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**) $\rightarrow$ higher confidence (**?**)

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)

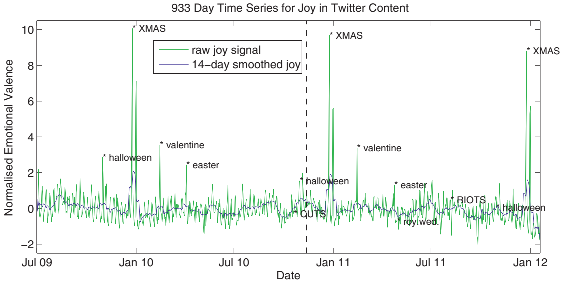
In this talk

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)

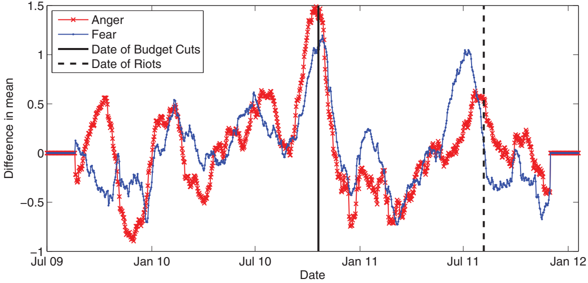
Proof of concept and a little more: extracting collective mood patterns

Time series of joy and anger based on UK tweets



joy

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

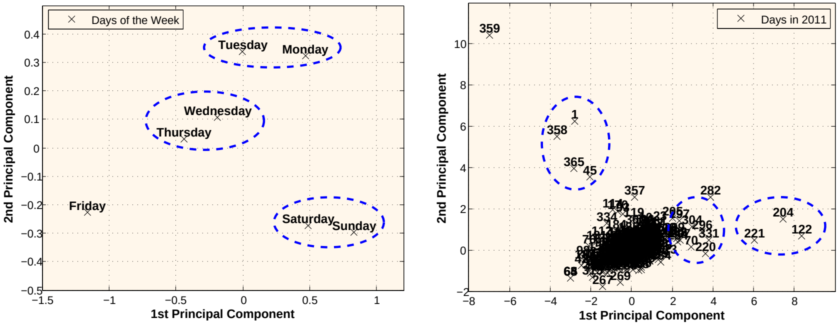


derivative of **anger & fear**

(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**)

(Lampos, 2012), (Strapparava, Valitutti, 2004) → WordNet Affect

Supervised learning

Primary outcomes (linear methods)

Regression basics — Ordinary Least Squares

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

Ordinary Least Squares (OLS)

$$\underset{\boldsymbol{w}_*}{\operatorname{argmin}} \|\boldsymbol{X}_* \boldsymbol{w}_* - \boldsymbol{y}\|_{\ell_2}^2 \Rightarrow \boldsymbol{w}_* = \left(\boldsymbol{X}_*^{\mathrm{T}} \boldsymbol{X}_*\right)^{-1} \boldsymbol{X}_*^{\mathrm{T}} \boldsymbol{y}$$

Why not?

– $\boldsymbol{X}_*^{\mathrm{T}} \boldsymbol{X}_*$ 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 $\boldsymbol{x}_i \in \mathbb{R}^m$, $i \in \{1, ..., n\}$ — $\boldsymbol{X}$
- responses $y_i \in \mathbb{R}$, $i \in \{1, ..., n\}$ — $\boldsymbol{y}$
- weights, bias $w_j, \beta \in \mathbb{R}$, $j \in \{1, ..., m\}$ — $\boldsymbol{w}_* = [\boldsymbol{w}; \beta]$

Ridge Regression (RR)

$$\underset{\boldsymbol{w}_*}{\operatorname{argmin}} \left\{ \|\boldsymbol{X}_* \boldsymbol{w}_* - \boldsymbol{y}\|_{\ell_2}^2 + \boxed{\lambda \|\boldsymbol{w}\|_{\ell_2}^2} \right\}$$

