: Internet-based emerging infectious disease intelligence and the HealthMap project." PLoS Med 5, no. 7 (2008): e151.
3. Burnap, Pete, Rachel Gibson, Luke Sloan, Rosalynd Southern, and Matthew Williams. "140 characters to victory?: Using Twitter to predict the UK 2015 General Election." Electoral Studies (2015).
4. Census Block Conversions API. Computer software. Federal Communications Commission. Accessed February 2016. https://www.fcc.gov/general/census-block-conversions-api.
5. Chang, Chih-Chung, and Chih-Jen Lin. "LIBSVM: a library for support vector machines." ACM Transactions on Intelligent Systems and Technology (TIST) 2, no. 3 (2011): 27.
6. Cheng, Zhiyuan, James Caverlee, and Kyumin Lee. "You are where you tweet: a content-based approach to geo-locating twitter users." In Proceedings of the 19th ACM international conference on Information and knowledge management, pp. 759-768. ACM, 2010.
7. Cho, Eunjoon, Seth A. Myers, and Jure Leskovec. "Friendship and mobility: user movement in location-based social networks." In Proceedings of the 17th ACM SIGKDD

international conference on Knowledge discovery and data mining, pp. 1082-1090. ACM, 2011.

8. Collier, Nigel, and Son Doan. "Syndromic classification of twitter messages." In Electronic Healthcare, pp. 186-195. Springer Berlin Heidelberg, 2011.
9. Collier, Nigel, Nguyen Son, and Ngoc Nguyen. "OMG U got flu? Analysis of shared health messages for bio-surveillance." *Journal of Biomedical Semantics*2.Suppl 5 (2011): S9.
10. Corley, Courtney D., Diane J. Cook, Armin R. Mikler, and Karan P. Singh. "Text and structural data mining of influenza mentions in web and social media." International journal of environmental research and public health 7, no. 2 (2010): 596-615.
11. Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20, no. 3 (1995): 273-297.
12. Culotta, Aron. "Towards detecting influenza epidemics by analyzing Twitter messages." In Proceedings of the first workshop on social media analytics, pp. 115-122. ACM, 2010.
13. Eysenbach, Gunther. "Infodemiology: tracking flu-related searches on the web for syndromic surveillance." In AMIA Annual Symposium Proceedings, vol. 2006, p. 244. American Medical Informatics Association, 2006.
14. Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. "Detecting influenza epidemics using search engine query data." Nature 457, no. 7232 (2009): 1012-1014.
15. Herman Tolentino, M. D., M. D. Raoul Kamadjeu, M. P. H. Michael Matters PhD, M. D. Marjorie Pollack, and M. D. Larry Madoff. "Scanning the emerging infectious diseases

Horizon-visualizing ProMED emails using EpiSPIDER." Advances in disease surveillance 2 (2007): 169.

16. Hossain, Nabil, Tianran Hu, Roghayeh Feizi, Ann Marie White, Jiebo Luo, and Henry Kautz. "Inferring fine-grained details on user activities and home location from social media: Detecting drinking-while-tweeting patterns in communities." arXiv preprint arXiv:1603.03181 (2016).
17. Keller, Mikaela, Michael Blench, Herman Tolentino, Clark C. Freifeld, Kenneth D. Mandl, Abla Mawudeku, Gunther Eysenbach, and John S. Brownstein. "Use of unstructured event-based reports for global infectious disease surveillance." Emerg Infect Dis 15, no. 5 (2009): 689-695.
18. Victor, L. Yu, and Lawrence C. Madoff. "ProMED-mail: an early warning system for emerging diseases." Clinical infectious diseases 39, no. 2 (2004): 227-232.
19. Lamb, Alex, Michael J. Paul, and Mark Dredze. "Separating Fact from Fear: Tracking Flu Infections on Twitter." In HLT-NAACL, pp. 789-795. 2013.
20. Lampos, Vasileios, Tijl De Bie, and Nello Cristianini. "Flu detector-tracking epidemics on Twitter." In Machine Learning and Knowledge Discovery in Databases, pp. 599-602. Springer Berlin Heidelberg, 2010.
21. Lansley, Guy, and Paul A. Longley. "The geography of Twitter topics in London." Computers, Environment and Urban Systems 58 (2016): 85-96.
22. Mahmud, Jalal, Jeffrey Nichols, and Clemens Drews. "Where Is This Tweet From? Inferring Home Locations of Twitter Users." ICWSM 12 (2012): 511-514.
23. MarkerClustererPlus for Google Maps V3 computer program. Version 2.0. Gary Little Accessed Aug 2013 https://github.com/mahnunchik/markerclustererplus

