OMG U got flu? Analysis of shared health messages for bio-surveillance

Nigel Collier
National Institute of Informatics
1-2-1 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
collier@nii.ac.jp

Nguyen Truong Son and **Ngoc Mai Nguyen**
VietNam National University at HCMC
HoChiMinh City, VietNam
ntson@fit.hcmus.edu.vn and maintn@uit.edu.vn

Abstract

Micro-blogging services such as Twitter offer the potential to crowdsource epidemics in real-time. However, Twitter posts ('tweets') are often ambiguous and reactive to media trends. In order to ground user messages in epidemic response we focused on tracking self-protective behaviour such as avoiding public gatherings or increased sanitation as the basis for further risk analysis. In initial experiments on influenza tracking we report results for unigrams, bigrams and regular expressions employed in two supervised classifiers (SVM and Naive Bayes) to classify tweets into 4 self-reported protective behaviour categories plus a self-reported diagnosis. We report moderately strong Spearman's Rho correlation for the classifiers against WHO/NREVSS laboratory data for A(H1N1) in the USA during the 2009-2010 influenza season.

1 Introduction

Rising awareness of infectious disease outbreaks and the high costs of extending traditional sensor networks means that we have an opportunity to harness new forms of social communication for crisis surveillance. The trend is already underway with automatic map generation from Twitter reports for earthquakes and typhoons (Earle, 2010; Sakaki et al., 2010), the symptom-based influenza tracking portal Flutracking (Dalton et al., 2009) as well as the humanitarian portal Ushahidi (Okolloh, 2009). Despite a risk of high false reporting rates there is nevertheless strong potential in having multiple sensor sources for verification, robustness and redundancy. In the case of earthquake detection, Earle notes that Twitter messages (tweets) can be available up to 20 minutes before the official report from the US Geological Survey. With epidemics too the time period from signal to detection is critical. Recent studies such as (Cheng et al., 2009) estimate that the average delay in receiving and disseminating data from traditional sentinel physician networks is about two weeks.

A small but growing number of early warning systems have already developed to mine event information from low cost Web sources mainly focussing on edited newswire reports (see (Hartley et al., 2010) for a survey). Success in operationalizing such systems has crucially de-

pended on building close collaborations with government and international public health agencies in order to perform detailed verification and risk assessment.

Recent studies on alerting from newswire reports (Collier, 2010) are beginning to make clear the operational boundaries in terms of their selectivity, volume and timeliness. In earlier work Collier noted the issue of late warnings, i.e. where there is a known outbreak in a country but true alerts at the province or city level are occluded by the aggregated system data for the country as a whole. To overcome this problem micro-blogging might have a role to play. Micro-blogs may be able to help also with very early epidemic detection, i.e. at the pre-diagnostic stage where there is maximum scientific uncertainty about symptoms, transmission routes and infectivity rates. Automatic geo-coding and the ability to send messages from many types of mobile device are a key advantage in this respect.

2 Background

In micro-blogging services such as Twitter, users describe their experiences directly in near-real time in short 140 character tweets. As of April 2010 it was estimated that Twitter had approximately 106 million registered users with 300,000 new users being added each month. Despite their potential coverage, timeliness and low overhead, tweets present their own unique challenges: pre-diagnostic unedited reports mean that there is a large trust issue to resolve within the modeling technique; also social media can reflect a high degree of reactivity to risk perception as seen during the H1N1 pandemic in 2009 - redistributing links or requests for information rather than generating user experience. To a degree this reflects newswire coverage and the amount of uncertainty readers feel. Re-tweets in themselves may provide useful signal but their role has yet to be quantified. Despite these obvious challenges we believe there is potential for using very short messages to detect epidemic trends, as hinted at by the success of Google Flu Trends (Ginsberg et al., 2009) which harness user's search queries.

In order to do this we propose to employ aberration detection for detecting sharp rises in the features that signal epidemics. A precursor to this is in identifying reliable features themselves. In this study we started by looking at precautionary actions as identified by Jones and Salathé in their behaviour response survey (Jones and Salathé, 2009) to A(H1N1). Modeling individual risk perception based on local health information appears to be an understudied area in event alerting which may add signal to early detection models.

