Combining Twitter and Media Reports on Public Health Events in MedISys

Erik van der Goot
Institute for the Protection and Security of the Citizen
Joint Research Centre of the European Commission
21027 Ispra (VA), Italy
erik.van-der-goot@jrc.ec.europa.eu

Hristo Tanev
Institute for the Protection and Security of the Citizen
Joint Research Centre of the European Commission
21027 Ispra (VA), Italy
hristo.tanev@jrc.ec.europa.eu

Jens P. Linge
Institute for the Protection and Security of the Citizen
Joint Research Centre of the European Commission
21027 Ispra (VA), Italy
jens.linge@jrc.ec.europa.eu

ABSTRACT

We describe the harvesting and subsequent analysis of tweets that are linked to media reports on public health events in order to identify which Internet resources are being referred to in these tweets. The aim was to automatically detect resources that are traditionally not considered mainstream media, but play a role in the discussion of public health events on the Internet. Interestingly, our initial evaluation of the results showed that most references related to public health events lead to traditional news media sites, even though URLs to non-traditional media receive a higher rank. We will briefly describe the Medical Information System (MedISys) and the methodology used to obtain and analyse tweets.

Categories and Subject Descriptors

I.2.7 [Natural Language Processing], H.4.m [Miscellaneous]

Keywords

Epidemic Intelligence; Event-based Surveillance; Twitter Analysis; MedISys

1. INTRODUCTION

With the use of social media, and specifically Twitter, publishing on the Internet is no longer the exclusive domain of traditional media. Although this trend is not new (blogs and forums have been around for a long time), the use of Twitter has changed the Internet publishing landscape dramatically. The two main reasons are the relative ease with which this microblogging can be performed (almost any modern mobile phone will do) and the open access to tweets (Twitter messages of up to 140 characters) through a number of application programming interfaces (APIs). Although this form of information dissemination is increasingly being used by traditional media as well (creating significant overlap between social media and traditional media), the information space defined by Twitter contains a large volume of information generated by the general public.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

2. MedISys

MedISys (http://MedISys.newsbrief.eu) is an open source information monitoring and analysis system for public health events [1]. The system is based on the Europe Media Monitoring engine (EMM, http://emm.newsbrief.eu), but has been customised to monitor health related sources (e.g. official government websites, health news) in addition to general media, and to perform categorisation for a wide range of public health threats. The system has been fully operational since 2005 and currently harvests and analyses around 175.000 new articles per day in more than 40 languages. The system performs a variety of analysis tasks, including geo-location, entity recognition, quote extraction, categorisation, sentiment/tonality analysis and topic-based clustering. The system furthermore performs semantic analysis in order to extract event metadata for public health related events. Not all analysis is performed in all languages, although many of the categories are defined for all EU languages plus Russian, Turkish, Arabic and Chinese. The system maintains statistics for all defined categories and co-occurrence statistics for categories and countries. By applying normalisation and statistics on a sliding time window the system can detect sudden (meaningful) increase in reporting on any category for any country and thus functions as a breaking news detection system. This functionality is used for the early detection of reporting on disease outbreaks for event-based surveillance [2].

3. MINING TWEETS

Using the MedISys processing chain, we identify the main stories on public health events in the media. For each of these events, we identify the geo-location and threat name. There are two ways to identify relevant tweets: First, we build a term vector from each story, where terms are scored using log-likelihood weighting. Then, we search for tweets which have similar lexical content. To do so, we calculate the vector similarity between the story and each tweet obtained from the search. Another method is to generate a Twitter geo-search in order to obtain tweets from the place of the event that mention the public health threat. In our experiments, we used a radius of 300 km around the place of the event. We currently use both the Twitter search API and the Twitter streaming API to collect tweets [3]. The collected tweets are then given a score based on several parameters, e.g. number of retweets and number of answers. In addition, we promote tweets that express opinions according to a subjectivity filter. In this manner, we determine the initial relevance. From these tweets, we then collect uniform resource locator (URLs) and user profiles.

Each detected URL receives a score which considers the number of times the URL is mentioned in the relevant tweets, the occurrence of the URL in conversation threads, the type of site to which the URL points, etc. User profiles also receive a score, which is calculated by considering the number of relevant tweets the user wrote.

