Mining the Social Web: A series of statistical NLP case studies

Vasileios Lampos
Department of Computer Science
University College London

March, 2015
Key assumptions about social media

- a significant **sample of the population** uses them — biases exist
- a significant amount of the published content is **geo-located**
- reflect on **collective** portions of real-life (e.g., opinions, events)
  - usually forming a **real-time** relationship
- it is **easy** to collect, store and process this content (?)
- more data (**big data**) → higher confidence (?)

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

- Reply  Retweet  Favorite

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

i think i have the flu but i still look fabulous

- Reply  Retweet  Favorite

And what about the statistical significance of the computed statistical significance?
#inception_in_statistics

- Reply  Delete  Favorite

v.lampos@ucl.ac.uk

Twitter in one slide

• 140 characters per published status (tweet)
• users can follow others and can be followed
• embedded usage of topics (#rbnews, #inception_in_statistics)
• user interaction: re-tweets, @replies, @mentions, favourites
• real-time nature
• biased demographics (13-15% of UK’s population)
In this talk

Case studies where we harness social media information to:

- extract simplified collective **mood patterns**
  (Lansdall et al., 2012)

- **nowcast** phenomena (an infectious disease or rainfall rates)
  (Lampos, Cristianini, 2010 & 2012)

- model **voting intention**
  (Lampos et al., 2013)

- estimate **user impact** and explore user characteristics related to it
  (Lampos et al., 2014)
Proof of concept and a little more: extracting collective mood patterns
Time series of joy and anger based on UK tweets

Joy
- happy, enjoy, love, glad, joyful, elated...

Anger & Fear

Derivative of

(Lansdall et al., 2012), (Strapparava, Valitutti, 2004) → WordNet Affect
Mood projections via PCA

Projection of 4-dimensional mood score signals (joy, sadness, anger and fear) on their top-2 principal components (2011 Twitter data)

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

O.B. Laden’s death (122), Winehouse’s death & Breivik (204), UK riots (221)

Supervised learning
Primary outcomes (linear methods)
Regression basics — Ordinary Least Squares

- observations \( x_i \in \mathbb{R}^m, \quad i \in \{1, \ldots, n\} \) — \( X \)
- responses \( y_i \in \mathbb{R}, \quad i \in \{1, \ldots, n\} \) — \( y \)
- weights, bias \( w_j, \beta \in \mathbb{R}, \quad j \in \{1, \ldots, m\} \) — \( w_* = [w; \beta] \)

**Ordinary Least Squares (OLS)**

\[
\arg\min_{w_*} \|X_*w_* - y\|_2^2 \Rightarrow w_* = \left(X_*^T X_*\right)^{-1} X_*^T y
\]

**Why not?**

- \( X_*^T X_* \) may be singular (thus difficult to invert)
- high-dimensional models become difficult to interpret
- unsatisfactory prediction accuracy (estimates have large variance)
Regression basics — Ridge Regression

- observations $x_i \in \mathbb{R}^m$, $i \in \{1, ..., n\}$ — $X$
- responses $y_i \in \mathbb{R}$, $i \in \{1, ..., n\}$ — $y$
- weights, bias $w_j, \beta \in \mathbb{R}$, $j \in \{1, ..., m\}$ — $w_* = [w; \beta]$

**Ridge Regression (RR)**

$$\text{argmin}_{w_*} \left\{ \|X_*w_* - y\|_2^2 + \lambda \|w\|_2^2 \right\}$$

+ size constraint on the weight coefficients (**regularisation**)
  → resolves problems caused by collinear variables
+ less degrees of freedom, better predictive accuracy than OLS
- does **not** perform feature selection (nonzero coefficients)

(Hoerl, Kennard, 1970)
Regression basics — Lasso

- observations \( x_i \in \mathbb{R}^m, \ i \in \{1, \ldots, n\} \) — \( X \)
- responses \( y_i \in \mathbb{R}, \ i \in \{1, \ldots, n\} \) — \( y \)
- weights, bias \( w_j, \beta \in \mathbb{R}, \ j \in \{1, \ldots, m\} \) — \( w_* = [w; \beta] \)

\( \ell_1 \)-norm regularisation or lasso (Tibshirani, 1996)

\[
\text{argmin}_{w_*} \left\{ \|X_*w_* - y\|_2^2 + \lambda\|w\|_1 \right\}
\]

