Tracking Disease Outbreaks Using Twitter

David J. Marchette* Elizabeth Hohman*

Abstract
In this paper we describe preliminary work aimed at utilizing social media data to enhance biosurveillance. Social media applications such as Twitter can be used to detect and track disease outbreaks in near real-time, provided that:

1. The data can be collected.
2. The observations can be tied to a geographic area.
3. It can be determined whether the person (or someone geographically close to them) is sick.
4. The "sick" individuals can be identified as suffering from the same disease or at least have similar symptoms.

We discuss our efforts in detection and tracking using Twitter data collected from January 2013 to the present, and discuss the various issues that arise in using Twitter data. In particular, we will discuss various keyword and topic-based methods, as well as methods for classifying a tweet or a user as "sick". We discuss some of our successes and failures and provide some insight into the utility and limitations of micro-blog data such as that provided by Twitter. We will discuss variations on the basic surveillance theme such as watching for a known disease (measles), a known set of symptoms (fever, stomach ache), and the more general (and difficult) problem of detecting an unusual number of sick individuals within a constrained geographic region (county).

Key Words: Social media, twitter, disease outbreak, text analysis, biosurveillance

1. Introduction

Social media is of considerable interest as a sensor into the thoughts, interests and health of a population (Collier et al. [2011], Corley et al. [2010], Dredze [2012]). By analyzing the postings on social media sites, one can learn peoples' reactions to events in the news, new products, politics, celebrities, and their overall satisfaction, happiness or anger. In this work we investigate the extent to which social media, in particular Twitter, can be used to detect an outbreak of a disease or illness.

We were given the task of determining whether Twitter could be used either for the early detection of outbreaks of some disease, or to better understand the outbreak or the evolving situation of the outbreak. We term these outbreaks "events", and the first question we address (and the focus of this paper) is can we detect events?

We must first define what we mean by "event". There are several possible definitions, including:

- Given a known disease, such as MERS, SARS, Measles, etc., an event corresponds to individuals contracting the disease.
- Given a set of symptoms (fever, stomach pain, etc.), an event is an unusual number of individuals[1] complaining of the symptoms.

[*] Naval Surface Warfare Center, 18444 Frontage Rd, Suite 327, Dahlgren, VA 22448. This work was funded through an Interagency Agreement with the Department of Homeland Security National Biosurveillance Integration Center (NBIC). NSWCDD-PN-14-00498 Distribution Statement A: Approved for Public Release. Distribution is Unlimited.

[1] We will always take "unusual" to mean as compared to the past, which we will make explicit in Section 4.

- Most generally: an event is an unusually large group of individuals who can be identified as being effected by some personal illness.

Note that to detect an "unusual number" of something, we need to count the indicators of the event, and we need to compare the current count with past counts. Further, we are generally interested in geographically constrained events, and so for this work we will focus on county-based counts. We will count the number of items (tweets or individuals) expressing the event indicator (a disease name, symptom, or classified as "personal health related" as indicated by our classifier).

As discussed below, our approach to detecting health related events is: **filter** → **classify** → **detect**. We first filter out tweets that contain no "health related" terms, then apply a classifier to each tweet. This classifier is designed to flag a tweet as being about "personal health" or not. We then aggregate the positive instances per day at the county level and detect as an event any county/day pair with an unusually high count (as compared to the recent past).

This paper is organized as follows. In Section 2 we will discuss the Twitter data we collect. We then discuss the classifier in Section 3 and the event detection methodology in Section 4. We show results and conclusions in Sections 5 and 6.

2. The Twitter Data

We use Twitter data primarily because of its availability and volume. Twitter is a service whereby users can broadcast short messages (constrained to be no more than 140 characters). The messages (called "tweets") can contain links to other content (usually represented as tiny URLs to reduce characters) as well as mentions of other users. These mentions guarantee that the tweet shows up in the mentioned user's notifications page, and are a way of sending messages directly to a user. In this work we focus exclusively on the text of the tweet, and do not utilize the content of links or the social network defined by the mentions, or by who is following whom.

A tweet can be tagged with a geographic location (latitude and longitude). This can happen through a variety of mechanisms, the most common (and accurate) is to use a smart phone with location enabled. While it is true that the location can be spoofed, and can be set via other mechanisms such as clicking the "tweet this" button on a web site, in this work we will assume that the reported location is correct, at least to the resolution of the county.

We collect all tweets within a rectangle which covers the continental United States. The Twitter API[2] provides free access to a subset of the tweets, with a limit on the number of tweets it provides within any time window. However, experiments we have run (tweeting at various times and locations) indicates that except for a few high-volume times, we do collect all tweets that have a geographical location within the rectangle. [3]

Figure 1 shows the tweets for one day in the continental United States, compared with a map of the county-level population.[4] Note that the overall pattern of the twitter volume is the same as the population density, giving some evidence of the coverage of the social media signal.

The Pew Research Center has collected statistics on twitter users[5]. They report that 19% of on-line adults use Twitter, with the 18–29 year olds accounting for the largest

[2] `https://dev.twitter.com/docs/api/streaming`

[3] There is a further caveat that we occasionally lose tweets due to power and network outages.

[4] The county level population map came from the web site `http://www.mapofusa.net/us-population-density-map.htm`.

[5] `http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/` accessed August 13, 2014.

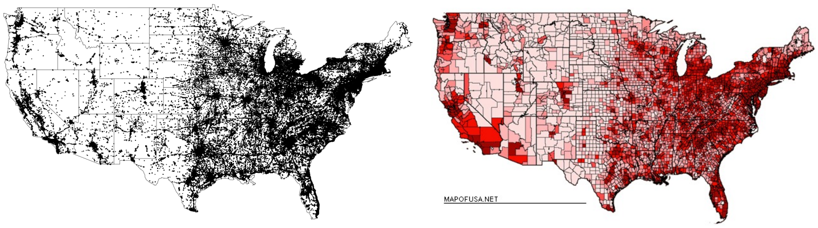

Figure 1: One day's worth of tweets in the continental United States (left) compared with the county-wide population density (right). There are 1.7 million tweets in this set, collected on May 2. 2013.

proportion. While they do not report statistics on the number of people who turn geo-location on, they report a large percentage (40%) of mobile phone users use some social networking site on their phones.

