Can Social Media tell us something about our lives?

Vasileios Lampos

Computer Science Department
University of Sheffield

March, 2013

Outline

- $\perp$ **Motivation, Aims** [Facts, Questions]
- $\perp$ **Data**
- $\dashv$ **Nowcasting Events**
- $\dashv$ **Extracting Mood Patterns**
- $\dashv$ **TrendMiner – Extracting Political Opinion**
- $\models$ **Conclusions**

Facts

We started to work on those ideas *back* in 2008, when...

- **Web** contained **1 trillion** unique pages (Google)

- **Social Networks** were rising, *e.g.*
  - *Facebook*: 100m (2008) → >**1 billion** active users (October, 2012)
  - *Twitter*: 6m (2008) → **500m** active users (July, 2012)

- **User behaviour** was changing
  - Socialising via the Web
  - Giving up privacy (Debatin *et al.*, 2009)

Some general 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, *Psychology*, *Epidemiology*

The Data (1/3)

Why Twitter?

- Has a lot of content that is **publicly accessible**
- Provides a well-documented **API** for several types of data collection
- **Opinions** and **personal statements** on various domains
- Connection with current affairs (usually in **real-time**)
- Some content is **geo-located**
- Option for **personalised modelling**
- *... and we got good results from the very first, simple experiment!*

The Data (2/3)

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

The Data (3/3)

Data Collection & Preprocessing

- The easiest part of the process...
  - **not true**! → Storage space, crawler implementation, parallel data processing, new technologies (*e.g.*, Map-Reduce) (Preotiuc *et al.*, 2012)
- Data collected via **Twitter's Search API**:
  - **collective sampling**
  - tweets geo-located in 54 urban centres in the UK
  - periodical crawling (every 3 or 5 minutes per urban centre)
- Data collected via **Twitter's REST API**:
  - **user-centric sampling**
  - preprocessing to approximate user's location (city & country)
  - ... or manual user selection from domain experts
  - get their latest tweets (3,000 or more)
- Several forms of **ground truth** (flu/rainfall rates, polls)

Nowcasting Events from the Social Web

‘Nowcasting’?

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

i.e. the very recent past

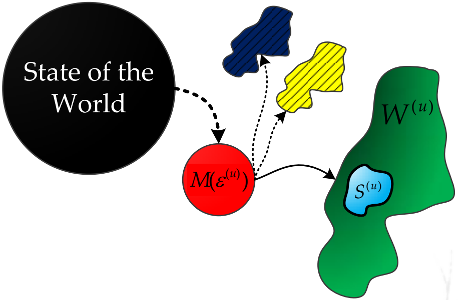


Figure 2 : Nowcasting the magnitude of an event (ε) emerging in the real world from Web information

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

What do we get in the end?

This is a **regression** problem (*text regression* in NLP)
i.e. $\forall$ time interval i we aim to infer $y_i \in \mathbb{R}$ using text input $\boldsymbol{x}_i \in \mathbb{R}^n$

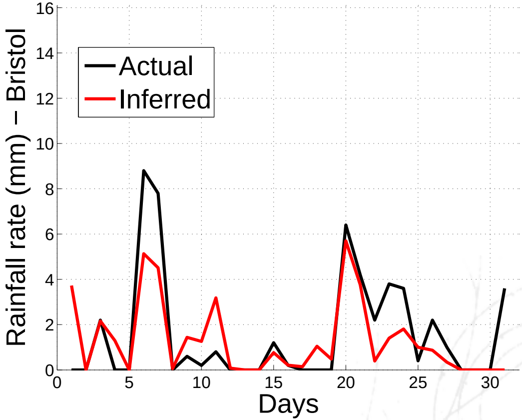


Figure 3 : Inferred **rainfall rates** for Bristol, UK (October, 2009)

Methodology (1/5) — Text in Vector Space

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 $\mathcal{P}^{(u)}$:

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

Methodology (2/5)

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

Time series of candidate features **scores**:

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

where

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

Target variable (event):

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

Methodology (3/5) — Feature selection

Solve the following **optimisation problem**:

$$\min_{w} \quad \|X^{(\mathcal{U})}\boldsymbol{w} - \boldsymbol{y}^{(\mathcal{U})}\|_{\ell_2}^2$$
$$\text{s.t.} \quad \|\boldsymbol{w}\|_{\ell_1} \leq t,$$
$$t = \alpha \cdot \|\boldsymbol{w}_{\text{OLS}}\|_{\ell_1}, \ \alpha \in (0, 1].$$

