Detecting Events and Patterns in the Social Web with Statistical Learning

Vasileios Lampos

Computer Science Department
University of Sheffield

Outline

- $\perp$ **Motivation, Aims**
- $\perp$ **Data**
- $\dashv$ **Nowcasting Events from the Social Web**
- $\dashv$ **Extracting Mood Patterns from the Social Web**
- $\models$ **Conclusions**

Facts

We started to work on this idea in 2008, when...

- **Web** contained **1 trillion** unique pages (Google)
- **Social Networks** were rising, *e.g.*
  - *Facebook*: 100m users in 2008, **955m** in 2012 (June)
  - *Twitter*: 6m users in 2008, **500m** active users in 2012 (April)
- **User behaviour** was changing
  - Socialising via the Web
  - Giving up privacy (Debatin *et al.*, 2009)

Questions

- Does user generated text posted on Social Web platforms include **useful information**?

- How can we **extract** this useful information...

  ... **automatically**? Therefore, not we, but a **machine**.

- Practical / real-life **applications**?

- Can those large samples of human input **assist studies in other scientific fields**?
  Social Sciences, *Psychiatry*...

One slide on @Twitter. What does a 'tweet' look like?

Figure 1: Some biased and anonymised examples of tweets (limit of **140 characters**/tweet, # denotes a **topic**)

Why do I feel so happy today hihi. Bedtimeeee, good night. Yey thank You Lord for everything. Answered prayer ♥

Reply Retweet Favorite

(a) (user will remain anonymous)

RT if you love Justin Bieber. Delete ur account if you don't.

Reply Retweet Favorite

50 RETWEETS **1** FAVORITE

(b) they live around us

another demo covered by citizens today in Thessaloniki int'l fair. Citizen journalism on a speed rise in #Greece. check #deth and #rbnews

Reply Retweet Favorite

(c) citizen journalism

i think i have the flu but i still look fabulous

Reply Retweet Favorite

(d) flu attitude

Data Collection

- Considered to be the **easiest part** of the process...
  ... **not true**!
  - Storage space
  - Crawler implementation, parallel data processing
  - Equipment, new technologies (*e.g.* Map-Reduce)

- **Data** collected and used in the following experiments
  - **tweets** *geo-located* in 54 urban centres in the UK
  - **collected periodically** (every 3 or 5 minutes per urban centre)
  - approx. **0.5** billion tweets by **10** million users (06/2009 to 01/2012)
  - **ground truth** (regional flu & local rainfall rates)

Nowcasting Events from the Social Web

'Nowcasting'?

We do not predict the future, but **infer the present** $-\ \delta$

i.e. the very recent past

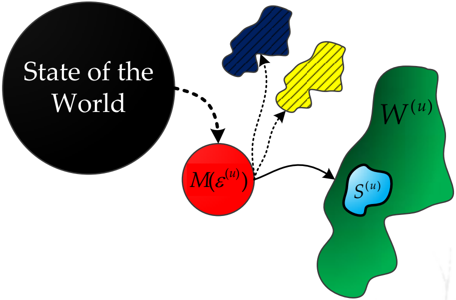


Figure 2: Nowcasting the magnitude of an event (ε) emerging in the real world from Web information

Our case studies: nowcasting (a) **flu rates** & (b) **rainfall rates (?!)**

What do we get in the end?

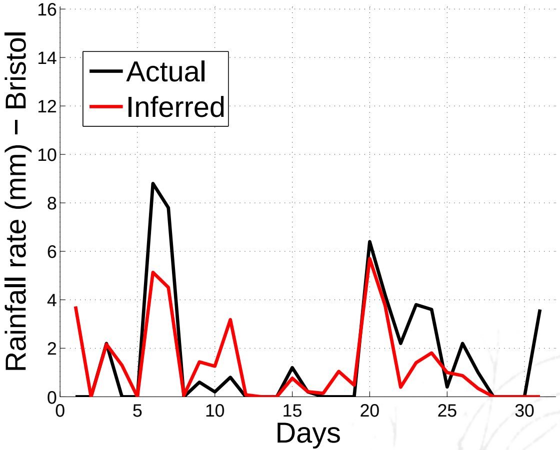


Figure 3: Inferred **rainfall rates** for Bristol, UK (October, 2009)

Core Methodology (1/3) – Turning text into numbers

Candidate features (n-grams): $\mathcal{C} = \{c_i\}$

Set of **Twitter posts** for a time interval u: $\mathcal{P}^{(u)} = \{p_j\}$

Frequency of c_i in p_j:

$$g(c_i, p_j) = \begin{cases} \varphi & \text{if } c_i \in p_j, \\ 0 & \text{otherwise.} \end{cases}$$

– g Boolean, maximum value for φ is 1 –

Score of c_i in $\mathcal{P}^{(u)}$:

$$s\left(c_i, \mathcal{P}^{(u)}\right) = \frac{\sum_{j=1}^{|\mathcal{P}^{(u)}|} g(c_i, p_j)}{|\mathcal{P}^{(u)}|}$$

Core Methodology (2/3)

Set of **time intervals**: $\mathcal{U} = \{u_k\} \sim$ 1 hour, 1 day, ...