24. Morse, Stephen S., Barbara Hatch Rosenberg, Jack Woodall, and ProMED Steering Committee Drafting Subgroup. "ProMED global monitoring of emerging diseases: design for a demonstration program." Health Policy 38, no. 3 (1996): 135-153.
25. Morse, Stephen S. "Global infectious disease surveillance and health intelligence." Health Affairs 26, no. 4 (2007): 1069-1077.
26. Mykhalovskiy, Eric, and Lorna Weir. "The Global Public Health Intelligence Network and early warning outbreak detection: a Canadian contribution to global public health." Canadian Journal of Public Health/Revue Canadienne de Sante'e Publique (2006): 42-44.
27. Paul, Michael J., and Mark Dredze. "You are what you Tweet: Analyzing Twitter for public health." ICWSM 20 (2011): 265-272.
28. Oliver, G. Gmapify. Computer software. Gmapify. Accessed August 2013. blog.gmapify.fr.
29. Sadilek, Adam, Henry A. Kautz, and Vincent Silenzio. "Predicting Disease Transmission from Geo-Tagged Micro-Blog Data." In *AAAI*. 2012.
30. Sadilek, Adam, Henry Kautz, Lauren DiPrete, Brian Labus, Eric Portman, Jack Teitel, and Vincent Silenzio. "Deploying nEmesis: Preventing Foodborne Illness by Data Mining Social Media." (2016).
31. Williams, Matthew L., Pete Burnap, and Luke Sloan. "Crime Sensing with Big Data: The Affordances and Limitations of using Open Source Communications to Estimate Crime Patterns." British Journal of Criminology (2016): azw031.
32. Wold, Henning Moberg, and Linn Christina Vikre. "Online News Detection on Twitter." (2015).

33. Yu, H., C. Ho, Y. Juan, and Chih-Jen Lin. "Libshorttext: A library for short-text classification and analysis." Rapport interne, Department of Computer Science, National Taiwan University. Software available at http://www.csie.ntu.edu.tw/~cjlin/libshorttext (2013).

Appendix 1: Twitter Streaming API JSON data format.


```
{
  "coordinates": null,
  "created_at": "Thu Oct 21 16:02:46 +0000 2010",
  "favorited": false,
  "truncated": false,
  "id_str": "28039652140",
  "entities": {
    "urls": [
      {
        "expanded_url": null,
        "url": "http://gnip.com/success_stories",
        "indices": [
          69,
          100
        ]
      }
    ],
    "hashtags": [

    ],
    "user_mentions": [
      {
        "name": "Gnip, Inc.",
        "id_str": "16958875",
        "id": 16958875,
        "indices": [
          25,
          30
        ],
        "screen_name": "gnip"
```

```
            }
          ]
        },
        "in_reply_to_user_id_str": null,
        "text": "what we've been up to at @gnip -- delivering data to happy
customers http://gnip.com/success_stories",
        "contributors": null,
        "id": 28039652140,
        "retweet_count": null,
        "in_reply_to_status_id_str": null,
        "geo": null,
        "retweeted": false,
        "in_reply_to_user_id": null,
        "user": {
          "profile_sidebar_border_color": "C0DEED",
          "name": "Gnip, Inc.",
          "profile_sidebar_fill_color": "DDEEF6",
          "profile_background_tile": false,
          "profile_image_url":
"http://a3.twimg.com/profile_images/62803643/icon_normal.png",
          "location": "Boulder, CO",
          "created_at": "Fri Oct 24 23:22:09 +0000 2008",
          "id_str": "16958875",
          "follow_request_sent": false,
          "profile_link_color": "0084B4",
          "favourites_count": 1,
          "url": "http://blog.gnip.com",
          "contributors_enabled": false,
          "utc_offset": -25200,
          "id": 16958875,
          "profile_use_background_image": true,
          "listed_count": 23,
```

```
            "protected": false,
            "lang": "en",
            "profile_text_color": "333333",
            "followers_count": 260,
            "time_zone": "Mountain Time (US & Canada)",
            "verified": false,
            "geo_enabled": true,
            "profile_background_color": "C0DEED",
            "notifications": false,
            "description": "Gnip makes it really easy for you to collect social data
for your business.",
            "friends_count": 71,
            "profile_background_image_url":
"http://s.twimg.com/a/1287010001/images/themes/theme1/bg.png",
            "statuses_count": 302,
            "screen_name": "gnip",
            "following": false,
            "show_all_inline_media": false
          },
          "in_reply_to_screen_name": null,
          "source": "web",
          "place": null,
          "in_reply_to_status_id": null
        }
```