+ size constraint on the weight coefficients (**regularisation**)
  $\rightarrow$ 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 $\boldsymbol{x}_i \in \mathbb{R}^m$, $i \in \{1, ..., n\}$ — $\boldsymbol{X}$
- responses $y_i \in \mathbb{R}$, $i \in \{1, ..., n\}$ — $\boldsymbol{y}$
- weights, bias $w_j, \beta \in \mathbb{R}$, $j \in \{1, ..., m\}$ — $\boldsymbol{w}_* = [\boldsymbol{w}; \beta]$

$\boldsymbol{\ell_1}$–norm regularisation or **lasso** (Tibshirani, 1996)

$$\underset{\boldsymbol{w}_*}{\operatorname{argmin}} \left\{ \|\boldsymbol{X}_* \boldsymbol{w}_* - \boldsymbol{y}\|_{\ell_2}^2 + \lambda \|\boldsymbol{w}\|_{\ell_1} \right\}$$

− no closed form solution — quadratic programming problem

+ Least Angle Regression (LAR) → entire reg. path (Efron et al., 2004)

+ **sparse $\boldsymbol{w}$**, interpretability, better performance (Hastie et al., 2009)

− if $m > n$, at most n variables can be selected

− **co-linear** predictors → **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"

• n-gram frequencies	$\boldsymbol{x}_i \in \mathbb{R}^m$,	$i \in \{1, ..., n\}$	—	$\boldsymbol{X}$
• target phenomenon	$y_i \in \mathbb{R}$,	$i \in \{1, ..., n\}$	—	$\boldsymbol{y}$
• weights, bias	$w_j, \beta \in \mathbb{R}$,	$j \in \{1, ..., m\}$	—	$\boldsymbol{w}_* = [\boldsymbol{w}; \beta]$

lasso (*for text regression*)

$$\underset{\boldsymbol{w}_*}{\operatorname{argmin}} \left\{ \|\boldsymbol{X}_* \boldsymbol{w}_* - \boldsymbol{y}\|_{\ell_2}^2 + \lambda \|\boldsymbol{w}\|_{\ell_1} \right\}$$

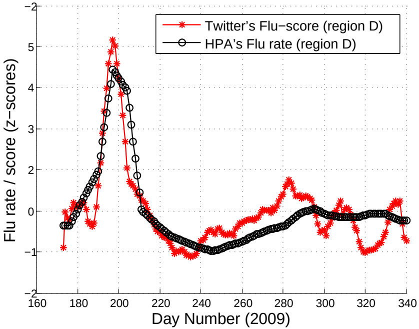
Nowcasting ILI rates from Twitter (1/2)

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 aggregate keyphrase frequency vs. official ILI rates



England & Wales (region D)

$r = .856$

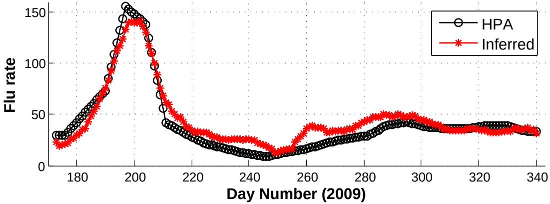
(Lampos, Cristianini, 2010)

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', ...

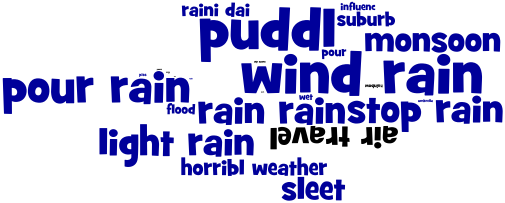


England & Wales
$r = .968$

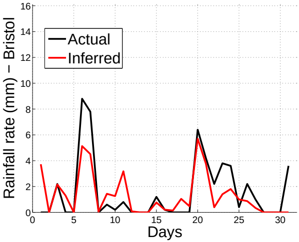
(Lampos, Cristianini, 2010)

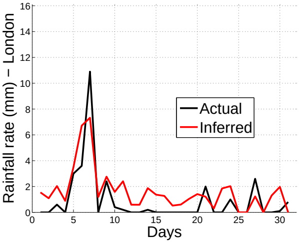
Nowcasting rainfall rates — a generalisation

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



Bristol





London

(Lampos, Cristianini, 2012)

