A user-centric model of voting intention from Social Media

Vasileios Lampos, Daniel Preoţiuc-Pietro and **Trevor Cohn**
Computer Science Department
University of Sheffield, UK
{v.lampos,d.preotiuc,t.cohn}@dcs.shef.ac.uk

Abstract

Social Media contain a multitude of user opinions which can be used to predict real-world phenomena in many domains including politics, finance and health. Most existing methods treat these problems as linear regression, learning to relate word frequencies and other simple features to a known response variable (*e.g.*, voting intention polls or financial indicators). These techniques require very careful filtering of the input texts, as most Social Media posts are irrelevant to the task. In this paper, we present a novel approach which performs high quality filtering automatically, through modelling not just words but also users, framed as a bilinear model with a sparse regulariser. We also consider the problem of modelling groups of related output variables, using a structured multi-task regularisation method. Our experiments on voting intention prediction demonstrate strong performance over large-scale input from Twitter on two distinct case studies, outperforming competitive baselines.

1 Introduction

Web Social Media platforms have ushered a new era in human interaction and communication. The main by-product of this activity is vast amounts of user-generated content, a type of information that has already attracted the interest of both marketeers and scientists because it offers – for the first time at a large-scale – unmediated access to peoples' observations and opinions.

One exciting avenue of research concentrates on mining interesting signals automatically from this stream of text input. For example, by exploiting Twitter posts, it is possible to infer time series that correlate with financial indicators (Bollen et al., 2011), track infectious diseases (Lampos and Cristianini, 2010; Lampos et al., 2010; Paul and Dredze, 2011) and, in general, nowcast the magnitude of events emerging in real-life (Sakaki et al., 2010; Lampos and Cristianini, 2012). Other studies suggest ways for modelling opinions encapsulated in this content in order to forge branding strategies (Jansen et al., 2009) or understand various socio-political trends (Tumasjan et al., 2010; O'Connor et al., 2010; Lansdall-Welfare et al., 2012). The main theme of the aforementioned works is linear regression between word frequencies and a real-world quantity. They also tend to incorporate hand-crafted lists of search terms to filter irrelevant content and use sentiment analysis lexicons for extracting opinion bias. Consequently, they are quite often restricted to a specific application and therefore, generalise poorly to new data sets (Gayo-Avello et al., 2011).

In this paper, we propose a generic method that aims to be independent of the characteristics described above (use of search terms or sentiment analysis tools). Our approach is able to explore not only word frequencies, but also the space of users by introducing a **bilinear** formulation for this learning task. Regularised regression on both spaces allows for an automatic selection of the most important terms and users, performing at the same time an improved noise filtering. In addition, more advanced regularisation functions enable **multi-task learning** schemes that can exploit shared structure in the feature space. The latter property becomes very useful in **multi-output regression** scenarios, where selected features are expected to have correlated as well as anti-correlated impact on each output (*e.g.*, when inferring voting intentions for competing political parties).

We evaluate our methods on the domain of politics using data from the microblogging service of Twitter to infer voting trends. Our pro-

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 993–1003,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

posed framework is able to successfully predict voting intentions for the top-3 and top-4 parties in the United Kingdom (UK) and Austria respectively. In both case studies – bound by different characteristics (including language, time-span and number of users) – the average prediction error is smaller than 1.5% for our best model using multi-task learning. Finally, our qualitative analysis shows that the models uncover interesting and semantically interpretable insights from the data.

2 Data

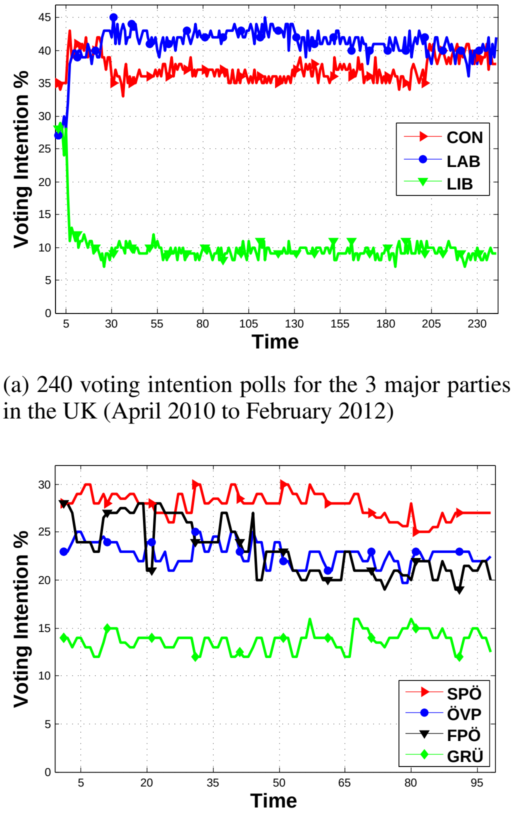
For the evaluation of the proposed methodologies we have created two data sets of Social Media content with different characteristics based in the UK and Austria respectively. They are used for performing regression aiming to infer voting intention polls in those countries. Data processing is performed using the TrendMiner architecture for Social Media analysis (Preoţiuc-Pietro et al., 2012).