According to various reports (see for example Leetaru et al. [2013]) somewhere between 1% and 3% of tweets contain a geo-reference (with quite a bit of variance amongst the different populations). We are aware of no study of the demographics of users who allow this geographic location in their tweets, however a perusal of the tweets that we collect leads us to believe that the overall demographics discussed above apply to this subset as well.

Obviously (from Figure 1) Twitter usage is higher in urban areas rather than rural, and there is some empirical evidence that this is more than simply a population effect – that Twitter is used by a higher percentage of urban dwellers than rural. [6]

2.1 Processing Tweets

Below are some example tweets. These have been slightly modified to remove mentions of user names and some non-ascii characters. Note the different uses of the word "sick", the shortness of the tweets (we have chosen shorter tweets so that they fit on a single line), the spelling and jargon, the hash tags (indicated by the # symbol) and the links.

```
not today. I'm sick
I'm hungry but the thought of food makes me sick..
Sore is the new sexy #truestory
them oxys got me like .... #oxy #drugs #beans #offthembeans
ion know im hurtin lol
I hate hospitals and the waiting game!! Grr...
I know ... Why u shaking ur head though
Day 2 of cramps. Yay me!
I'm talking about physical painful rain
Lol nat has mono
Panera Bread sounds sooo good rn.
It's a sunglasses and Advil kinda day
Lhh I'm weak text me mf
My stomach hurts so bad :(
My teeth are sooo sore omg #crying
http://t.co/zuP5kasrNl Jumping Jesus why does this country sick so badly.
I'm physically sick right now. But I'm even more emotionally sick because of all thesw kids with NO GUIDANCE.
I feel badddd
cramps will literally have u like http://t.co/wORN7k7sHI me last night
Coworker is sick at work because pregnant with twins. I'm sick at work because pregnant with beer. Help.
I wants these piercings but I know my mama gone be sick.
Clear cases are so sick
God, he makes me sick.
Free At Last is such a sick band.
On time to finish this house today #sick #jomo3 #rebuildjoplin
Mehico two weeks from today. Sickkkkk
```

A number of cleaning procedures are run on each tweet to produce a "cleaned" version of the text. All tweets are mapped to lower case, mentions (username) and links are removed as are punctuation and non-ascii characters. Hashtags (for example: #crying) are

[6] http://www.pewinternet.org/2013/02/14/social-networking-site-users/

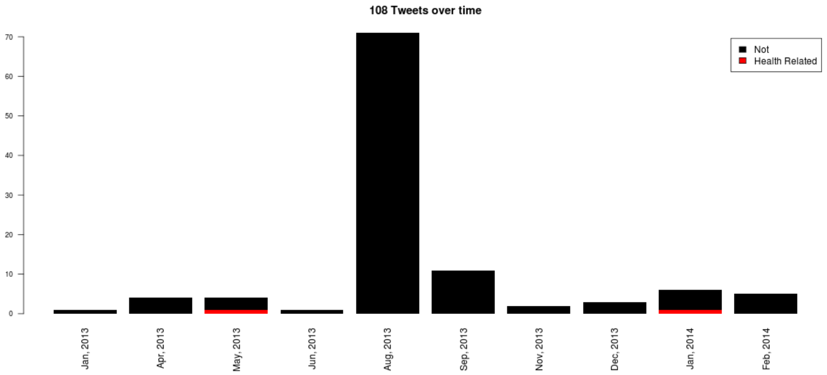

Figure 2: Number of tweets per month that are geographically located within the state of Texas and contain the word "measles". Tweets that are considered to be about personal health by the classifier (see Section 3) are indicated in read.

split on case (prior to lower case mapping): "GoTigers" becomes "go tigers". Finally, any letter repeated more than two times is reduced to a single instance: "im soooooo happpppyyyyy" becomes "im so hapy". Note that we made a choice here: instead of a single instance, we could have gone to a double – "im soo happyy"; we could have mapped end-of-words to one, otherwise two – "im so happy" (but then "im soooooo happppppyyyyy tooooo" becomes "im so happy to") or we could have done a dictionary search/spelling correction. All of the above have problems, and none will work universally, so we chose the fastest (arbitrarily mapping to one instead of two).

The above cleaning allows us to use a bag-of-words vector space model.[7] Since tweets are so small, we do not weight, instead we construct a binary term-document matrix. We will return to this in Section 3.

Each tweet is assigned a county based on its latitude and longitude. This is done by indexing into a matrix of county numbers that is pre-computed, rather than by using the county polygons directly. The tweets whose position is outside the United States [8] are dropped, as are retweets.[9]

2.2 Tracking a Named Disease

Suppose one knows a priori what diseases are of interest. We can search the tweets for words associated with a disease and process these to determine whether there is an event of interest. See Lampos et al. [2010], Aramaki et al. [2011], Achrekar et al. [2011] for some examples looking at influenza.

In 2013 there was an outbreak of measles in several counties in Texas, due in large part to a failure to immunize children to the disease (Silverman [2013]). In Figure 2 we see the number of tweets matching the word "measles" in Texas. Note the large number of tweets in August, which is when most of the news organizations reported the story.

To get an idea of what these tweets are about, we can look at who the users are who are tweeting. Figure 3 shows a bar plot of the number of tweets by user. Most of the screen

[7] `http://en.wikipedia.org/wiki/Bag-of-words_model`, `http://en.wikipedia.org/wiki/Vector_space_model`

[8] Recall that the API gives us all tweets inside a rectangle, so some will be in the parts of Canada and Mexico covered by the rectangle.

[9] A retweet is essentially the equivalent of a forwarded message in email. These are unlikely to be an indication that the person forwarding the tweet is personally ill.

