Twitter Informatics: Tracking and Understanding Public Reaction during the 2009 Swine Flu Pandemic

Martin Szomszor
City eHealth Research Centre
City University
London, EC1V 0HB, UK
martin.szomszor.1@city.ac.uk

Patty Kostkova
City eHealth Research Centre
City University
London, EC1V 0HB, UK
patty@soi.city.ac.uk

Connie St Louis
Department of Journalism
City University
London, EC1V 0HB, UK
c.stlouis@city.ac.uk

***Abstract* — Much attention has been focused on Twitter because it serves as a central hub for the publishing, dissemination, and discovery of online media. This is true for both traditional news outlets and user generated content, both of which can vary widely in their journalistic and scientific quality. The recent Swine Flu pandemic of 2009 highlighted this aspect perfectly; global events that created a large online buzz, with some dubious medical facts leaking into public opinion. This paper presents an investigation into how online resources relating to Swine Flu were discussed on Twitter, with a focus on identifying and analyzing the popularity of trusted information sources (e.g. from quality news outlets and official health agencies). Our findings indicate that reputable sources are more popular than untrusted ones, but that information with poor scientific merit can still leak into to the network and potentially cause harm.**

Keywords-component; Twitter; Swine Flu, H1N1, Data Mining; News; Social Media; eHealth; Web Science

I. INTRODUCTION

The Internet has revolutionised the way information is published and consumed. In the past, news organisations and governments had control over the media that was published and how the population accessed it. With the relatively recent invention of the Web and increasingly popular social media websites, any individual can create and post material online, potentially connecting with the global population without editorial comment or moderation. While editors or a peer-review process moderates traditional news sources and scientific outlets, online publishing such as blogs and podcasting has enabled untrusted sources of information to be published and consumed in large quantities.

The explosion in popularity of social media has subsequently raised concerns about the quality of information that is present online and how it reaches various members of the population. Using popular search engines, it is easy to find a variety of different beliefs surrounding a particular topic, some of which would contradict the popular consensus. Consumers of information on the Web now have the arduous tasks of assessing the quality of the information they see without any specific training or guidance. As a result, many of the prominent technologies that enabled this social media revolution have come under attack, such as the popular microblogging site Twitter, or the widespread social networking site Facebook. Claims are often made that these forms of communication may promote the consumption of fringe beliefs and scientifically unsound information.

In this paper, we use the recent Swine Flu pandemic of 2009 as a use case to investigate the prevalence and uptake of online media. This topic serves as an excellent example because it received widespread attention during 2009 and was covered extensively in the press and social media. In particular, we seek to answer whether Twitter provides any insight into the popularity and consumption of online resources, and in particular, to answer the following two questions about the role Twitter plays:

- Do Twitter users have a preference when promoting online materials, for example, for official government health bodies over those of untrusted blogs?
- What are the dynamics of information dissemination during important global events? In particular, how does timeliness ultimately affect the popularity of online content?

II. RELATED WORK

Twitter provides an excellent way to sample large populations. In terms of epidemic intelligence, Twitter can be used to both track [i,ii] and even predict [iii] the spread of infectious diseases. In a recent study of antibiotic understanding on Twitter [iv], social media was shown to be a useful may to disseminate medical information, but that it is also prone to abuse.

Twitter has also proved to have excellent real-time benefits. Earthquake detection [v] is possible by examining the tweets of users in the local area containing terms related to earthquakes. When natural disasters strike, Twitter can help coordinates rapid responses [vi] and increase situational awareness with users providing important information on local conditions (such as weather, visibility, road conditions, etc).

Much research, such as [vii] has focused on understanding how information cascades through the Twitter network. Since Twitter users usually follow other Twitter users to stay up-to-date with what people in the social network are doing, the connections that people make are intrinsic to the dynamics of information flow. Various studies [viii,ix] show that influential people in the Twitter network (i.e. those with large numbers of followers) are the main hubs and control the spread of information. However, large numbers of followers does not guarantee that information will propagate through the network – other factors such as timeliness, accuracy, entertainment and so on, play an import role.

III. DATASET

We searched Twitter 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. Overall, tweets containing links are the most prominent, accounting ~65% of all tweets that contain the term 'flu'. Tweets containing hashtags are uncommon (~10%) and retweeting is rare (only 1% of tweets in this sample were retweets).