- no closed form solution — quadratic programming problem
+ Least Angle Regression (LAR) \( \rightarrow \) entire reg. path (Efron et al., 2004)
+ sparse \( w \), interpretability, better performance (Hastie et al., 2009)
- if \( m > n \), at most \( n \) variables can be selected
- co-linear predictors \( \rightarrow \) unable to select true model (Zhao, Yu, 2009)
Lasso for text regression

- **n-gram**: set of n words or tokens
- **n-gram frequency**: count (often normalised) in a corpus
- **target variable**: numerical representation of an “event”

\[
\text{n-gram frequencies } \mathbf{x}_i \in \mathbb{R}^m, \quad i \in \{1, \ldots, n\} \quad \rightarrow \quad X
\]

\[
\text{target phenomenon } y_i \in \mathbb{R}, \quad i \in \{1, \ldots, n\} \quad \rightarrow \quad y
\]

\[
\text{weights, bias } w_j, \beta \in \mathbb{R}, \quad j \in \{1, \ldots, m\} \quad \rightarrow \quad w_* = [w; \beta]
\]

**lasso** (for text regression)

\[
\arg\min_{w_*} \left\{ \|X_* w_* - y\|_2^2 + \lambda \|w\|_1 \right\}
\]
Nowcasting ILI rates from Twitter (1/2)

Assumptions

- Twitter users post about their health condition
- We can turn this information into an influenza-like-illness (ILI) rate

Is there a signal in the data?

- 41 illness related keyphrases (e.g., flu, fever, sore throat, headache)
- z-scored aggregate keyphrase frequency vs. official ILI rates

England & Wales (region D)

\[ r = .856 \]

(Lampos, Cristianini, 2010)
Nowcasting ILI rates from Twitter (2/2)

• create a pool of 1-gram features (approx. 1600) by indexing relevant web pages (e.g., Wikipedia, NHS, health forums)
• stop-words removed, Porter-stemming applied
• automatic 1-gram selection and weighting via lasso

**Selected uni-grams**

'unwel', 'temperatur', 'headach', 'appetit', 'symptom', 'diarrhoea', 'muscl', 'feel', 'flu', 'cough', 'nose', 'vomit', 'diseas', 'sore', 'throat', 'fever', 'ach', 'runni', 'sick', 'ill', ...

England & Wales
\[ r = .968 \]

(Lampos, Cristianini, 2010)
Nowcasting rainfall rates — a generalisation

- fix lasso’s model selection with **bootstrap lasso** (Bach, 2008)
- include **2-grams** and perform hybrid combination with 1-grams

(Lampos, Cristianini, 2012)

v.lampos@ucl.ac.uk

Back to regression basics — Elastic Net

- observations $x_i \in \mathbb{R}^m, \quad i \in \{1, \ldots, n\}$ — $X$
- responses $y_i \in \mathbb{R}, \quad i \in \{1, \ldots, n\}$ — $y$
- weights, bias $w_j, \beta \in \mathbb{R}, \quad j \in \{1, \ldots, m\}$ — $w_\ast = [w; \beta]$

**linear Elastic Net (LEN)**

$$\arg\min_{w_\ast} \|X_\ast w_\ast - y\|_2^2 + \lambda_1 \|w\|_2^2 + \lambda_2 \|w\|_1$$

+ **combination** of RR (co-linear predictors) and lasso (sparsity)
+ entire reg. path can be explored by modifying LAR
+ if $m > n$, number of selected variables not limited to $n$
  - may select redundant variables!

*(Zhou, Hastie, 2005)*
Supervised learning

Bilinear approaches
Bilinear text regression — The general idea (1/2)

Linear regression: \( f(x_i) = x_i^T w + \beta \)

- observations \( x_i \in \mathbb{R}^m, \quad i \in \{1, \ldots, n\} \) — \( X \)
- responses \( y_i \in \mathbb{R}, \quad i \in \{1, \ldots, n\} \) — \( y \)
- weights, bias \( w_j, \beta \in \mathbb{R}, \quad j \in \{1, \ldots, m\} \) — \( w_\star = [w; \beta] \)

Bilinear regression: \( f(Q_i) = u^T Q_i w + \beta \)