2.1 Tweets from users in the UK

The first data set (we refer to it as $\mathbf{C_{uk}}$) used in our experimental process consists of approx. 60 million tweets produced by approx. 42K UK Twitter users from 30/04/2010 to 13/02/2012. We assumed each user to be from the UK, if the location field in their profile matched with a list of common UK locations and their time zone was set to G.M.T. In this way, we were able to extract hundreds of thousands of UK users, from which we sub-sampled 42K users to be distributed across the UK geographical regions proportionally to their population figures.[1]

2.2 Tweets for Austria

The second data set ($\mathbf{C_{au}}$) is shorter in terms of the number of users involved (1.1K), its time span (25/01 to 01/12/2012) and, consequently, of the total number of tweets considered (800K). However, this time the selection of users has been made by Austrian political experts who decided which accounts to monitor by subjectively assessing the value of information they may provide towards political-oriented topics. Still, we assume that the different users will produce information of varying quality, and some should be eliminated entirely. However, we emphasise that there may be smaller



(a) 240 voting intention polls for the 3 major parties in the UK (April 2010 to February 2012)

(b) 98 voting intention polls for the 4 major parties in Austria (January to December 2012)

Figure 1: Voting intention polls for the UK and Austria.

potential gains from user modelling compared to the UK case study. Another important distinction is language, which for this data set is primarily German with some English.

2.3 Ground Truth

The ground truth for training and evaluating our regression models is formed by voting intention polls from YouGov (UK) and a collection of Austrian pollsters[2] – as none performed high frequency polling – for the Austrian case study. We focused on the three major parties in the UK, namely Conservatives (CON), Labour (LAB) and Liberal Democrats (LBD) and the four major parties in Austria, namely the Social Democratic Party (SPÖ), People's Party (ÖVP), Freedom Party (FPÖ) and the Green Alternative Party (GRÜ). Matching with the time spans of the data sets described in the previous sections, we have acquired 240 unique polls for the UK and 65 polls for Austria. The latter have been expanded to 98 polls by replicating the poll of day i for day

[1] Data collection was performed using Twitter API, `http://dev.twitter.com/`, to extract all posts for our target users.

[2] Wikipedia, `http://de.wikipedia.org/wiki/Nationalratswahl_in_\%D6sterreich_2013`.

$i - 1$ where possible.[3] There exists some interesting variability towards the end for the UK polls (Fig. 1a), whereas for the Austrian case, the main changing point is between the second and the third party (Fig. 1b).

3 Methods

The textual content posted on Social Media platforms unarguably contains valuable information, but quite often it is hidden under vast amounts of unstructured user generated input. In this section, we propose a set of methods that build on one another, which aim to filter the non desirable noise and extract the most informative features not only based on word frequencies, but also by incorporating users in this process.

3.1 The bilinear model

There exist a number of different possibilities for incorporating user information into a regression model. A simple approach is to expand the feature set, such that each user's effect on the response variable can be modelled separately. Although flexible, this approach would be doomed to failure due to the sheer size of the resulting feature set, and the propensity to overfit all but the largest of training sets. One solution is to group users into different types, such as journalist, politician, activist, etc., but this presupposes a method for classification or clustering of users which is a non-trivial undertaking. Besides, these naïve approaches fail to account for the fact that most users use similar words to express their opinions, by separately parameterising the model for different users or user groups.

We propose to account for individual users while restricting all users to share the same vocabulary. This is formulated as a bilinear predictive model,

$$f(X) = \boldsymbol{u}^{\mathrm{T}} X \boldsymbol{w} + \beta \,, \tag{1}$$

where X is an $m \times p$ matrix of user-word frequencies and $\boldsymbol{u}$ and $\boldsymbol{w}$ are the model parameters. Let $\mathcal{Q} \in \mathbb{R}^{n \times m \times p}$ be a tensor which captures our training inputs, where n, m and p denote the considered number of samples (each sample usually refers to a day), terms and users respectively; $\mathcal{Q}$ can simply be interpreted as n versions of X (denoted by $\mathcal{Q}_i$ in the remainder of the script), a different one for each day, put together. Each element $\mathcal{Q}_{ijk}$ holds the frequency of term j for user k during the day i in our sample. If a user k has posted $c_{i \cdot k}$ tweets during day i, and $c_{ijk} \leq c_{i \cdot k}$ of them contain a term j, then the frequency of j for this day and user is defined as $\mathcal{Q}_{ijk} = \frac{c_{ijk}}{c_{i \cdot k}}$.

Aiming to learn sparse sets of users and terms that are representative of the voting intention signal, we formulate our optimisation task as follows:

$$\{\boldsymbol{w}^*, \boldsymbol{u}^*, \beta^*\} = \underset{\boldsymbol{w}, \boldsymbol{u}, \beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(\boldsymbol{u}^{\mathrm{T}} \mathcal{Q}_i \boldsymbol{w} + \beta - y_i\right)^2 + \psi(\boldsymbol{w}, \rho_1) + \psi(\boldsymbol{u}, \rho_2) \,, \tag{2}$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is the response variable (voting intention), $\boldsymbol{w} \in \mathbb{R}^m$ and $\boldsymbol{u} \in \mathbb{R}^p$ denote the term and user weights respectively, $\boldsymbol{u}^{\mathrm{T}} \mathcal{Q}_i \boldsymbol{w}$ expresses the bilinear term, $\beta \in \mathbb{R}$ is a bias term and $\psi(\cdot)$ is a regularisation function with parameters ρ_1 or ρ_2. The first term in Eq. 2 is the standard regularisation loss function, namely the sum squared error over the training instances.[4]

