#swineflu: Twitter predicts swine flu outbreak in 2009

Martin Szomszor[1], Patty Kostkova[1], Ed de Quincey[2]

[1]City eHealth Research Centre, City University, London, UK
[2]School of Computing and Mathematics, University of Greenwich, UK
{martin.szomszor.1@city.ac.uk, patty@soi.city.ac.uk}

Abstract. Early warning systems for the identification and tracking of infections disease outbreaks have become an important tool in the field of epidemiology. While government lead initiatives to increase the sharing of surveillance data have improved early detection and control, along with advanced web monitoring and analytics services, the recent swine flu outbreak of 2009 demonstrated the important role social media has and the wealth of data it exposes. In this paper, we present an investigation into Twitter, using around 3 Million tweets gathered between May and December 2009, as a possible source of surveillance data and its feasibility to serve as an early warning system. By performing simple filtering and normalization, we demonstrate that Twitter can serve as a self-reporting tool, and hence, provide indications of increased infection spreading. Our initial findings indicate that Twitter can detect such events up to one week before conventional GP reported surveillance data.

Keywords: Epidemic Intelligence, Twitter, H1N1, Pandemic Flu

1 Introduction

Social media, such as blogging, social networking, Wikis, etc., has attracted much interest recently as a possible source of data for epidemic intelligence (EI). The real-time nature of micro-blogging and status updating presents unique opportunities to gather information on large numbers of individuals and offers the opportunity to enhance early warning outbreak detection systems. During the 2009 swine flu outbreak, Twitter (a popular micro-blogging website) received a substantially increased amount of traffic related to swine flu, with many individuals reporting that they contracted the virus. While traditional EI systems, such as GPHIN and Medisys are well established and used routinely by the European Centre for Disease Control (ECDC) and the World Health Organisation (WHO), new sources of data are constantly under review. Recent work [1] by companies such as Google has demonstrated that online search queries for keywords relating to flu and its symptoms can serve as a proxy for the number of individuals who are sick. However, such search data remains proprietary and therefore not useful for research or the construction of non-commercial applications. However, Twitter data is publicly available and offers a highly accessible view into people's online and offline real time activity. In this paper, we present our analysis of Twitter data from May until December 2009, and demonstrate its potential as a data source for early warning systems.

2 Background

2.1 Epidemic Intelligence

Epidemic Intelligence (EI) is an automated early identification of health threats and disease outbreaks, their verification and risk assessment and investigation to inform health authorities about the required measure to protect the citizens [2, 3, 4]. This is of a particular concern in situations of mass gatherings (e.g. sport events such as World Cups and Olympics, festivals, etc.) and humanitarian emergencies [5]. European, national and regional level surveillance systems produce routine reports and can provide indications of potential risks and abnormal events. However, more dynamic data collection is needed to identify threats early enough to asses the risk and launch an appropriate response. This process has been strengthen by the International Health Regulations (IHR) [6], coordinated by WHO, and signed by all UN member states, requiring states to report incidents of infectious diseases to facilitate outbreak prevention at the source, rather than at borders. ECDC [7] in Europe has proposed an improved epidemic intelligence framework bringing together an indicator-based surveillance and event-based surveillance.

This new legal framework is further changing the reporting culture which in the past relied on health authorities in the countries and therefore was often subject to under-reporting due to fear from economical sanctions imposed by the EC and other states. However, in a world with the large-scale blogging, social networks and Web 2.0, an outbreak is often discovered sooner through EI tools than health authorities of a concerned country might even know through traditional reporting channels.

2.2 The Role of the Internet and Social Media in EI

Epidemic intelligence has been relying on automated news media searching systems for over a decade. Tools such as Global Public Health Intelligence Network (GPHIN), developed by Health Canada and in use by WHO and Medisys, developed by the JRC, gather news from global media to identify disease outbreaks threats using multi-lingual natural language processing and appropriately weighted set of keywords, categories and taxonomies [8, 9]. In addition, the email-based system ProMED-mail has been an informal source of upcoming emergencies [10].