- Least Absolute Shrinkage and Selection Operator (**LASSO**)

$$\underset{\boldsymbol{w}}{\operatorname{argmin}} \|X^{(\mathcal{U})}\boldsymbol{w} - \boldsymbol{y}^{(\mathcal{U})}\|_{\ell_2}^2 + \lambda\|\boldsymbol{w}\|_{\ell_1}$$

(Tibshirani, 1996)

- Expect a **sparse *w*** (feature selection)
- Least Angle Regression (**LARS**) – computes entire regularisation path (***w***'s for different values of λ) (Efron *et al.*, 2004)

Methodology (4/5)

LASSO is **model-inconsistent**:

- inferred sparsity pattern may deviate from the true model, *e.g.*, when predictors are highly correlated (Zhao and Yu, 2006)

- bootstrap [**?**] LASSO (**Bolasso**) performs a more robust feature selection (Bach, 2008)
  ?:
  - in each bootstrap, input space is sampled with replacement
  - apply LASSO (LARS) to select features
  - select features with nonzero weights in all bootstraps

- better alternative — **soft-Bolasso**:
  - a less strict feature selection
  - select features with nonzero weights in p% of bootstraps
  - (learn p using a separate validation set)

- **weights** of selected features determined via OLS regression

Methodology (5/5) — Simplified summary

Observations: $X \in \mathbb{R}^{m \times n}$ (m time intervals, n features)

Response variable: $\boldsymbol{y} \in \mathbb{R}^{m}$

For $i = 1$ to *number of bootstraps*
  Form $X_i \subset X$ by sampling X with replacement
  Solve LASSO for X_i and $\boldsymbol{y}$, *i.e.* learn $\boldsymbol{w}_i \in \mathbb{R}^{n}$
  Get the $k \leq n$ features with nonzero weights
End_For

Select the $v \leq n$ features with nonzero weight in $p\%$ of the bootstraps

Learn their weights with OLS regression on $X^{(v)} \in \mathbb{R}^{m \times v}$ and $\boldsymbol{y}$

How do we form candidate features?

- Commonly formed by indexing the **entire corpus** (Manning, Raghavan and Schütze, 2008)

- We extract them from Wikipedia, Google Search results, Public Authority websites (*e.g.*, NHS)

  Why?

  - reduce **dimensionality** to bound the error of LASSO

    $$\mathcal{L}(\boldsymbol{w}) \leq \mathcal{L}(\hat{\boldsymbol{w}}) + \mathcal{Q}, \text{ with } \mathcal{Q} \sim \min\left\{\frac{W_1^2}{N} + \frac{p}{N}, \frac{W_1^2}{N} + \frac{W_1}{\sqrt{N}}\right\}$$

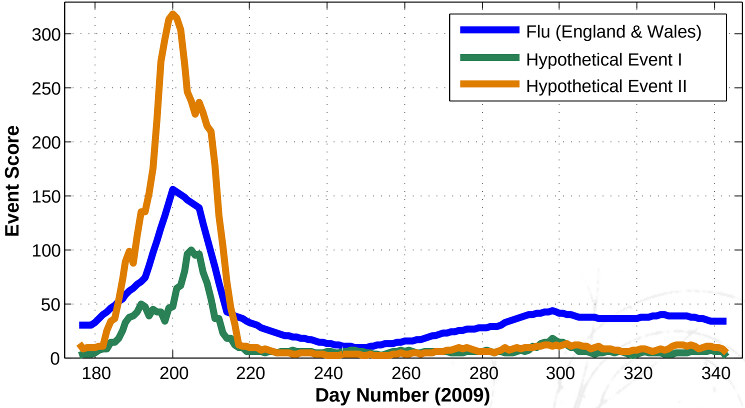
    p candidate features, N samples, empirical loss $\mathcal{L}(\hat{\boldsymbol{w}})$ and

    $\|\hat{\boldsymbol{w}}\|_{\ell_1} \leq W_1$ (Bartlett, Mendelson and Neeman, 2011)

  - **Harry Potter Effect!**

The 'Harry Potter' effect (1/2)

Figure 4 : Events co-occurring (*correlated*) with the inference target may affect feature selection, especially when the sample size is small.