In the main formulation of our bilinear model, as the regularisation function $\psi(\cdot)$ we use the **elastic net** (Zou and Hastie, 2005), an extension of the well-studied ℓ_1-norm regulariser, known as the LASSO (Tibshirani, 1996). The ℓ_1-norm regularisation has found many applications in several scientific fields as it encourages sparse solutions which reduce the possibility of overfitting and enhance the interpretability of the inferred model (Hastie et al., 2009). The elastic net applies an extra penalty on the ℓ_2-norm of the weight vector, and can resolve instability issues of LASSO which arise when correlated predictors exist in the input data (Zhao and Yu, 2006). Its regularisation function $\psi_{\textbf{el}}(\cdot)$ is defined by:

$$\psi_{\textbf{el}}(\boldsymbol{w}, \lambda, \alpha) = \lambda \left(\frac{1-\alpha}{2} \|\boldsymbol{w}\|_2^2 + \alpha \|\boldsymbol{w}\|_1 \right), \tag{3}$$

where $\lambda > 0$ and $\alpha \in [0, 1)$; setting parameter α to its extremes transforms elastic net to ridge regression ($\alpha = 0$) or vanilla LASSO ($\alpha = 1$).

Eq. 2 can be treated as a biconvex learning task (Al-Khayyal and Falk, 1983), by observing that for a fixed $\boldsymbol{w}$, learning $\boldsymbol{u}$ is a convex problem and vice versa. Biconvex functions and possible applications have been well studied in the optimisation literature (Quesada and Grossmann, 1995;

[3] This has been carried out to ensure an adequate number of training points in the experimental process.

[4] Note that other loss functions could be used here, such as logistic loss for classification, or more generally bilinear variations of Generalised Linear Models (Nelder and Wedderburn, 1972).

Pirsiavash et al., 2009). Their main advantage is the ability to solve efficiently non-convex problems by a repeated application of two convex processes, *i.e.*, a form of coordinate ascent. In our case, the bilinear technique makes it possible to explore both word and user spaces, while maintaining a modest training complexity.

Therefore, in our bilinear approach we divide learning in two phases, where we learn word and user weights respectively. For the first phase we produce the term-scores matrix $\mathcal{V} \in \mathbb{R}^{n \times m}$ with elements given by:

$$\mathcal{V}_{ij} = \sum_{z=1}^{p} u_z \mathcal{Q}_{ijz}. \tag{4}$$

$\mathcal{V}$ contains weighted sums of term frequencies over all users for the considered set of days. The weights are held in $\boldsymbol{u}$ and are representative of each user. The initial optimisation task is formulated as:

$$\begin{aligned}\{\boldsymbol{w}^*, \beta^*\} = \underset{\boldsymbol{w},\beta}{\operatorname{argmin}} \|\mathcal{V}\boldsymbol{w} + \beta - \boldsymbol{y}\|_2^2 \\ + \psi_{\text{el}}(\boldsymbol{w}, \lambda_1, \alpha_1) \ ,\end{aligned} \tag{5}$$

where we aim to learn a sparse but consistent set of weights w^* for the terms of our vocabulary.

In the second phase, we are using $\boldsymbol{w}^*$ to form the user-scores matrix $\mathcal{D} \in \mathbb{R}^{n \times p}$:

$$\mathcal{D}_{ik} = \sum_{z=1}^{m} w_z^* \mathcal{Q}_{izk} \, , \tag{6}$$

which now contains weighted sums over all terms for the same set of days. The optimisation task becomes:

$$\begin{aligned}\{\boldsymbol{u}^*, \beta^*\} = \underset{\boldsymbol{u},\beta}{\operatorname{argmin}} \|\mathcal{D}\boldsymbol{u} + \beta - \boldsymbol{y}\|_2^2 \\ + \psi_{\text{el}}(\boldsymbol{u}, \lambda_2, \alpha_2) \, .\end{aligned} \tag{7}$$

This process continues iteratively by inserting the weights of the second phase back to phase one, and so on until convergence. We cannot claim that a global optimum will be reached, but biconvexity guarantees that our global objective (Eq. 2) will decrease in each step of this iterative process. In the remainder of this paper, we refer to the method described above as Bilinear Elastic Net (**BEN**).

3.2 Exploiting term-target or user-target relationships

The previous model assumes that the response variable $\boldsymbol{y}$ holds information about a single inference target. However, the task that we are addressing in this paper usually implies the existence of several targets, *i.e.*, different political parties or politicians. An important property, therefore, is the ability to perform multiple output regression. A simple way of adapting the model to the multiple output scenario is by framing a separate learning problem for each output, but tying together some of the parameters. Here we consider tying together the user weights $\boldsymbol{u}$, to enforce that the same set of users are relevant to all tasks, while learning different term weights. Note that the converse situation, where $\boldsymbol{w}$'s are tied and $\boldsymbol{u}$'s are independent, can be formulated in an equivalent manner.