Back to regression basics — Elastic Net

• observations	$\boldsymbol{x}_i \in \mathbb{R}^m$,	$i \in \{1, ..., n\}$	—	$\boldsymbol{X}$
• responses	$y_i \in \mathbb{R}$,	$i \in \{1, ..., n\}$	—	$\boldsymbol{y}$
• weights, bias	$w_j, \beta \in \mathbb{R}$,	$j \in \{1, ..., m\}$	—	$\boldsymbol{w}_* = [\boldsymbol{w}; \beta]$

linear **Elastic Net** (LEN)

$$\underset{\boldsymbol{w}_*}{\operatorname{argmin}} \left\{ \underbrace{\|\boldsymbol{X}_* \boldsymbol{w}_* - \boldsymbol{y}\|_{\ell_2}^2}_{\text{OLS}} + \underbrace{\lambda_1 \|\boldsymbol{w}\|_{\ell_2}^2}_{\text{RR reg.}} + \underbrace{\lambda_2 \|\boldsymbol{w}\|_{\ell_1}}_{\text{Lasso reg.}} \right\}$$

+ **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: $f(\boldsymbol{x}_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{w} + \beta$

• observations	$\boldsymbol{x}_i \in \mathbb{R}^m$,	$i \in \{1, ..., n\}$	—	$\boldsymbol{X}$
• responses	$y_i \in \mathbb{R}$,	$i \in \{1, ..., n\}$	—	$\boldsymbol{y}$
• weights, bias	$w_j, \beta \in \mathbb{R}$,	$j \in \{1, ..., m\}$	—	$\boldsymbol{w}_* = [\boldsymbol{w}; \beta]$

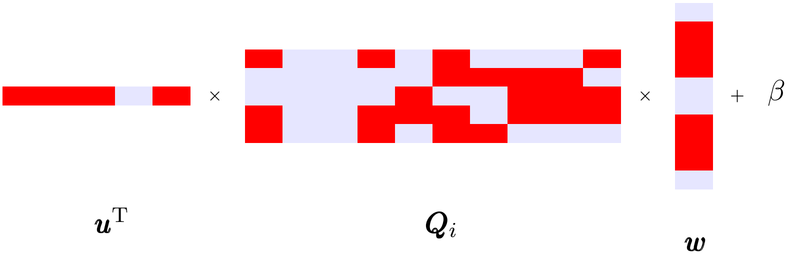
Bilinear regression: $f(\boldsymbol{Q}_i) = \boldsymbol{u}^{\mathrm{T}}\boldsymbol{Q}_i\boldsymbol{w} + \beta$

• users	$p \in \mathbb{Z}^+$			
• observations	$\boldsymbol{Q}_i \in \mathbb{R}^{p \times m}$,	$i \in \{1, ..., n\}$	—	$\boldsymbol{\mathcal{X}}$
• responses	$y_i \in \mathbb{R}$,	$i \in \{1, ..., n\}$	—	$\boldsymbol{y}$
• weights, bias	$u_k, w_j, \beta \in \mathbb{R}$,	$k \in \{1, ..., p\}$ $j \in \{1, ..., m\}$	—	$\boldsymbol{u}$, $\boldsymbol{w}$, β

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

users	$p \in \mathbb{Z}^+$			
observations	$\boldsymbol{Q}_i \in \mathbb{R}^{p \times m}$,	$i \in \{1, ..., n\}$	—	$\boldsymbol{\mathcal{X}}$
responses	$y_i \in \mathbb{R}$,	$i \in \{1, ..., n\}$	—	$\boldsymbol{y}$
weights, bias	$u_k, w_j, \beta \in \mathbb{R}$,	$k \in \{1, ..., p\}$ $j \in \{1, ..., m\}$	—	$\boldsymbol{u}$, $\boldsymbol{w}$, β

$$f\left(\boldsymbol{Q}_i\right) = \boldsymbol{u}^{\mathrm{T}} \boldsymbol{Q}_i \boldsymbol{w} + \beta$$