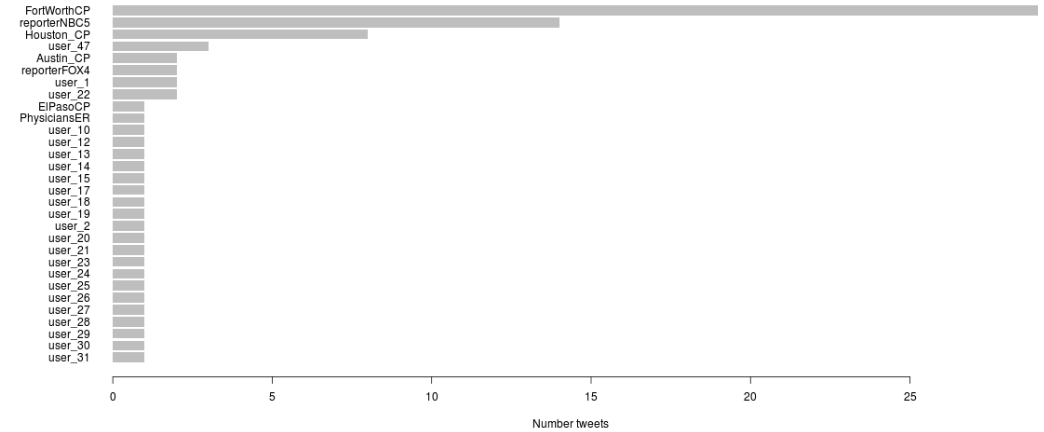

Figure 3: Number of tweets per user. The names of the form "user_#" have been obfuscated to protect their privacy.

names have been obfuscated to preserve privacy,[10] but the ones that we have left allow for a good idea of what these tweets are. The end "CP" on a user name generally means "City and Press", and we see news organizations and PhysiciansER, which in this case is warning about measles, not reporting a specific case.

Further analysis of individual tweets shows that there are at most two tweets about individuals personally affected by the measles.[11] So, in this case we see that the twitter "buzz" is primarily a result of the news reports of the outbreak, either news organizations reporting on the event, or of people discussing these news reports. One interesting fact we learned from this exercise is that during the same time period that people were discussing measles in humans, there was an outbreak of a measles-like disease in dolphins, which also generated some discussion. There were also reports of measles in Europe, so even though the tweeters were located in Texas, they are discussing events from around the world.

In a subsequent study, we obtained all tweets that Twitter was able to locate to Texas (not just the ones that reported a latitude and longitude) and we found a few more instances of people tweeting about individuals in their family contracting the disease. This indicates that there is hope for this type of surveillance for specific diseases, but it works best when one can obtain all the tweets, not just those with a geographic location attached. In order to determine the place of an outbreak we require the ability to accurately locate each tweet (user) at least to the county-level. This is an area of future research.

2.3 Tracking a Symptom

Instead of looking for a known disease, with all the issues discussed above, we could look for a set of symptoms. As discussed above, looking for symptoms is a bit problematic, since the number of ways that people might refer to a given symptom can be large (specially in twitter), but perhaps one can select symptoms that are fairly universal. In Figure 4 we show the number of tweets containing several different words: fever, doctor, headache, sick, stomach. The strong diurnal (weekly) pattern is easy to see, particularly in the counts for

[10] Although twitter is a microblogging service whose main purpose is to broadcast information to the world at large, we prefer to protect the privacy of individuals to the maximal extent, and so except for the institutional or professional users, all screen names are replaced with generic names.

[11] It is not always possible to determine from the text of a tweet what the individual is actually talking about: one says "she got the measles lol" which may or may not be about an actual case of the measles; the other tweet cannot be reproduced here without great risk of causing offense.

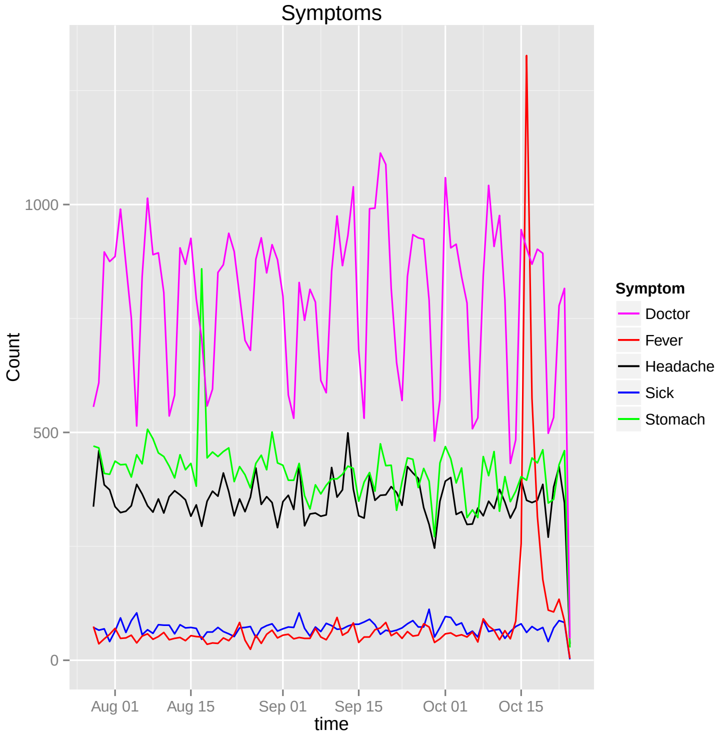

Figure 4: Counting the number of tweets containing the words: doctor, fever, headache, sick and stomach. These tweets were collected in North Carolina in 2012. The spike in the fever curve is the release of the moving "Cabin Fever II".

the word "doctor".