(Lampos, 2012a)

The 'Harry Potter' effect (2/2)

Table 1 : Top 1-grams correlated with flu rates in England/Wales (06–12/2009)

1-gram	Event	Corr. Coef.
latitud	Latitude Festival	0.9367
flu	Flu epidemic	0.9344
swine	▲	0.9212
harri	Harry Potter Movie	0.9112
slytherin	▲	0.9094
potter	▲	0.8972
benicassim	Benicàssim Festival	0.8966
graduat	Graduation (?)	0.8965
dumbledor	Harry Potter Movie	0.8870
hogwart	▲	0.8852
quarantin	Flu epidemic	0.8822
gryffindor	Harry Potter Movie	0.8813
ravenclaw	▲	0.8738
princ	▲	0.8635
swineflu	Flu epidemic	0.8633
ginni	Harry Potter Movie	0.8620
weaslei	▲	0.8581
hermion	▲	0.8540
draco	▲	0.8533

Solution: ground truth with some degree of variability

(Lampos, 2012a)

About n-grams

1-grams

- decent (dense) representation in the Twitter corpus
- unclear semantic interpretation
  Example: "*I am not sick. But I don't feel great either!*"

2-grams

- very sparse representation in tweets
- sometimes clearer semantic interpretation

Experimental process indicated that...

a **hybrid combination*** of **1**-grams and **2**-grams
delivers the best inference performance

* refer to (Lampos, 2012a)

Flu 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)

Rainfall rates – Example of selected features

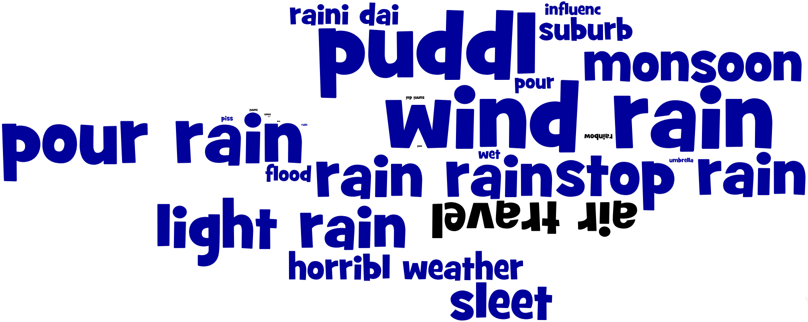
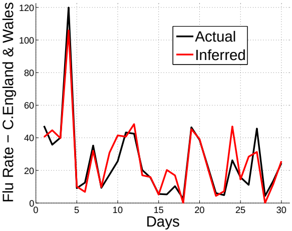


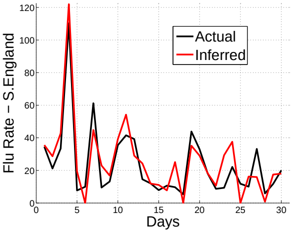
Figure 6 : Font size is proportional to the weight of each feature; flipped n-grams are negatively weighted. All words are stemmed (Porter, 1980).

(Lampos and Cristianini, 2012)

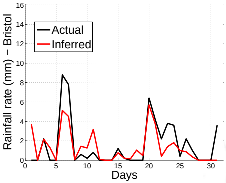
Examples of inferences



(a) Central England/Wales (flu)



(b) South England (flu)



(c) Bristol (rain)

Figure 7 : Examples of flu and rainfall rates **inferences** from Twitter content (Lampos and Cristianini, 2012)

Performance figures

Table 2 : RMSE for **flu rates** inference (5-fold cross validation), 50m tweets, 21/06/2009–19/04/2010

Method	**1-grams**	**2-grams**	**Hybrid**
Baseline*	12.44±2.37	13.81±3.29	11.62±1.58
Bolasso	**11.14±2.35**	**12.64±2.57**	**10.57±2.2**
CART ensemble**	**9.63**±5.21	13.13±4.72	**9.4**±4.21

Table 3 : RMSE (in *mm*) for **rainfall rates** inference (6-fold cross validation), 8.5m tweets, 01/07/2009–30/06/2010