However, with the ever increase user activity on the Internet and Web 2.0 and social networks, a valuable real-time source of data to assist this process has become available. Unlike Google's Flu Trends research that has estimated an upcoming flu epidemics sooner than CDC surveillance data evaluated online search queries for keywords relating to flu [11]. A similar study, on a smaller scale, was conducted by the NeLI/NRIC portal identifying user information needs during the swine flu pandemics in 2009 [12]. However, in order to use user searchers to assist EI system a global search portals receiving billions of queries a day to analyse a sufficient volume of data, however, a drawback for public health is that the information is stored in weblogs at commercial servers, which cannot be accessed and made available for EI systems. The increase in Web 2.0 and user-generated content via social networking tools such as Facebook and Twitter, however provides EI systems with a highly accessible source of real-time online activity. Facebook's privacy setting allow users to restrict their profile content and activity, however, Twitter [13], a micro-blogging

service that allows people to post and read other users' 140 character messages, called "tweets", is available in public domain and therefore freely searchable and analyzable using a provided API [14]. The information posted on twitter, currently used by over 15 million unique users per month [15], is describing a real time activity due to the social nature of the service, unlike search queries collected by search engines. Therefore, utilizing this increasingly popular freely available data source has a potential for EI and other rapid information intelligence systems.

In addition to using Twitter for outbreak detection, which is the aim of our study, Twitter has been successfully used to demonstrate it could track an earthquake or typhoon [16] and both Facebook and Twitter are becoming increasingly more popular for raising awareness and raising funds for global relief [17].

3. Twitter based Surveillance

Twitter, a micro-blogging service that allows people to send and receive messages otherwise known as *tweets*. Tweets are limited to 140 characters and are displayed on the user's profile page. Individuals have their own personalised feed, displaying the recent tweets made by anyone they follow. Users are free to follow any other users and use this facility to build networks that support social, business and academic activities. Current usage estimates place the number of tweets made per day at 65 million.

3.1 Data Collection

We searched for the term 'flu' and collected over 3 million tweets in the period from May 7th until December 22nd 2009 and carry on collecting them on a 1 minute basis. We found just less than 3 million tweets containing the keyword "flu", including individuals reporting flu symptoms or self-diagnosing; sharing links to news articles, websites, and blogs; and generally commenting on the topic. The most popular words in these tweets and their frequencies are show in Table 1.

Freq	Word	Freq	Word
2,993,022	flu	92,999	#swineflu
1,6217,82	swine	88,801	cases
264,903	rt	82,130	#h1n1
223,876	h1n1	71,323	today
195,163	vaccine	69,071	shots
156,658	shot	66,167	hope
109,995	health	64,271	feel
107,675	sick	63,732	school
97,889	news	61,004	:(

Table 1 – Top 20 most frequently occurring words

3.2 Classification of Tweets

To investigate the use of Twitter as a mechanism for self-reporting of flu, we first classify the tweets using the following classes (it is possible for a tweet to be placed in more than one class):

1. **Tweets containing a Link** A popular activity in Twitter is to post a link to a website. Many use this mechanism to link their followers to online news articles, blogs, videos, images, etc. Because of the 140 character limit of tweets, and the typical long length of urls, url shortening services (such as bit.ly and tinyurl.com) are often used.
2. **Retweets** Another popular Twitter behaviour is to *Retweet* a message. In essence, users who see an interesting tweet will pass it onto their followers by reposting the original message and quoting the original author. Retweets themselves often contain links. We search for *"rt @<hyperlink to user>"* to find retweets.
3. **Self-Reporting Flu** We check the text of each tweet and search for phrases that indicate the user has the flu. These include the phrases *"have flu"*, *"have the flu"*, *"have swine flu"*, *"have the swine flu"* in present and past tenses.

Figure 1 contains a time-series plot for the total number of tweets recorded during the period 11-05-2009 until 20-12-2009. A 7-day moving window average is applied to smooth the data. The plot shows the total number of tweets containing the keyword flu (labelled "All Flu Tweets") for each day, the total number of tweets containing a link ("Contains Link"), the total number of tweets reporting flu ("Self Reporting via Twitter"), and the total number of retweets ("Retweets"). Due to technical problems, a section of data is missing for the period 30/08/2010 to 14/09/2010.