What is obvious in this figure is the spike in mid October for the word "fever". This is not an outbreak of some fever-causing disease, but rather the release of a popular movie with the word "fever" in its title.[12] Further analysis of tweets containing the word fever shows that people use the word "fever" in many different ways.[13] For example, one sees: spring fever, bieber fever, football fever, and many other "<sport> fever", "<celebrity> fever" and other cultural references that use the word "fever" to indicate positive emotion or approval. One can obviously filter these out, provided one knows about them, but it is difficult to do so for new usages. One also sees "<expletive> fever", "degree fever" and "103 fever" and variations, which (usually) are health related.

The word "stomach" is another word that is potentially problematic. People say

- stomach ache
- stomachache
- stomach is killing me
- stomach hurts
- stomach pain
- stomach cramps

and variations, which are likely related to illness, but also talk about "stomach growling" and similar phrases, which may or may not be about illness (and are more likely about hunger).

[12] Cabin Fever II

[13] This is obvious in retrospect.

One might attempt to use linguistic methods, such as part-of-speech tagging, to eliminate such“non-health” usage. Our preliminary experiments with this were not encouraging. Tweets tend to be ungrammatical, with frequent misspellings and liberal use of expletives, lacking or incorrect use of punctuation, jargon, colloquialisms and acronyms. As a result, part-of-speech tagging is difficult and error prone without specific training for Twitter. There has been some work in this area (for example see ARK) and so further investigation is warranted.

We consider a different approach to symptom tracking. We first use a large (about 300-phrases) set of terms that are indicative of health (such as symptom terms, remedy terms, and disease terms), then pass the tweets matching one or more terms through a classifier trained to classify tweets as “about personal health” or not. The next section discusses this classifier.

Our list of phrases contains, for example:

- Disease terms: flu, salmonella, salmonella, strep, strepto$*$, pertussis, pertusis, a cold.
- Symptom terms: headache, bellyache, fever, burning up, feel awful (and several variations).
- Remedy terms: tylenol, vaccine, cold med$*$, antibiotic.

These have word-boundaries set as appropriate, and allow for partial matching as indicated by the wild-card $*$. As indicated, we have spelling variations, and variations on phrasing for some of the terms. We add to the list as we discover new diseases and new ways that people refer to (or misspell) various symptoms or otherwise indicate illness.

One word that is not in our list is “ill”. This is because the vast majority of the times this word appears in a tweet, it corresponds to “I’ll” (with the apostrophe missing and usually not capitalized). We have added “feeling ill” and variations to catch these types of usage, so the word does appear, just not by itself. We have similar issues with “er” (short for emergency room, but also used as an interjection; “rn” for registered nurse, but more commonly used as an acronym for “right now”, “dr” for doctor, but also used to denote “drive” in an address.

We retain the word “sick” in our list, even though it has a high false alarm rate, due to the use of “sick” to mean:

- Good: “I just heard a sick band”.
- Bad: “Hunting animals for fun is just sick”.
- Feeling bad (not illness): “Watching him treat her like that makes me sick”.
- A verb: “Next time I see him I’m going to sick my dog on him”.
- A noun (possibly illness?): “I hate it when I have to clean up her sick”.
- An expression of surprise or reaction: “sick!”

To navigate through all these meanings and find just the ones that are about personal health is difficult, and even our classifier (discussed in Section 3) has difficulty with some of these. While we can mitigate this to some degree with the approach we discuss below, it may be reasonable to take the position that the word “sick” alone in a tweet is not enough to retain it as a possible indication of illness. However, tweets like: “i am sooooo sick rn” are common enough that we choose to retain such tweets and use a classifier to attempt to distinguish between real “personal health” tweets and others.

3. Personal Illness Classifier

In order to determine whether a tweet is relevant and should be counted, we constructed a classifier trained to distinguish between “personal health” and not. So a tweet of the form

Table 1: Confusion matrix for the random forest classifier trained on 10,000 and tested on 3,342 withheld tweets.

True Class	Predicted Class		
	Class	0	1
	0	1895	157
	1	191	1099

"i am really sick today" would be class 1 (personal health) while a tweet such as "be sure to vaccinate against the flu" would be class 0.

First we hand-tagged a set of about 2000 tweets according to our two classes. We processed the tweets as discussed in Section 2.1. In addition to words we also counted bigrams (adjacent pairs of words), whether there as a link or user mention in the tweet, and whether there was a positive or negative emoticon. We then ran a number of classifiers (using leave-one-out cross validation to determine accuracy) including k-nearest neighbors, naive Bayes, support vector machines and random forests. Random forests performed the best on these data, and so we decided to use that for our classifier.

We then tagged a much larger set of tweets (over 13 thousand), and used these to train the random forest classifier. We ran the random forest classifier on a subset of the data and looked at the importance of the variables. This allowed us to reduce the dimensionality of the input to about 700 terms that were most useful for the classification task.

The vector space model we used was a binary document-term matrix, where each document is a row and there is a 1 in each column (corresponding to a word or a bigram) if that term was in the cleaned version of the tweet. No weighting is performed on these values, nor are stop words removed – words like 'i' and 'have' are useful for the task of determining whether the tweet is about an illness that the tweeter or someone close to them is suffering from. The terms included where only those contained in the 700 terms that the previous analysis had indicated were useful for classification.

We tested the algorithm by running a number of training/testing runs, as indicated in Figure 5. As can be seen, we observe an error of about 10%, once the number of training observations gets sufficiently large. The selection of the 700 most important terms was not done independently on the training/testing runs, and so it may have introduced a slight bias into our results. However, we have anecdotal evidence on new data that indicates that we still get over 80% correct classification, even though the twitter data is constantly changing, as people discuss new issues and new jargon is invented. It is clear that for a "deployed" system some mechanism for updating the classifier should be considered.

Table 1 shows the confusion matrix for the classifier trained on 10,000 tweets and classified on the remaining 3,342. As can be seen, the classifier makes fairly symmetric errors. Given the ambiguity of the data and the number of ways people have of saying things, plus the 140 character constraints of twitter, we believe that the performance is quite good, and is adequate to perform the next step, which is counting the number of health related ("sick") tweets or users within a constrained geographic region.