- users \( p \in \mathbb{Z}^+ \)
- observations \( Q_i \in \mathbb{R}^{p \times m}, \quad i \in \{1, \ldots, n\} \) — \( X \)
- responses \( y_i \in \mathbb{R}, \quad i \in \{1, \ldots, n\} \) — \( y \)
- weights, bias \( u_k, w_j, \beta \in \mathbb{R}, \quad k \in \{1, \ldots, p\}, \quad j \in \{1, \ldots, m\} \) — \( u, w, \beta \)

v.lampos@ucl.ac.uk
Bilinear text regression — The general idea (2/2)

- **users** \( p \in \mathbb{Z}^+ \)
- **observations** \( Q_i \in \mathbb{R}^{p \times m}, \quad i \in \{1, \ldots, n\} \)
- **responses** \( y_i \in \mathbb{R}, \quad i \in \{1, \ldots, n\} \)
- **weights, bias** \( u_k, w_j, \beta \in \mathbb{R}, \quad k \in \{1, \ldots, p\}, \quad j \in \{1, \ldots, m\} \)

\[
f(Q_i) = u^T Q_i w + \beta
\]
Bilinear text regression — Regularisation

- users \( p \in \mathbb{Z}^+ \)
- observations \( Q_i \in \mathbb{R}^{p \times m}, \quad i \in \{1, \ldots, n\} \)
- responses \( y_i \in \mathbb{R}, \quad i \in \{1, \ldots, n\} \)
- weights, bias \( u_k, w_j, \beta \in \mathbb{R}, \quad k \in \{1, \ldots, p\}, \quad j \in \{1, \ldots, m\} \)

\[
\underset{u, w, \beta}{\text{argmin}} \left\{ \sum_{i=1}^{n} (u^T Q_i w + \beta - y_i)^2 + \psi(u, \theta_u) + \psi(w, \theta_w) \right\}
\]

\( \psi(\cdot) \): regularisation function with a set of hyper-parameters (\( \theta \))

- if \( \psi(v, \lambda) = \lambda \|v\|_{\ell_1} \) Bilinear Lasso
- if \( \psi(v, \lambda_1, \lambda_2) = \lambda_1 \|v\|_{\ell_2}^2 + \lambda_2 \|v\|_{\ell_1} \) Bilinear Elastic Net (BEN)

(Lampos et al., 2013)
Learning the parameters of BEN

\[
\arg\min_{u,w,\beta} \left\{ \sum_{i=1}^{n} \left( u^T Q_i w + \beta - y_i \right)^2 + \lambda_u \|u\|_2^2 + \lambda_u \|u\|_1 + \lambda_w \|w\|_2^2 + \lambda_w \|w\|_1 \right\}
\]

Bi-convexity: fix \(u\), learn \(w\) and vice versa
Iterating through convex optimisation tasks: convergence

FISTA (Beck, Teboulle, 2009)
implemented in SPAMS (Mairal et al., 2010)
Large-scale optimisation solver,
quick convergence

RMSE on held-out data
vs Obj. function through iterations
Supervised learning

Bilinear approaches

for modelling voting intention

(*based on social media content*)
Political opinion/voting intention mining — Brief recap

Primary papers

• predict the result of an election via Twitter (Tumasjan et al., 2010)
• model socio-political sentiment polls (O’Connor et al., 2010)
• above 2 failed in 2009 US congr. elections (Gayo-Avello, 2011)
• desired properties of such models (Metaxas et al., 2011)

Features used

• lexicon-based, e.g. using LIWC (Tausczik, Pennebaker, 2010)
• task-specific keywords (names of parties, politicians)
• tweet volume

reviewed in (Gayo-Avello, 2013)

However...

– political **descriptors change** in time, differ per country
– personalised (**user**) modelling missing (present in actual polls)
– **multi-task** learning? a user who likes party A, may dislike party B

v.lampos@ucl.ac.uk

Voting intention modelling — Data (UK)

- **42K users** distributed proportionally to regional population figures
- **60 million tweets** from 30/04/2010 to 13/02/2012
- **80,976 1-grams** → (Preștiuc-Pietro et al., 2012)
- **240 voting intention polls** (YouGov)
- **3 parties**: Conservatives (**CON**), Labour Party (**LAB**), Liberal Democrats (**LIB**)
- main language: English

voting intention for the UK
Voting intention modelling — Data (Austria)