Inspection of tweets containing links to online resources reveals little sentiment or user reaction – in almost all cases, only the title of the article and a URL are included in the tweet text. Therefore, it is not possible to analyse the public reaction by mining the text alone. Instead, we classify the types of resources linked and how popular they are and determine how the timeliness and reputation of sources effects uptake of articles posted. In the following sections, two experiments are presented, each describing the motivation (what is the question we are hoping to answer), the method employed (e.g. how data was processed), and the results obtained.

IV. LINKED RESOURCES ANALYSIS

A significant portion of the Twitter traffic we sampled contained a link. Twitter users post links to a variety of online resources, such as news articles, blogs, videos, etc, usually because they have some interest in them and / or they want to advertise them to their followers. Therefore, analysis of the links posted on Twitter provides some insight into the interests of the Twitter population. To investigate whether Twitter favours the dissemination of trusted information sources over untrusted ones, we conduct a classification of the most popular web resources found in our sample dataset to find out what types of resource are the most popular.

A complete index of all hyperlinks posted to Twitter was constructed, including the total number of times the URL appears as well as the total number of distinct authors. Tracking the total number of distinct authors allows us to easily distinguish spammers and to factor out excessive self-promotion (when a user repeats the same tweet). Because of the 140 character tweet limit, many use URL shortening services (such as bit.ly) to obtain shortened version of the URLs that they wish to link to. Since there are many services available to accomplish this, a large number of different URLs can point to the same resource. Hence, any URL found was retrieved programmatically (using the cURL[1] tool) to determine whether the URL posted is the final destination, or if a redirection exists.

After creating an index of all resources linked, a classification task was conducted (by an experienced journalism grad-student) on the most popular 769 resources posted between 02/06/09 and 29/08/09 placing each item in one of the following categories: Blog, News, Medical Organisation, Spam, Video, Poll, Comic, Aggregator, Game, or Sales, Download, Campaign or Suspended Account.

Table 1 contains the total number of distinct authors and total number of resources for each classification category. The most widely represented in terms of number of distinct resource linked is spam (40%). In majority of cases, this was simple to verify because the user's Twitter account has been suspended, or the redirection link registered with URL shortening services has been disabled.

In terms of the number of distinct authors that tweeted a reference to a resource (and hence a direct measure of its popularity), blogs are the most widely linked (26%), closely followed by official news articles (21%), and pages from official medical organisations

[1] http://curl.haxx.se/

(15%). Since blogs represent a possible source of untrusted information, we analyse these more closely.

Category	Total Authors	Total Resources
Blog	7573	162
News	6151	117
Medical Organisation	4388	38
Spam	4231	312
Video	3897	72
Poll	741	5
Comic	484	8
Aggregator	318	10
Game	294	4
Sales	288	31
Download	248	8
Campaign	63	1
Suspended account	5	1

Table 1 - Categories of flu related resources posted to Twitter from 02/06/09 to 29/08/09

URL	Total Authors
http://www.theonion.com/articles/obamas-declaration-of-swine-flu-emergency-prompts,6952/	547
http://www.benckenstein.com/digital-media/swine-flu-susan-boyle-and-the-network-multiplier-effect/	468
http://mashable.com/2009/11/10/google-flu-shot-map/	319
http://mashable.com/2009/11/14/swine-flu-appointments/	262
http://www.theatlantic.com/magazine/archive/2009/11/does-the-vaccine-matter/7723/	185
http://techcrunch.com/2009/10/26/harvard-medical-school-launches-swine-flu-iphone-app/	180
http://www.fannation.com/si_blogs/grant_wahl/posts/74041-landon-donovan-has-h1n1-flu-virus	147
http://articles.mercola.com/sites/articles/archive/2009/10/06/Why-You-Should-NOT-Vaccinate-Your-Children-Against-the-Flu-This-Season.aspx	138
http://www.informationisbeautiful.net/2009/is-the-h1n1-swine-flu-vaccine-safe/	134
http://pitchfork.com/news/35776-jens-lekman-contracts-swine-flu/	121

Table 2 - The most popular flu related blog articles posted on Twitter from 02/06/09 to 29/08/09

Table 2 contains the top 10 most popular (in terms of the number of distinct authors that posted the link) blog resources found in our sample dataset. The most popular is a satirical piece by the popular parody newspaper 'The Onion;. Other popular resources are technology related (such as Mashable and TechCrunch). However, one story that contains information that is contrary to the current scientific consensus did receive attention from 138 users. The article "Do NOT Let Your Child Get Flu Vaccine" is representative of the type of article that official health agencies don't want published online since it is not evidence based and is authored by someone with no medical qualifications.

