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.

1. What Is This All About?

2. DataSets

From 23/06 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 the daily vector space representations, we compute the Twitter corpus for region r.

3. Notation

- Set of textual markers: M = {m_i}, i∈[1,K]
- Their respective weights: W = {w_i}, i∈[1,K]
- Set of tweets: T = {t_j}, j∈[1,n]
- Function for forming vector space representations:
  
  \[ u = \max(0,1 - \text{flu-score}(t_j)) \]
  
  - Flu-score of a set of tweets T:
    
    \[ f_T(M) = \sum_{i,j} w_i m_i(t_j) \]
  
  - Flu-subscore of a marker m_i:
    
    \[ f_{m_i}(T) = \sum_{i,j} m_i(t_j) \]

4. The Starting Point

Using a small set of 41 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.

5. 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 M_r = {m_i}, i∈[1,K].

- Compute their daily, regional, and unweighted flu-subscores \( f(T_r, m_i) \) given \( T_r \) which denotes the Twitter corpus for region r.

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

  \[ F_{d,r} = [f(T_r, m_1), \ldots, f(T_r, m_K)]^T. \]

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:

\[ X_r = [F_{d_1,r}, \ldots, F_{d_ℓ,r}]^T. \]

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 \( X_r^{(7)}. \)

- 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 \( h_r^{(7)}. \)

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

  \[
  \min_w \|X_r^{(7)}w - h_r^{(7)}\|^2 \quad \text{s.t.} \quad \|w\|_1 \leq t,
  \]

  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

  \[ t = \alpha \times \|w^{(0)}\|_1, \]

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

VI. Validation & Results

- Train on \( X_r^{(7)}, \ r \in \{A-E\} \), validate LASSO’s shrinkage parameter on \( X_r^{(7)}, \ r \in \{A-E\} - r \), and test on the remaining regional time series.

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:

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.

### 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 (FP7-216686). V. Lampos is supported by EPSRC (DTA/ SB1826) and NOKIA Research. N. Cristianini is supported by a Royal Society Wolfson Merit Award.

### References