Suppose that our target variable $\boldsymbol{y} \in \mathbb{R}^{\tau n}$ refers now to τ political entities, $\boldsymbol{y} = \left[\boldsymbol{y}_1^{\mathrm{T}} \boldsymbol{y}_2^{\mathrm{T}} ... \boldsymbol{y}_\tau^{\mathrm{T}}\right]^{\mathrm{T}}$; in this formation the top n elements of $\boldsymbol{y}$ match to the first political entity, the next n elements to the second and so on. In the first phase of the bilinear model, we would have to solve the following optimisation task:

$$\begin{aligned}\{\boldsymbol{w}^*, \beta^*\} = \underset{w,\beta}{\operatorname{argmin}} \sum_{i=1}^{\tau} \|\mathcal{V}\boldsymbol{w_i} + \beta_i - y_i\|_2^2 \\ + \sum_{i=1}^{\tau} \psi_{\text{el}}(\boldsymbol{w}_i, \lambda_1, \alpha_1) \ ,\end{aligned} \tag{8}$$

where $\mathcal{V}$ is given by Eq. 4 and $\boldsymbol{w}^* \in \mathbb{R}^{\tau m}$ denotes the vector of weights which can be sliced into τ sub-vectors $\{\boldsymbol{w}_1^*, ..., \boldsymbol{w}_\tau^*\}$ each one representing a political entity. In the second phase, sub-vectors $\boldsymbol{w}_i^*$ are used to form the input matrices $\mathcal{D}_i$, $i \in \{1, ..., \tau\}$ with elements given by Eq. 6. The input matrix $\mathcal{D}'$ is formed by the vertical concatenation of all $\mathcal{D}_i$ user score matrices, *i.e.*, $\mathcal{D}' = \left[\mathcal{D}_1^{\mathrm{T}} \ ... \ \mathcal{D}_\tau^{\mathrm{T}}\right]^{\mathrm{T}}$, and the optimisation target is equivalent to the one expressed in Eq. 7. Since $\mathcal{D}' \in \mathbb{R}^{\tau n \times p}$, the user weight vector $\boldsymbol{u}^* \in \mathbb{R}^p$ and thus, we are learning a single weight per user and not one per political party as in the previous step.

The method described above allows learning different term weights per response variable and then binds them under a shared set of user weights. As mentioned before, one could also try the opposite (*i.e.*, start by expanding the user space); both those models can also be optimised in an iterative process. However, our experiments revealed that those approaches did not improve on the performance of BEN. Still, this behaviour could be problem-specific, *i.e.*, learning different words from a shared set of users (and the opposite) may not be a good modelling practice for the domain of politics. Nevertheless, this observation served as a motivation for the method described in the next section, where we extract a consistent set of words and users that are weighted differently among the considered political entities.

3.3 Multi-task learning with the ℓ_1/ℓ_2 regulariser

All previous models – even when combining all inference targets – were not able to explore relationships across the different task domains; in our case, a task domain is defined by a specific political label or party. Ideally, we would like to make a sparse selection of words and users but with a regulariser that promotes inter-task sharing of structure, so that many features may have a positive influence towards one or more parties, but negative towards the remaining one(s). It is possible to achieve this multi-task learning property by introducing a different set of regularisation constraints in the optimisation function.

We perform multi-task learning using an extension of group LASSO (Yuan and Lin, 2006), a method known as $\boldsymbol{\ell_1/\ell_2}$ regularisation (Argyriou et al., 2008; Liu et al., 2009). Group LASSO exploits a predefined group structure on the feature space and tries to achieve sparsity in the group-level, *i.e.*, it does not perform feature selection (unlike the elastic net), but group selection. The ℓ_1/ℓ_2 regulariser extends this notion for a τ-dimensional response variable. The global optimisation target is now formulated as:

$$\{W^*, U^*, \boldsymbol{\beta}^*\} = \underset{W,U,\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{t=1}^{\tau}\sum_{i=1}^{n}\left(\boldsymbol{u}_t^{\mathrm{T}} \mathcal{Q}_i \boldsymbol{w}_t + \beta_t - y_{ti}\right)^2 + \lambda_1 \sum_{j=1}^{m} \|W_j\|_2 + \lambda_2 \sum_{k=1}^{p} \|U_k\|_2, \quad (9)$$

where the input matrix $\mathcal{Q}_i$ is defined in the same way as earlier, $W = [\boldsymbol{w}_1 \ldots \boldsymbol{w}_\tau]$ is the term weight matrix (each $\boldsymbol{w}_t$ refers to the t-th political entity or task), equivalently $U = [\boldsymbol{u}_1 \ldots \boldsymbol{u}_\tau]$, W_j and U_j denote the j-th rows of weight matrices W and U respectively, and vector $\boldsymbol{\beta} \in \mathbb{R}^\tau$ holds the bias terms per task. In this optimisation process, we aim to enforce sparsity in the feature space but in a structured manner. Notice that we are now regularising the $\ell_{2,1}$ mixed norm of W and U, which is defined as the sum of the row ℓ_2-norms for those matrices. As a result, we expect to encourage the activation of a sparse set of features (corresponding to the rows of W and U), but with nonzero weights across the τ tasks (Argyriou et al., 2008). Consequently, we are performing filtering (many users and words will have zero weights) and, at the same time, assign weights of different magnitude and sign on the selected features, something that suits a political opinion mining application, where pro-A often means anti-B.