V. PANDEMIC STATUS CHANGE

When the WHO upgraded the status of H1N1 to 'pandemic' (11th June 2009), a significant amount of reaction was captured by Twitter. The key questions under investigation in this part of the study are: i) Do the articles that enter Twitter first ultimately more successful than those that enter later (i.e. is it first come, first served)? ii) Did public health organisation make a significant impact on the Twitter population?

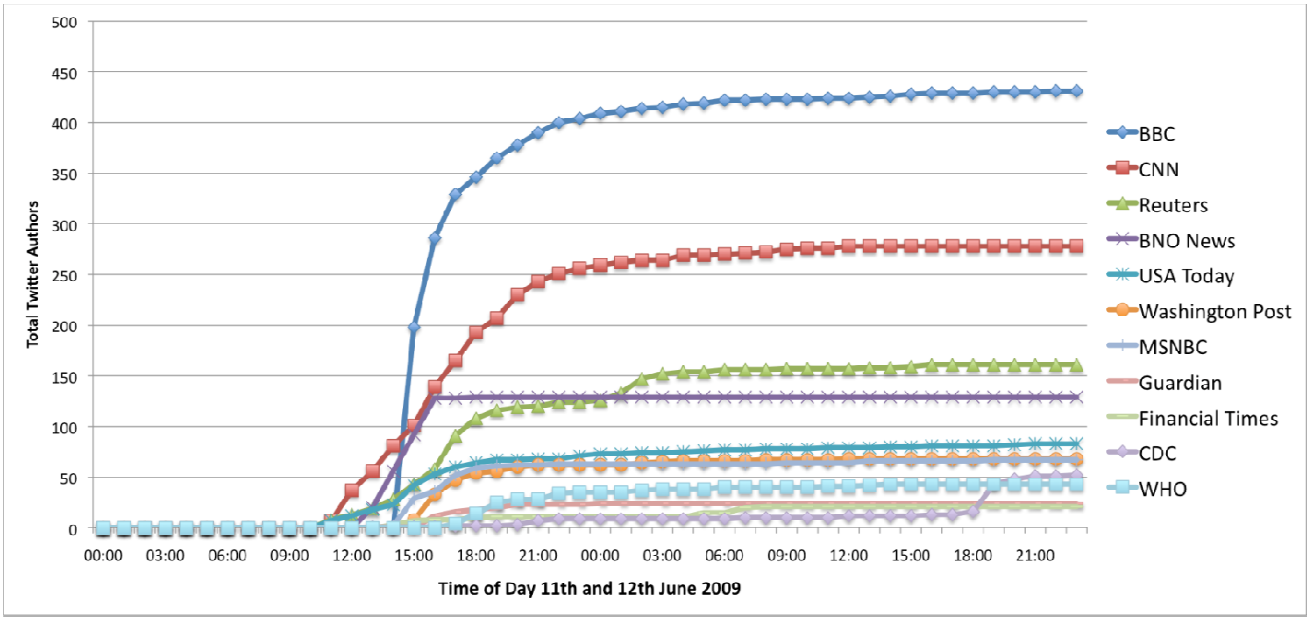
All links found in our sample dataset on the 11th and 12th June 2009 were examined. URLs were programmatically harvested to determine if they are still active (i.e. they have not been disabled because they were spam), and if they are redirected via a URL shortening service. Each resource was inspected manually to determine whether it was a direct reference to the WHO announcement. Articles from the most popular news organisations (both UK and US) were shortlisted, along with those from two official health agencies – the World Health Organization (WHO) and the Centers for Disease Prevention and Control (CDC).