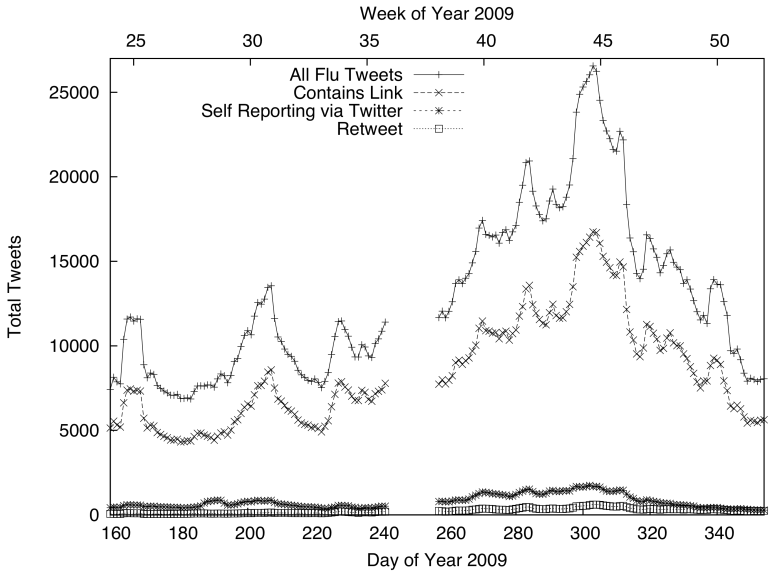

Fig. 1 - A time series showing all tweets containing the keyword flu, those containing links, those reporting flu, and retweets. A 7-day moving window average has been applied to smooth the data.

The time-series indicates significant increases in activity around week 30 (20/07/2010), and again around week 40 (28/08/2010). Posting of links constitutes the most significant percentage of tweets - around 67%, the number of self-reporting tweets is around 5%, and the number of retweets is approximately 2%.

3.3 Distribution of Links and Retweets

Since the posting of links makes up a significant proportion of flu related tweets, we decided to perform further analysis of these cases to identify any global trends. An increase in the posting of links could indicate an increased reaction to news and other online media. Figure 2 plots the percentage of tweets for each day that contain a link (using left axis), and the percentage of tweets that are retweets (right axis). The plot shows that the posting of links remains relatively constant over time (around 67%). The percentage of retweets displays an overall increase from approximately 0.75% in week 25, to around 3% in week 52. It is not clear from the data we have gathered whether this increase in retweeting is a trend specific to flu related tweets or a trend across the whole of twitter. The latter seems more likely since individuals have become more aware of the retweeting practice in Twitter since the beginning of 2009.

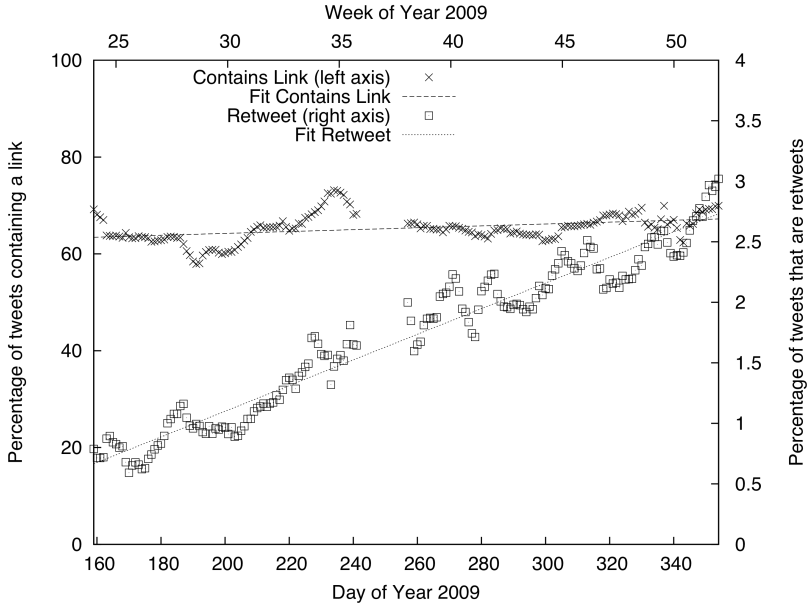

Fig. 2 - A plot showing the proportion of links each day that contain a link and the those that are retweets.

4 Experiment

4.1 Correlation with UK national surveillance data

To test the accuracy of Twitter as a mechanism for self-reporting flu, and hence it's potential to provide early warning detection, we collected official surveillance data from the UK Health Protection Agency (HPA) [18]. The HPA provide weekly reports on the RCGP influenza-like illness (ILI) consultation rate for England and Wales,

Scotland, and Northern Ireland. For comparison, we calculate the percentage of tweets that are self-reporting flu for each day in our investigation period. This normalization process means that global trends in Twitter activity (e.g. spam, increased retweeting, and increased posting of links) are not factored in. Instead, the data here shows the number of individuals self diagnosing as a percentage of all flu related Twitter activity. The plot shown in Figure 3 contains the HPA RCGP ILI consultation rate for England and Wales (square points, right axis), and the percentage of Twitter activity reporting flu (crossed points, left axis). First impressions reveal a strong correlation between the two data sources: a sharp peak in activity on twitter (around week 28, 6/07/2009) corresponds to the rapid increase in the number of consultations.