Bilinear text regression — Regularisation

• users	$p \in \mathbb{Z}^+$			
• observations	$\boldsymbol{Q}_i \in \mathbb{R}^{p \times m}$,	$i \in \{1, ..., n\}$	—	$\boldsymbol{\mathcal{X}}$
• responses	$y_i \in \mathbb{R}$,	$i \in \{1, ..., n\}$	—	$\boldsymbol{y}$
• weights, bias	$u_k, w_j, \beta \in \mathbb{R}$,	$k \in \{1, ..., p\}$ $j \in \{1, ..., m\}$	—	$\boldsymbol{u}$, $\boldsymbol{w}$, β

$$\underset{\boldsymbol{u},\boldsymbol{w},\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(\boldsymbol{u}^{\mathrm{T}} \boldsymbol{Q}_i \boldsymbol{w} + \beta - y_i \right)^2 + \psi(\boldsymbol{u}, \theta_u) + \psi(\boldsymbol{w}, \theta_w) \right\}$$

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

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

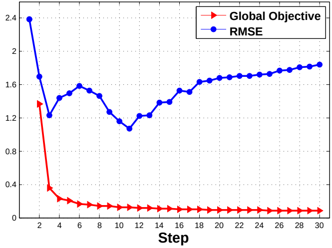
(Lampos et al., 2013)

Learning the parameters of BEN

$$\underset{\boldsymbol{u},\boldsymbol{w},\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(\boldsymbol{u}^{\mathrm{T}} \boldsymbol{Q}_i \boldsymbol{w} + \beta - y_i \right)^2 + \lambda_{u_1} \|\boldsymbol{u}\|_{\ell_2}^2 + \lambda_{u_2} \|\boldsymbol{u}\|_{\ell_1} \right.$$
$$\left. + \lambda_{w_1} \|\boldsymbol{w}\|_{\ell_2}^2 + \lambda_{w_2} \|\boldsymbol{w}\|_{\ell_1} \right\}$$

Bi-convexity: fix $\boldsymbol{u}$, learn $\boldsymbol{w}$ and vice versa
Iterating through convex optimisation tasks: **convergence**
(Al-Khayyal, Falk, 1983; Horst, Tuy, 1996)

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

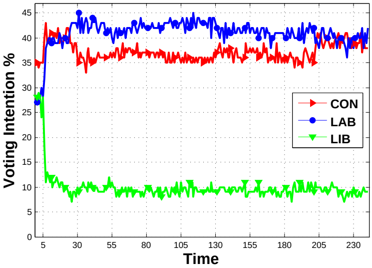
reviewed in (Gayo-Avello, 2013)

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

Voting intention modelling — Data (UK)

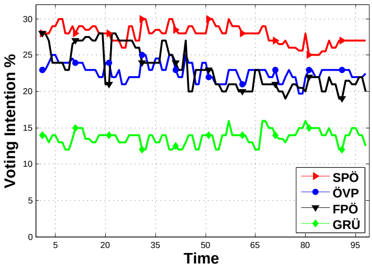
- **42K users** distributed proportionally to regional population figures
- **60 million tweets** from 30/04/2010 to 13/02/2012
- **80,976 1-grams** → (Preţiuc-Pietro et al., 2012)
- **240** voting intention **polls** (YouGov)
- **3 parties**: Conservatives (**CON**), Labour Party (**LAB**), Liberal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

- $\mathbf{B}_{\boldsymbol{\mu}}$: constant prediction based on $\mu(\boldsymbol{y})$ in the training set
- $\mathbf{B}_{\mathbf{last}}$: constant prediction based on $\mathrm{last}(\boldsymbol{y})$ 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

	CON	LAB	LIB	μ
B_{μ}	2.272	1.663	**1.136**	1.69
B_{last}	2	2.074	1.095	1.723
LEN	3.845	2.912	2.445	3.067
BEN	**1.939**	**1.644**	**1.136**	**1.573**

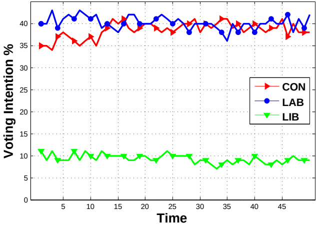
'Austria' case study

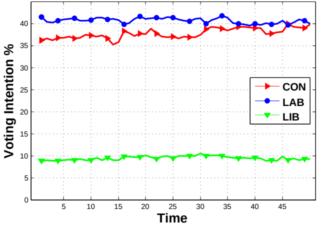
	SPÖ	ÖVP	FPÖ	GRÜ	μ
B_{μ}	1.535	1.373	3.3	1.197	1.851
B_{last}	**1.148**	1.556	**1.639**	1.536	1.47
LEN	1.291	**1.286**	2.039	**1.152**	**1.442**
BEN	1.392	1.31	2.89	1.205	1.699