Time series of candidate features **scores**:

$$\mathcal{X}^{(\mathcal{U})} = \left[x^{(u_1)} \; \dots \; x^{(u_{|\mathcal{U}|})} \right]^T,$$

where

$$x^{(u_i)} = \left[s\left(c_1, \mathcal{P}^{(u_i)}\right) \; \dots \; s\left(c_{|\mathcal{C}|}, \mathcal{P}^{(u_i)}\right) \right]^T$$

Target variable (event):

$$y^{(\mathcal{U})} = \left[y_1 \; \dots \; y_{|\mathcal{U}|} \right]^T$$

Core Methodology (3/3) – Feature selection

Solve the following **optimisation problem**:

$$\min_{w} \quad \|\mathcal{X}^{(\mathcal{U})} w - y^{(\mathcal{U})}\|_{\ell_2}^2$$

$$\text{s.t.} \quad \|w\|_{\ell_1} \leq t,$$

$$t = \alpha \cdot \|w_{\mathsf{OLS}}\|_{\ell_1}, \; \alpha \in (0, 1].$$

- Least Absolute Shrinkage and Selection Operator (**LASSO**) (Tibshirani, 1996)
- Enforce **sparsity** on w (feature selection)
- Least Angle Regression (**LARS**) – computes entire regularisation path (Efron *et al.*, 2004)

Flu rates – Example of selected features

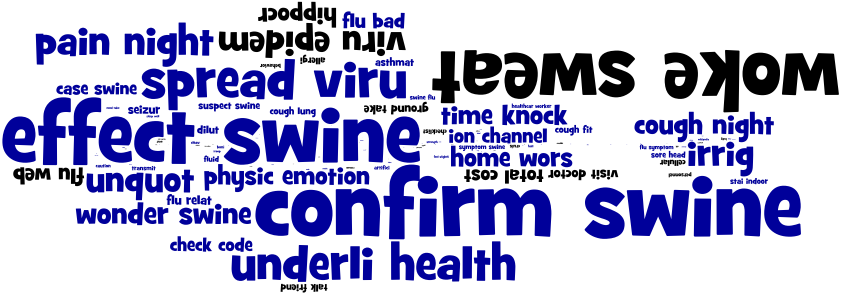


Figure 4: Font size is proportional to the weight of each feature; flipped n-grams are negatively weighted. All words are stemmed (Porter, 1980).

(Lampos and Cristianini, 2012)

Rainfall rates – Example of selected features

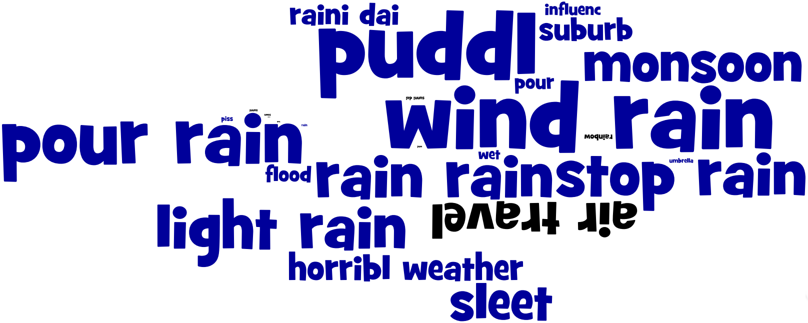
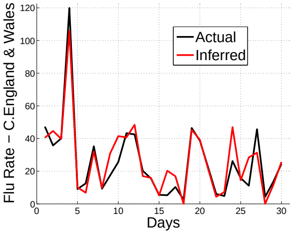


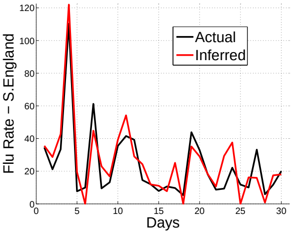
Figure 5: Font size is proportional to the weight of each feature; flipped n-grams are negatively weighted. All words are stemmed (Porter, 1980).

