Tracking the flu pandemic by monitoring the Social Web

Vasileios Lampos, Nello Cristianini

Intelligent Systems Laboratory
Faculty of Engineering
University of Bristol, UK
{*bill.lampos*, *nello.cristianini*}*@bristol.ac.uk*

***Abstract*—Tracking the spread of an epidemic disease like seasonal or pandemic influenza is an important task that can reduce its impact and help authorities plan their response. In particular, early detection and geolocation of an outbreak are important aspects of this monitoring activity. Various methods are routinely employed for this monitoring, such as counting the consultation rates of general practitioners. We report on a monitoring tool to measure the prevalence of disease in a population by analysing the contents of social networking tools, such as *Twitter*. Our method is based on the analysis of hundreds of thousands of tweets per day, searching for symptom-related statements, and turning statistical information into a flu-score. We have tested it in the United Kingdom for 24 weeks during the H1N1 flu pandemic. We compare our flu-score with data from the Health Protection Agency, obtaining on average a statistically significant linear correlation which is greater than 95%. This method uses completely independent data to that commonly used for these purposes, and can be used at close time intervals, hence providing inexpensive and timely information about the state of an epidemic.**

I. Introduction

Monitoring the diffusion of an epidemic in a population is an important and challenging task. Information gathered from the general population can provide valuable insight to health authorities about the location, timing and intensity of an epidemic, or even alert the authorities of the existence of a health threat. Gathering this information, however, is a difficult as well as resource-demanding procedure.

Various methods can be used to estimate the actual number of patients affected by a given illness, from school and workforce absenteeism figures [1], to phone calls and visits to doctors and hospitals [2]. Other methods include randomised telephone polls, or even sensor networks to detect pathogens in the atmosphere or sewage ([3], [4]). All of these methodologies require an investment in infrastructure and have various drawbacks, such as the delay due to information aggregation and processing times (for some of them).

A recent article reported on the use of search engine data to detect geographic clusters with a heightened proportion of health-related queries, particularly in the case of "Influenza-like Illness" (ILI) [5]. It demonstrated that timely and reliable information about the diffusion of an illness can be obtained by examining the content of search engine queries.

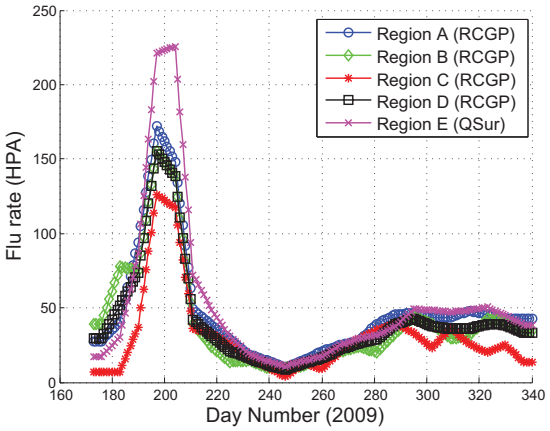


Fig. 1: Flu rates from the Health Protection Agency (HPA) for regions A-E (weeks 26-49, 2009). The original weekly HPA's flu rates have been expanded and smoothed in order to match with the daily data stream of Twitter (see section III-B).

This paper extends that concept, by monitoring the content of social-web tools such as *Twitter*[1], a micro-blogging website, where users have the option of updating their status with their mobile phone device. These updates (referred to as *tweets*) are limited to 140 characters only, similarly to the various schemes for mobile text messaging. Currently there are approximately 5.5 million Twitter users in the United Kingdom (UK). We analyse the stream of data generated by Twitter in the UK and extract from it a score that quantifies the diffusion of ILI in various regions of the country. The score, generated by applying machine learning technology, is compared with official data from the Health Protection Agency (HPA)[2], with which it has a statistically significant linear correlation coefficient that is greater than 95% on average.