Voting intention modelling — BEN's performance (2/2)

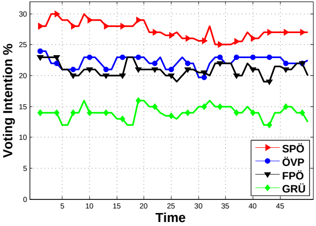
Polls | BEN

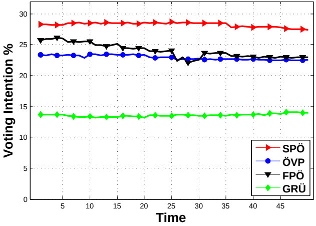
UK





Austria





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

$$\|\boldsymbol{W}\|_{2,1} = \sum_{j=1}^{m} \|\boldsymbol{W}_j\|_{\ell_2}, \text{ where } \boldsymbol{W}_j \text{ denotes the } j\text{-th row}$$

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

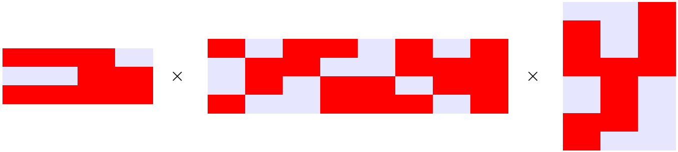
$$\underset{\boldsymbol{W},\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \|\boldsymbol{X}\boldsymbol{W} - \boldsymbol{Y}\|_{\ell_F}^2 + \lambda \sum_{j=1}^{m} \|\boldsymbol{W}_j\|_{\ell_2} \right\}$$

- **multi-task learning**: instead of $\boldsymbol{w} \in \mathbb{R}^m$, learn $\boldsymbol{W} \in \mathbb{R}^{m \times \tau}$, where τ is the number of tasks
- $\ell_{2,1}$-norm regularisation $\rightarrow$ sum of $\boldsymbol{W}$'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 τ-dimensional rows of $\boldsymbol{W}$

Bilinear multi-task learning

- tasks $\tau \in \mathbb{Z}^+$
- users $p \in \mathbb{Z}^+$
- observations $\boldsymbol{Q}_i \in \mathbb{R}^{p \times m}$, $i \in \{1, ..., n\}$ — $\boldsymbol{\mathcal{X}}$
- responses $\boldsymbol{y}_i \in \mathbb{R}^{\tau}$, $i \in \{1, ..., n\}$ — $\boldsymbol{Y}$
- weights, bias $\boldsymbol{u}_k, \boldsymbol{w}_j, \boldsymbol{\beta} \in \mathbb{R}^{\tau}$, $k \in \{1, ..., p\}$, $j \in \{1, ..., m\}$ — $\boldsymbol{U}$, $\boldsymbol{W}$, $\boldsymbol{\beta}$

$$f(\boldsymbol{Q}_i) = \operatorname{tr}\left(\boldsymbol{U}^{\mathrm{T}} \boldsymbol{Q}_i \boldsymbol{W}\right) + \boldsymbol{\beta}$$



$\boldsymbol{U}^{\mathrm{T}}$ $\boldsymbol{Q}_i$ $\boldsymbol{W}$

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

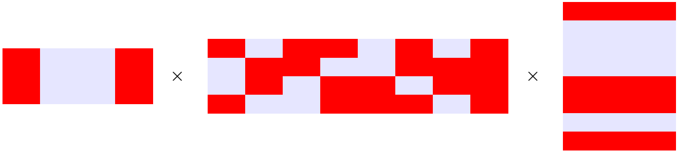
- tasks $\tau \in \mathbb{Z}^+$
- users $p \in \mathbb{Z}^+$
- observations $\boldsymbol{Q}_i \in \mathbb{R}^{p \times m}$, $i \in \{1, ..., n\}$ — $\boldsymbol{\mathcal{X}}$
- responses $\boldsymbol{y}_i \in \mathbb{R}^{\tau}$, $i \in \{1, ..., n\}$ — $\boldsymbol{Y}$
- weights, bias $\boldsymbol{u}_k, \boldsymbol{w}_j, \boldsymbol{\beta} \in \mathbb{R}^{\tau}$, $k \in \{1, ..., p\}$, $j \in \{1, ..., m\}$ — $\boldsymbol{U}, \boldsymbol{W}, \boldsymbol{\beta}$