(Lampos and Cristianini, 2012)

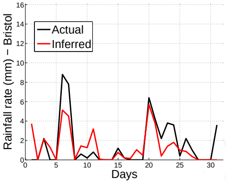
Examples of inferences



(a) Central England/Wales (flu)



(b) South England (flu)



(c) Bristol (rain)

Figure 6: Examples of flu and rainfall rates **inferences** from Twitter content (Lampos and Cristianini, 2012)

Flu Detector

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

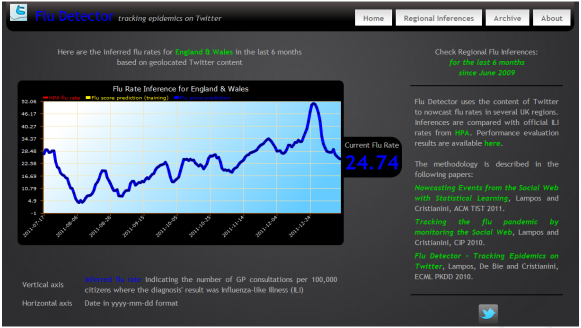


Figure 7: Flu Detector uses the content of Twitter to nowcast flu rates in several UK regions

(Lampos, De Bie and Cristianini, 2010)

Extracting Mood Patterns from the Social Web

Computing a mood score

Table 1: **Mood terms** from **WordNet Affect**

Fear	**Sadness**	**Joy**	**Anger**
afraid	depressed	admire	angry
fearful	discouraged	cheerful	despise
frighten	disheartened	enjoy	enviously
horrible	dysphoria	enthousiastic	harassed
panic	gloomy	exciting	irritate
...	...	...	...
(*92 terms*)	(*115 terms*)	(*224 terms*)	(*146 terms*)

Mood score computation for a **time interval** u using n **mood terms** and a sample of D **days**:

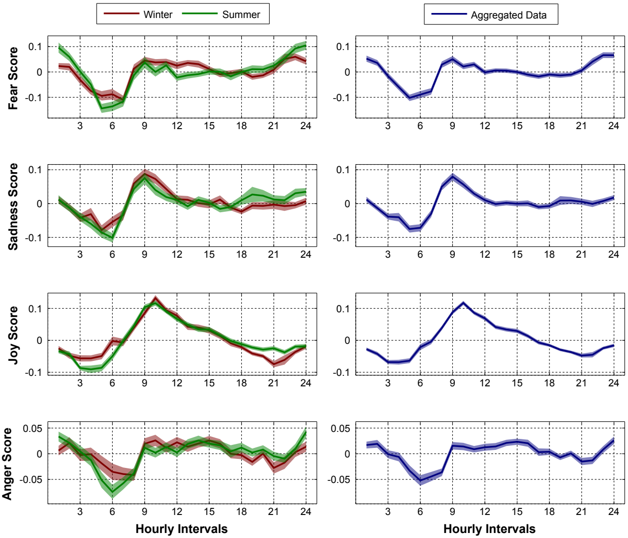
$$\mathcal{M}_s(u) = \frac{1}{|D|} \sum_{j=1}^{|D|} \left(\frac{1}{n} \sum_{i=1}^{n} sf_i^{(t_{j,u})} \right)$$

$$sf_i^{(t_{d,u})} = \frac{f_i^{(t_{d,u})} - \bar{f}_i}{\sigma_{f_i}}, \; i \in \{1, ..., n\}.$$

$f_i^{(t_{d,u})}$: normalised frequency of a mood term i during time interval u in day $d \in D$

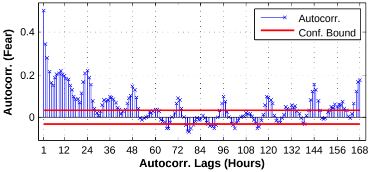
Circadian mood patterns (1/2)

Figure 8: Circadian (**24-hour**) mood patterns based on UK Twitter content

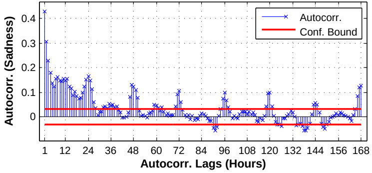


Circadian mood patterns (2/2)

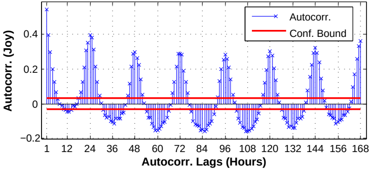
Figure 9: **Autocorrelation** of circadian mood patterns based on **hourly lags** revealing periodicities



