Linear Regression

- Standard linear regression framework:

\[ f(X) = w^T X + \beta \]

- The optimization objective is:

\[ \{w^*, \beta^*\} = \arg \min \sum_{i=1}^{n} (w^T x_i + \beta - y_i)^2 + \Psi(w, \rho) \]

where \(\Psi\) is the regularisation (e.g. L1, L2, EN)
• Task: predict real valued outputs based on textual variables (e.g. word counts)

Lampos V., Cristianini N. (2010)
http://geopatterns.enm.bris.ac.uk/epidemics/

• Other examples: voting intention, financial indicators, weather, etc.
Bilinear Regression

• Bilinear regression framework:

\[ f(X) = u^T Xw + \beta \]

• The optimization objective is:

\[ \{w^*, u^* \beta^*\} = \arg\min \sum_{i=1}^{n} (u^T x_i w + \beta - y_i)^2 + \psi_1 (u, \rho_1) + \psi_2 (w, \rho_2) \]

• Goal: the model `divides’ the variables into two disjoint sets, each meant to model different aspects
Outline

• Use case
• Motivation
• Data
• 2 models: BEN, BGL
• Learning
• Results
• Current and future work
Trendminer project

• ‘Large scale, cross-lingual trend mining and summarization of real time media streams’
• 7 organisations; we work with University of Southampton and SORA on machine learning
• application to predicting political polls and financial indicators

www.trendminer-project.eu
Use case

• predicting political polls (not elections!)
• strong baselines, realistic evaluation
• 2 different use cases (U.K. and Austria)

UK polls, 04/2010 – 02/2012
Ö. polls, 01/2012 – 12/2012
Motivation

- Twitter and real population demographics are different
- Social media has biased opinions, not the most mentioned/positive sentiment party is indicative of real world trends
- More similar setup to traditional polls
- Most of the users are not informative for our task and all their tweets represent noise
Motivation

• only a few words are informative of the task
• we want to obtain a model of sparse users & sparse words
• tune based on existing polls
• regression learns weights for features without using prior knowledge, making models more portable
Data

• collection focused on **all** the data from users of Twitter
  
  40000 U.K. (random)
  
  60 m. tweets
  
  1200 Austrian (selected by pol. scientists)
  
  800k tweets
Model

• Bilinear predictive model:

\[ f(X) = u^T Xw + \beta \]

\[ u \rightarrow \text{user weights}, \ w \rightarrow \text{word weights}, \ X \rightarrow \text{word/user counts} \]

• The optimization objective is:

\[ \{w^*, u^*, \beta^*\} = \text{argmin} \sum_{i=1}^{n} (u^T x_i w + \beta - y_i)^2 \]

\[ + \psi_1 (u, \rho_1) + \psi_2 (w, \rho_2) \]

\[ Y \rightarrow \text{response variable}, \ \psi_{1,2} \rightarrow \text{regularisers for users and words} \]
BEN (Bilinear Elastic Net)

- Regularizers are both Elastic Nets
- a BEN model for predicting each party’s score

**Drawback:** expect shared information between the tasks (e.g. + LAB is likely to be – CON)
Model

• build a bilinear model that learns multiple tasks and shares strength across them
• we use the Group LASSO inside the bilinear framework
• features inside a group have to be all zero/non-zero for all the tasks
• each group is the same word/user for each task
BGL (Bilinear Group Lasso)

• the tasks are predicting each party’s score
• optimisation task is:

$$\{w^*, u^*, \beta^*\} = \argmin \sum_{t=1}^{T} \sum_{i=1}^{n} (u_t Q_i w_t + \beta_t - y_{ti})^2 +$$

$$+ \lambda_1 \sum_{j=1}^{m} \|W_j\|_2 + \lambda_2 \sum_{k=1}^{p} \|U_k\|_2$$
• Biconvex learning task: solved by a repeated application of 2 convex processes
• Regulariser parameters are fixed and found using grid search on validation
• Empirically choose to stop after 4 steps
Results – U.K.