$$\underset{\boldsymbol{U},\boldsymbol{W},\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{t=1}^{\tau} \sum_{i=1}^{n} \left(\boldsymbol{u}_t^{\mathrm{T}} \boldsymbol{Q}_i \boldsymbol{w}_t + \beta_t - y_{ti} \right)^2 + \lambda_u \sum_{k=1}^{p} \|\boldsymbol{U}_k\|_2 + \lambda_w \sum_{j=1}^{m} \|\boldsymbol{W}_j\|_2 \right\}$$

- **Learning**: 2 convex tasks $\rightarrow$ first learn $\{\boldsymbol{W}, \boldsymbol{\beta}\}$, then $\{\boldsymbol{U}, \boldsymbol{\beta}\}$ and vice versa; iterate through this process

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

$$\underset{\boldsymbol{U},\boldsymbol{W},\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{t=1}^{\tau} \sum_{i=1}^{n} \left(\boldsymbol{u}_t^{\mathrm{T}} \boldsymbol{Q}_i \boldsymbol{w}_t + \beta_t - y_{ti} \right)^2 + \lambda_u \sum_{k=1}^{p} \|\boldsymbol{U}_k\|_2 + \lambda_w \sum_{j=1}^{m} \|\boldsymbol{W}_j\|_2 \right\}$$



$\boldsymbol{U}^{\mathrm{T}}$ $\boldsymbol{Q}_i$ $\boldsymbol{W}$

- 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

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

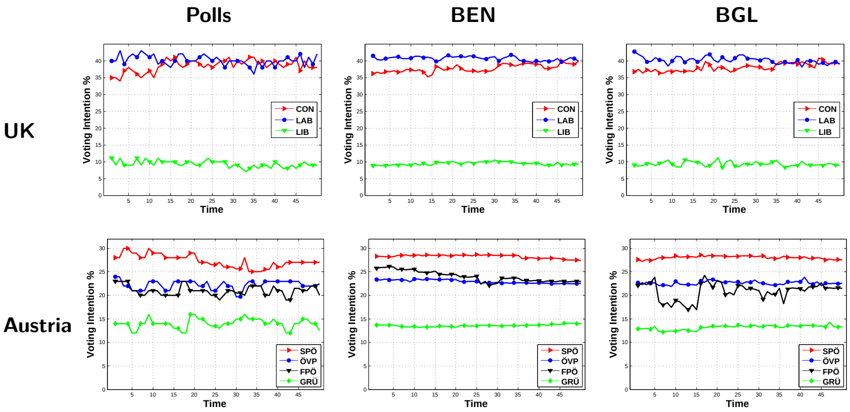
'UK' case study

	CON	LAB	LIB	μ
B_μ	2.272	1.663	1.136	1.69
B_{last}	2	2.074	1.095	1.723
LEN	3.845	2.912	2.445	3.067
BEN	1.939	1.644	1.136	1.573
BGL	**1.785**	**1.595**	**1.054**	**1.478**

'Austria' case study

	SPÖ	ÖVP	FPÖ	GRÜ	μ
B_μ	1.535	1.373	3.3	1.197	1.851
B_{last}	**1.148**	1.556	**1.639**	1.536	1.47
LEN	1.291	1.286	2.039	**1.152**	1.442
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)