The advantage of using Twitter to monitor the diffusion of ILI is that it can reveal the situation on the ground by utilising a stream of data (tweets) created within a few hours, whereas the HPA releases its data with a delay of 1 to 2 weeks. Furthermore, since the source of the data is entirely

[1]**Twitter**, http://twitter.com/.
[2]**Health Protection Agency**, http://www.hpa.org.uk/.

independent of search engine query-logs or any of the standard approaches, our method can also be used in combination with them, to improve accuracy.

II. Data Collection and Preprocessing

A. Retrieving data from Twitter

We were collecting a daily average of 160,000 tweets over a period of 24 weeks from 22/06/2009 (start of week 26, day 173) to 06/12/2009 (end of week 49, day 340). We focused our collection process on the 54 most populated urban centres in the UK. Twitter's geolocation feature enables us to track the location of a user posting a message, and consequently to label each tweet with a set of geographical coordinates, *i.e* longitude and latitude. We assign a tweet to an urban centre, when it falls inside a 10 Km radius from its geographical centre. To create vector space representations of specific subsets of the Twitter corpus, we remove stop words, and then stem by applying Porter's algorithm [6].

B. Official health reports

As our ground truth basis, we use weekly reports from the HPA related to the H1N1 epidemic. HPA provides regional statistics for the UK, based on rates gathered by the Royal College of General Practitioners (RCGP) [7], and the more general QSurveillance scheme (QSur)[3]. RCGP and QSur metrics express the number of GP consultations per 10^5 citizens, where the result of the diagnosis was ILI. RCGP data are gathered for four UK regions: Central England & Wales (region A), South England (region B), North England (region C), and England & Wales (region D), whereas QSur covers England, Wales & Northern Ireland (region E). Figure 1 shows the time series of HPA's influenza rates for regions A-E during weeks 26-49 in 2009.

III. Data Analysis

A. Computing a flu-score from Twitter corpus

We compile a set of textual markers $\mathcal{M} = \{m_i\}$, where $i \in [1, k]$, and look for them in the Twitter corpus of a day. The daily set of tweets is denoted as $\mathcal{T} = \{t_j\}$, where $j \in [1, n]$. If a marker m_i appears in a tweet t_j, we set $m_i(t_j) = 1$; otherwise $m_i(t_j) = 0$. The flu-score of a tweet $s(t_j)$ is equal to the number of markers it contains divided by the total number of markers used:

$$s(t_j) = \frac{\sum_i m_i(t_j)}{k}, \quad (1)$$

where k denotes the total number of markers. We compute, the flu-score of the daily Twitter corpus $f(\mathcal{T}, \mathcal{M})$, as the sum of all the tweet flu-scores divided by the total number of tweets:

$$f(\mathcal{T}, \mathcal{M}) = \frac{\sum_j s(t_j)}{n} = \frac{\sum_j \sum_i m_i(t_j)}{k \times n}, \quad (2)$$

where n denotes the total number of tweets for one day.

[3]**QSurveillance**, University of Nottingham and Egton Medical Information Systems Ltd, http://www.qresearch.org/Public/QSurveillance.aspx.

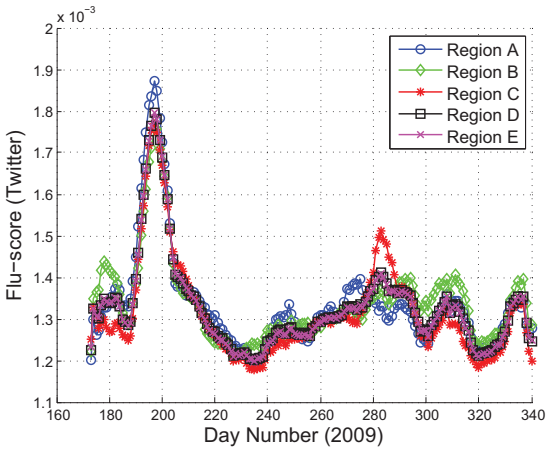


Fig. 2: Twitter's flu-scores based on our choice of markers for regions A-E (weeks 26-49, 2009). Smoothing with a 7-point moving average (the length of a week) has been applied.

TABLE I: Correlation Coefficients between Twitter's flu-scores and HPA's rates based on our choice of markers (for weeks 26-49, 2009).