4. TWITTER ANALYSIS

The aim of the analysis is to establish what users refer to when they tweet about a certain topic. In this context, the most important part of the tweet is the mentioned URL. Most of the URLs are in a short form, made possible by URL shorteners (e.g. bit.ly, goo.gl) that provide a simple redirect to the original location of the resource on the Internet. Since it is impossible to know whether two different short URLs point to the same resource, short URLs need to be followed (resolved) until the original resource is located in order to perform the analysis. We attempt to exclude non-relevant references, e.g. to gambling websites etc. This quality check is important and works well; however, some unwanted references are still present.

Once the URLs are collected, they are scored according to the following formula:

Score(URL) = URLCoefficient (Mentions + 1.3 Retweeted + 4 Favourited).(InConversations + 2 StartedConversations + 1)

where:

URLCoefficient reflects the relevance of the target site or the content to which the URL refers. URL to photo-sharing services, blogs, social media such as Facebook, are scored higher. In this way we give preference to user-generated content and media rather than mainstream news, which we already monitor through the MedISys system. We consider the domain and the file extension, when defining the value of this coefficient. *URLCoefficient* is 4, when the URL refers to a Youtube video, points to a Facebook page or a page which is hosted in any of the well-known blog or media-sharing sites. The *URLCoefficient* is also 4 when it points to an image file (recognizable by the ending - '.jpeg', '.jpg' or '.png'). *URLCoefficient* is 2.5, when it contains one of the following strings: 'photo', 'gallery', 'picture', 'video' or 'forum'.

Mentions is the number of retrieved tweets which contain the URL.

Retweeted is the number of times the URL was retweeted by other users. Retweeting reflects the dynamic of the spread of information inside Twitter, which makes this parameter important.

Favourited is the number of users which chose a tweet with this URL as favourite. Message (tweet) can be marked as favourite by a Twitter user. This means that the tweet was considered interesting by this user;

InConversations is the number of times in which the URL was used in a conversation thread, but not in a tweet, starting the thread. We consider conversation threads to be important criterion for scoring URLs, since a conversation expresses more interest than a simple retweet or marking a tweet as favourite;

StartedConversation is the number of conversation threads, which were started by a tweet, containing this URL.

Our monitoring system searches (using Twitter Search API) or reads tweets in real time (using the Twitter Stream API) with specific keywords.

We also provide a score for Twitter users:

- Users who report more relevant URL are scored higher;
- Users who publish tweets loosely similar to the news cluster get a higher score; and
- Users who publish tweets which contain more subjective language or are likely to contain reports from the field are scored higher.

The use of subjective language is established by using a simple subjectivity classifier, based on the Support Vector Machine learning model. As features we use phrases which people typically use when expressing personal experiences and opinions - "I", "me", "my", "I saw", "I heard", etc.

We experimented with location-restricted Twitter search. Twitter allows to detect tweets which come from a specific area defined as a circle with given centre and a radius or as a set of rectangular boxes. In detecting the place of the tweets, Twitter Search and Streaming API use the precise GPS co-ordinates of the tweet s (provided for only about 2% of all the tweets [4]) as well as the location, declared in the user profile, if more precise information cannot be found. In this way, one can detect tweets about specific symptoms and diseases, which come from a predefined area. In this way the spread of a disease can be tracked on the map.

In order to visualize the location-specific information, we generated for each area, for which we performed location-specific search, a KML file in which individual tweets are provided with geographical co-ordinates. Then, the KML file is loaded in Google Earth, which shows the tweets on the map of the area, in which we are interested. This visualization proved to be a simple, yet efficient method for tracking health-related topics.

5. RESULTS & DISCUSSION

The Twitter space is obviously very large and noisy. The semantic analysis of tweet content can be difficult due to non-traditional language, e.g. slang, abbreviations, particular use of Twitter vocabulary and the use of emoticons.

As an example, we cite here the coronavirus outbreak in the UK in February 2013 [5]. Using patterns related to "coronavirus" and "SARS", we identified several thousand individual tweets geolocated in the UK. As top Twitter user, we identified BBCscience; the top Twitter links were hosted on:

www.manchestereveningnews.co.uk
www.bbc.co.uk
www.mirror.co.uk
www.guardian.co.uk
www.reuters.com
www.krtpro.org
www.themarketingblog.co.uk
www.facebook.com
instagram.com

Examples of individual geo-located tweets were:

> "SARS In Manchester. No one better cough on me. It's infectious when you're symptomatic! Not before. #lovely virus",
>
> "Everyone in the office is sick! Save me from the germs! Tempted by a SARS/Michael Jackson facemask", and
>
> "Getting bored of the horse meat news stories now, can we bring back SARS or something? Sars was cool.."