Voting intention modelling — Qualitative insight

Party	Tweet	Score	Author
CON	PM in friendly chat with top EU mate, Sweden's Fredrik Reinfeldt, before family photo	1.334	Journalist
LAB	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	-0.552	Politician (Labour)
LBD	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*)	0.874	LibDem MP
SPÖ	Inflationsrate in Ö. im Juli leicht gesunken: von 2,2 auf 2,1%. Teurer wurde Wohnen, Wasser, Energie. **Translation:** *Inflation rate in Austria slightly down in July from 2,2 to 2,1%. Accommodation, Water, Energy more expensive.*	0.745	Journalist
ÖVP	kann das buch "res publica" von johannes #voggenhuber wirklich empfehlen! so zum nachdenken und so... #europa #demokratie **Translation:** *can really recommend the book "res publica" by johannes #voggenhuber! Food for thought and so on #europe #democracy*	-2.323	User
GRÜ	Protestsong gegen die Abschaffung des Bachelor-Studiums Internationale Entwicklung: *<link>* #IEbleibt #unibrennt #uniwut **Translation:** *Protest songs against the closing-down of the bachelor course of International Development: <link> #IDremains #uniburns #unirage*	1.45	Student Union

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)

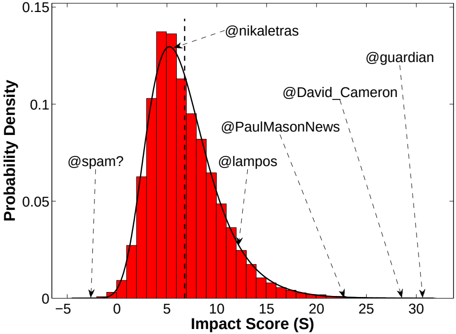
(Lampos et al., 2014)

User impact — a simplified definition

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

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

Histogram of the user impact scores in our data set
$\mu(S) = 6.776$



User activity features

a_1	# of tweets
a_2	proportion of retweets
a_3	proportion of non-duplicate tweets
a_4	proportion of tweets with hashtags
a_5	hashtag-tokens ratio in tweets
a_6	proportion of tweets with @-mentions
a_7	# of unique @-mentions in tweets
a_8	proportion of tweets with @-replies
a_9	links ratio in tweets
a_{10}	# of favourites the account made
a_{11}	total # of tweets (entire history)
a_{12}	using default profile background (binary)
a_{13}	using default profile image (binary)
a_{14}	enabled geolocation (binary)
a_{15}	population of account's location
a_{16}	account's location latitude
a_{17}	account's location longitude
a_{18}	proportion of days with nonzero tweets

User participation in topic-specific discussions

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

Label	Cluster's words ranked by centrality
Weather (τ_1)	mph, humidity, barometer, gust, winds, hpa, temperature, kt
Healthcare, Finance, Housing (τ_2)	nursing, nurse, rn, registered, bedroom, clinical, #news, estate, #hospital, rent, healthcare, therapist, condo, investment, furnished, medical, #nyc, occupational, investors, #ny
Politics (τ_3)	senate, republican, gop, police, arrested, voters, robbery, democrats, presidential, elections, charged, election, charges, #religion, arrest, repeal, dems, #christian, reform
Showbiz, Movies (τ_4)	damon, potter, #tvd, harry, elena, kate, portman, pattinson, hermione, jennifer, kristen, stefan, robert, catholic, stewart, katherine, lois, jackson, vampire, natalie, #vampirediaries
Commerce (τ_5)	chevrolet, inventory, coupon, toyota, mileage, sedan, nissan, adde, jeep, 4x4, 2002, #coupon, enhanced, #deal, dodge
Twitter hashtags (τ_6)	#teamfollowback, #500aday, #tfb, #instantfollowback, #ifollowback, #instantfollow, #followback
Social unrest (τ_7)	#egypt, #tunisia, #iran, #israel, #palestine, tunisia, arab, #jan25, iran, israel, protests, egypt, #yemen, #iranelection, israeli, #jordan, regime, yemen, #gaza, protesters, #lebanon
...	...