- **1.1K users** manually selected by political analysts (SORA)
- **800K tweets** from 25/01 to 01/12/2012
- **22,917 1-grams** → (Prețiu-Pietro et al., 2012)
- **98 voting intention polls** from various pollsters
- **4 parties**: Social Democratic Party (SPÖ), People’s Party (ÖVP), Freedom Party (FPÖ), Green Alternative Party (GRÜ)
- main language: German

![Voting Intention for Austria](attachment:graph.png)
Voting intention modelling — Evaluation

- 10-fold (not cross) validation
  - train a model using data based on a set of contiguous polls $\mathcal{A}$
  - test on the next $\mathcal{D} = 5$ polls
  - expand training set to $\mathcal{A} \cup \mathcal{D}$, test on the next $|\mathcal{D}'| = 5$ polls

- realistic scenario: train on past, predict future polls

- overall test predictions on 50 polls (in each case study)

Baselines

- $B_\mu$: constant prediction based on $\mu(y)$ in the training set
- $B_{\text{last}}$: constant prediction based on $\text{last}(y)$ in the training set
- LEN: (linear) Elastic Net prediction (using word frequencies)
Voting intention modelling — BEN’s performance (1/2)

Average RMSEs on the voting intention percentage predictions in the 10-step validation process

### ‘UK’ case study

<table>
<thead>
<tr>
<th></th>
<th>CON</th>
<th>LAB</th>
<th>LIB</th>
<th>μ</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B_\mu$</td>
<td>2.272</td>
<td>1.663</td>
<td>1.136</td>
<td>1.69</td>
</tr>
<tr>
<td>$B_{\text{last}}$</td>
<td>2</td>
<td>2.074</td>
<td>1.095</td>
<td>1.723</td>
</tr>
<tr>
<td>LEN</td>
<td>3.845</td>
<td>2.912</td>
<td>2.445</td>
<td>3.067</td>
</tr>
<tr>
<td>BEN</td>
<td>1.939</td>
<td>1.644</td>
<td>1.136</td>
<td>1.573</td>
</tr>
</tbody>
</table>

### ‘Austria’ case study

<table>
<thead>
<tr>
<th></th>
<th>SPÖ</th>
<th>ÖVP</th>
<th>FPÖ</th>
<th>GRÜ</th>
<th>μ</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B_\mu$</td>
<td>1.535</td>
<td>1.373</td>
<td>3.3</td>
<td>1.197</td>
<td>1.851</td>
</tr>
<tr>
<td>$B_{\text{last}}$</td>
<td>1.148</td>
<td>1.556</td>
<td>1.639</td>
<td>1.536</td>
<td>1.47</td>
</tr>
<tr>
<td>LEN</td>
<td>1.291</td>
<td>1.286</td>
<td>2.039</td>
<td>1.152</td>
<td>1.442</td>
</tr>
<tr>
<td>BEN</td>
<td>1.392</td>
<td>1.31</td>
<td>2.89</td>
<td>1.205</td>
<td>1.699</td>
</tr>
</tbody>
</table>
Voting intention modelling — BEN’s performance (2/2)

UK

Polls

BEN

Austria

Voting Intention %

Time

CON
LAB
LIB

SPÖ
ÖVP
FPÖ
GRÜ

good, but probably not good enough?
Supervised learning

MULTI-TASK Bilinear approaches

for modelling voting intention

*(based on social media content)*
Multi-task learning

What

- Instead of learning/optimising a single task (one target variable)
- ... optimise multiple tasks jointly

Why (Caruana, 1997)

- improves generalisation performance exploiting domain-specific information of related tasks
- a good choice for under-sampled distributions
  - knowledge transfer
- application-driven reasons
  - e.g., explore interplay between political parties

How

- Multi-task regularised regression
Linear multi-task learning: the $\ell_{2,1}$-norm regularisation

$$\|W\|_{2,1} = \sum_{j=1}^{m} \|W_j\|_2,$$
where $W_j$ denotes the $j$-th row

$l_{2,1}$-norm regularisation

$$\text{argmin}_{W,\beta} \left\{ \|XW - Y\|_F^2 + \lambda \sum_{j=1}^{m} \|W_j\|_2 \right\}$$