The classifier is only applied to tweets containing at least one of our 300 (approximately) phrases, and it is worth asking how many legitimate tweets we miss by performing this pre-filtering. We looked at a set of 8000 tweets that were randomly selected from those that did not contain one of the terms, and found 4 tweets that (arguably) should have been tagged as "health related".

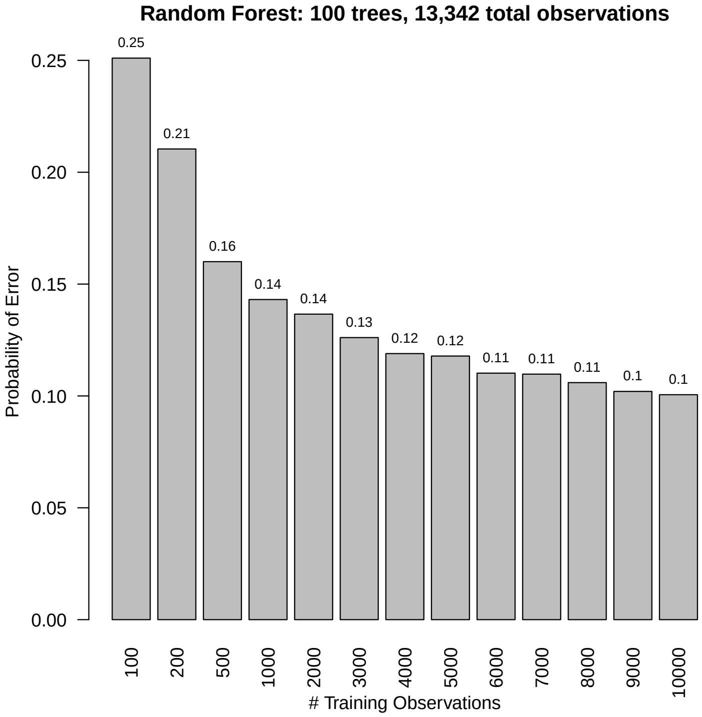

Figure 5: Performance of the random forest classifier. On the x-axis is the number of observations used to train the classifier, while on the y-axis is the probability of error for the classifier applied to the rest of the data.

4. Events

As noted above, tracking known events and word-based symptoms is problematic, and doesn't solve the problem of detecting an outbreak of an unknown or unexpected disease. Our procedure is to classify each tweet using our random forest classifier as "health related" or not, then count the number of class 1 tweets, or the number of users with at least one class 1 tweet, per day within each county.

Given a time series of counts per day (c_t) for a given county, we determine an event as an unusual number of tweets on a day defined by:

$$z_t^W = \frac{c_t - \mu_t^W}{\sigma_t^W}. \tag{1}$$

Here μ_t^W and σ_t^W are the mean and standard deviation for the counts in a window W of time prior to t. For this paper we will define W to be the seven days prior to t. Alternatively, we use the median and inter quartile range as a more robust version. An event is a county/day pair for which $z_t^W > 3$.

To illustrate this approach, consider Figure 6. Here is a case with an extreme z-score, that is essentially the result of a single hour on the day of the detected event. In this case we are counting "health related" tweets, not users, and this is the cause of what turns out to be a spurious detection. It is caused by a single individual tweeting nearly 100 tweets of the form "You know I should be asleep bc it's 4am and I'm sick . . . NUMBER", where "NUMBER" runs from 1 to 96.

This illustrates the necessity to track users rather than individual tweets, however this alone is not sufficient since we also see nearly identical tweets that are tweeted out by a

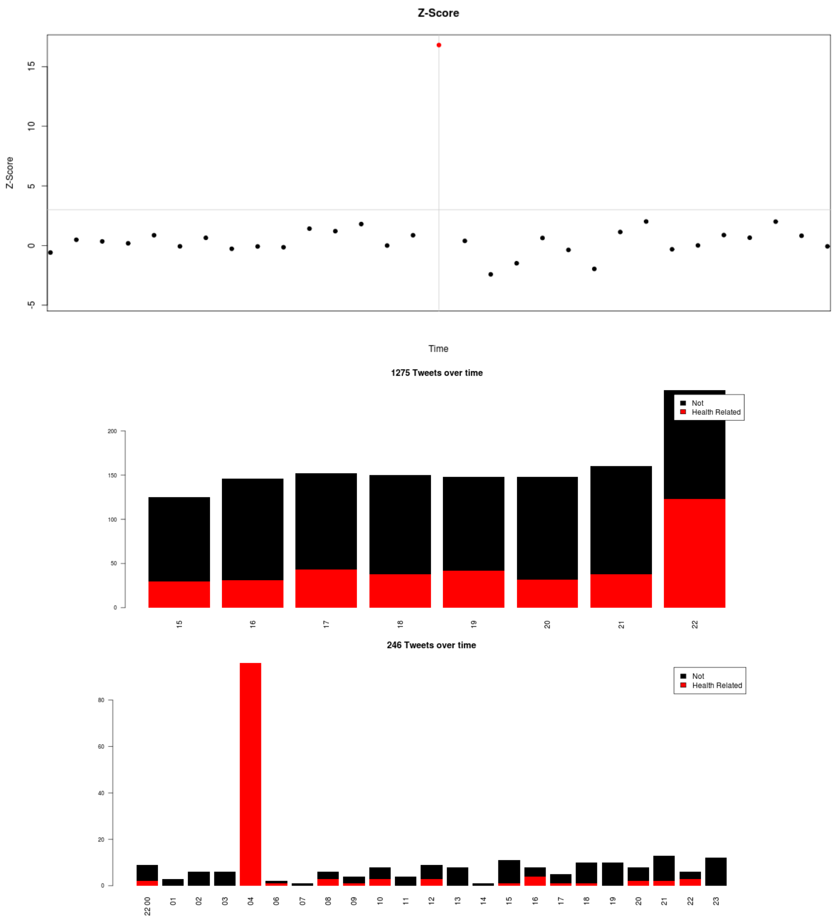

Figure 6: An event with a high z-score, on December 22, 2013, in Cobb County, GA. The top plot shows the z-score with a dotted line indicating the 3-standard deviation threshold and the event in red, the middle plot the tweets per day for the event and the six days prior, and the bottom plot the tweets per hour on the day of the detected event. Red bars indicate tweets that have been identified by the classifier as being about personal illness, while black bars are tweets containing "health related" phrases but are considered by the classifier to not be about personal illness.

large number of individuals (either because their accounts have been hacked or because of some fad or flash mob event to which we ourselves are not privy). In any event, while the extreme z-score events would seem to be the ones we want to focus on, this example shows that in reality these may not be caused by an illness of particular interest at all, but rather by some other phenomenon.