Ground truth

BGL

BEN

<table>
<thead>
<tr>
<th></th>
<th>CON</th>
<th>LAB</th>
<th>LBD</th>
<th>μ</th>
</tr>
</thead>
<tbody>
<tr>
<td>B_µ</td>
<td>2.272</td>
<td>1.663</td>
<td>1.136</td>
<td>1.69</td>
</tr>
<tr>
<td>B_{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>
## Results – U.K.

<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>Have Liberal Democrats broken electoral rules? Blog on Labour complaint to cabinet secretary</td>
<td>-0.991</td>
<td>Journalist</td>
</tr>
<tr>
<td>LAB</td>
<td>Blog Post Liverpool: City of Radicals Website now Live &lt;link&gt; #liverpool #art</td>
<td>1.954</td>
<td>Art Fanzine</td>
</tr>
<tr>
<td></td>
<td>I am so pleased to head 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>Political (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 MP</td>
</tr>
<tr>
<td></td>
<td>Blog Post Liverpool: City of Radicals 2011 – More Details Announced #liverpool #art</td>
<td>-0.521</td>
<td>Art Fanzine</td>
</tr>
</tbody>
</table>
Results – Austria

**Ground truth**

**BEN**

<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><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>
<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><strong>1.005</strong></td>
<td>1.757</td>
<td>1.374</td>
<td><strong>1.430</strong></td>
</tr>
<tr>
<td>Party</td>
<td>Tweet</td>
<td>Score</td>
<td>Author</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>--------</td>
<td>---------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPO</td>
<td>Inflationsrate in O¨ . im Juli leicht gesunken: von 2,2 auf 2,1%. Teurer wurde Wohnen, Wasser, Energie.</td>
<td>0.745</td>
<td>Journalist</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hans Rauscher zu Felix #Baumgartner “A klaner Hitler” &lt;link&gt;</td>
<td>-1.711</td>
<td>Journalist</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OVP</td>
<td>#IchPirat setze mich dafu¨r ein, dass eine große Koalition mathematisch verhindert wird! 1.Geige: #Gruene + #FPOe + #OeVP</td>
<td>4.953</td>
<td>User</td>
<td></td>
<td></td>
</tr>
<tr>
<td></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>
<td></td>
<td></td>
</tr>
<tr>
<td>FPO</td>
<td>Neue Kampagne der #Krone zur #Wehrpflicht: “GIB BELLO EINE STIMME!”</td>
<td>7.44</td>
<td>Political Satire</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kampagne der Wiener SPO “zum Zusammenleben” spielt Rechtspopulisten in die H¨ande &lt;link&gt;</td>
<td>-3.44</td>
<td>Human Rights</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRU</td>
<td>Protestsong gegen die Abschaffung des Bachelor-Studiums Internationale Entwicklung: &lt;link&gt; #IEbleibt #unibrennt #uniwu</td>
<td>1.45</td>
<td>Student Union</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pilz “ich will in dieser Republik weder kriminelle Asylwerber, noch kriminelle orange Politiker” - BZO¨ -Abschiebung ok, aber wohin? #amPunkt</td>
<td>-2.172</td>
<td>User</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Current work

- classification
- financial applications
- online implementation
- use clusters of features
Future work

• regional analysis

• include other user features (e.g. location)

• explore other pairs of variables for different tasks

• non-stationarity
Team

Bill Lampos
Sheffield

Trevor Cohn
Sheffield

Sina Samangooei
Southampton
Publications

A user centric model of voting intention from Social Media
Lampos V., Preotiuc-Pietro D., Cohn T.
ACL 2013, www.preotiuc.ro

Regression models of trends. Tools for mining non-stationary data: functional protoype
Samangooei S., Lampos V., Cohn T., Gibbins N., Niranjan M.
Public deliverable, www.trendminer-project.eu
Thank you !