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/`

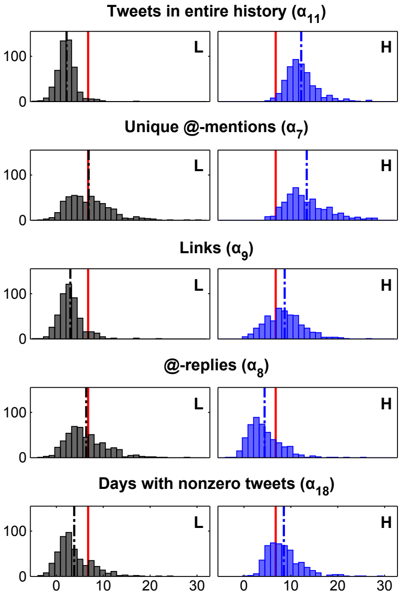
Performance estimates

Model	Linear (RR)		Nonlinear (GP)	
	r	RMSE	r	RMSE
A	.667	2.642	.759	2.298
AW	.712	2.529	.768	2.263
AC, $\|\tau\| = 50$	.703	2.518	.774	2.234
AC, $\|\tau\| = 100$	.714	2.480	**.780**	**2.210**

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)

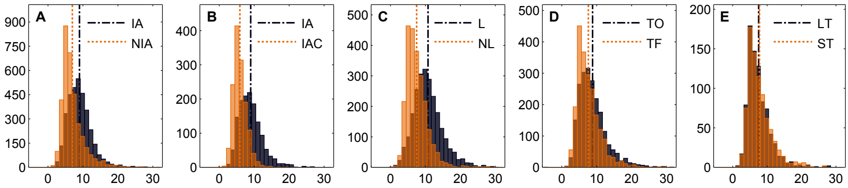
User impact — Qualitative analysis (1/2)



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

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

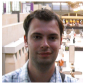
– embed such derivations into real-world systems and enhance decision making (i.e., epidemiological surveillance tasks)

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

In collaboration with

Trevor Cohn, University of Melbourne

Nello Cristianini, University of Bristol



Daniel Preoţiuc-Pietro, University of Pennsylvania

Nikolaos Aletras, University College London

Thomas Lansdall-Welfare, University of Bristol

i-sense
EPSRC IRC in Early Warning Sensing
Systems for Infectious Diseases

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

Thank you

Any questions?

Download the slides from

http://www.lampos.net/research/talks-posters

References I

Al-Khayyal and Falk. **Jointly Constrained Biconvex Programming**. MOR, 1983.

Argyriou, Evgeniou and Pontil. **Convex multi-task feature learning**. Machine Learning, 2008.

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

Beck and Teboulle. **A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems**. J. Imaging Sci., 2009.

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

Caruana. **Multitask Learning**. Machine Learning, 1997.

Efron, Hastie, Johnstone and Tibshirani. **Least Angle Regression**. The Annals of Statistics, 2004.

Gayo-Avello. **A Meta-Analysis of State-of-the-Art Electoral Prediction From Twitter Data**. SSCR, 2013.

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

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

Hoerl and Kennard. **Ridge regression: Biased estimation for nonorthogonal problems**. Technometrics, 1970.

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

References II

Lampos and Cristianini. **Tracking the flu pandemic by monitoring the Social Web**. CIP, 2010.

Lampos and Cristianini. **Nowcasting Events from the Social Web with Statistical Learning**. ACM TIST, 2012.

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

Lampos, Aletras, Preoţiuc-Pietro and Cohn. **Predicting and Characterising User Impact on Twitter**. EACL, 2014.

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

von Luxburg. **A tutorial on spectral clustering**. Statistics and Computing, 2007.

Mairal, Jenatton, Obozinski and Bach. **Network Flow Algorithms for Structured Sparsity**. NIPS, 2010.

Metaxas, Mustafaraj and Gayo-Avello. **How (not) to predict elections**. SocialCom, 2011.

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

Preoţiuc-Pietro, Samangooei, Cohn, Gibbins and Niranjan. **Trendminer: An architecture for real time analysis of social media text**. ICWSM, 2012.

Rasmussen and Williams. **Gaussian Processes for Machine Learning**. MIT Press, 2006.

References III

Strapparava and Valitutti. **Wordnet-Affect: An affective extension of WordNet**. LREC, 2004.

Tausczik and Pennebaker. **The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods**. JLSP, 2010.

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

Tumasjan, Sprenger, Sandner and Welpe. **Predicting elections with Twitter: What 140 characters reveal about political sentiment**. ICWSM, 2010.

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

Zhao and Yu. **On model selection consistency of LASSO**. JMLR, 2006.

Zhou and Hastie. **Regularization and variable selection via the elastic net**. JRSS, 2005.