Eq. 9 can be broken into two convex tasks (following the same notion as in Eqs. 5 and 7), where we individually learn $\{W, \boldsymbol{\beta}\}$ and then $\{U, \boldsymbol{\beta}\}$; each step of the process is a standard linear regression problem with an ℓ_1/ℓ_2 regulariser. Again, we are able iterate this bilinear process and in each step convexity is guaranteed. We refer to this method as Bilinear Group ℓ_1/ℓ_2 (**BGL**).

4 Experiments

The proposed models are evaluated on $\mathrm{C_{uk}}$ and $\mathrm{C_{au}}$ which have been introduced in Section 2. We measure predictive performance, compare it to the performance of several competitive baselines, and provide a qualitative analysis of the parameters learned by the models.

4.1 Data preprocessing

Basic preprocessing has been applied on the vocabulary index of $\mathrm{C_{uk}}$ and $\mathrm{C_{au}}$ aiming to filter out some of the word features and partially reduce the dimensionality of the problem. Stop words and web links were removed in both sets, together with character sequences of length <4 and <3 for $\mathrm{C_{uk}}$ and $\mathrm{C_{au}}$ respectively.[5] As the vocabulary size of $\mathrm{C_{uk}}$ was significantly larger, for this data set we have additionally merged Twitter hashtags (*i.e.*, words starting with '#') with their exact non topic word match, where possible (by dropping the '#' when the word existed in the index). After performing the preprocessing routines described above, the vocabulary sizes for $\mathrm{C_{uk}}$ and $\mathrm{C_{au}}$ were set to 80,976 and 22,917 respectively.

4.2 Predictive accuracy

To evaluate the predictive accuracy of our methods, we have chosen to emulate a real-life scenario

[5] Most of the times those character sequences were not valid words. This pattern was different in each language and thus, a different filtering threshold was applied in each data set.

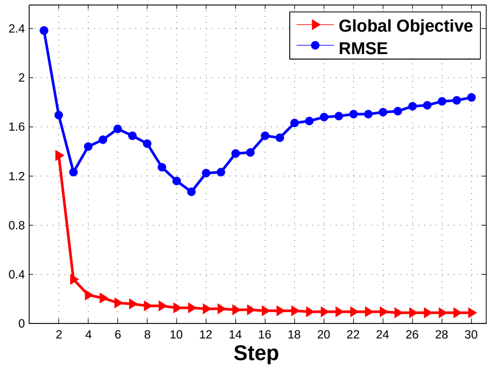


Figure 2: Global objective function and RMSE on a validation set for **BEN** in 15 iterations (30 steps) of the model.

of voting intention prediction. The evaluation process starts by using a fixed set of polls matching to consecutive time points in the past for training and validating the parameters of each model. Testing is performed on the following δ (unseen) polls of the data set. In the next step of the evaluation process, the training/validation set is increased by merging it with the previously used test set (δ polls), and testing is now performed on the next δ unseen polls. In our experiments, the number of steps in this evaluation process is set to 10 and in each step the size of the test set is set to $\delta = 5$ polls. Hence, each model is tested on 50 unseen and consecutive in time samples. The loss function in our evaluation is the standard Mean Square Error (**MSE**), but to allow a better interpretation of the results, we display its root (**RMSE**) in tables and figures.[6]

The parameters of each model (α_i for BEN and λ_i for BEN and BGL, $i \in \{1, 2\}$) are optimised using a held-out validation set by performing grid search. Note that it may be tempting to adapt the regularisation parameters in each phase of the iterative training loop, however this would change the global objective (see Eqs. 2 and 9) and thus convergence will not be guaranteed. A key question is how many iterations of training are required to reach convergence. Figure 2 illustrates how the BEN global objective function (Eq. 2) converges during this iterative process and the model's performance on an unseen validation set. Notice that there is a large performance improvement after the first step (which alone is a linear solver), but overfitting occurs after step 11. Based on this result, for subsequent experiments we run the training process for two iterations (4 steps), and take the best performing model on the held-out validation set.

	CON	LAB	LBD	μ
$\mathbf{B}_{\mu}$	2.272	1.663	1.136	1.69
$\mathbf{B}_{\text{last}}$	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**

Table 1: UK case study — Average RMSEs representing the error of the inferred voting intention percentage for the 10-step validation process; $\boldsymbol{\mu}$ denotes the mean RMSE across the three political parties for each baseline or inference method.

	SPÖ	ÖVP	FPÖ	GRÜ	μ
$\mathbf{B}_{\mu}$	1.535	1.373	3.3	1.197	1.851
$\mathbf{B}_{\text{last}}$	**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**