Recently a number of studies have appeared looking at the effectiveness of search queries and social media. (Lampos et al., 2010) studied tweets in the 49 most populated urban centres of the UK and found a strong linear correlation with Health Protection Agency influenza like illness (ILI) data from general practitioner (GP) consultations during the 2009-2010 influenza season. Studies on user query data from Google Flu Trends has also shown strong correlations with sentinel network data. (Valdivia et al., 2010) showed for the 2009 Influenza A(H1N1) pandemic there was a strong Spearman's Rho correlation with ILI and acute respiratory infection (ARI) data from sentinel networks in Europe.

Nevertheless challenges in interpreting query and social networking data remain. (Ortiz et al., 2010) for example discuss the potential for confusion in Google Flu Trends between ILI and non-influenza illnesses. Influenza data was

compared from Google Flu, the CDC outpatient surveillance network and the US influenza virologica surveillance system. Whilst correlation with ILI was found to be high, it was found that correlation with actual influenza test positive results was lower. This result highlights the fact that both social media and user queries are secondary indicators that should be correlated with patient reported symptoms. Significant deviation between user's searching behaviour and ILI rates was noted for the 2003-2004 influenza season when influenza activity, pediatric deaths and news media coverage of influenza were particularly high. This highlights another understudied issue: that we need to work hard to remove elements of reporting bias by users during media storms.

3 Method

3.1 Annotation

Taking Jones and Salathé's behaviour responses as a starting point we surveyed potential messages in Twitter in relation to H1N1 influenza topics. From an initial group of thirteen categories we decided, due to low frequency counts, to conflate several into a final grouping of four. e.g. avoiding people who cough/sneeze, avoiding large gatherings of people, avoiding school/work and avoiding travel to infected areas were joined into a general 'avoidance behaviour' category. To this we added a final category for direct reporting of influenza. The final list of categories is: (A) Avoidance Behavior - behaviours which avoid agents thought to be at risk of infection; (I) Increased sanitation - sanitation measures to promote individual health and prevent infection; (P) Seeking pharmaceutical intervention - seeking clinical advice or using medicine or vaccines; (W) Wearing a mask; and (S) Self reported diagnosis - reporting that one has influenza.

As expected there are a number of caveats to each of these broad classes. We list up only a representative sample here: (1) A message is only tagged positive if the user or a close family member is the subject of the tweet; (2) If the message indicates that the action is hypothetical then the classification is negative; (3) The time of the reported event should be within one week of the current time; (4) Messages can belong to more than one category. Examples of (anonymized) messages are shown in Table 1.

At a practical level the problem of identifying self protection messages can be characterised as classifying very biased data. In order to handle this we adopted two stages of filtering. The first stage used a bag of 7 keywords to select tweets on topics related to influenza (*flu, influenza, H1N1, H5N1, swine flu, pandemic, bird flu*). For 1st March 2010 to April 30th 2010 this resulted in a pool of about 225,000 tweets. This first stage of filtering was also designed to reduce the ambiguity of keywords such as 'fever' and 'cough' which occur in a wide variety of contexts.

The second stage used hand built patterns to select a total of 14,508 tweets. From these we randomly chose 7,412 tweets spread across the five classes. All duplicates were removed leaving 5,283 messages and the resulting data was then classified by hand using a single annotator as detailed in Table 2. Results for mean character length and standard deviation showed no category-specific trend except to illustrate the wide variety of message lengths.