Region	HPA Scheme	Corr. Coef.	P-value
A	RCGP	0.8471	1.95e-47
B	RCGP	0.8293	8.37e-44
C	RCGP	0.8438	9.84e-47
D	RCGP	0.8556	2.39e-49
E	QSur	0.8178	1.10e-41

B. Correlations between Twitter flu-scores and HPA flu rates

Using a small set of 41 textual markers (or n-grams) expressing illness symptoms or relevant terminology, *e.g.* 'fever', 'temperature', 'sore throat', 'infection', 'headache', and so on, we compute the Twitter flu-score time series for regions A-E. We smooth each time series with a 7-point moving average in order to express a weekly tendency in our data. Figure 2 shows the time series of Twitter's flu-scores for days 173 to 340 in 2009.

In order to retrieve an equal representation between the weekly HPA flu rates and the daily Twitter flu-scores, we expand each point of the former over a 7-day period; in fact, each weekly point of an original HPA flu rate is assigned on every day of the respective week. After expanding the HPA flu rates, we perform smoothing on them with a 7-point moving average (as in Twitter's flu-scores). Figure 1 shows the expanded and smoothed HPA's flu rates.

For each one of the five regions, we compute the linear correlation coefficients between Twitter's and HPA's flu-score time series. The correlations and their respective p-values are presented in Table I. The largest correlation of 85.56% (with a p-value of 2.39e-49) is found for region D (see Figure 3), whereas the smallest reported correlation is 81.78% for region E.

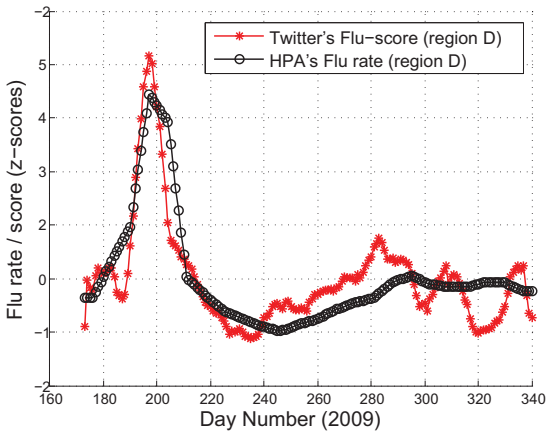


Fig. 3: Comparison of the unweighted Twitter's flu score (based on our choice of markers) and the respective HPA rates for region D (England & Wales) using their z-scores. Their linear correlation is equal to 85.56%.

C. Learning HPA's flu rates from Twitter flu-scores

We extend our previous scheme in order to form a model for predicting the HPA flu rate by observing the flu-score on Twitter. In the new scheme, we attach a weight w_i to each textual marker m_i. The weighted flu-score of a tweet is equal to:

$$s_w(t_j) = \frac{\sum_i w_i \times m_i(t_j)}{k}, \quad (3)$$

where k denotes the number of markers. Similarly, the weighted flu-score based on Twitter's daily corpus $\mathcal{T}$ is computed by:

$$f_w(\mathcal{T}, \mathcal{M}) = \frac{\sum_j s_w(t_j)}{n} = \frac{\sum_j \sum_i w_i \times m_i(t_j)}{k \times n}, \quad (4)$$

where n denotes the total number of tweets for this day. The contribution of each marker m_i in f_w can be considered as a flu-subscore and is equal to:

$$f_{w_i}(\mathcal{T}, m_i) = w_i \times \frac{\sum_j m_i(t_j)}{k \times n}. \quad (5)$$

Therefore, a daily Twitter's flu-score can be represented as a vector $\mathcal{F}_w$ of k elements $\mathcal{F}_w = [f_{w_1}(\mathcal{T}, m_1), ..., f_{w_k}(\mathcal{T}, m_k)]^T$ each one corresponding to Twitter's flu-subscore for marker m_i.