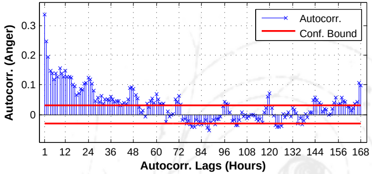
(a) Fear



(b) Sadness



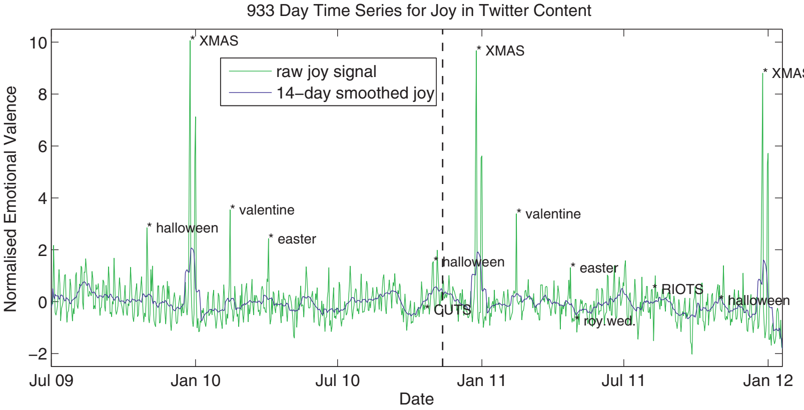
(c) Joy



(d) Anger

The mood of the nation (1/4)

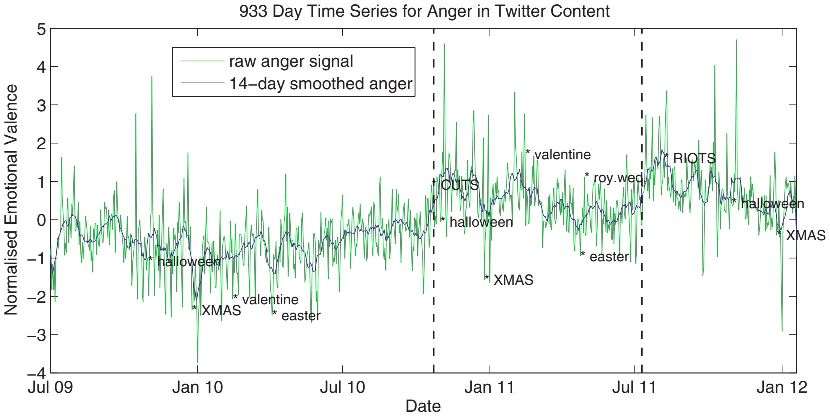
Figure 10: Daily time series for the mood of **Joy** based on Twitter content geo-located in the **UK**



(Lansdall, Lampos and Cristianini, 2012)

The mood of the nation (2/4)

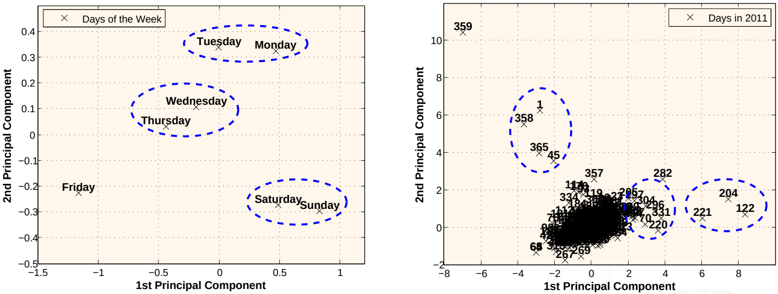
Figure 11: Daily time series for the mood of **Anger** based on Twitter content geo-located in the **UK**



(Lansdall, Lampos and Cristianini, 2012)

The mood of the nation (3/4)

Figure 12: Projections of 4-dimensional mood signals on their top-2 principal components (based on 2011 Twitter content)



(a) Days of the week (2011)

(b) Days of the year (2011)

Days **1**/**45**/**358**/**365**: New Year's / Valentine's / Christmas Eve / New Year's Eve

Days **122**/**204**/**221**: O.B. Laden's death / Winehouse's death, Breivik / UK riots

(Lampos, 2012a)

The mood of the nation (4/4)