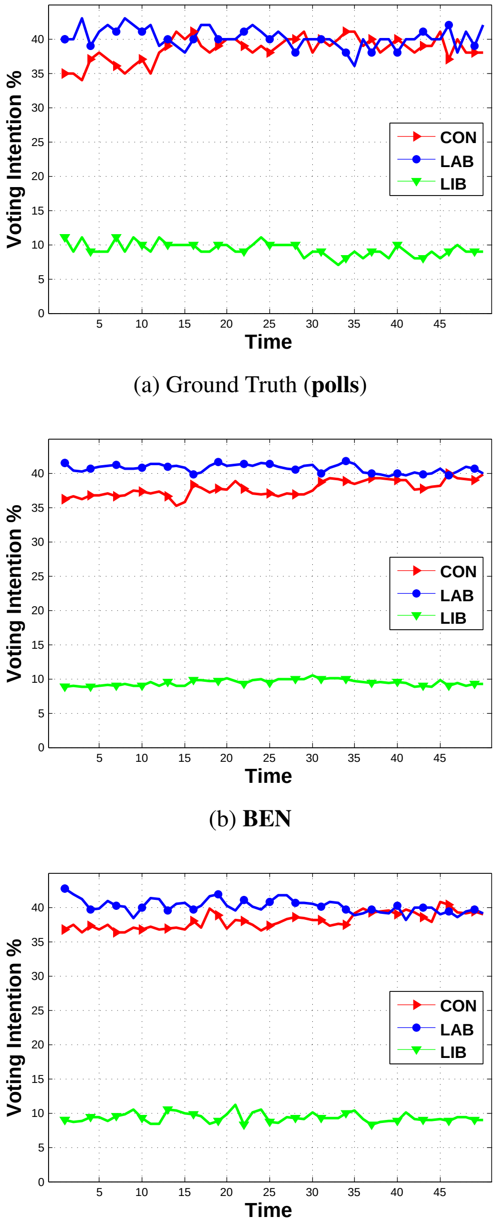
Table 2: Austrian case study — Average RMSEs for the 10-step validation process.

We compare the performance of our methods with three baselines. The first makes a constant prediction of the mean value of the response variable $\boldsymbol{y}$ in the training set ($\mathbf{B}_{\mu}$); the second predicts the last value of $\boldsymbol{y}$ ($\mathbf{B}_{\text{last}}$); and the third baseline (**LEN**) is a linear regression over the terms using elastic net regularisation. Recalling that each test set is made of 5 polls, B_{last} should be considered as a hard baseline to beat[7] given that voting intentions tend to have a smooth behaviour. Moreover, improving on LEN partly justifies the usefulness of a bilinear approach compared to a linear one.

Performance results comparing inferred voting intention percentages and polls for C_{uk} and C_{au} are presented in Tables 1 and 2 respectively. For the UK case study, both BEN and BGL are able to beat all baselines in average performance across all parties. However in the Austrian case study, LEN performs better that BEN, something that could be justified by the fact that the users in C_{au} were selected by domain experts, and consequently there was not much gain to be had by filtering them further. Nevertheless, the difference in performance was rather small (approx. 0.26% error) and the in-

[6] RMSE has the same metric units as the response variable.

[7] The last response value could be easily included as a feature in the model, and would likely improve predictive performance.

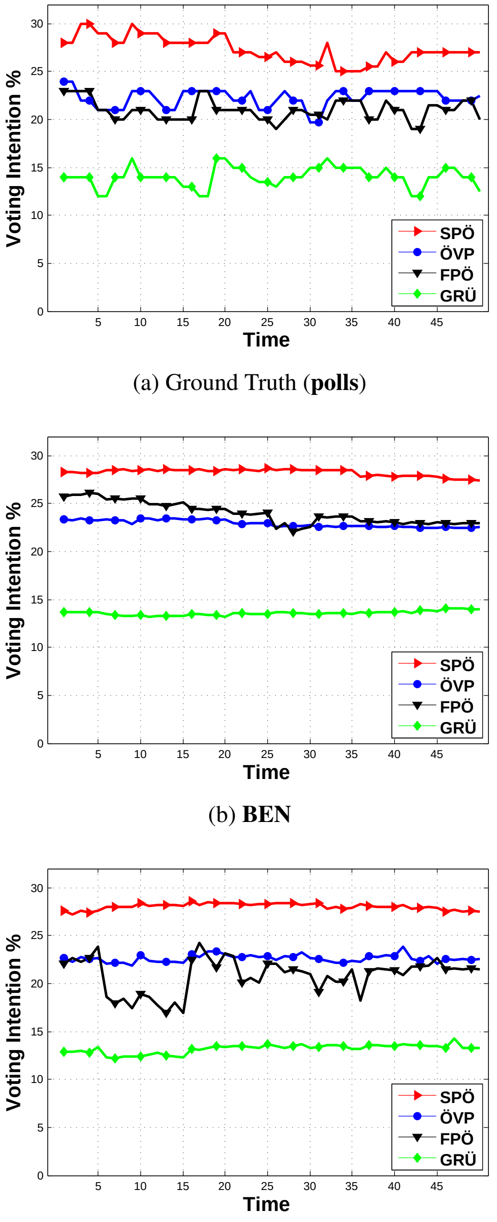


(a) Ground Truth (**polls**)

(b) **BEN**

(c) **BGL**

Figure 3: UK case study — Voting intention inference results (50 polls, 3 parties). Sub-figure 3a is a plot of ground truth as presented in voting intention polls (Fig. 1a).



(a) Ground Truth (**polls**)

(b) **BEN**

(c) **BGL**

Figure 4: Austrian case study — Voting intention inference results (50 polls, 4 parties). Sub-figure 4a is a plot of ground truth as presented in voting intention polls (Fig. 1b).

ferences of LEN and BEN followed a very similar pattern ($\bar{\rho} = .94$ with $p < 10^{-10}$).[8] Multi-task learning (BGL) delivered the best inference performance in both case studies, which was on average smaller than 1.48% (RMSE).

