Tracking the flu pandemic by monitoring the Social Web

Vasileios Lampos and Nello Cristianini

Intelligent Systems Laboratory, University of Bristol, UK

✓I. What Is This All About?

We present a method for **measuring 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** that quantifies the diffusion of influenza-like illness (ILI) in various regions of the UK. This method uses completely independent data to that commonly used for these purposes, and can be applied at close time intervals, hence providing inexpensive and timely information about the state of an epidemic.

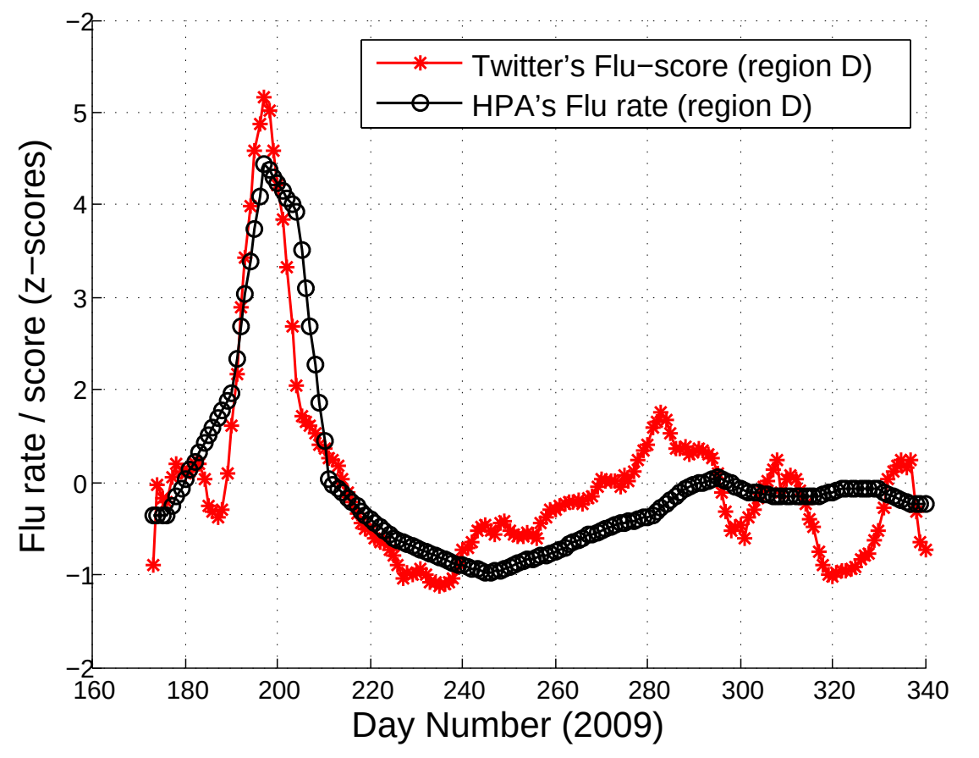
✓II. DataSets

From 22/06 to 06/12 (weeks 26-49, 2009) we were collecting:

▶ a daily average of 160,000 **tweets** geolocated in the 54 most populated urban centres in the UK

▶ weekly **reports from the Health Protection Agency** (HPA) for 5 UK's regions (denoted by r), where r∈{A-E}. The reports express the number of GP consultations per 10^5 citizens, where the result of the diagnosis was ILI. For retrieving an equal representation between the weekly HPA flu rates and Twitter's daily vector space representations, we expand each point of the former over a 7-day period. After expanding the HPA flu rates, we perform smoothing on them with a 7-point moving average.

✓IV. The Starting Point



Twitter *vs.* HPA flu scores (region D)

Using a small set of 41 textual markers expressing illness related terminology, *e.g.* fever, sore throat, headache, cough, infection, etc., we compute the Twitter flu-score time series for regions A-E. The linear correlation coefficients between Twitter's and HPA flu-score time series are between 80% and 86%, ∀r.

Acknowledgements

We would like to thank **Twitter Inc.** for making its data publicly available. This work is partially supported by European Commission through the **PASCAL2** NoE (FP7-216866). V. Lampos is supported by **EPSRC** (DTA/ SB1826) and **NOKIA Research**. N. Cristianini is supported by a **Royal Society Wolfson Merit Award**.