URL: `http://geopatterns.enm.bris.ac.uk/mood`

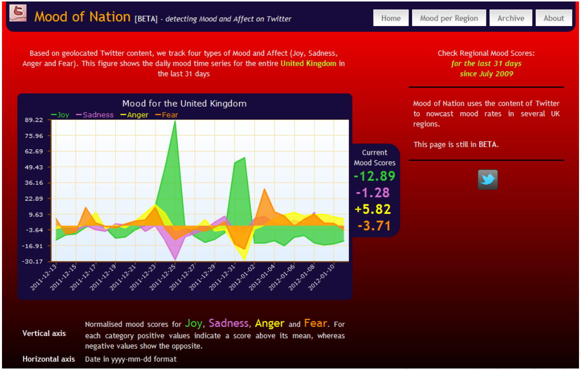
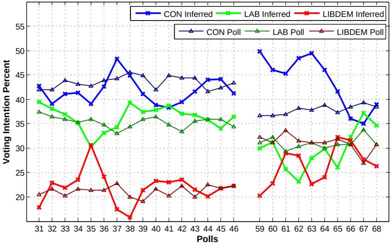


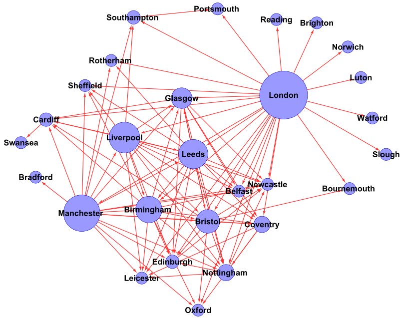
Figure 13: Mood of the Nation uses the content of Twitter to nowcast mood rates in several UK regions

(Lampos, 2012a)

More applications (snapshots)



(a) Inferences of voting intention polls prior to the UK 2010 General Election



(b) Content similarity network

Figure 14: Further information extraction examples from Twitter content

(Lampos, 2012a & 2012b)

Not covered

Amongst the things you **didn't see**:

- how the **model inconsistency** problems of LASSO are resolved
- different schemes for **combining 1**-grams and **2**-grams
- **performance metrics** and comparison with baseline techniques or other **nonlinear, nonparametric** learners
- further **statistical analysis** and **psychiatric viewpoint** of circadian mood patterns
- comparison of different **scoring functions** for mood signals

Conclusions

- **Social Web** holds **valuable information**
- **interesting inferences** can be made by applying **statistical methods** on Twitter (user-generated) content
- machines can extract portions of this information **automatically**
  - **nowcasting events** (flu and rainfall case studies)
  - extraction of **collective mood** patterns

Currently participating in the **TrendMiner** EU-FP7 project.
How **user-generated web content** can be used to...

- model **political opinion**
- infer **voting intention** polls, **election**/**referendum** outcome
- nowcast/predict **financial indicators**

Last Slide!

The end.

Any questions?

Download the slides from
`http://goo.gl/KZRke`

References

1. B. Debatin, J.P. Lovejoy, A.M.A. Horn, B.N. Hughes. **Facebook and Online Privacy: Attitudes, Behaviors, and Unintended Consequences**. Journal of Computer-Mediated Communication 15, pp. 83–108, 2009.
2. V. Lampos and N. Cristianini. **Nowcasting Events from the Social Web with Statistical Learning**. ACM TIST 3(4), n. 72, 2012.
3. R. Tibshirani. **Regression Shrinkage and Selection via the LASSO**. Journal of the Royal Statistical Society, series B, 58(1), pp. 267–288, 1996.
4. B. Efron, T. Hastie, I. Johnstone and R. Tibshirani. **Least Angle Regression**. The Annals of Statistics 32(2), pp. 407–499, 2004.
5. M.F. Porter. **An algorithm for suffix stripping**. Program 14(3), pp. 130–137, 1980.
6. V. Lampos and N. Cristianini. **Tracking the flu pandemic by monitoring the Social Web**. Proceedings of CIP '10, pp. 411–416, 2010.
7. V. Lampos, T. De Bie and N. Cristianini. **Flu Detector – Tracking Epidemics on Twitter**. Proceedings of ECML PKDD '10, pp. 599–602, 2010.
8. T. Lansdall-Welfare, V. Lampos and N. Cristianini. **Nowcasting the mood of the nation**. Significance 9(4), pp. 26–28, 2012.
9. V. Lampos. **Detecting Events and Patterns in Large-Scale User Generated Textual Streams with Statistical Learning Methods**. PhD Thesis, University of Bristol, p. 243, 2012.(a)
10. V. Lampos. **On voting intentions inference from Twitter content: a case study on UK 2010 General Election**. CoRR, 2012.(b)