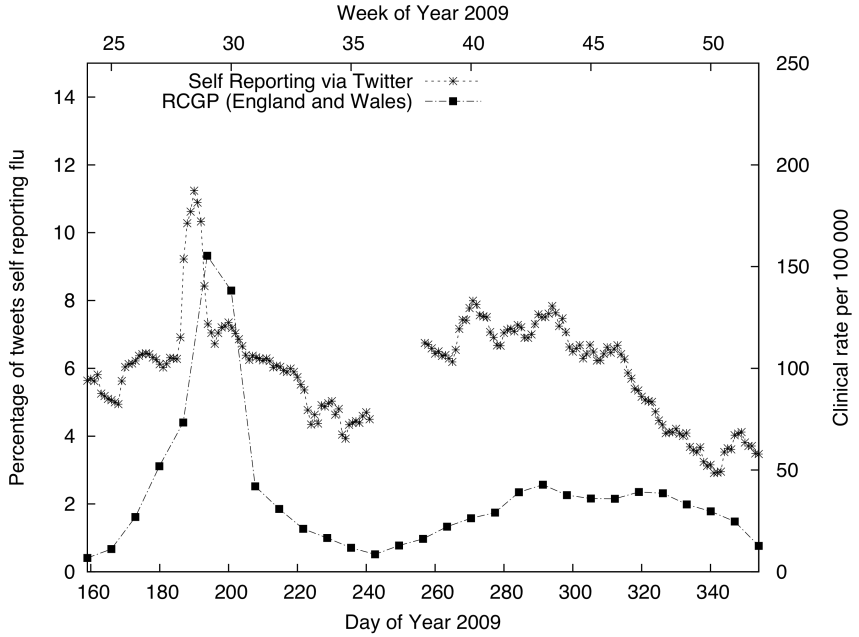

Fig. 3 - A plot showing the RCGP ILI rate for England vs the number of self-reported cases on Twitter.

4.1 Normalized Cross-Correlation: Twitter predicts

To provide some indication of the correlation between Twitter and the official UK surveillance data, we calculate the normalized cross-correlation ratio between various signals from Twitter and the official HPA surveillance data. Since the HPA data is gathered on a weekly basis, we perform the comparison using a weekly aggregation of Twitter data. Equation 1 gives the normalized cross-correlation function we use, where $x(t)$ is the total number of tweets during week t, and $y(t-i)$ is the number of reported cases according to the HPA during week $(t-i)$. We calculate r across all flu tweets, those that are self reporting, those that contain links, and those that are retweets for values of i between -4 and 4.

$$r = \frac{\sum_t (x(t) - \bar{x}) * (y(t-i) - \bar{y})}{\sqrt{\sum_t (x(t) - \bar{x})^2 * \sum_t (y(t-i) - \bar{y})^2}}$$

Eq. 1 - Normalised Cross-Correlation

Figure 4 displays the various values of r for weekly offsets between $i = -4$ and $i = 4$. The cross-correlation ratio (or sliding dot product) is a measure of how similar two signals against a moving time lag. This means that values of *r* for *i=0* represent how much two signals a correlated, when *i=-1*, it represents how much the first signal predicts the second signal. The higher the value of r, the stronger the correlation. Figure 4 shows that the self reporting tweets have a strong correlation with the HPA data – the signals for all flu tweets, those containing links, and retweets do not. This would indicate that our filtering and normalization process has been successful, allowing us to discriminate messages that indicate someone has the flu from the general noise on Twitter. Although the strongest correlation occurs at *i=0* (when *r=0.1140*), indicating a co-occurrence of tweets and surveillance data, there is still a strong correlation at *i=-1* (when *r=0.0750*) indicating that the HPA surveillance data could be predicted by Twitter up to 1 week in advance, and therefore demonstrates the potential of twitter for early warning and outbreak detection.