Method	**1-grams**	**2-grams**	**Hybrid**
Baseline*	2.91±0.6	3.1±0.57	4.39±2.99
Bolasso	**2.73±0.65**	2.95±0.55	**2.60**±0.68
CART ensemble**	**2.71**±0.69	**2.72±0.72**	**2.64±0.63**

* As implemented in (Ginsberg *et al.*, 2009)

** Classification and Regression Tree (Breiman *et al.*, 1984) & (Sutton, 2005)

Flu Detector

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

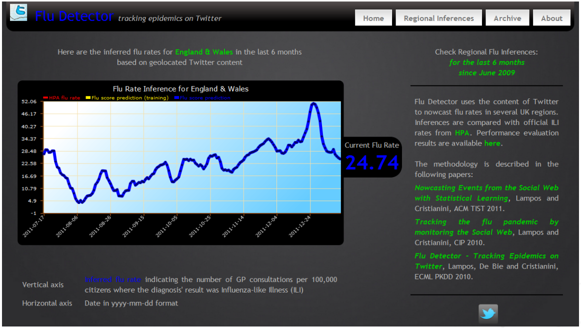


Figure 8 : Flu Detector uses the content of Twitter to nowcast flu rates in several UK regions

(Lampos, De Bie and Cristianini, 2010)

Extracting Mood Patterns from the Social Web

Computing a mood score

Table 4 : **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 d using n mood terms

$$\mathrm{ms}_d = \frac{1}{n} \sum_{i=1}^{n} \frac{c_i^{(t_d)}}{N(t_d)}$$

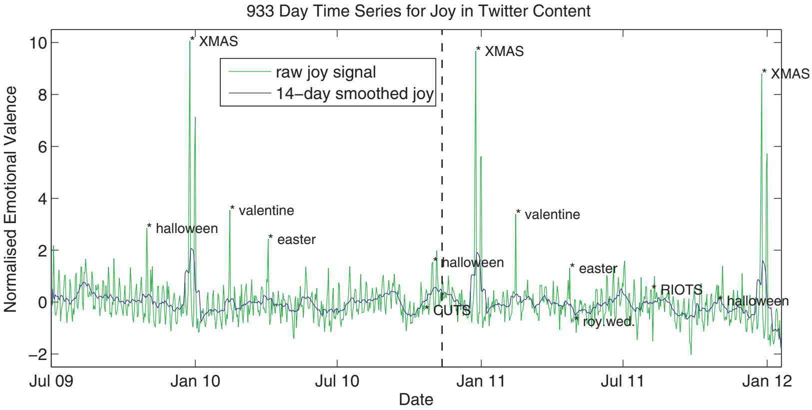
$c_i^{(t_d)}$: count of term i in the Twitter corpus of day d
$N(t_d)$: number of tweets for day d

Using the sample of d days, compute a standardised mood score:

$$\mathrm{ms}_d^{\mathrm{std}} = \frac{\mathrm{ms}_d - \mu_{\mathrm{ms}}}{\sigma_{\mathrm{ms}}}$$

The mood of the nation (1/5)

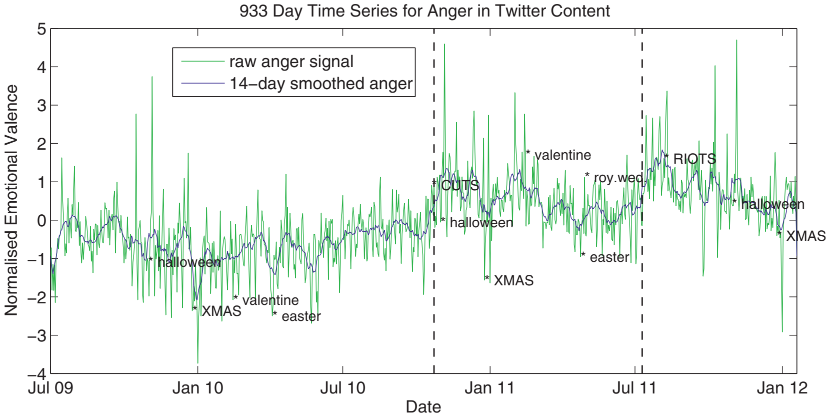
Figure 9 : Daily time series (actual & their 14-point moving average) for the mood of **Joy** based on Twitter content geo-located in the **UK**



(Lansdall, Lampos and Cristianini, 2012a&b)

The mood of the nation (2/5)