Initially, we retrieve from the Twitter corpus of a day an unweighted flu-score vector $\mathcal{F} = [f(\mathcal{T}, m_1), ..., f(\mathcal{T}, m_k)]^T$. The unweighted time series of each term's flu-subscores ($f(\mathcal{T}, m_i)$ for all the days) are smoothed with a 7-point moving average. We perform least squares linear regression between the time series of $\mathcal{F}$'s smoothed version and the expanded and smoothed HPA's flu rates in order to learn the weights w_i for the terms m_i. We use as a training set the data that correspond to one region, and then we test the predictability of the inferred weights on the remaining four regions. We perform this training/testing method for all possible (five) training choices.

TABLE II: Linear regression using the markers of our choice - An element (i, j) denotes the correlation coefficient between the weighted flu-scores time series and HPA's flu rates on region j, after training the weights on region i. The p-value for all the correlations is $<$ 10e-32.

Train/Test	A	B	C	D	E	Avg.
A	-	0.8389	0.9605	0.9539	0.9723	**0.9314**
B	0.7669	-	0.8913	0.9487	0.8896	0.8741
C	0.8532	0.702	-	0.8887	0.9445	0.8471
D	0.8929	0.9183	0.9388	-	0.9749	0.9312
E	0.9274	0.8307	0.9204	0.9749	-	0.9134
					Total Avg.	**0.8915**

The linear correlation coefficient between the inferred and the official time series for the HPA flu rates is used as the performance indicator. The results are presented in Table II; the correlation coefficients that were retrieved after training on a region A-E are presented in the row A-E respectively. The average performance over all possible training and testing choices is equal to 89.15%. The p-values of all the presented correlations indicate strong statistical significance (all of them are $<$ 10e-32). The maximum average performance is achieved when using tweets from region A (Central England & Wales) for training, and is equal to 93.14%; the linear correlation of the flu-scores' time series and the HPA flu rates for region E by applying the weights learnt from region A is equal to 97.23%.

To assess the predictive power of the former result differently, we perform linear regression on the aggregated time series of the flu-scores and HPA's flu rates using the data from all the regions. The data which belong in weeks 28 and 41 (during the peak and the stabilised period of the epidemic respectively) form the test set; the remaining data are used for training the weights. This results to a linear correlation of 92.34% with a p-value of 5.61e-30 on the test set. Additionally, we perform a 10-fold cross validation (1000 repeats, where the folds are randomly decided each time) using again linear regression for learning. On average, we obtain a linear correlation of 94.12% with a standard deviation equal to 1.54%.

While in all cases the score was tested on unseen data, in the first set of experiments we trained on data gathered on one region, and then tested on the remaining regions, but on the same period of time; in the last two experiments, using an aggregation of our data sets, we carried out training and testing on different times. Together these two sets of experiments, provide strong support to the predictive power of the flu-score we developed.

D. Automatic extraction of ILI textual markers

In the previous sections, we made use of hand crafted ILI related textual markers. In this section, we present a method for extracting weighted markers (or features) automatically. The method selects a subset of keywords and their weights to maximise the correlation with the HPA flu rates, while also

minimising the size of the keyword set. It is formed of 2 parts: creating a set of candidate features, and then selecting the most informative ones.

At first, we create a pool of candidate markers from web articles related to influenza. We use an encyclopedic reference[4] as well as a more informal reference where potential flu patients discuss their personal experiences[5]. After preprocessing (tokenisation, stop-word removal), we extract a set of $K = 1560$ stemmed candidate markers (1-grams). The latter is denoted by $\mathcal{M}_C = \{m_{ci}\}$, $i \in [1, K]$. $\mathcal{M}_C$ contains words which form a very good description of the topic as well as many irrelevant ones.