- multi-task learning: instead of $w \in \mathbb{R}^m$, learn $W \in \mathbb{R}^{m \times \tau}$, where $\tau$ is the number of tasks
- $l_{2,1}$-norm regularisation $\rightarrow$ sum of $W$’s row $l_2$-norms (Argyriou et al., 2008; Liu et al., 2009) extends group lasso (Yuan, Lin, 2006)
  - group lasso: instead of single variables, selects groups of variables
- ‘groups’ now become the $\tau$-dimensional rows of $W$
Bilinear multi-task learning

- **tasks** $\tau \in \mathbb{Z}^+$
- **users** $p \in \mathbb{Z}^+$
- **observations** $Q_i \in \mathbb{R}^{p \times m}, \quad i \in \{1, \ldots, n\}$ — $X$
- **responses** $y_i \in \mathbb{R}^\tau, \quad i \in \{1, \ldots, n\}$ — $Y$
- **weights, bias** $u_k, w_j, \beta \in \mathbb{R}^\tau, \quad k \in \{1, \ldots, p\}$ — $U, W, \beta$

$$f(Q_i) = \text{tr}(U^T Q_i W) + \beta$$

$v.lampos@ucl.ac.uk$

Bilinear Group $\ell_{2,1}$ (BGL) (1/2)

- **tasks** $\tau \in \mathbb{Z}^+$
- **users** $p \in \mathbb{Z}^+$
- **observations** $Q_i \in \mathbb{R}^{p \times m}$, $i \in \{1, \ldots, n\}$ — $\mathbf{X}$
- **responses** $y_i \in \mathbb{R}^{\tau}$, $i \in \{1, \ldots, n\}$ — $\mathbf{Y}$
- **weights, bias** $u_k, w_j, \beta \in \mathbb{R}^{\tau}$, $k \in \{1, \ldots, p\}$ — $\mathbf{U}, \mathbf{W}, \beta$
  
  $j \in \{1, \ldots, m\}$

$$\arg\min_{U, W, \beta} \left\{ \sum_{t=1}^{\tau} \sum_{i=1}^{n} (u_t^T Q_i w_t + \beta_t - y_{ti})^2 
  + \lambda_u \sum_{k=1}^{p} \|U_k\|_2 + \lambda_w \sum_{j=1}^{m} \|W_j\|_2 \right\}$$

- **Learning**: 2 convex tasks $\rightarrow$ first learn $\{W, \beta\}$, then $\{U, \beta\}$ and vice versa; iterate through this process
Bilinear Group $\ell_{2,1}$ (BGL) (2/2)

$$\arg\min_{U,W,\beta} \left\{ \sum_{t=1}^{\tau} \sum_{i=1}^{n} \left( u_t^T Q_i \omega_t + \beta_t - y_{ti} \right)^2 + \lambda_u \sum_{k=1}^{p} \|U_k\|_2 + \lambda_w \sum_{j=1}^{m} \|W_j\|_2 \right\} \times U^T \times Q_i \times W$$

- A feature (user or word) is activated (selected) for all tasks with different weights.
- Especially useful in the domain of politics, e.g., user pro party A, but against parties B and C.

v.lampos@ucl.ac.uk

Voting intention modelling — BGL’s performance (1/2)

### ‘UK’ case study

<table>
<thead>
<tr>
<th></th>
<th>CON</th>
<th>LAB</th>
<th>LIB</th>
<th>$\mu$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B_\mu$</td>
<td>2.272</td>
<td>1.663</td>
<td>1.136</td>
<td>1.69</td>
</tr>
<tr>
<td>$B_{\text{last}}$</td>
<td>2</td>
<td>2.074</td>
<td>1.095</td>
<td>1.723</td>
</tr>
<tr>
<td>LEN</td>
<td>3.845</td>
<td>2.912</td>
<td>2.445</td>
<td>3.067</td>
</tr>
</tbody>
</table>

| BEN | 1.939 | 1.644 | 1.136 | 1.573 |
| BGL | 1.785 | 1.595 | 1.054 | 1.478 |

### ‘Austria’ case study

<table>
<thead>
<tr>
<th></th>
<th>SPÖ</th>
<th>ÖVP</th>
<th>FPÖ</th>
<th>GRÜ</th>
<th>$\mu$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B_\mu$</td>
<td>1.535</td>
<td>1.373</td>
<td>3.3</td>
<td>1.197</td>
<td>1.851</td>
</tr>
<tr>
<td>$B_{\text{last}}$</td>
<td><strong>1.148</strong></td>
<td>1.556</td>
<td><strong>1.639</strong></td>
<td>1.536</td>
<td>1.47</td>
</tr>
<tr>
<td>LEN</td>
<td>1.291</td>
<td>1.286</td>
<td>2.039</td>
<td><strong>1.152</strong></td>
<td>1.442</td>
</tr>
</tbody>
</table>

| BEN | 1.392 | 1.31 | 2.89 | 1.205 | 1.699 |
| BGL | 1.619 | **1.005** | 1.757 | 1.374 | **1.439** |
Voting intention modelling — BGL’s performance (2/2)