Figure 10 : Daily time series (actual & their 14-point moving average) for the mood of **Anger** based on Twitter content geo-located in the **UK**



(Lansdall, Lampos and Cristianini, 2012a&b)

The mood of the nation (3/5)

Window of **100 days**: 50 before & after the point of interest

$$\mathrm{ms}_i^{\mathrm{std}} = \mu\left(\mathbf{ms}^{\mathrm{std}}_{i+1\rightarrow i+50}\right) - \mu\left(\mathbf{ms}^{\mathrm{std}}_{i-50\rightarrow i-1}\right)$$

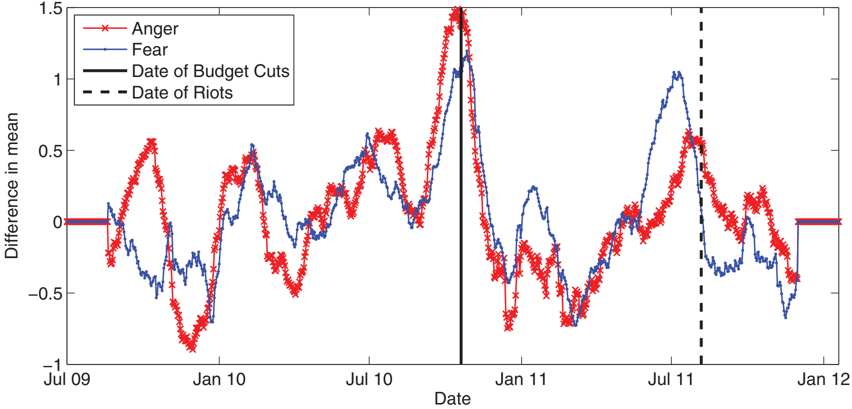
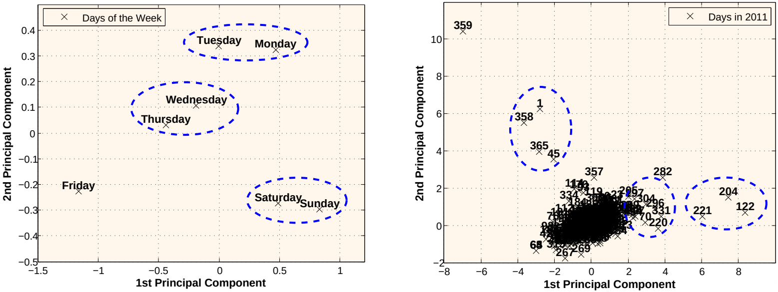


Figure 11 : Change point detection using a 100-day moving window

(Lansdall, Lampos and Cristianini, 2012a)

The mood of the nation (4/5)

Figure 12 : Projections of **4-dimensional mood score signals** (joy, sadness, anger and fear) on their **top-2 principal components** (PCA) – Twitter content from 2011



(a) Days of the week (2011) (b) Days of the year (2011)

Cluster I

New Year (**1**), Valentine's (**45**), Christmas Eve (**358**), New Year's Eve (**365**)

Cluster II

O.B. Laden's death (**122**), Winehouse's death + Breivik (**204**), UK riots (**221**)

(Lampos, 2012a)

The mood of the nation (5/5)

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

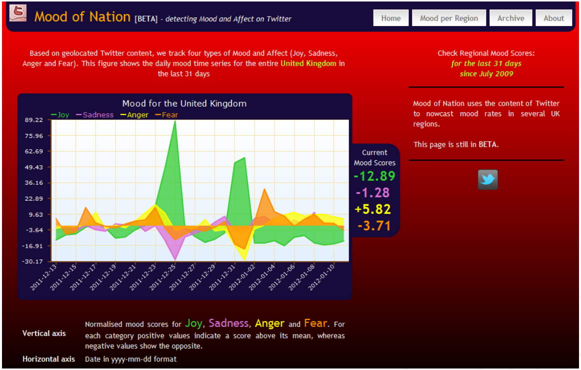


Figure 13 : Mood of the Nation uses the content of Twitter to nowcast mood rates in several UK regions

(Lampos, 2012a)

Circadian mood patterns (1/3)

Compute **24-h** mood score patterns

Mood score computation for a **time interval** $u = 24$hours using n **mood terms** (WordNet) 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 (2/3)