After forming the candidate features, we compute their daily, regional, and unweighted flu-subscores $f(\mathcal{T}_r, m_{ci})$ given $\mathcal{T}_r$ which denotes the Twitter corpus for region r, $r \in \{\text{A-E}\}$. For a day d, the flu score on Twitter is represented as a vector $\mathcal{F}_{d,r} = [f(\mathcal{T}_r, m_{c1}) \ ... \ f(\mathcal{T}_r, m_{cK})]^T$. Consequently, for a region r and a period of ℓ days, we can form an array with the time series of the flu-subscores for all the candidate features: $X_r = [\mathcal{F}_{1,r} \ ... \ \mathcal{F}_{\ell,r}]^T$, where ℓ denotes the total number of days considered. The columns of X_r, *i.e.* the time series of the flu-subscores of each candidate feature, are smoothed using a 7-point moving average (as in the previous cases); the resulting array is denoted as $X_r^{(s)}$.

The expanded and smoothed time series of the HPA's flu rates for region r and for the same period of ℓ days are denoted by the vector $h_r^{(s)}$. At this point, one could use the correlation coefficient between each column of $X_r^{(s)}$ and $h_r^{(s)}$ or other linear regression methods (least squares, rigde regression, etc.) in order to rank or learn weights for the candidate features. For this purpose, the LASSO method has been chosen as it has the advantage of producing sparse solutions, *i.e.* it will discard candidate features which are proven to be redundant in terms of predictability [8]. LASSO is an established method for estimating least squares parameters subject to an L1 penalty. It can be considered as a constrained optimisation task, which in our case is formulated as

$$\begin{aligned} \min_{w} \quad & \|X_r^{(s)} w - h_r^{(s)}\|_2^2 \\ \text{s.t.} \quad & \|w\|_1 \leq t, \end{aligned} \tag{6}$$

where vector w is the sparse solution, and t is the shrinkage parameter. The shrinkage parameter can be expressed as

$$t = \alpha \times \|w^{(ls)}\|_1, \tag{7}$$

where $w^{(ls)}$ denotes the least squares estimates for our regression problem, and $\alpha \in (0, 1)$ is the shrinkage percentage.

We use time series of a region $r_i \in \{\text{A-E}\}$ as the training set, the time series of a region $r_j \in \{\{\text{A-E}\} - r_i\}$ as the validation set for deciding the optimal shrinkage percentage α, and we test on the data of the remaining three regions. We repeat this procedure for all possible five training set choices. LARS algorithm is applied to compute LASSO's estimates [9]. The results of our method are captured in Table III. Most of the possible training/validating choices lead to high linear correlations. The average linear correlation over all possible settings is 92.56% indicating the robustness of our method. The experiments showed that the optimal choice was to train on region A and use region B for validating α, leading to an average correlation of 95.94% on the remaining three regions (C-E) (for a shrinkage percentage α equal to 87%). Figures 4(a), 4(b), and 4(c) show a comparison between the inferred and HPA's flu rates time series on regions C-E respectively (for the optimal choice). The learnt weights vector w had 97 non-zero values, *i.e.* we were able to extract 97 markers (or features), which, in turn, are presented in Table IV. The majority of the markers is pointing directly or indirectly to illness related vocabulary.

We also assess the performance of our method differently, following the same principle as in the previous section. We aggregate our regional data sets $X_r^{(s)}$ and $h_r^{(s)}$, and as before, we form a test set by using the data for weeks 28 and 41, a

[4]Influenza on Wikipedia, http://en.wikipedia.org/wiki/Influenza.

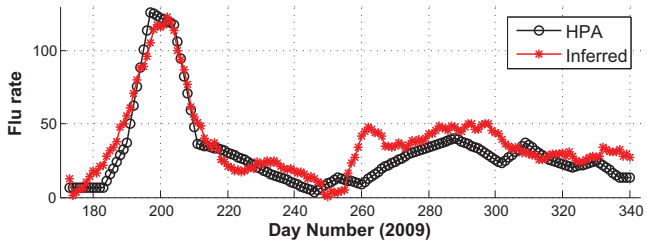
[5]Swine Flu on NHS (with potential patients comments), http://www.nhs.uk/Conditions/pandemic-flu/Pages/Symptoms.aspx.

TABLE III: Linear correlations on the test sets after performing the LASSO - An element (i, j) denotes the average correlation coefficient on the three remaining regions, after performing LASSO on region i in order to learn the weights, and validating the shrinkage parameter t on region j.