In order to test the stability of the annotation scheme and our assumptions about its reproducibility we calculated kappa for 2,116 messages balanced across all the classes. For this another an-

Table 1: Positive (+) and negative (-) examples of classified messages.

n	Message	A	I	P	W	S
e1	home this weekend? i've been off work all week with the flu	+	-	-	-	+
e2	there is alot more to preparing for Swine Flu than just washing your hands	-	-	-	-	-
e3	everyone wash your hands.. no one wants swine flu	-	+	-	-	-
e4	awl u need to go get to the doc so u dnt past da swine flu	-	-	-	-	-
e5	it's 2:10pm, I have flu and I'm still wearing my pajama	-	-	-	-	+
e6	I have the flu. I had a normal flu shot	-	-	+	-	+
e7	This guy has a nasty cough! Thank god he's sitting far away from me - the swine flu travels	+	-	-	-	-
e8	I'm sick too... cold or flu, I don't know... I couldn't go to work today...	+	-	-	-	+
e9	Trivia for tonight has been cancelled due to flu bug	+	-	-	-	-
e10	Feel like I've washed my hands a 1000 times Gotta loveworkin during cold & flu season	-	+	-	-	-
e11	overhyped public scare. I want to remove this mask	-	-	-	+	-
e12	i don't know. she just keeps getting sick, but it's not the flu. i hate keeping her off school	-	-	-	-	-
e13	i feel terrible, don't want to be at work, wish id never had the h1n1 jab	-	-	+	-	-
e14	" Some cleaning products were especially made to kill the H1N1 ...	-	-	-	-	-
e15	She has a surgical mask on in the movies I'm nervous hope it's not h1n1	-	-	-	-	-
e16	regretting not getting a flu shot this year	-	-	-	-	-

Table 2: Message frequency in the training/testing corpus for self-protection classes

	A	I	P	W	S
Positive	251	37	499	32	741
Negative	632	43	974	230	1873
Total	883	80	1443	262	2614
Mean length	109.2	118.8	107.0	117.3	100.9
Sd. length	28.9	21.9	30.6	27.7	33.4

notator was chosen who did not take part in the creation of the guidelines and was not a co-investigator in this study. The simple agreement ratio was 0.88 (the total number of matched class assignments divided by the total number of messages). Kappa was calculated as $\kappa = (pA - pE)/(1 - pE)$, where pA was 0.88 and pE was 0.12. κ was then found to be 0.86. Both results reveal a high level of agreement in the annotation scheme and give us confidence to move ahead with automated classification.

3.2 Models

We employed two widely used classification models implemented in the Weka Toolkit (Holmes et al., 1994), Naive Bayes and Support Vector Machines (SVMs) (Cristianini and Shawe-Taylor, 2000) to classifying five data sets into positive or negative. SVM used a RBF-kernel and grid search for finding the best parameter settings. Since we hypothesized that custom built regular expressions might have more traction for achieving precision we decided to use a freely available toolkit called the Simple Rule Language (McCrae et al., 2009) for this purpose. SRL comes with an interface for maintaining the rule base which can be run in testing mode to convert surface expressions into structured information.

SRL rules were built from a held out set of tweets not used in training. Rules consist of string literals, skip expressions, word lists, named entity classes and guard expressions for limiting the scope of matched entities. Rule building took approximately 10 hours of work. The rule book contains specialised synonym sets to recognize common and slang terms for medicines (e.g. *shot, vaccine, drug, tamiflu, jab, medicine, vacc*), physicians (e.g. *doctor, doc, dr, physician*) and other key domain entities. Verb lists are maintained for specialized lexical classes such as prescribe (e.g. *prescribe, perscribe**). Lists are also built for pronouns, common temporal adverbs, modal verbs and negations. Special rules were built to recognize past events. The exceptional class was I (increased sanitation) where we were not able to identify enough examples with confidence to build meaningful rules by hand. In this case only unigrams and bigrams were used to train the classifiers.

We found that the language used in tweets to express user's behaviour is very diverse and idiosyncratic so it is challenging to achieve a high degree of coverage in the rules with surety. With this in mind we combined features from the rules with unigrams and bigrams. If a rule matched a tweet its feature value was set to 1, otherwise to 0.

4 Results 1: Classification experiments

All test runs used 10-fold cross validation on each of the 5 test sets. We calculated recall, precision and F-score performance for each category. As we can see in Tables 3 and 4, SVM overall performs better on all categories except for W (wearing a mask). Both model's performance generally follows the amount of training data except for S (Self diagnosis) where the F-score is slightly lower than the trend in other classes despite large numbers of examples. The overall trend for Naive Bayes is to have stronger recall than precision whereas for SVM precision is generally higher than recall.