UK

Polls

BEN

BGL

Austria

Voting Intention %

Time

CON

LAB

LIB

SPÖ

ÖVP

FPÖ

GRÜ

v.lampos@ucl.ac.uk

## Voting intention modelling — Qualitative insight

<table>
<thead>
<tr>
<th>Party</th>
<th>Tweet</th>
<th>Score</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CON</strong></td>
<td>PM in friendly chat with top EU mate, Sweden’s Fredrik Reinfeldt, before family photo</td>
<td>1.334</td>
<td>Journalist</td>
</tr>
<tr>
<td><strong>LAB</strong></td>
<td>I am so pleased to hear Paul Savage who worked for the Labour group has been Appointed the Marketing manager for the baths hall GREAT NEWS</td>
<td>−0.552</td>
<td>Politician (Labour)</td>
</tr>
<tr>
<td><strong>LBD</strong></td>
<td>RT @user: Must be awful for TV bosses to keep getting knocked back by all the women they ask to host election night (via @user)</td>
<td>0.874</td>
<td>LibDem MP</td>
</tr>
<tr>
<td><strong>SPÖ</strong></td>
<td>Inflationsrate in Ö. im Juli leicht gesunken: von 2,2 auf 2,1%. Teurer wurde Wohnen, Wasser, Energie.</td>
<td>0.745</td>
<td>Journalist</td>
</tr>
<tr>
<td><strong>ÖVP</strong></td>
<td>kann das buch “res publica” von johannes #voggenhuber wirklich empfehlen! so zum nachdenken und so... #europa #demokratie</td>
<td>−2.323</td>
<td>User</td>
</tr>
<tr>
<td><strong>GRÜ</strong></td>
<td>Protestsong gegen die Abschaffung des Bachelor-Studiums Internationale Entwicklung: &lt;link&gt; #IEbleibt #unibrennt #uniwut</td>
<td>1.45</td>
<td>Student Union</td>
</tr>
<tr>
<td></td>
<td>Translation: Protest songs against the closing-down of the bachelor course of International Development: &lt;link&gt; #IDremains #uniburns #unirage</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What does content tell us about users?
User impact characterisation on Twitter
(with a nonlinear approach)
Predicting and characterising user impact on Twitter

Motivation

• **predict user impact** from user activity, including text
• use this prediction model as a guide to **qualitatively** investigate links between user impact and **user behaviour**

Data

• 48 million tweets posted by 38,020 UK users
  – from 14/04/2011 to 12/04/2012
  – subset of the data set used in *(Lampos et al., 2013)*
• 400 million tweets (from the Gardenhose stream — 10%)
  – from 02/01/2011 to 28/02/2011
  – for creating topic clusters
  – data processed via *(Prețiuc-Pietro et al., 2012)*

*(Lampos et al., 2014)*
User impact — a simplified definition

\[ S(\phi_{\text{in}}, \phi_{\text{out}}, \phi_{\lambda}) = \ln \left( \frac{(\phi_{\lambda} + \theta)(\phi_{\text{in}} + \theta)^2}{\phi_{\text{out}} + \theta} \right) \]

- \( \phi_{\text{in}} \): number of followers, \( \phi_{\text{out}} \): number of followees
- \( \phi_{\lambda} \): number of times the account has been listed
- \( \theta = 1 \), logarithm is applied on a positive number
- \( (\phi_{\text{in}}^2/\phi_{\text{out}}) = (\phi_{\text{in}} - \phi_{\text{out}}) \times (\phi_{\text{in}}/\phi_{\text{out}}) + \phi_{\text{in}} \)