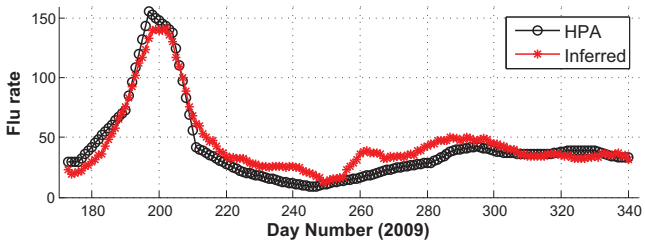
Train/Validate	**A**	**B**	**C**	**D**	**E**
A	-	**0.9594**	0.9375	0.9348	0.9297
B	0.9455	-	0.9476	0.9267	0.9003
C	0.9154	0.9513	-	0.8188	0.908
D	0.9463	0.9459	0.9424	-	0.9337
E	0.8798	0.9506	0.9455	0.8935	-
			Total Avg.		**0.9256**

TABLE IV: 97 stemmed markers extracted by applying LASSO regionally. The markers are sorted in a descending order based on their weights (read horizontally, starting from the top-left corner).

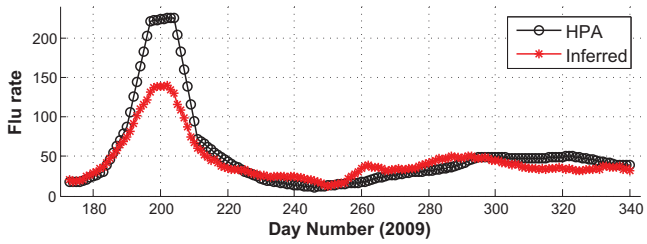
lung	unwel	temperatur	like	headach	season
unusu	chronic	child	dai	appetit	stai
symptom	spread	diarrhoea	start	muscl	weaken
immun	feel	liver	plenti	antivir	follow
sore	peopl	nation	small	pandem	pregnant
thermomet	bed	loss	heart	mention	condit
high	group	tired	import	risk	carefulli
work	short	stage	page	diseas	recognis
servic	wors	case	similar	term	home
increas	exist	ill	sens	counter	better
cough	vomit	earli	neurolog	catch	onlin
fever	concern	check	drink	long	far
consid	ach	breath	flu	member	kidnei
mild	number	sick	throat	famili	water
read	includ	swine	confirm	need	nose
medic	phone	cancer	disord	unsur	suddenli
runni					



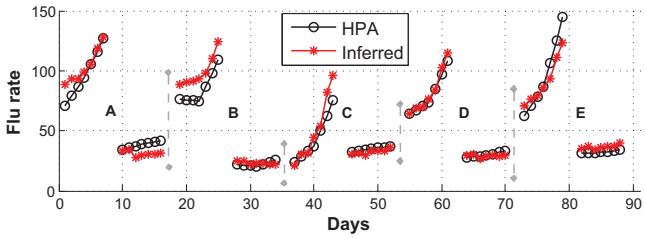
(a) Region C - Correlation: 93.49% (p-value: 1.39e-76)



(b) Region D - Correlation: 96.77% (p-value: 2.98e-101)



(c) Region E - Correlation: 97.55% (p-value: 3.85e-111)



(d) Inference on the aggregated data set for weeks 28 and 41 - Correlation: 97.13% (p-value: 3.96e-44)

Fig. 4: Comparison of the inferred versus the HPA's flu rates, after using LASSO method for learning. Figures 4(a), 4(b), and 4(c) present the results for the regional experiment (training on one region, validating on another, and testing on the remaining three), whereas 4(d) shows the results for the aggregated sets (5 partitions of a 14-day period each from left to right correspond to the results for regions A-E respectively). Note that for subfigure 4(d) the 'Days' axis does not denote yearly day numbers.

validation set (for deciding the optimal value of the shrinkage percentage α) by using weeks 36 and 49, and a training set with the remaining data sets. The outcome of the experiment indicates that the optimal value for α is 0.0049; for this value we retrieve a linear correlation of 97.13% (p-value is equal to 3.96e-44) on the test set. The corresponding vector of weights w has 73 non-zero features which are shown in Table V. Only one of them (the stemmed word 'pleas') was not included in the previously extracted set of features. Figure 4(d) presents a comparison of the inferred versus the HPA's flu rates for all the test points. Again, we have demonstrated how a list of markers can be automatically inferred from a large set of candidates by using a supervised learning algorithm and HPA's index as the target signal; this approach delivers a correlation greater than 97% with the target signal on unseen data.