5. Results

Figure 7 shows a legitimate detection, the Boston Marathon bombing. From the figure it is clear that this event is just barely above our significance threshold of 3 standard deviations from the mean, and it is clear that the tweets that match the health phrases are much more prevalent than those that are classified as "personal health". Note however that the detection starts at 3pm, which is just after the bombing. In all the plots we are only considering tweets that have a reported latitude and longitude in the county of Middlesex, MA.

Note that our procedure knows nothing about bombs and explosions and such terms

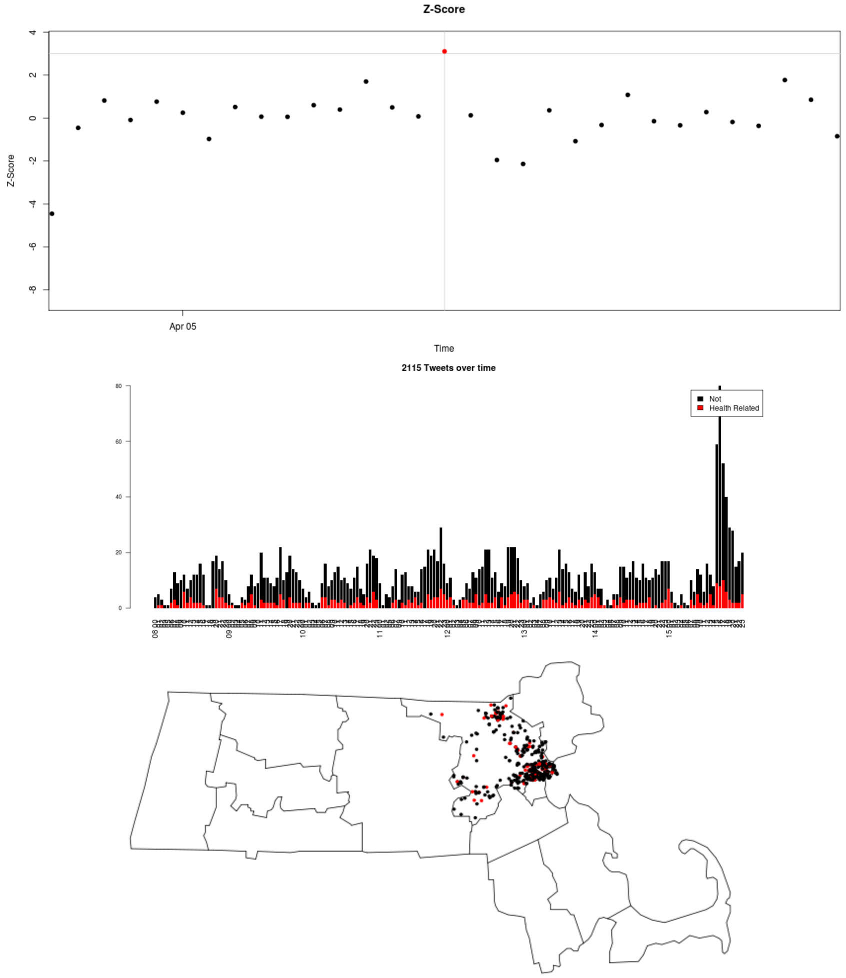

Figure 7: An event with a high z-score, on April 15, 2013, in Middlesex County, MA. The top plot shows the z-score, the middle plot shows the tweets per day and per hour for the event and the six days prior, and the bottom plot displays the tweets on a map showing their spatial distribution in the county.

that one would expect in an event like this. Perusal of the tweets that were classed as "personal health" found a number of tweets that were related to the marathon itself, and quite a few pain-related tweets, as well as a number of tweets containing the word "sick" in reference to the bombers and the bombing that were incorrectly classified as about personal health. So although this is a detection of a legitimate event, it is not of the type that we are primarily interested in (an outbreak of an infectious or communicable disease).

Figure 8 depicts a clustering of the data in the Boston Marathon bombing.[14] We have chosen 4 clusters to illustrate the methodology, and in this case the dendrogram strongly supports this choice. In the figure we depict the dendrogram, the "central" tweet of the cluster, and a word cloud indicating the top terms of the cluster. While the dendrogram indicates that the clusters are quite distinct, the word clouds fail to provide sufficient information to determine the "meaning" of the clusters. The red cluster has a large number of tweets containing "sick", and for the green cluster the word "hurt" is most frequent, but

[14] In this work we use hierarchical clustering on the document-term frequency matrix, although we believe that there are superior clustering methodologies for these data, and this is an area we will be exploring in the future.

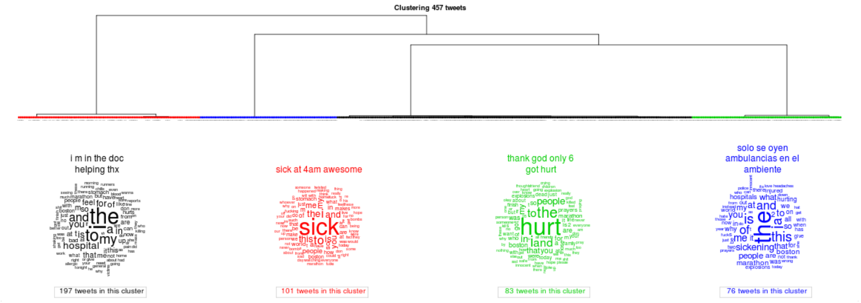

Figure 8: Clusters for the Boston Marathon bombing Twitter data.