✓III. Notation

▶ Set of textual **markers**: $\mathcal{M} = \{m_i\}$, $i \in [1,k]$

▶ Their respective **weights**: $\mathcal{W} = \{w_i\}$, $i \in [1,k]$

▶ Set of **tweets**: $\mathcal{T} = \{t_j\}$, $j \in [1,n]$

▶ Function for forming vector space representations:

$$m_i(t_j) = \begin{cases} 1 & \text{if } m_i \text{ appears in } t_j \\ 0 & \text{otherwise} \end{cases}$$

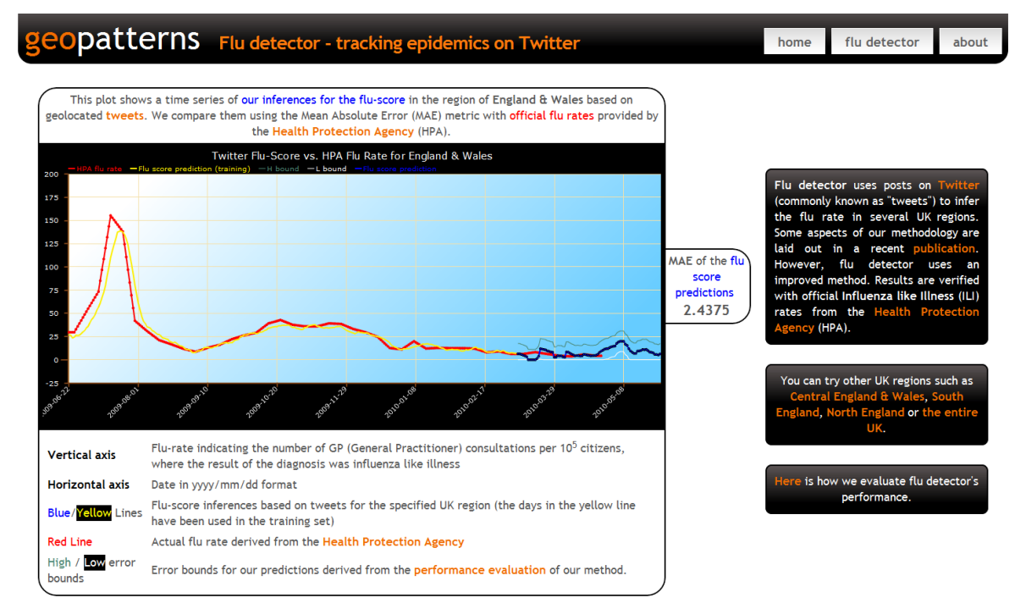
▶ **Flu-score** of a set of tweets $\mathcal{T}$:

$$f_{\mathcal{W}}(\mathcal{T}, \mathcal{M}) = \frac{\sum_j \sum_i w_i \times m_i(t_j)}{k \times n}$$

▶ **Flu-subscore** of a marker m_i:

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

Flu Detector



This plot shows a time series of our inferences for the flu-score in the region of England & Wales based on geolocated tweets. We compare them using the Mean Absolute Error (MAE) metric with official flu rates provided by the Health Protection Agency (HPA).

MAE of the flu score predictions 2.4375

Vertical axis	Flu-rate indicating the number of GP (General Practitioner) consultations per 10^5 citizens, where the result of the diagnosis was influenza like illness
Horizontal axis	Date in yyyy/mm/dd format
Blue/Yellow Lines	Flu-score inferences based on tweets for the specified UK region (the days in the yellow line have been used in the training set)
Red Line	Actual flu rate derived from the Health Protection Agency
High / Low error bounds	Error bounds for our predictions derived from the performance evaluation of our method.

Flu detector uses posts on Twitter (commonly known as "tweets") to infer the flu rate in several UK regions. Some aspects of our methodology are laid out in a recent publication. However, flu detector uses an improved method. Results are verified with official Influenza like Illness (ILI) rates from the Health Protection Agency (HPA).

You can try other UK regions such as Central England & Wales, South England, North England or the entire UK.

Here is how we evaluate flu detector's performance.

Flu detector is a tool we developed that uses an improved methodology (which includes the application of Bolasso [2], the bootstrap version of LASSO) to make **live**, daily predictions for the ILI rates in several UK regions based on the contents of Twitter (ECML PKDD 2010).

URL: geopatterns.enm.bris.ac.uk/epidemics/

References

[1] R. Tibshirani Regression shrinkage and selection via the lasso. In Journal of the Royal Statistical Society 58B, 267–288 (1996).