The results suggest that our SRL rule book seemed to offer substantial benefits when combined with unigrams but less certain improvements when combined with unigrams plus bigrams. Looking slightly deeper into the results we found a correlation between message length and

Table 3: F1 results for tweet classification using Naive Bayes. UNI = unigram, BI = bigram, SRL = Simple Rule Language regular expression

	P	R	F1
	A		
UNI	0.73	0.76	0.73
UNI+SRL	0.74	0.76	0.74
UNI+BI	0.73	0.77	0.73
UNI+BI+SRL	0.73	0.77	0.74
	I		
UNI	0.56	0.55	0.54
UNI+BI	0.49	0.49	0.48
	P		
UNI	0.74	0.76	0.74
UNI+SRL	0.75	0.78	0.75
UNI+BI	0.75	0.78	0.75
UNI+BI+SRL	0.76	0.79	0.76
	W		
UNI	0.59	0.68	0.58
UNI+SRL	0.63	0.76	0.63
UNI+BI	0.60	0.71	0.59
UNI+BI+SRL	0.60	0.71	0.59
	S		
UNI	0.70	0.73	0.71
UNI+SRL	0.74	0.77	0.75
UNI+BI	0.72	0.76	0.73
UNI+BI+SRL	0.74	0.77	0.74

Table 4: F1 results for tweet classification using SVM. UNI = unigram, BI = bigram, SRL = Simple Rule Language regular expression

	P	R	F1
	A		
UNI	0.79	0.77	0.78
UNI+SRL	0.80	0.79	0.79
UNI+BI	0.80	0.79	0.79
UNI+BI+SRL	0.81	0.80	0.80
	I		
UNI	0.66	0.65	0.63
UNI+BI	0.63	0.62	0.61
	P		
UNI	0.79	0.80	0.78
UNI+SRL	0.78	0.81	0.79
UNI+BI	0.77	0.79	0.78
UNI+BI+SRL	0.78	0.79	0.78
	W		
UNI	0.54	0.51	0.51
UNI+SRL	0.57	0.55	0.55
UNI+BI	0.54	0.51	0.51
UNI+BI+SRL	0.61	0.56	0.57
	S		
UNI	0.74	0.73	0.74
UNI+SRL	0.78	0.79	0.79
UNI+BI	0.79	0.73	0.75
UNI+BI+SRL	0.82	0.76	0.78

classification accuracy in Naive Bayes and SVM. For Naive Bayes, whilst the length of messages didn't seem to make much difference to the false negative rate which remained constant at about 0.2 to 0.25 on messages in the length range of 34 to 144 characters, it impacted to a greater degree on false positives (0.23 on shorter messages of length 34 to 56 down to 0.08 for messages of length 122 to 144). For SVM there appeared to be a general reduction in both false positives and false negatives as message length increased.

As expected, frequent misspellings, abbreviated word forms, slang and lack of punctuation complicated the classification task. Missing auxiliary verbs and articles need to be compensated for within the SRL rules in order to ensure successful matching.

Potentially issues of duplication through re-tweeting still remain which we have not modelled in this study. Clearly we should have more confidence in the alerting model if a larger number of independent sources report an event at the same time. This will form part of our future work. Future work will also need to ensure that the classification model remains relevant over time as the data content in tweets shifts.

5 Results 2: Comparison to CDC data

In order to provide a proof of concept we operationalised the classifiers and ran them on a corpus of Twitter data called the Edinburgh Corpus (Petrovic et al., 2010). The Edinburgh corpus holds 97 million tweets for the period November 11th 2009 to February 1st 2010 from 9 million users. This represents over 2 billion words from a variety of languages. Of these 12.5 million are reported as topic tags, 55 million are @ replies and 20 million are links.