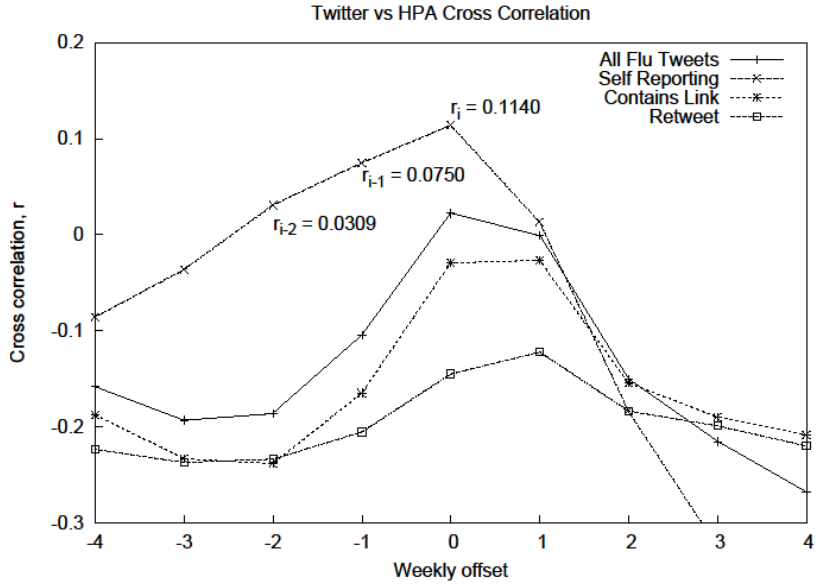

Fig 4. - The cross-correlation plot between Twitter and the HPA Surveillance Data

5 Conclusions and Future Work

In this paper, we have provided presented our analysis of Twitter data relating to the Pandemic Flu outbreak of 2009. We have shown that although Twitter contains quite a lot of noise in the form of spam, posting of links, and retweets, a simple filtering

method can be used to extract those tweets that indicate that a user has the flu. In the future, more advanced computational linguistics will be applied to identify individuals that are reporting flu-like symptoms, as well as directly reporting having the flu.

By comparing the data gather from Twitter to the official national surveillance data from the HPA, we have shown that Twitter could be used as an early warning detection system: Our initial findings indicate that HPA data could be predicted up to one week in advance. Clearly, the use of Twitter in an EI system would provide an even faster response since official data usually takes some time to collate and process.

A further piece of information that is vital to EI systems is that of location. Location awareness is becoming more popular in Twitter and is likely to become a core-part of the API in the future. This extra piece of information would provide even more motivation to exploit Twitter in EI systems.

6 References

1. http://www.google.org/flutrends/
2. R Kaiser, D Coulombier, M Maldari, D Morgan, C Paquet. What is epidemic intelligence, and how it is being improved in Europe? Eurosureillance 2006; 11(2): 060202
3. Kaiser R, Coulombier D.: Different approaches to gathering epidemic intelligence in Europe. Euro Surveillance. 2006;11(17):pii=2948 (2006)
4. Paquet C, Coulombier D, Kaiser R, Ciotti M. Epidemic intelligence: a new framework for strengthening disease surveillance in Europe. Euro Surveillance. 2006;11(12):pii=665 (2006)
5. Coulombier, D., Pinto, A., Valenciano, M.: Epidemiological surveillance during humanitarian emergencies. Médecine tropicale: revue du Corps de santé colonial 62(4):391-395. (2002)
6. http://www.who.int/ihr/en/
7. http://www.ecdc.europa.eu/en/Pages/home.aspx
8. WHO,http://www.who.int/csr/alertresponse/epidemicintelligence/en/index.html
9. Linge JP, Steinberger R, Weber TP, Yangarber R, van der Goot E, Al Khudhairy DH, Stilianakis NI. Internet surveillance systems for early alerting of health threats. Euro Surveill. 2009;14(13):pii=1916. (2009)
10. Madoff, L. C.,: ProMED-mail: An Early Warning System for Emerging Diseases. Clinical Infectious Diseases 39(2): 227 (2004)
11. Google Flu Trends, http://www.google.org/flutrends/
12. de Quicney, E., Kostkova, P., Wiseman, S.: An investigation into the potential of Web 2.0 websites to tracks disease outbreak. Poster at Infection 2009, Birmingham, UK. (2009)
13. Twitter, http://www.twitter.com
14. Williams, D.: API Overview, http://apiwiki.twitter.com/API-Overview
15. http://www.crunchbase.com/company/twitter
16. Sakaki, T., Okazaki, M., and Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: WWW '10: Proceedings of the 19th international conference on World wide web, pp. 851—860, Raleigh, North Carolina, USA, (2010)
17. http://www.fastcompany.com/blog/kit-eaton/technomix/facebook-twitter-turn-charity-efforts-11
18. http://www.hpa.org.uk/