[2] F.R. Bach Bolasso: model consistent Lasso estimation through the bootstrap. ICML 25, 33–40 (2008).

✓V. Methodology

▶ **Form a pool of candidate markers** (features) from web pages related to influenza – we use an encyclopedic reference from Wikipedia and a more informal reference from the NHS website where potential flu patients discuss their personal experiences. We extract a set of $K = 1560$ stemmed candidate markers denoted by $\mathcal{M}_C = \{m_{ci}\}$, $i \in [1, K]$.

▶ **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.

▶ For a day d, Twitter's regional flu-score is represented as a vector

$$\mathcal{F}_{d,r} = [f(\mathcal{T}_r, m_{c1}) \ \ldots \ f(\mathcal{T}_r, m_{cK})]^T.$$

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 markers:

$$X_r = [\mathcal{F}_{1,r} \ \ldots \ \mathcal{F}_{\ell,r}]^T.$$

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 - 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 $h_r^{(s)}$.

▶ **LASSO** [1] is an established method for estimating least squares parameters subject to an L1 penalty. In our case, it is formulated as

$$\min_{w} \ \|X_r^{(s)} w - h_r^{(s)}\|_2^2$$
$$\text{s.t.} \quad \|w\|_1 \leq t,$$

where vector w is expected to be a sparse solution (therefore *feature selection* is performed as well), and t is LASSO's **shrinkage parameter**. The shrinkage parameter can be expressed as

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

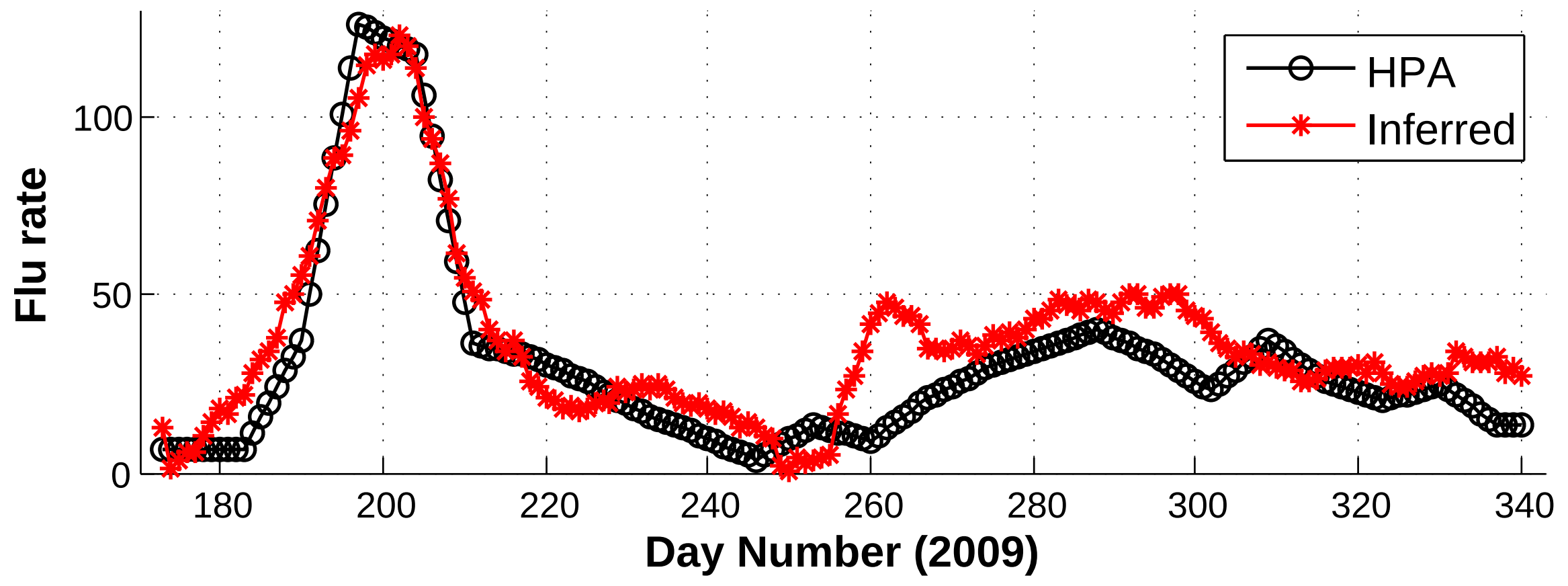
where $w^{(ls)}$ denotes the least squares estimates for our regression problem, and $\alpha \in (0, 1)$ is the shrinkage percentage.

