Mining the social web: A series of statistical NLP case studies

Vasileios Lampos
Department of Computer Science
University College London

May, 2014
Key assumptions about social media

• a significant **sample of the population** uses them
• a significant amount of the published content is **geo-located**
• this content reflects on collective portions of real-life (opinions, events, phenomena)
  ◦ usually forming a **real-time** relationship
• it is **easy (?)** to collect, store and process this content
• and everyone seems to know how to use this “**big data**”

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

i think i have the flu but i still look fabulous

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

And what about the statistical significance of the computed statistical significance?
#inception_in_statistics
• 140 characters per published status (tweet)
• users can follow and can be followed
• embedded usage of topics (#rbnews, #inception_in_statistics)
• retweets (RT), @replies, @mentions, favourites
• real-time nature
• biased user demographics (13-15% of UK’s population is now on Twitter)
In this talk

Ways for harnessing social media information...

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

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

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

- to understand characteristics related to user impact
  (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

(Lansdall et al., 2012), (Strapparava, Valitutti, 2004) → WordNet Affect

joy

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

derivative of

anger & fear
Mood projections

Projections 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
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 difficult to interpret
- unsatisfactory prediction accuracy (estimates have large variance)
Regression basics — Ridge Regression

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

**Ridge Regression (RR)**

\[
\arg\min_{\mathbf{w}_*} \left\{ \| \mathbf{X}_* \mathbf{w}_* - \mathbf{y} \|_2^2 + \lambda \| \mathbf{w} \|_2^2 \right\}
\]

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

\[ \ell_1 \text{-norm regularisation or lasso} \text{ (Tibshirani, 1996)} \]

\[
\arg\min_{w_\ast} \left\{ \|Xw_\ast - y\|_2^2 + \lambda \|w\|_1 \right\}
\]

- no closed form solution — quadratic programming problem
+ Least Angle Regression (LAR) explores 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
- strongly corr. predictors \( \rightarrow \) model-inconsistent (Zhao, Yu, 2009)
Lasso for text regression

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

\( \ell_1 \)-norm regularisation or lasso

\[
\text{argmin}_{w_*} \left\{ \|X w_* - y\|_{\ell_2}^2 + \lambda \|w\|_{\ell_1} \right\}
\]
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 cumulative frequency vs z-scored official ILI rates

England & Wales (region D)

\[ r = 0.856 \]

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

- create a pool of unigram features by indexing all words in relevant web pages (Wikipedia, NHS pages)
- stop-words removed, Porter-stemming
- automatic unigram selection and weighting via lasso

**Selected uni-grams**


![Flu rate graph](image)

**England & Wales**

\[ r = 0.968 \]

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

- including bi-grams — hybrid combination with uni-grams
- fixing lasso’s inconsistencies with bootstrap lasso (Bach, 2008)

(Lampos, Cristianini, 2012)

v.lampos@ucl.ac.uk

Slides: http://bit.ly/1v3Jeiy
Regression basics — Elastic Net

- **Observations** \(x_i \in \mathbb{R}^m, \quad i \in \{1, \ldots, n\} \quad \rightarrow \quad X\)
- **Responses** \(y_i \in \mathbb{R}, \quad i \in \{1, \ldots, n\} \quad \rightarrow \quad y\)
- **Weights, Bias** \(w_j, \beta \in \mathbb{R}, \quad j \in \{1, \ldots, m\} \quad \rightarrow \quad w_* = [w; \beta]\)

**Linear Elastic Net (LEN)**

\[
\text{argmin}_{w_*} \left\{ \left\| X w_* - y \right\|_{\ell_2}^2 + \lambda_1 \left\| w \right\|_{\ell_2}^2 + \lambda_2 \left\| w \right\|_{\ell_1} \right\}
\]

+ ‘compromise’ between ridge regression (handles collinear predictors) and lasso (favours 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 models
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 \( w_j, \beta \in \mathbb{R}, \quad j \in \{1, \ldots, m\} \) — \( w_* = [w; \beta] \)

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 \( u_k, w_j, \beta \in \mathbb{R}, \quad k \in \{1, \ldots, p\}, \quad j \in \{1, \ldots, m\} \) — \( u, w, \beta \)
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** \( u_k, w_j, \beta \in \mathbb{R}, \quad k \in \{1, \ldots, p\}, j \in \{1, \ldots, m\} \)