Inferences for both BEN and BGL have been plotted on Figures 3 and 4. They are presented as continuous lines of 50 inferred points (per party) which are created by concatenating the inferences on all test sets.[9] For the UK case study, one may observe that BEN (Fig. 3b) cannot register any change – with the exception of one test point – in the leading party fight (CON versus LAB); BGL (Fig. 3c) performs much better in that aspect. In the Austrian case study this characteristic becomes more obvious. BEN (Fig. 4b) consistently predicts the wrong ranking of ÖVP and FPÖ, whereas BGL (Fig. 4c) does much better. Most importantly, a

[8] Pearson's linear correlation averaged across the four Austrian parties.

[9] Voting intention polls were plotted separately to allow a better presentation.

Party	Tweet	Score	Author
CON	PM in friendly chat with top EU mate, Sweden's Fredrik Reinfeldt, before family photo	1.334	Journalist
	Have Liberal Democrats broken electoral rules? Blog on Labour complaint to cabinet secretary	−0.991	Journalist
LAB	Blog Post Liverpool: City of Radicals Website now Live <*link*> #liverpool #art	1.954	Art Fanzine
	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
	Blog Post Liverpool: City of Radicals 2011 – More Details Announced #liverpool #art	−0.521	Art Fanzine
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
	Hans Rauscher zu Felix #Baumgartner "A klaner Hitler" <*link*> **Translation:** *Hans Rauscher on Felix #Baumgartner "A little Hitler" <link>*	−1.711	Journalist
ÖVP	#IchPirat setze mich dafür ein, dass eine große Koalition _mathematisch_ verhindert wird! 1.Geige: #Gruene + #FPOe + #OeVP **Translation:** *#IPirate am committed to prevent a grand coalition mathematically! Calling the tune: #Greens + #FPO + #OVP*	4.953	User
	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
FPÖ	Neue Kampagne der #Krone zur #Wehrpflicht: "GIB BELLO EINE STIMME!" **Translation:** *New campaign by the #Krone on #Conscription: "GIVE WOOFY A VOICE!"*	7.44	Political satire
	Kampagne der Wiener SPÖ "zum Zusammenleben" spielt Rechtspopulisten in die Hände <*link*> **Translation:** *Campaign of the Viennese SPÖ on "Living together" plays right into the hands of right-wing populists <link>*	−3.44	Human Rights
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
	Pilz "ich will in dieser Republik weder kriminelle Asylwerber, noch kriminelle orange Politiker" - BZÖ-Abschiebung ok, aber wohin? #amPunkt **Translation:** *Pilz "i want neither criminal asylum-seekers, nor criminal orange politicians in this republic" - BZÖ-Deportation OK, but where? #amPunkt*	−2.172	User

Table 3: Examples of tweets amongst the ones with top positive and negative scores per party for both $\mathbf{C}_{\text{uk}}$ and $\mathbf{C}_{\text{au}}$ data sets (tweets in Austrian have been translated in English as well). Notice that weight magnitude may differ per case study and party as they are based on the range of the response variable and the total number of selected features.

general observation is that BEN's predictions are smooth and do not vary significantly with time. This might be a result of overfitting the model to a single response variable which usually has a smooth behaviour. On the contrary, the multi-task learning property of BGL reduces this type of overfitting providing more statistical evidence for the terms and users and thus, yielding not only a better inference performance, but also a more accurate model.

4.3 Qualitative Analysis

In this section, we refer to features that have been selected and weighted as significant by our bilinear learning functions. Based on the weights for the word and the user spaces that we retrieve after the application of BGL in the last step of the evaluation process (see the previous section), we compute a score (weighted sum) for each tweet in our training data sets for both C_{uk} and C_{au}. Table 3 shows examples of interesting tweets amongst the top weighted ones (positively as well as negatively) per party. Together with their text (anonymised for privacy reasons) and scores, we also provide an attribute for the author (if present). In the displayed tweets for the UK study, the only possible outlier is the '*Art Fanzine*'; still, it seems to register a consistent behaviour (positive towards

LAB, negative towards LBD) and, of course, hidden, indirect relationships may exist between political opinion and art. The Austrian case study revealed even more interesting tweets since training was conducted on data from a very active pre-election period (we made an effort to translate those tweets in English language as well). For a better interpretation of the presented tweets, it may be useful to know that '*Johannes Voggenhuber*' (who receives a positive comment for his book) and '*Peter Pilz*' (whose comment is questioned) are members of GRÜ, '*Krone*' (or Kronen Zeitung) is the major newspaper in Austria[10] and that FPÖ is labelled as a far right party, something that may cause various reactions from '*Human Rights*' organisations.

5 Related Work

The topic of political opinion mining from Social Media has been the focus of various recent research works. Several papers have presented methods that aim to predict the result of an election (Tumasjan et al., 2010; Bermingham and Smeaton, 2011) or to model voting intention and other kinds of socio-political polls (O'Connor et al., 2010; Lampos, 2012). Their common feature is a methodology based on a meta-analysis of word frequencies using off-the-shelf sentiment tools such as LIWC (Pennebaker et al., 2007) or Senti-WordNet (Esuli and Sebastiani, 2006). Moreover, the proposed techniques tend to incorporate posting volume figures as well as hand-crafted lists of words relevant to the task (*e.g.*, names of politicians or parties) in order to filter the content successfully.