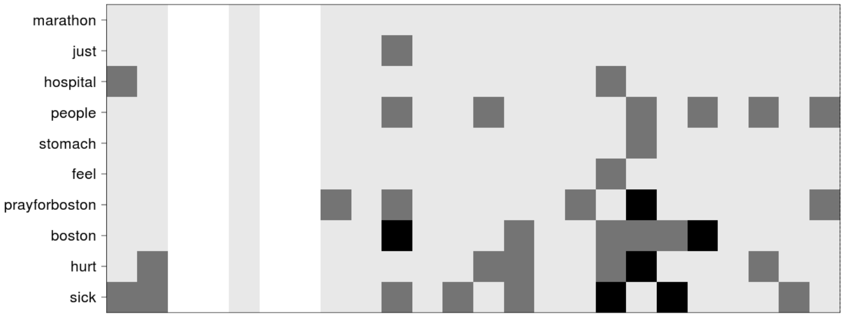

Figure 9: Top 10 words in time for the Boston bombing. Each of the 24 columns in this plot is an hour, with the gray-scale indicating the number of tweets containing the word in that hour. The plot is read from bottom to top, with the most frequent words on the bottom.

the other two clusters are less interpretable. Note that we have not removed stop words in this. We often don't, because these can be useful: for example, the words "i", "me", "my", etc. can be indicative of people talking about personal illness or personal opinions about an event. For this reason, we typically investigate several different views of the data, with different choices as to what set of words are considered in the analysis. This is best done using an interactive tool that we are currently developing, but the details of this tool are beyond the scope of this paper.

Figure 9 depicts the top words in time. The first near-black box for the word "sick" occurs at 3pm, just after the detonations. Note the words "marathon", "boston" and "prayforboston" (a popular hashtag after the bombing). It is easy to determine the cause of the event from these analyses of the data, even though the event is one for which our methodology is not specifically designed.

We also have a number of detections on January 1, 2014. These are complaints of headaches, stomach aches, hangovers, and other things likely to be related to festivities the night before. Figure 10 shows a typical event of this type. Note the slight increase in health-related tweets on the day of the event, with the bulk in the morning starting around 10AM (9AM local time).

This pattern is typical for many of the events we detect: an increased number of individuals complaining about headaches, stomach aches, and generally feeling sick. Note that the clusters in Figure 10 show these groups quite clearly. The top "health related" terms plot in the bottom of Figure 10 shows pretty clearly the characteristic of the event: the most common words are "headache", "hurt", "sick", "stomach" and "sore", and these persist as high frequency words through most of the day. Note also that "throw up" and "throwing up" are relatively high frequency words in the early hours of the morning.

Aside from the Boston Marathon bombing, we have been unable to unambiguously tie the detections to a specific outbreak or disease. This is in part because there have been few such outbreaks, although the measles outbreak discussed above is one such, and there have been several norovirus and food poisoning events. The fact that we can't (unambiguously) say that we detect these events is due in part to the fact that we are only collecting a fraction of the tweets, in part to the fact that some of these events are spread in time and over many counties, and in part to the limitations of Twitter – we can only observe things that people tweet about.

We do detect many events – up to 300 in 2013 (depending on parameters such as detection threshold, and the minimum county size that we consider), and most of them are very much like that depicted in Figure 10. We also see a number of events during the beginning of school, that appear to be the typical school related illnesses. Note that in this work we are not explicitly tracking the flu, for several reasons, the main being that it tends not to be localized geographically. However, as indicated in the discussion above on tracking named diseases and symptoms, the methods we use could easily be applied to investigate early detection and tracking of the flu.

6. Conclusions and Future Work

There are several ways that Twitter or other social media could be used for biosurveillance, and we have discussed three here:

- Tracking a known disease such as measles or ebola.
- Looking for an excess of certain symptoms such as fever or stomach ache.
- Looking for an excess of individuals complaining of being sick.

Our work has focused primarily on the last. This is the hardest of the three, but the methods we have investigated to solve this problem are applicable to the others as well. In particular, the approach of **filter** $\rightarrow$ **classify** $\rightarrow$ **detect**, wherein we first consider only those tweets matching certain "health related" terms, then classify the tweets as being about "personal health" or not, and finally detect anomalies, is a very powerful method for processing these data.

Our future work will focus on assessing the cause of an event. We need automated methods to provide the analyst with an assessment of what the event is about, and provide her with the information necessary to determine whether the event is of interest or requires further investigation. We have illustrated some methods for this, such as clustering the tweets, using word frequencies and word clouds to show the most common terms. We have also investigated topic models (Blei et al. [2003]), with limited success.

One reason for choosing to look at 4 clusters of the data is that most events tend to have a "headache" cluster, a "stomach ache" cluster, and a generic "sick" cluster. This leaves a fourth cluster to account for anything specific to the event. This points to one area of future research. Once we have determined which tweets are health related, instead of using a purely unsupervised clustering method as is done in this paper, we can investigate a partially specified clustering scheme: group the tweets into clusters based on whether the tweet is mostly about a headache, a stomach ache, is general unspecified illness, or other. This allows us to characterize whether the event is a result primarily of an increase in one or more of the three common complaints, and separates out tweets that might contain information specific to the event and how it differs from common complaints.

Alternatively, if one were more focused on symptoms and looking for serious disease outbreaks, one could filter out the "generic sick" tweets altogether prior to generating the counts and detecting the event. This would have the advantage of only detecting events that are likely to be serious.

We have investigated aggregating symptoms into higher level categories such as *respiratory*, *gastrointestinal*, etc., and this might offer another approach similar to the clustering approach discussed above. In our investigations of this approach, we have discovered that events are almost always swamped by the *nonspecific illness* category (which we term *malaise*), and so further work is needed to determine the best way to proceed with this type of analysis.