Histogram of the user impact scores in our data set

\[ \mu(S) = 6.776 \]
### User activity features

<table>
<thead>
<tr>
<th>$a_1$</th>
<th># of tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_2$</td>
<td>proportion of retweets</td>
</tr>
<tr>
<td>$a_3$</td>
<td>proportion of non-duplicate tweets</td>
</tr>
<tr>
<td>$a_4$</td>
<td>proportion of tweets with hashtags</td>
</tr>
<tr>
<td>$a_5$</td>
<td>hashtag-tokens ratio in tweets</td>
</tr>
<tr>
<td>$a_6$</td>
<td>proportion of tweets with @-mentions</td>
</tr>
<tr>
<td>$a_7$</td>
<td># of unique @-mentions in tweets</td>
</tr>
<tr>
<td>$a_8$</td>
<td>proportion of tweets with @-replies</td>
</tr>
<tr>
<td>$a_9$</td>
<td>links ratio in tweets</td>
</tr>
<tr>
<td>$a_{10}$</td>
<td># of favourites the account made</td>
</tr>
<tr>
<td>$a_{11}$</td>
<td>total # of tweets (entire history)</td>
</tr>
<tr>
<td>$a_{12}$</td>
<td>using default profile background (binary)</td>
</tr>
<tr>
<td>$a_{13}$</td>
<td>using default profile image (binary)</td>
</tr>
<tr>
<td>$a_{14}$</td>
<td>enabled geolocation (binary)</td>
</tr>
<tr>
<td>$a_{15}$</td>
<td>population of account’s location</td>
</tr>
<tr>
<td>$a_{16}$</td>
<td>account’s location latitude</td>
</tr>
<tr>
<td>$a_{17}$</td>
<td>account’s location longitude</td>
</tr>
<tr>
<td>$a_{18}$</td>
<td>proportion of days with nonzero tweets</td>
</tr>
</tbody>
</table>
User participation in topic-specific discussions

NPMI (Bouma, 2009) + Spectral Clustering (von Luxburg, 2007)

<table>
<thead>
<tr>
<th>Label</th>
<th>Cluster’s words ranked by centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weather ($\tau_1$)</td>
<td>mph, humidity, barometer, gust, winds, hpa, temperature, kt</td>
</tr>
<tr>
<td>Healthcare, Finance,</td>
<td>nursing, nurse, rn, registered, bedroom, clinical, #news, estate, #hospital, rent, healthcare,</td>
</tr>
<tr>
<td>Housing ($\tau_2$)</td>
<td>therapist, condo, investment, furnished, medical, #nyc, occupational, investors, #ny</td>
</tr>
<tr>
<td>Politics ($\tau_3$)</td>
<td>senate, republican, gop, police, arrested, voters, robbery, democrats, presidential, elections,</td>
</tr>
<tr>
<td></td>
<td>charged, election, charges, #religion, arrest, repeal, dems, #christian, reform</td>
</tr>
<tr>
<td>Showbiz, Movies ($\tau_4$)</td>
<td>damon, potter, #tvd, harry, elena, kate, portman, pattinson, hermione, jennifer, kristen, stefan,</td>
</tr>
<tr>
<td></td>
<td>robert, catholic, stewart, katherine, lois, jackson, vampire, natalie, #vampirediaries</td>
</tr>
<tr>
<td>Commerce ($\tau_5$)</td>
<td>chevrolet, inventory, coupon, toyota, mileage, sedan, nissan, adde, jeep, 4x4, 2002, #coupon,</td>
</tr>
<tr>
<td></td>
<td>enhanced, #deal, dodge</td>
</tr>
<tr>
<td>Twitter hashtags ($\tau_6$)</td>
<td>#teamfollowback, #500aday, #tfb, #instantfollowback,</td>
</tr>
<tr>
<td></td>
<td>#ifollowback, #instantfollow, #followback</td>
</tr>
<tr>
<td>Social unrest ($\tau_7$)</td>
<td>#egypt, #tunisia, #iran, #israel, #palestine, tunisia, arab, #jan25, iran, israel, protests,</td>
</tr>
<tr>
<td></td>
<td>egypt, #yemen, #iranelection, israeli, #jordan, regime, yemen, #gaza, protesters, #lebanon</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
User impact modelling as a regression task

**Feature sets**
- user activity only ($A$)
- $A$ and top 1-grams ($AW$)
- $A + |\tau|$ topic clusters ($AC$)

**Regression via**
- Ridge Regression (RR)
- **Gaussian Process** (GP) using a Squared Exponential kernel with Automatic Relevance Determination (ARD)

