Early Warning and Outbreak Detection Using Social Networking Websites: The Potential of Twitter

Ed de Quincey and Patty Kostkova

City eHealth Research Centre, City University
Northampton Square, London EC1V 0HB
ed.de.quincey@city.ac.uk

Abstract. Epidemic Intelligence is being used to gather information about potential diseases outbreaks from both formal and increasingly informal sources. A potential addition to these informal sources are social networking sites such as Facebook and Twitter. In this paper we describe a method for extracting messages, called "tweets" from the Twitter website and the results of a pilot study which collected over 135,000 tweets in a week during the current Swine Flu pandemic.

Keywords: Epidemic Intelligence, social networking, swine flu.

1 Introduction

Epidemic Intelligence (EI) is being used by public health authorities to gather information regarding disease activity, early warning and infectious disease outbreak [1, 2, 3, 4]. EI systems systematically gather official reports and rumours of suspected outbreaks from a wide range of formal and increasingly, informal sources[1] [5]. Tools such as the Global Public Health Intelligence Network (GPHIN) and Medisys gather data from global media sources such as news wires and web sites to identify information about disease outbreaks [5, 6].

A potential improvement to these systems has been demonstrated by Google's Flu Trends research that has estimated flu activity via aggregating live online search queries for keywords relating to flu [7]. The drawback however is that the information stored in commercial search query logs, which could be integrated into EI systems is not freely available.

The increase in user-generated content on the web via social networking services such as Facebook and Twitter, however provides EI systems with a highly accessible source of real-time online activity. Twitter [8], a micro-blogging service that allows people to post and read other users' 140 character messages, called "tweets", currently has over 15 million unique users per month [9]. Twitter allow third parties to search user messages and return the text along with information about the poster, such as their location, in a format that can be easily stored and analysed.

[1] Informal sources account for more than 60% of the initial outbreak reports [5].

P. Kostkova (Ed.): eHealth 2009, LNICST 27, pp. 21–24, 2010.

In this paper we detail the information that is accessible via services such as Twitter, a process that can be used to access it and present the results of a pilot study into identifying trends of flu activity in May 2009, present in messages sent via Twitter.

2 Methodology

Twitter allows access to users' tweets via Application Programming Interfaces (APIs): a REST API and the Search API. The REST API method allows developers to "access core Twitter data" [10] such as user profile information, ability to post tweets etc.. The Search API, which is utilised in this paper, allows developers to query tweets in real-time using any combination of keywords. This is via making a request to a url in the following format:

```
http://search.twitter.com/search?q=keyword
```

A number of other parameters can also be passed via the querystring such as number of results to return e.g. `rpp=100`. The matching tweets, containing the text of the tweet, user information and a timestamp, are returned in either atom (an xml format) or json (a computer data interchange format). This data can then be parsed programmatically using PHP, Ruby, C etc..

2.1 Use of Twitter in This Study

For this preliminary study, the Search API was utilised to return the last one hundred tweets that contained instances of the word "flu". PHP code was then written to parse the returned tweets (in atom format) and save them to a MYSQL database, comprised of one table. Records collected comprised of the following fields:

```
id, published, link, title, content, author, terms
```

A batch file was created that ran the PHP code every minute with new tweets being saved[2] in the database. The program was started at 14:00 on Thursday 7th May 2009 and has been running continuously since then. The results presented in this paper are taken from the following week, i.e. until 14:00 on Thursday 14th May 2009.

3 Preliminary Results

During the week, there were a total of 135,438 tweets, posted by 70,756 unique users that contained the word "flu" with the following table showing their daily distribution.

The lowest number of tweets was recorded on Sunday the 10th of May (discounting the 14th as only 14 hours of tweets was collected) and the highest on Friday the 8th of May.

[2] It was found that this rate was sufficient as it was unlikely that there were more than one hundred new tweets in a minute.