TABLE V: 73 stemmed markers extracted by applying LASSO on the aggregated data set of regions A-E. The markers are sorted in a descending order based on their weights (read horizontally, starting from the top-left corner). 72 of them have also occurred in Table IV but here the order (*i.e.* the corresponding weights) is different.

muscl	like	appetit	read	unwel	child
work	follow	season	page	throat	nose
check	suddenli	pleas	immun	phone	swine
sick	dai	symptom	consid	sens	breath
cough	loss	recognis	peopl	number	mild
home	condit	mention	servic	runni	member
wors	diseas	diarrhoea	high	short	onlin
pregnant	small	exist	headach	unsur	cancer
stai	concern	fever	earli	tired	carefulli
import	weaken	nation	famili	similar	temperatur
feel	ach	flu	case	sore	unusu
spread	vomit	ill	thermomet	pandem	increas
stage	far				

IV. Related Work

Similar studies have been performed using the content of web search queries. In [10] the frequency of influenza-related queries on Yahoo! search engine has been proven to be correlated with influenza and mortality rates in the United States, whereas in [5] a representative set of user search queries has been extracted by applying a linear regression fit with official health reports, achieving on average a linear correlation of 90%. Similarly, user queries on a Swedish medical website have been used to learn ILI rates in Sweden [11].

Our study makes use of independent data, and hence can be used both as an alternative source of information, or combined with other data sources (or methods), to achieve an even higher accuracy. Used in a stand-alone manner, our method can be considered as more reliable since it performs better than the aforementioned ones.

Other recent studies have attempted to extract epidemic information from textual data streams originating in social media, mostly "traditional" blogs, but also Twitter. The study reported in [12] deals with a similar question to ours, but with the important difference that it is based on blogs (where geolocation is not very accurate) and simply counts occurrences

of only 2 keywords achieving a linear correlation of 76.7% with the official ILI rates. In [13], where the authors mainly present an infrastructure for collecting data from Twitter, it is reported that the increasing numbers of tweets with illness oriented vocabulary might indicate an emerging epidemic, but there is no comparison with any ground truth in order to validate this. Neither of these methods uses machine learning to automatically select the features and learn a flu-scoring function.

Various online tools[6,7], created to detect "chatter" about a disease are somewhat related to our work, in that they do collect twitter or web media data, but again there is no use of ground truth or machine learning, and hence they also act mostly as news aggregators.

V. Conclusions and Future Directions

We have presented a method for tracking the flu epidemic in the UK by using the contents of Twitter; our approach could give an early warning in various situations, but mostly can give timely and free information to health agencies to plan health care. This method is based on the textual analysis of micro-blog contents, is calibrated with the ground truth provided by the HPA, and could also be extended to the contents of text messages sent via mobile devices (besides privacy concerns), as the latter are of the same format with tweets (140 characters limitation).

Our approach can automatically learn the inputs (textual markers and their weights) of a scoring function that correlates very highly with the HPA flu score ($> 95\%$). This method only requires access to the time series of geolocated blog contents and a ground truth. It is entirely language independent, and we are planning to extend this to other languages soon.

Differently from [5], where only data from traditional flu seasons have been used for a 5-year period, this paper is based on 6 months of data (June to December, 2009). Like that study our method can potentially be affected by panic or other factors that push people to post about illness-related symptoms. Differently from search engine logs, in this type of data, we can isolate more informative "self diagnostic" statements, that are less likely to be caused by general panic or discussions about the flu.