(Rasmussen and Williams, 2006)

*GPs offer a very interesting (and well established) framework for performing regression [and classification] tasks in a nonlinear, kernelised fashion — intro at: [http://videolectures.net/gpip06_mackay_gpb/](http://videolectures.net/gpip06_mackay_gpb/)*
### Performance estimates

<table>
<thead>
<tr>
<th>Model</th>
<th>Linear (RR)</th>
<th>Nonlinear (GP)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$r$</td>
<td>RMSE</td>
</tr>
<tr>
<td>A</td>
<td>.667</td>
<td>2.642</td>
</tr>
<tr>
<td>AW</td>
<td>.712</td>
<td>2.529</td>
</tr>
<tr>
<td>AC, $</td>
<td>\tau</td>
<td>= 50$</td>
</tr>
<tr>
<td>AC, $</td>
<td>\tau</td>
<td>= 100$</td>
</tr>
</tbody>
</table>

Most **valuable / relevant** features

1. default profile image
2. # of historical tweets
3. # of unique @-mentions
4. # of tweets (last year)
5. links (ratio)
6. topic: *weather*
7. topic: *healthcare-finance*
8. topic: *politics*
9. : days with nonzero tweets (ratio)
10. @-replies (ratio)
User impact — Qualitative analysis (1/2)

Impact score distribution for user accounts with high (H) or low (L) values for the most relevant user attributes

solid line: $\mu(S')$ in our data
dashed line: $\mu(S)$ in user class
User impact — Qualitative analysis (2/2)

- **A**: Interactive (IA) vs non Interactive (NIA) users
  - interactive: tweet regularly, do many @-mentions and @-replies, mention many different users

- **B**: IA vs clique-Interactive (IAC)
  - IAC: interactive but not mentioning many different users

- **C**: Use links (L) vs does not (NL) when discussing the most prediction relevant topics (i.e., Politics and Showbiz)

- **D**: Topic focused (TF) vs topic overall (TO)

- **E**: ‘Serious’ (ST) vs ‘light’ (LT) topics

v.lampos@ucl.ac.uk

Summary

You’ve seen:

+ how user-generated data can be used to make inferences about
  ◦ collective mood / emotions
  ◦ real-world phenomena — flu, rainfall rates
  ◦ political preference — voting intention

+ a new class of bilinear models adaptive to the nature of social media content

+ how a simplified notion of impact is connected to the usage of social media platforms

Simple future challenges

– embed such derivations into real-world systems and enhance decision making (i.e., epidemiological surveillance tasks)

– further improvements on the applied supervised modelling (predictive models)
In collaboration with

**Trevor Cohn**, University of Melbourne

**Nello Cristianini**, University of Bristol

**Daniel Preoțiuc-Pietro**, University of Pennsylvania

**Nikolaos Aletras**, University College London

**Thomas Lansdall-Welfare**, University of Bristol

http://www.i-sense.org.uk/
Thank you

Any questions?

Download the slides from
http://www.lampos.net/research/talks-posters
References


Bach. **Bolasso: Model Consistent Lasso Estimation through the Bootstrap.** ICML, 2008.


Bouma. **Normalized (pointwise) mutual information in collocation extraction.** GSCL, 2009.


Gayo-Avello, Metaxas and Mustafaraj. **Limits of Electoral Predictions using Twitter.** ICWSM, 2011.

Hastie, Tibshirani and Friedman. **The Elements of Statistical Learning.** 2009.


Horst and Tuy. **Global Optimization: Deterministic Approaches.** 1996.
References II


Lampos, Preoțiuc-Pietro and Cohn. **A user-centric model of voting intention from Social Media.** ACL, 2013.


Liu, Ji and Ye. **Multi-task feature learning via efficient \( \ell_{2,1} \)-norm minimization.** UAI, 2009.


O’Connor, Balasubramanyan, Routledge and Smith. **From Tweets to polls: Linking text sentiment to public opinion time series.** ICWSM, 2010.


v.lampos@ucl.ac.uk  Slides: http://bit.ly/1GrxI8j


Tibshirani. **Regression Shrinkage and Selection via the LASSO.** JRSS, 1996.


Yuan and Lin. **Model selection and estimation in regression with grouped variables.** JRSS, 2006.