We applied the same keyword filtering method used on the Edinburgh corpus for the first set of experiments and obtained 52,193 tweets for the period of study. Following this we applied the SVM UNI model and then compared the week by week volumes against laboratory results for weeks 47 to 5 of the 2009-2010 influenza season in the USA (Division, 2010). Counts are shown in Table 5. Several intersting trends can be observed: (a) The total volume of positively identified Tweets was relatively small compared to the volume of Tweets as a whole; (b) Wearing a mask was totally absent from our classified data; (c) The aggregated counts for self protection (A+I+P, data not shown) seem to have a close correlation to CDC results (data not shown).

To measure correlation we calculated the Spearman's Rho [1] between counts of positive messages in each class and the CDC laboratory data for A(H1N1). Table 6 shows moderately strong correlations. The strongest correlation appeared when A,I and P were combined. Besides W which failed to provide any data, the weakest evidence came from Increased saniation (I). Differences could be due to (a) the global geographic coverage of

[1] See http://en.wikipedia.org/wiki/Spearmanś_rank_correlation_coefficient

Table 5: Positively identified Tweets in the Edinburgh corpus shown against Influenza Positive tests reported to CDC by U.S. WHO/NREVSS collaborating laboratories, National Summary, 2009-2010. Counts for W were zero throughout and are therefore not shown. [A] For week 46 we only have partial Twitter data available in the Edinburgh corpus.

Wk	A	S	I	P	CDC
46[A]	49	48	22	222	2715
47	32	72	30	258	1408
48	24	49	9	181	997
49	35	41	10	199	610
50	35	39	10	154	480
51	21	35	12	150	251
52	19	26	4	37	285
1	25	32	6	63	266
2	25	32	5	81	261
3	29	31	7	73	317
4	29	20	7	62	268
5	29	23	6	46	290

Table 6: Correlation between CDC AH1N1 laboratory data frequency for Influenza 2009-2010 and aggregated self protection behaviour counts and self reported diagnosis from Tweets. [a] Spearman's rank-order correlation coefficient. [b] p values are reported for a two-tailed test. Calculations were done using VassarStats (http://faculty.vassar.edu/lowry-/corr_rank.html)

Category	Spearman's Rho[a]	p-value[b]
A	0.66	0.020
S	0.66	0.021
I	0.58	0.048
P	0.67	0.017
A+I+P	0.68	0.008
A+I+P+S	0.67	0.017

tweets in our collection; and (b) the syndromes covered in our self protection behaviour and self reporting messages are wider than A(H1N1) and could actually be other diseases such as common colds, strep throat, adenovirus infection and so on.

Drill down analysis reveals that we still need to do more to remove false positives by strengthening the linguistic features within the limits of the 140 character length. Examples of false positives include interogative sentences, hypothetical sentences, reports on events that took place in the distant past, comments on influenza advice from others, etc.

6 Conclusion

In this paper we have made the first steps towards classifying Twitter messages according to self reported risk behaviour. The results have shown moderately strong correlation with CDC A(H1N1) data but we still need to make further progress in order to achieve the high degrees of correlation reported bewteen Google Flu trends and sentinal influenza data. The next step will be to extend our training data, strengthen the linguistic features and see if we can use these signals to detect emerging disease outbreaks. It was shown in Jones and Salathé that after an initial peak in levels of risk concern, anxiety faded once the immediate threat of the A(H1N1) pandemic had passed. In follow up work we intend to look at how closely these signals track epidemic case data.

We also believe that the signals we have modelled make them applicable to a wide range of diseases within the respiratory syndrome and we intend to explore how these features can be used to detect diseases other than influenza.

Besides disaster alerting, results from analysis of behavioural responses may also help in the future to evaluate the success of official prevention campaigns. For example, it is known that little notice was taken of antiviral therapies, goggle or mask wearing advice in the Netherlands after the Avian Influenza epidemic was introduced to Europe (De Zwart et al., 2007). Conversely, empirical studies of individual risk perception in diesase severity and susceptibility may help official agencies in the future to avoid 'over-hyping' epidemic threats and tune risk communication strategies more effectively.