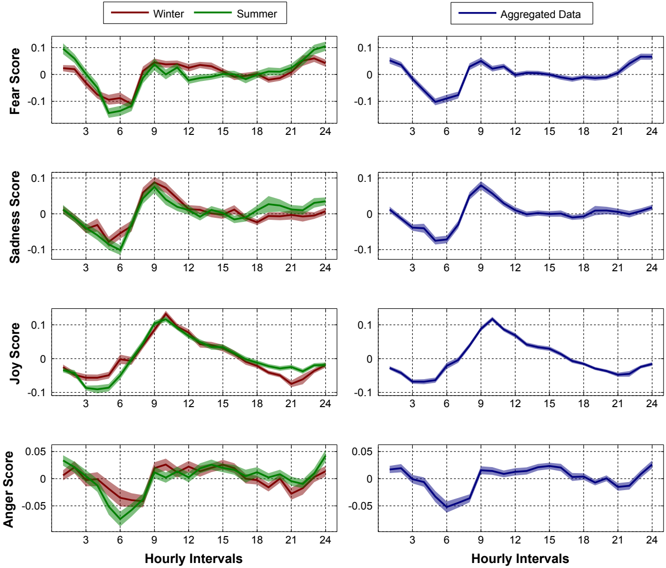
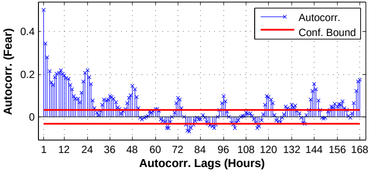


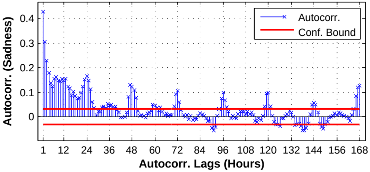
Figure 14 : Circadian (**24-hour**) mood patterns based on UK Twitter content

Circadian mood patterns (3/3)

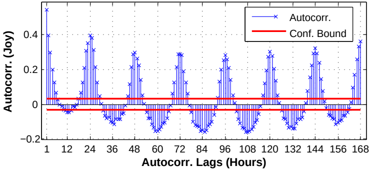
Figure 15 : **Autocorrelation** of circadian mood patterns based on **hourly lags** revealing daily and weekly periodicities



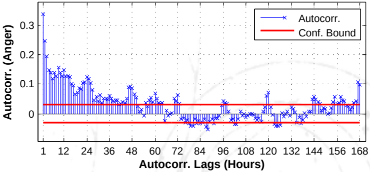
(a) Fear



(b) Sadness



(c) Joy



(d) Anger

... ***further analysis*** *on those patterns (in collab. with domain experts) under submission*

TrendMiner Project

Extracting political opinion from Social Media

A few words about the project...

- **TrendMiner** is an EU-FP7 project
- Several participants incl. the Univ. of Sheffield & Southampton (UK) and DFKI (Germany)
- Aims to form **methods for interpreting** the vast stream of **online information**
- Our focus on analysis of Twitter content $\rightarrow$ **political opinion**, **financial indicators**
- *Work in progress and under submission process* $\rightarrow$ ***cannot*** *go into much detail!*

Some new challenges

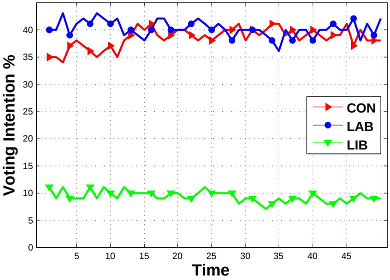
- Aim: **model voting intention**
  - regression task
  - multiple outputs
- Overcome **limitations** of previous methods
  - use of sentiment analysis taxonomies → language specific, restrictive
  - **combined modelling** of word frequencies and the domain of users?
  - **multi-task learning** → exploit correlations in the feature space
  - multi-task & **multi-domain** learning
    → model political opinion + financial indicators jointly
- Proper **evaluation**
  - *k*-fold cross-validation may sometimes be misleading
  - can we actually predict future values?