✓VI. Validation & Results

▶ Train on $X_{r_i}^{(s)}$, r_i∈{A-E}, validate LASSO's shrinkage parameter on $X_{r_j}^{(s)}$, $r_j \in \{\{\text{A-E}\} - r_i\}$, and test on the remaining regional time series.

Train/Validate	**A**	**B**	**C**	**D**	**E**
A	-	**0.95**	0.93	0.93	0.92
B	0.94	-	0.94	0.92	0.90
C	0.91	0.95	-	0.81	0.90
D	0.94	0.94	0.94	-	0.93
E	0.87	0.95	0.94	0.89	-
			Total Avg.		**0.92**

Here is a comparison of the inferred flu scores with the official flu rates; region A is used for training, region B for validating the shrinkage parameter, and testing is done on region C:

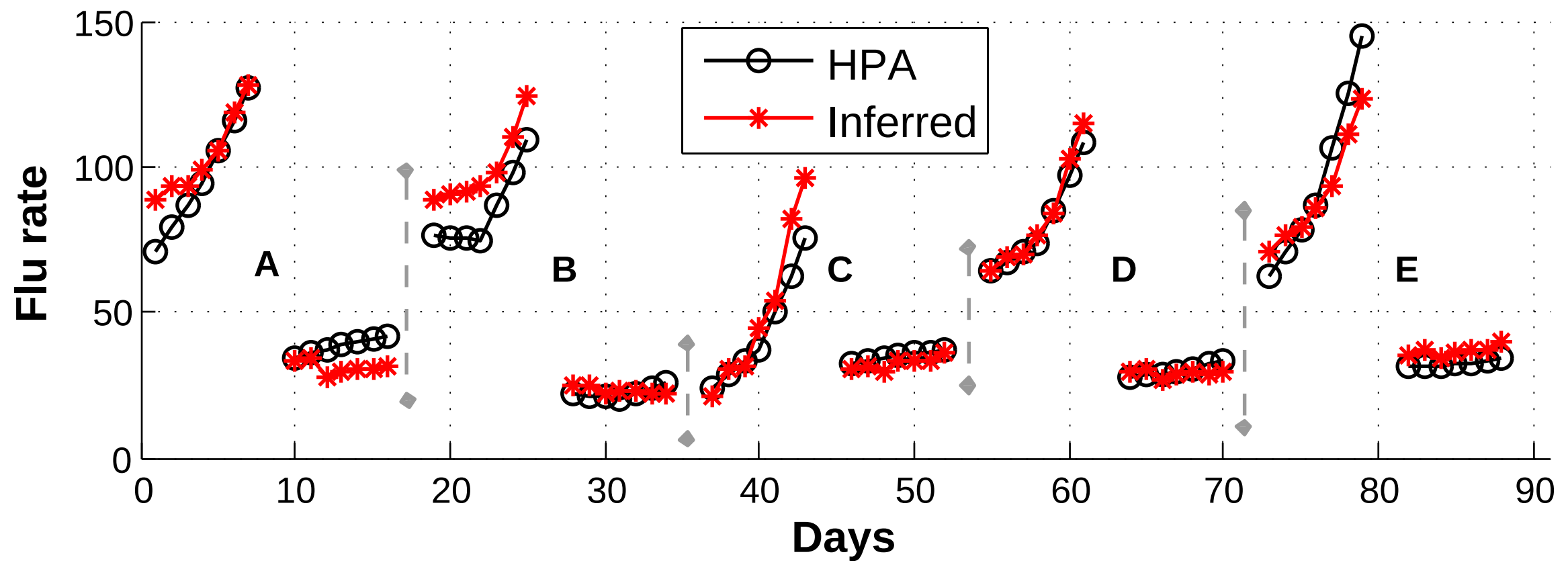


Correlation: 93.49% (p-value: 1.39e-76)

▶ Aggregate all regional time series, use 2 weeks (28 & 41) per region for testing, 2 weeks (36 & 49) per region for validating, and the remaining weeks for training. Our method selects the following 73 stemmed markers (in a descending weight order):

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

Here is a comparison of the inferred flu scores with the official flu rates for all five regions (2 weeks are tested per region):



Correlation: 97.13% (p-value: 3.96e-44)