Acknowledgement

We gratefully acknowledge support from the National Institute of Informatics Global Liason Office for internship funding for this study. The study was conceived and directed by NC, corpus collection and analysis was done by NM and the machine learning experiments were done by NS.

References

Cheng, C. K., E. H. Lau, D. K. Ip, A. S. Yeung, L. M. Ho, and B. J. Cowling. 2009. A profile of the online dissemination of national influenza surveillance data. *BMC Public Health*, 9(1):339.

Collier, N. 2010. What's unusual in online disease outbreak news? *Biomedical Semantics*, 1(1), March. doi:10.1186/2041-1480-1-2.

Cristianini, N. and J. Shawe-Taylor. 2000. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, Cambridge, England; ISBN 0521780195.

Dalton, C., D. Durrheim, J. Fejsa, L. Francis, S. Carlson, and E. et al. ursan d'Espaignet. 2009. Flutracking: A weekly australian community online survey of influenza-like illness in 2006, 2007 and 2008. *Communicable Disease Intelligence*, 33(3):316–322. doi:10.1038/ngeo832.

De Zwart, O., I. K. Veldhuijzen, G. Elam, A. R. Aro, T. Abraham, G. D. Bishop, J. H. Richardus, and J. Brug. 2007. Avian influenza risk perception, europe and asia. *Emerging Infectious Diseases*, 13(2):290–293.

Division, CDC Influenza. 2010. Fluview: 2009-2010 influenza season week 20 ending may 22, 2010. Technical report, Centers for Disease Control and Prevention, May. Available at http://www.cdc.gov/flu/weekly/.

Earle, P. 2010. Earthquake twitter. *Nature Geoscience*, 3(4):221–222. doi:10.1038/ngeo832.

Ginsberg, J., M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, and L. Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature*, 457:1012–1014.

Hartley, D., N. Nelson, R. Walters, R. Arthur, R. Yangarber, L. Madoff, J. Linge, A. Mawudeku, N. Collier, J. Brownstein, G. Thinus, and N. Lightfoot. 2010. The landscape of international biosurveillance. *Emerging Health Threats J.*, 3(e3), January. doi:10.1093/bioinformatics/btn534.

Holmes, G., A. Donkin, and I. H. Witten. 1994. WEKA: a machine learning workbench. Technical report, Department of Computer Science, Waikato University, New Zealand, September.

Jones, J. and M. Salathé. 2009. Early assessment of anxiety and behavioral response to novel swine-origin influenza a(h1n1). *PLoS One*, 4(12):e8032.

Lampos, V., T. De Bie, and N. Cristianini. 2010. Flu detector - tracking epidemics on twitter. In *Machine Learning and Knowledge Discovery in Databases*, volume 6223/2010, pages 599–602. Lecture Notes in Computer Science.

McCrae, J., M. Conway, and N. Collier. 2009. Simple rule language editor. Google code project, September. Available from: http://code.google.com/p/srl-editor/.

Okolloh, O. 2009. Ushahidi, or 'testimony': Web 2.0 tools for crowdsourcing crisis information. *Participatory Learning and Action*, 59(1):65–70, June.

Ortiz, J. R., H. Zhou, D. K. Shay, K. M. Neuzil, and C. H. Goss. 2010. Does google influenza tracking correlate with laboratory tests positive for influenza? In *Proc. , USA*.

Petrovic, S., M. Osborne, and V. Lavrenko. 2010. The edinburgh twitter corpus. In *Proc. #SocialMedia Workshop: Computational Linguistics in a World of Social Media, LA, USA*, pages 25–26, June.

Sakaki, T., M. Okazaki, and Y. Matsuo. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proc. of the 19th International World Wide Web Conference, Raleigh, NC, USA*.

Valdivia, A., J. Lopez-Alcalde, M. Vicente, M. Pichiule, M. Ruiz, and M. Ordobas. 2010. Monitoring influenza activity in europe with google flu trends: comparison with the findings of sentinel physician networks - results for 2009-10. *Eurosurveillance*, 15(29):2–7.