A snapshot of the results

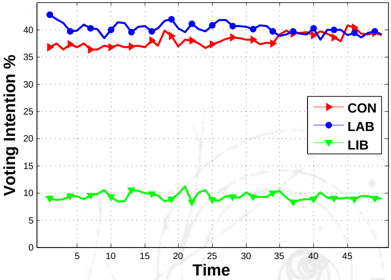
$$\mathbf{vi} = \boldsymbol{u}^{\mathsf{T}} X \boldsymbol{w} + \beta$$

(plus multi-task learning)

Figure 16 : 50 voting intention polls (YouGov) and their respective inferred values for the Conservative (RMSE: **1.78**%), Labour (**1.59**%) and Liberal Democrat (**1.05**%) parties (Nov. 2011 to Feb. 2012)



(a) Voting intention polls



(b) Voting intention inferences

Qualitative evaluation is also essential...

- Some domains may be represented by **smooth** trends (*e.g.*, political domain)

- Predictions could be easy in that context
  $\rightarrow$ how do we know we are not **overfitting**?

- Perform qualitative analysis using the selected features (words, users and tweets)
  - Do the selected words and users make some sense?
  - Does their combination make sense? $\rightarrow$ score single tweets

- Possibly better models when increasing the statistical evidence (multi-task learning)

Conclusions – Did *they* tell us anything?

- **Social Media** hold **valuable information**
- We can develop **methods** to extract portions of this information **automatically**
  - **detect, quantify, nowcast events**
  - extract **collective mood** patterns
  - model other domains (such as **politics**)
- User generated input + other features
  → tell/reveal **something** about the users & their context
- Side effect: what about our **privacy**? ...

In collaboration with...

Prof. Nello Cristianini, University of Bristol (*Ph.D. Advisor*)

Prof. Ricardo Araya, University of Bristol (*Psychiatry*)

Dr. Tijl De Bie, University of Bristol

Thomas Lansdall-Welfare, University of Bristol

Dr. Trevor Cohn, University of Sheffield (*TrendMiner*)

Daniel Preotiuc-Pietro, University of Sheffield (*TrendMiner*)

The end.

Any questions?

Download the slides from
http://www.lampos.net/research/presentations-and-posters

References

1. B. Debatin, J.P. Lovejoy, A.M.A. Horn and B.N. Hughes. **Facebook and Online Privacy: Attitudes, Behaviors, and Unintended Consequences**. Journal of Computer-Mediated Communication 15, pp. 83–108, 2009.
2. D. Preotiuc-Pietro, S. Samangooei, T. Cohn, N. Gibbins and M. Niranjan. **TrendMiner: An Architecture for Real Time Analysis of Social Media Text**. Proceedings of ICWSM '12, pp. 38–42, 2012.
3. V. Lampos and N. Cristianini. **Nowcasting Events from the Social Web with Statistical Learning**. ACM TIST 3(4), n. 72, 2012.
4. R. Tibshirani. **Regression Shrinkage and Selection via the LASSO**. Journal of the Royal Statistical Society, series B, 58(1), pp. 267–288, 1996.
5. B. Efron, T. Hastie, I. Johnstone and R. Tibshirani. **Least Angle Regression**. The Annals of Statistics 32(2), pp. 407–499, 2004.
6. C.D. Manning, P. Raghavan and H. Schütze. **Introduction to Information Retrieval**. Cambridge University Press, p. 544, 2008.
7. P.L. Bartlett, S. Mendelson and J. Neeman. **L1-regularized linear regression: persistence and oracle inequalities**. Probability Theory and Related Fields, pp. 1–32, 2011.
8. M.F. Porter. **An algorithm for suffix stripping**. Program 14(3), pp. 130–137, 1980.
9. V. Lampos and N. Cristianini. **Tracking the flu pandemic by monitoring the Social Web**. Proceedings of CIP '10, pp. 411–416, 2010.
10. V. Lampos, T. De Bie and N. Cristianini. **Flu Detector – Tracking Epidemics on Twitter**. Proceedings of ECML PKDD '10, pp. 599–602, 2010.
11. T. Lansdall-Welfare, V. Lampos and N. Cristianini. **Effects of the Recession on Public Mood in the UK**. Proceedings of WWW '12, pp. 1221–1226, 2012.(a)
12. T. Lansdall-Welfare, V. Lampos and N. Cristianini. **Nowcasting the mood of the nation**. Significance 9(4), pp. 26–28, 2012.(b)
13. 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)
14. V. Lampos. **On voting intentions inference from Twitter content: a case study on UK 2010 General Election**. CoRR, 2012.(b)