If we aim at predicting HPA's rates, we may still need to separate media hype and discussions about the flu from reporting of actual flu cases, which is what we are trying to count. In this case, it is likely that just counting the word "flu" (like most existing systems do) will be much more subject to hype than detecting statements about symptoms, *e.g.* "I have a fever". This is why a learning system should be trained to automatically discover which keywords are more useful to predict the ground truth rates. Indeed, in our system most words relate to symptoms, not just to discussions of flu.

Future work will involve the exploitation of geographic information and also include the integration of other data sources, for example weather, to improve the accuracy of predictions. A generalised version of this method can also be applied to generate automatically the most informative "diagnostic markers", that can allow us to monitor more than one epidemic at once (if their symptoms are different) in various countries independently of their language. The general concept of this work is an application of open source intelligence [14], which could also be applied for learning tendencies in different types of contexts such as politics, finance, and public opinion.

[6]Monitoring US health using Twitter, http://www.cs.uiowa.edu/~asignori/projects/twitter-monitor-us-health/.

[7]Medical Information System (MedISys), http://medusa.jrc.it/medisys/.

Acknowledgments

Vasileios Lampos wishes to thank Nokia Research and EPSRC (DTA/SB1826) for various degrees of support of his research. Nello Cristianini is supported by a Royal Society Wolfson Merit Award. Both authors would like to thank Mark Schmidt for providing an implementation of LASSO LARS algorithm for MATLAB as well as Tijl de Bie and Elena Hensinger for their constructive feedback.

References

[1] K. Neuzil, C. Hohlbein, and Y. Zhu, "Illness among schoolchildren during influenza season: effect on school absenteeism, parental absenteeism from work, and secondary illness in families," *Archives of pediatrics & adolescent medicine*, vol. 156, no. 10, pp. 986–991, 2002.

[2] A. Elliot, C. Powers, A. Thornton, C. Obi, C. Hill, I. Simms, P. Waight, H. Maguire, D. Foord, E. Povey *et al.*, "Monitoring the emergence of community transmission of influenza A/H1N1 2009 in England: a cross sectional opportunistic survey of self sampled telephone callers to NHS Direct," *British Medical Journal*, vol. 339, p. b3403, 2009.

[3] D. Ivnitski, I. Abdel-Hamid, P. Atanasov, and E. Wilkins, "Biosensors for detection of pathogenic bacteria," *Biosensors and Bioelectronics*, vol. 14, no. 7, pp. 599–624, 1999.

[4] T. Metcalf, J. Melnick, and M. Estes, "Environmental virology: from detection of virus in sewage and water by isolation to identification by molecular biology-a trip of over 50 years," *Annual Reviews in Microbiology*, vol. 49, no. 1, pp. 461–487, 1995.

[5] J. Ginsberg, M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, pp. 1012–1014, 2008.

[6] M. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[7] D. Fleming and A. Elliot, "Lessons from 40 years' surveillance of influenza in England and Wales," *Epidemiology and Infection*, vol. 136, no. 7, pp. 866–875, 2007.

[8] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[9] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of statistics*, vol. 32, no. 2, pp. 407–451, 2004.

[10] P. Polgreen, Y. Chen, D. Pennock, F. Nelson, and R. Weinstein, "Using internet searches for influenza surveillance," *Clinical Infectious Diseases*, vol. 47, no. 11, pp. 1443–1448, 2008.

[11] A. Hulth, G. Rydevik, and A. Linde, "Web queries as a source for syndromic surveillance," *PLoS ONE*, vol. 4, no. 2, p. e4378, 2009.

[12] C. Corley, A. Mikler, K. Singh, and D. Cook, "Monitoring influenza trends through mining social media," *International Conference on Bioinformatics & Computational Biology*, pp. 340–346, 2009.

[13] L. Lopes, J. Zamite, B. Tavares, F. Couto, F. Silva, and M. Silva, "Automated social network epidemic data collector," *INForum - Simpósio de Informática*, pp. 263–272, 2009.

[14] F. Stalder and J. Hirsh, "Open source intelligence," *First Monday*, vol. 7, no. 6, 2002.