Table 1. Number of tweets containing the word "flu

Date	Number of Tweets
Thursday 7th May 2009[3]	16,422
Friday 8th May 2009	24,692
Saturday 9th May 2009	18,484
Sunday 10th May 2009	15,213
Monday 11th May 2009	19,140
Tuesday 12th May 2009	19,353
Wednesday 13th May 2009	14,370
Thursday 14th May 2009[3]	7,764
Total	135,438

The use of the word "flu" however varied greatly in the tweets with users utilising the term to refer to themselves, a friend, a news story, a link etc.. Further analysis into the actual meaning of the tweets is currently being planned but to identify any immediate trends, the content of all the tweets was analysed using concordance software. The following table shows a selection of the top words present in all of the tweets (common words such as "a", "the", "to" etc. have been removed).

Table 2. Most popular words found in all tweets

Word	Frequency	Word	Frequency
Flu	138,260	New	7,668
Swine	99,179	News	6,498
Have	13,534	Confirmed	6,456
Cases	13,300	Just	6,373
H1N1	9,134	People	5820
Has	8,010	Case	5647

In the majority of tweets the word "swine" was present along with "flu" (which would perhaps be expected with the current swine flu pandemic). Although the word "have" is considered to be a common word in the English language (24th most common [11]), it has been included in this list because it might be an indication of people tweeting that they "have flu"[4]. For a similar reason the use of the word "has" may indicate that the tweet contains information about someone else having flu e.g. "he has flu". The words "confirmed" and "case(s)" perhaps indicate a number of tweets that are publicising "confirmed cases of swine flu". Further investigation into this is being conducted using collocation analysis, a sample of which is shown in the following table (again excluding common words).

[3] Recording of tweets began at 14:00 on 7/5/2009 and stopped at 14:00 14/5/2009.
[4] Interestingly the phrase "have flu" was found only 137 times.

Table 3. Collocation of words, one word to the right and the left of the word "flu"

1 word to the left		1 word to the right	
Word	**Frequency**	**Word**	**Frequency**
Swine	96,651	Http	6,598
The	5,701	Cases	6,194
H1n1	5,225	Case	2,210
Bird	1,425	Death	2,001
New	1,304	Virus	1,411
Pig	1,164	Outbreak	1,321
Man	720	H1n1	1,147
Stomach	510	Spreads	927
Regular	426	Lol	924
Flu-bird	319	Deaths	912

4 Conclusion

The results described in the previous section highlight the potential for twitter to be used in conjunction with pre-existing EI tools. Although a potential explanation of the number of tweets collected could be due to the current swine flu pandemic, the amount of real-time information present on twitter, either with regards to users reporting their own illness, the illness of others or reporting confirmed cases from the media, is both rich and highly accessible. Further work is planned into the data already collected and the system is continually retrieving and storing tweets to be analysed in relation to users' geographical location and the semantic syntax of tweets.

References

1. Linge, J.P., Steinberger, R., Weber, T.P., Yangarber, R., van der Goot, E., Al Khudhairy, D.H., Stilianakis, N.: Internet surveillance systems for early alerting of health threats. Euro Surveill. 14(13), pii=19162 (2009)
2. Kaiser, R., Coulombier, D.: Different approaches to gathering epidemic intelligence in Europe. Euro Surveill. 11(17), pii=2948 (2006)
3. Paquet, C., Coulombier, D., Kaiser, R., Ciotti, M.: Epidemic intelligence: a new framework for strengthening disease surveillance in Europe. Euro Surveill. 11(12), pii=665 (2006)
4. Coulombier, D., Pinto, A., Valenciano, M.: Epidemiological surveillance during humanitarian emergencies. Médecine tropicale: revue du Corps de santé colonial 62(4), 391–395 (2002)
5. WHO, http://www.who.int/csr/alertresponse/epidemicintelligence/en/index.html
6. Linge, J.P., Steinberger, R., Weber, T.P., Yangarber, R., van der Goot, E., Al Khudhairy, D.H., Stilianakis, N.: Internet surveillance systems for early alerting of health threats. Euro Surveill. 14(13), pii=1916 (2009)
7. Google Flu Trends, http://www.google.org/flutrends/
8. Twitter, http://www.twitter.com
9. http://www.crunchbase.com/company/twitter
10. Williams, D.: API Overview, http://apiwiki.twitter.com/API-Overview
11. http://www.world-english.org/english500.htm