Figure 1 is a plot showing the popularity of links posted to Twitter (in terms of the number of distinct authors) on an hour-by-hour basis. Ultimately, the most popular resource is the BBC article, but this is not the first to make an appearance in Twitter. CNN, Reuters and USA Today were the first to arrive in Twitter – 4 hours before the BBC article was picked up. Both WHO and CDC also have articles that appear in Twitter (CDC arriving much sooner than WHO), but their uptake is relatively small compared to the BBC and CNN articles. A link to the website of the European Centre for Disease Prevention and Control (ECDC) appeared only once.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we have performed a systematic analysis of a sample of Twitter data that we collected during 2009 surrounding the topic of 'flu'. We recall the questions listed in Section I, and briefly summarise the results presented in the paper:

- **Do Twitter users have a preference when promoting online materials, for example, for official government health bodies over those of untrusted blogs?** The articles related to Swine Flu that became popular where often from quality news sites (such as the BBC) or official medical organisations (such as the WHO, CDC, or ECDC). However, in some cases blog posts with poor scientific merit did become popular.



om the major news organizations

- **What are the dynamics of information dissemination during important global events? In particular, how does timeliness ultimately affect the popularity of online content?** Twitter reacts extremely fast to online media. Within the space of a few hours, most major news organisations had published on the topic and those articles were propagated through Twitter. It would seem that timeliness isn't a good predictor of overall success: between 11th-12th June 2009, the BBC article became the most popular even those it appears in Twitter 4 hours later than other news agencies.

To develop this study further, we plan to perform a systematic analysis of all articles appearing in Twitter on the 11th and 12th June 2009 to assess their journalistic merit, categorising them in terms of the number of properly used terms (e.g. the correct use of the term pandemic), whether accurate definitions of the terms or the virus in question are included, if the proper figures were used, whether sources of information are properly quoted, whether appropriate and up-to-date sources of information are cited, etc.

VII. ACKNOWLEDGMENTS

We should like to acknowledge Achintya Rao for his assistance classifying the resources found in Twitter. We would also like to thank Ed de Quincey for his foresight in setting up the Twitter data collection in May 2009 – this work would not be possible without it.

VIII. REFERENCES

[i] Lamos, V., de Bie, T., and Cristianini, N. "Flu Detector – Tracking Epidemics on Twitter". In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2010), pp. 599-602.

[ii] Lampos, V. and Cristianini, N. "Tracking the flu pandemic by monitoring the Social Web". In Proceedings of the 2nd IAPR Workshop on Cognitive information Processing (CIP2010), pp 411—416, 2010

[iii] Szomszor, M., Kostkova, P., de Quincey, E. "#swineflu : Twitter predicts swine flu outbreak in 2009". In 3rd International ICST Conference on Electronic Healthcare for the 21st Century (eHealth2010), Casablanca, Morocco, December 2010.

[iv] Scanfeld, D., Scanfeld, V., Larson, E. L. "Dissemination of health information through social networks: Twitter and antibiotics". In AJIC: American Journal of Infection Control 3(8) pp. 182—188, April 2010.

[v] Sakaki, T., Okazaki, M., and Matsuo, Y.: "Earthquake shakes Twitter users: real-time event detection by social sensors". In Proceedings of the 19th international conference on World wide web, pp. 851—860, Raleigh, North Carolina, USA, 2010.

[vi] Vieweg, S., Hughes, A. L., Starbird, K., and Palen, L. "Microblogging during two natural hazards events: what twitter may contribute to situational awareness". In Proceedings of the 28th International Conference on Human Factors in Computing Systems (CHI2010), pp 1079—1088, Atlanta, Georgia, USA, 2010.

[vii] Kwak, H., Changhyun, L., Hosung, P., and Moon, S. "What is Twitter, a Social Network or a News Media?". In Proceedings of the 19th International Conference on World Wide Web (WWW2010), pp 591—600, Raleigh, USA, 2010.

[viii] Cha, M., Hamed, H., Fabricio, B., Krishna, P. G. "Measuring Influence in Twitter: The Million Follower Fallacy". In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM2010), Washington DC, US, May 2010.

[ix] Lee, C., Kwak, H., Park, H, and Moon, S. "Finding influentials based on the temporal order of information adoption in twitter". In Proceedings of the 19th International Conference on World Wide Web (WWW2010), pp 591—600, Raleigh, USA, 2010.