References

Ark: Twitter nlp and part-of-speech tagging. URL `www.ark.cs.cmu.edu/TweetNLP/`.

Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, and Benyuan Liu. Predicting flu trends using twitter data. In *The First International Workshop on Cyber-Physical Networking Systems*, pages 702–707, 2011.

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1576, 2011.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Nigel Collier, Nguyen Truong Son, and Ngoc Mai Nguyen. OMG u got flu? analysis of shared health messages for bio-surveillance. *Journal of Biomedical Semantics*, 2:(Suppl 5):S9, 2011. URL `http://www.jbiomedsem.com/content/2/S5/S9`.

Courtney D. Corley, Diane J. Cook, Armin R. Mikler, and Karan P. Singh. Text and structural data mining of influenza mentions in web and social media. *Int. J. Environ. Res. Public Health*, 7, 2010. doi: 10.3390/ijerph7020596.

Mark Dredze. How social media will change public health. *IEEE Intelligent Systems*, 27 (4):81–84, July/August 2012. doi: 10.1109/MIS.2012.76.

Vasileios Lampos, Tijl De Bie, and Nello Cristianini. Flu detector – tracking epidemics on twitter. *Machine Learning and Knoledge Discovery in Databases*, 6323:599–602, 2010.

Kalev H. Leetaru, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, and Eric Shook. Mapping the global twitter heartbeat: The geography of twitter. *First Monday*, 18(5), May 2013. doi: doi:10.5210/fm.v18i5.4366.

Lauren Silverman. Texas megachurch at center of measles outbreak. *National Public Radio*, 2013. URL `http://www.npr.org/2013/09/01/217746942/texas-megachurch-at-center-of-measles-outbreak`.

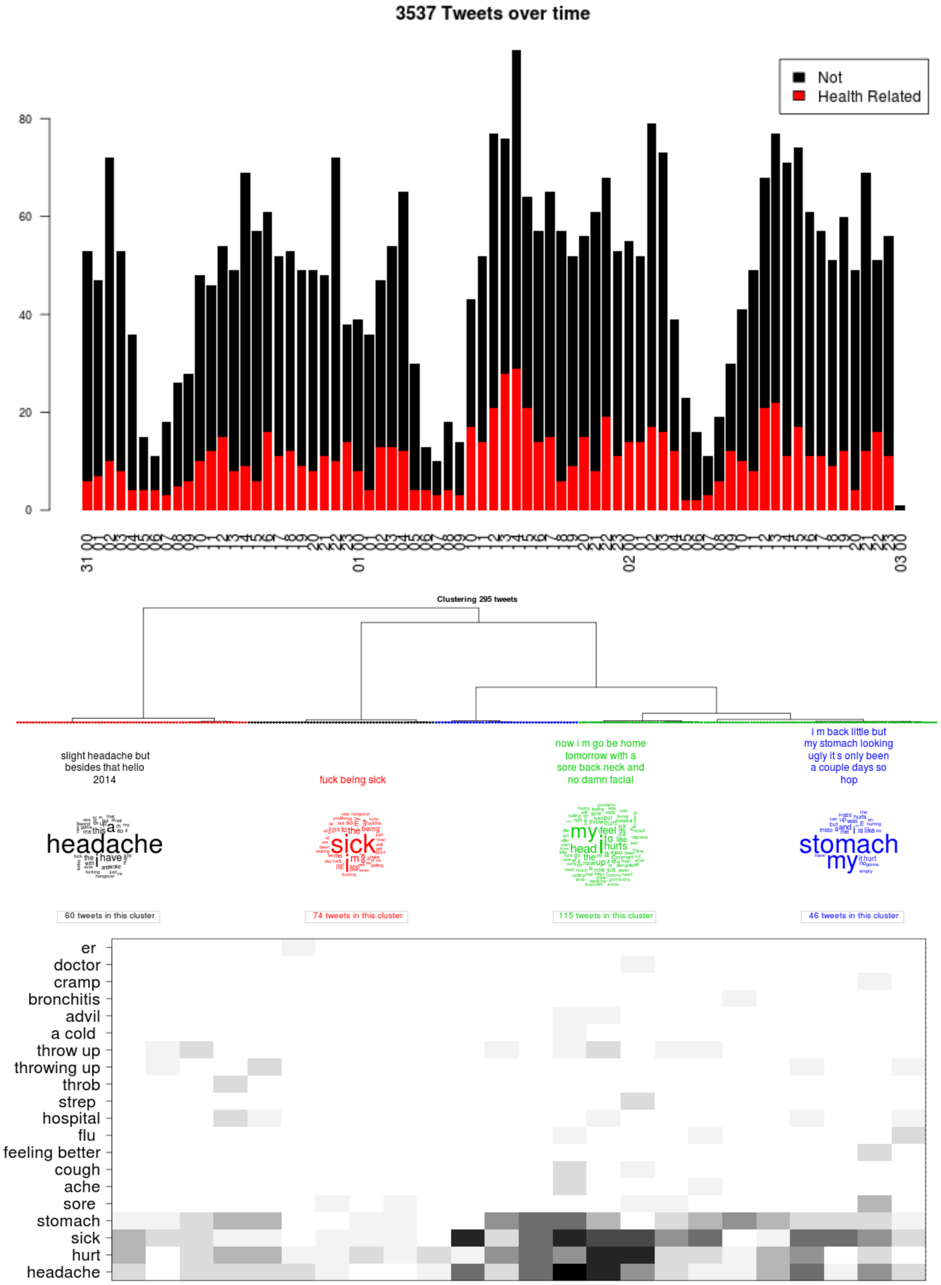

Figure 10: An event with a high z-score on January 1, 2014, in Cook County, IL (Chicago). This top plot shows the number of tweets from the day prior to the event through the day after the event. The middle plot shows the clustering for the day of the event. The bottom plot shows the top “health related” terms for the day of the event (most frequent terms on the bottom of the plot).