As expected, the individual tweets themselves were of limited value for tracking the disease (too many tweets, poor signal to noise ratio). However, the top Twitter user and top Twitter links provided very useful information on the outbreak. Although media reports from local, regional and national newspapers had already been captured by MedISys, the top Twitter links identified additional sources both in traditional media (which use Twitter as a distribution channel) as well as on blogs and forums. In addition, links to video- and photo-sharing sites such as Youtube, Instagram, and TwitPic were often referred to, thereby pointing to content published by the general public. Automatic qualitative analysis of these resources is not in the scope of the current work; the results are merely used to present an analyst with an indication of the social complement of the news.

In our approach, subjectivity classifiers promote user-generated content, thereby favouring tweets from individual users. This does not mean that references to traditional media are excluded, but we rank references to non-traditional media higher. When looking at topics that are widely discussed in the public in a more controversial manner, e.g. vaccination (MMR vaccine, HPV vaccine), this is more evident due to a high volume of tweets with strong opinions (Alexandra Balahur-Dobrescu, personal communication, 1 March 2013).

6. OUTLOOK

The method for Twitter analysis has been operational in the EMM engine for some time, but has only recently been deployed to MedISys.

The initial evaluation of the results leads to the interesting conclusion that most references related to public health events are to traditional news media sites, despite the fact that our method gives URLs to non-traditional media a higher rank.

For the strategy described above, the actual content of the tweet is not of great relevance, other than that the tweet is about the story that is analysed. However, having collected the tweets, additional analysis can be performed and we are currently developing a sentiment analysis model for tweet content. Another criterion for the ranking of tweets can be the geo-location information; more research is necessary to identify how to make best use of this information in the analysis process. Furthermore, we are also refining the automatically generated queries and optimising the use of the streaming API as part of the operational deployment of this system.

Initially, we focused primarily on English language tweets, although we apply the same methodology to other languages as well, notably French, Spanish and Portuguese. However, more work will be necessary to extend language coverage.

The Twitter analysis application runs 24/7 and has been tested in-house. It will be made publicly available soon to allow analysts to assess its usefulness in daily operations. The additional information extracted from Twitter will be automatically presented alongside the traditional analysis on the public MedISys website. Analysts will thus benefit from news items from traditional media sites (either monitored by MedISys directly or identified by MedISys from Twitter data) and from references to blogs, forums and video- and photo-sharing sites.

7. REFERENCES

[1] Linge J., Belyaeva, J., Steinberger, R., Gemo, M., Fuart, F., Al-Khudhairy, D., Bucci, S., Yangarber, R. and van der Goot, E. (2010). MedISys: Medical Information System. In: Eleana Asimakopoulou & Nik Bessis (eds). Advanced ICTs for Disaster Management and Threat Detection: Collaborative and Distributed Frameworks, pp. 131-142. IGI Global.

[2] Linge, J.P., Mantero, J. Fuart, F., Belyaeva, J., Atkinson, M., van der Goot, E. Tracking Media Reports on the Shiga toxin-producing Escherichia coli O104:H4 outbreak in Germany. eHealth conference, Malaga. P. Kostkova, M. Szomszor, and D. Fowler (Eds.): eHealth 2011, LNICST 91, pp. 178–185, 2012.

[3] Tanev, H., Ehrmann, M., Piskorski, J., and Zavarella, V.Enhancing Event Descriptions through Twitter Mining, Sixth International AAAI Conference on Weblogs and Social Media, 2012, Dublin

[4] Burton,S., Tanner,K., Giraud-Carrier,C., West,J., Barnes,M. "Right Time, Right Place" Health Communication on Twitter: Value and Accuracy of Location Information, in Journal of Medical Internet Research 14(6), 2012

[5] ECDC. Rapid Risk Assessment Severe respiratory disease associated with a novel coronavirus, 19 February 2013, retrieved from http://ecdc.europa.eu on 1 March 2013.