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

- **users** $p \in \mathbb{Z}^+$
- **observations** $Q_i \in \mathbb{R}^{p \times m}$, $i \in \{1, \ldots, n\}$ — $\mathbf{X}$
- **responses** $y_i \in \mathbb{R}$, $i \in \{1, \ldots, n\}$ — $\mathbf{y}$
- **weights, bias** $u_k, w_j, \beta \in \mathbb{R}$, $k \in \{1, \ldots, p\}$ — $\mathbf{u}, \mathbf{w}, \beta$
  \hspace{0.5cm} $j \in \{1, \ldots, m\}$

$$
\argmin_{\mathbf{u}, \mathbf{w}, \beta} \left\{ \sum_{i=1}^{n} \left( \mathbf{u}^T Q_i \mathbf{w} + \beta - y_i \right)^2 + \psi(\mathbf{u}, \theta_u) + \psi(\mathbf{w}, \theta_w) \right\}
$$

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

- if $\psi(\mathbf{v}, \lambda) = \lambda \|\mathbf{v}\|_{\ell_1}$  
  Bilinear Lasso
- if $\psi(\mathbf{v}, \lambda_1, \lambda_2) = \lambda_1 \|\mathbf{v}\|_{\ell_2}^2 + \lambda_2 \|\mathbf{v}\|_{\ell_1}$  
  Bilinear Elastic Net (BEN)  
  \hspace{1cm} (Lampos et al., 2013)
Bilinear Elastic Net (BEN)

\[
\text{argmin}_{u,w,\beta} \left\{ \sum_{i=1}^{n} \left( u^T Q_i w + \beta - y_i \right)^2 + \lambda_{u1} \|u\|_2^2 + \lambda_{u2} \|u\|_1 \\
+ \lambda_{w1} \|w\|_2^2 + \lambda_{w2} \|w\|_1 \right\}
\]

**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
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)
- 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

But:

- 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
Voting intention modelling — Data (UK)

- 42K users distributed proportionally to regional population figures
- 60m tweets from 30/04/2010 to 13/02/2012
- 80,976 uni-grams (word features) → (Prețiu-Pietro et al., 2012)
- 240 voting intention polls (YouGov)
- 3 parties: Conservatives (CON), Labour Party (LAB), Liberal Democrats (LIB)
- main language: English
Voting intention modelling — Data (Austria)

- 1.1K users manually selected by Austrian political analysts (SORA)
- 800K tweets from 25/01 to 01/12/2012
- 22,917 unigrams (word features) → (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
Voting intention modelling — Evaluation

- 10-fold validation
  - train a model using data based on a set of contiguous polls $A$
  - test on the next $D = 5$ polls
  - expand training set to $\{A \cup D\}$, test on the next $|D'| = 5$ polls
- realistic scenario: train on past, predict future polls
- overall we 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 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>$\mu$</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>
<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>$\mu$</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>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

Austria

maybe multi-task learning will do better?
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
The $\ell_{2,1}$-norm regularisation

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

$\ell_{2,1}$-norm regularisation

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

- 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
- $\ell_{2,1}$-norm regularisation, i.e. the sum of $W$’s row $\ell_2$-norms (Argyriou et al., 2008; Liu et al., 2009) extends the notion of 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** $Q_i \in \mathbb{R}^{p \times m}$, $i \in \{1, \ldots, n\}$ — $X$
- **responses** $y_i \in \mathbb{R}^\tau$, $i \in \{1, \ldots, n\}$ — $Y$
- **weights, bias** $u_k, w_j, \beta \in \mathbb{R}^\tau$, $k \in \{1, \ldots, p\}$, $j \in \{1, \ldots, m\}$ — $U, W, \beta$

$$f (Q_i) = \text{tr} \left( U^T Q_i W \right) + \beta$$
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** $u_k, w_j, \beta \in \mathbb{R}^\tau$, $k \in \{1, \ldots, p\}$ — $\mathbf{U}, \mathbf{W}, \beta$

$$\arg\min_{\mathbf{U}, \mathbf{W}, \beta} \left\{ \sum_{t=1}^{\tau} \sum_{i=1}^{n} \left( u_t^T Q_i w_t + \beta_t - y_{ti} \right)^2 ight\}$$

$$+ \lambda_u \sum_{k=1}^{p} \| U_k \|_2 + \lambda_w \sum_{j=1}^{m} \| W_j \|_2$$

- BGL can be broken into 2 convex tasks: first learn $\{ \mathbf{W}, \beta \}$, then $\{ \mathbf{U}, \beta \}$ and vice versa + iterate through this process
Bilinear Group $\ell_{2,1}$ (BGL) (2/2)

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

- a feature (user/word) is selected for **all tasks** (not just one), but possibly with different weights
- especially useful in the **domain of politics** (e.g. user pro party A, against party B)
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>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>
<tr>
<td>BGL</td>
<td>1.785</td>
<td>1.595</td>
<td>1.054</td>
<td>1.478</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>
<tr>
<td>BGL</td>
<td>1.619</td>
<td>1.005</td>
<td>1.757</td>
<td>1.374</td>
<td>1.439</td>
</tr>
</tbody>
</table>
Voting intention modelling — BGL’s performance (2/2)

Polls

UK

BEN

BGL

Austria
<table>
<thead>
<tr>
<th>Party</th>
<th>Tweet</th>
<th>Score</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>CON</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>LAB</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</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(Labour)</td>
</tr>
<tr>
<td>LBD</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</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MP</td>
</tr>
<tr>
<td>SPÖ</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>ÖVP</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>GRÜ</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>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>

Translation: Inflation rate in Austria slightly down in July from 2.2 to 2.1%. Accommodation, Water, Energy more expensive.

Translation: can really recommend the book “res publica” by johannes #voggenhuber! Food for thought and so on #europe #democracy

Translation: Protest songs against the closing-down of the bachelor course of International Development: <link> #IDremains #uniburns #unirage
What does content tell us about users?

Predicting and characterising user impact on Twitter
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 activity

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 02/01/2011 to 28/02/2011 (Gardenhose stream — 10%) for creating “topic” clusters
  — data processed via (Prețiu-Pietro et al., 2012)
User impact — a simplified definition

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

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

Histogram of the user impact scores in our data set

\[ \mu(S) = 6.776 \]
## User activity features (1/2)

<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 activity features (2/2)

**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 (τ₁)</td>
<td>mph, humidity, barometer, gust, winds, hpa, temperature, kt nursing, nurse, rn, registered, bedroom, clinical, #news, estate, #hospital, rent, healthcare, therapist, condo, investment, furnished, medical, #nyc, occupational, investors, #ny</td>
</tr>
<tr>
<td>Healthcare, Finance, Housing (τ₂)</td>
<td>...</td>
</tr>
<tr>
<td>Politics (τ₃)</td>
<td>senate, republican, gop, police, arrested, voters, robbery, democrats, presidential, elections, charged, election, charges, #religion, arrest, repeal, dems, #christian, reform</td>
</tr>
<tr>
<td>Showbiz, Movies (τ₄)</td>
<td>damon, potter, #tvd, harry, elena, kate, portman, pattinson, hermione, jennifer, kristen, stefan, robert, catholic, stewart, katherine, lois, jackson, vampire, natalie, #vampirediaries</td>
</tr>
<tr>
<td>Commerce (τ₅)</td>
<td>chevrolet, inventory, coupon, toyota, mileage, sedan, nissan, adde, jeep, 4x4, 2002, #coupon, enhanced, #deal, dodge</td>
</tr>
<tr>
<td>Twitter hashtags (τ₆)</td>
<td>#teamfollowback, #500aday, #tfb, #instantfollowback, #ifollowback, #instantfollow, #followback</td>
</tr>
<tr>
<td>Social unrest (τ₇)</td>
<td>#egypt, #tunisia, #iran, #israel, #palestine, tunisia, arab, #jan25, iran, israel, protests, 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

• 3 models
  — user attributes (A), A + top-words (AW), A + n clusters (AC)
• Ridge Regression, Gaussian Process (GP)
• GP using a Squared Exponential (SE) kernel with Automatic Relevance Determination (ARD) (Rasmussen and Williams, 2006)

<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>τ</td>
<td>= 50</td>
</tr>
<tr>
<td>AC,</td>
<td>τ</td>
<td>= 100</td>
</tr>
</tbody>
</table>

Most **predictive / relevant** features
default profile image, # of historical tweets, # of unique @-mentions, # of tweets (last year), links (ratio), topic: *weather*, topic: *healthcare-finance*, topic: *politics*, days with nonzero tweets (ratio), @-replies (ratio)
Impact score distribution for user accounts with high (H) or low (L) values for the most relevant user attributes:

- Tweets in entire history ($\alpha_{11}$)
- Unique @-mentions ($\alpha_7$)
- Links ($\alpha_9$)
- @-replies ($\alpha_8$)
- Days with nonzero tweets ($\alpha_{18}$)

Impact score distribution:
- 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 (CIA)
— CIA: interactive but not mentioning many different users

C: Use links (L) vs does not (NL) when discussing most prevalent topics (Politics, Showbiz)

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

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

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

Nello Cristianini, University of Bristol

Trevor Cohn, University of Melbourne

Daniel Preoțiuc-Pietro, University of Pennsylvania

Nikolaos Aletras, University of Sheffield

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 I


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.