Such papers have been criticised as their methods do not generalise when applied on different data sets. According to the work in (Gayo-Avello et al., 2011), the methods presented in (Tumasjan et al., 2010) and (O'Connor et al., 2010) failed to predict the result of US congressional elections in 2009. We disagree with the arguments supporting the statement "you cannot predict elections with Twitter" (Gayo-Avello, 2012), as many times in the past actual voting intention polls have also failed to predict election outcomes, but we agree that most methods that have been proposed so far were not entirely generic. It is a fact that the majority of sentiment analysis tools are English-specific (or even American English) and, most importantly, political word lists (or ontologies) change in time, per country and per party; hence, generalisable methods should make an effort to limit reliance from such tools.

Furthermore, our work – indirectly – meets the guidelines proposed in (Metaxas et al., 2011) as we have developed a framework of "well-defined" algorithms that are "Social Web aware" (since the bilinear approach aims to improve noise filtering) and that have been tested on two evaluation scenarios with distinct characteristics.

6 Conclusions and Future Work

We have presented a novel method for text regression that exploits both word and user spaces by solving a bilinear optimisation task, and an extension that applies multi-task learning for multi-output inference. Our approach performs feature selection – hence, noise filtering – on large-scale user-generated inputs automatically, generalises across two languages without manual adaptations and delivers some significant improvements over strong performance baselines ($< 1.5\%$ error when predicting polls). The application domain in this paper was politics, though the presented methods are generic and could be easily applied on various other domains, such as health or finance.

Future work may investigate further modelling improvements achieved by applying different regularisation functions as well as the adaptation of the presented models to classification problems. Finally, in the application level, we aim at an in-depth analysis of patterns and characteristics in the extracted sets of features by collaborating with domain experts (*e.g.*, political analysts).

Acknowledgments

This work was funded by the TrendMiner project (EU-FP7-ICT n.287863). All authors would like to thank the political analysts (and especially Paul Ringler) from SORA[11] for their useful insights on politics in Austria.

[10] "Accused of abusing its near monopoly to manipulate public opinion in Austria", Wikipedia, 19/02/2013, `http://en.wikipedia.org/wiki/Kronen_Zeitung`.

[11] SORA – Institute for Social Research and Consulting, `http://www.sora.at`.

References

Faiz A Al-Khayyal and James E Falk. 1983. Jointly Constrained Biconvex Programming. *Mathematics of Operations Research*, 8(2):273–286.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. 2008. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, January.

Adam Bermingham and Alan F Smeaton. 2011. On using Twitter to monitor political sentiment and predict election results. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, pages 2–10, November.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, March.

Andrea Esuli and Fabrizio Sebastiani. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceeding of the 5th Conference on Language Resources and Evaluation (LREC)*, pages 417–422.

Daniel Gayo-Avello, Panagiotis T Metaxas, and Eni Mustafaraj. 2011. Limits of Electoral Predictions using Twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 490–493.

Daniel Gayo-Avello. 2012. No, You Cannot Predict Elections with Twitter. *IEEE Internet Computing*, 16(6):91–94, November.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer.

Bernard J Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. 2009. Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 60(11):2169–2188.

Vasileios Lampos and Nello Cristianini. 2010. Tracking the flu pandemic by monitoring the Social Web. In *2nd IAPR Workshop on Cognitive Information Processing*, pages 411–416. IEEE Press.

Vasileios Lampos and Nello Cristianini. 2012. Nowcasting Events from the Social Web with Statistical Learning. *ACM Transactions on Intelligent Systems and Technology*, 3(4):1–22, September.

Vasileios Lampos, Tijl De Bie, and Nello Cristianini. 2010. Flu Detector - Tracking Epidemics on Twitter. In *Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, pages 599–602. Springer.

Vasileios Lampos. 2012. On voting intentions inference from Twitter content: a case study on UK 2010 General Election. *CoRR*, April.

Thomas Lansdall-Welfare, Vasileios Lampos, and Nello Cristianini. 2012. Effects of the recession on public mood in the UK. In *Proceedings of the 21st international conference companion on World Wide Web*, WWW '12 Companion, pages 1221–1226. ACM.

Jun Liu, Shuiwang Ji, and Jieping Ye. 2009. Multi-task feature learning via efficient l2,1-norm minimization. pages 339–348, June.

Panagiotis T Metaxas, Eni Mustafaraj, and Daniel Gayo-Avello. 2011. How (Not) To Predict Elections. In *IEEE 3rd International Conference on Social Computing (SocialCom)*, pages 165 – 171. IEEE Press.

John A Nelder and Robert W M Wedderburn. 1972. Generalized Linear Models. *Journal of the Royal Statistical Society - Series A (General)*, 135(3):370.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, pages 122–129. AAAI Press.

Michael J Paul and Mark Dredze. 2011. You Are What You Tweet: Analyzing Twitter for Public Health. *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, pages 265–272.

James W Pennebaker, Cindy K Chung, Molly Ireland, Amy Gonzales, and Roger J Booth. 2007. The Development and Psychometric Properties of LIWC2007. Technical report, Universities of Texas at Austin & University of Auckland, New Zealand.

Hamed Pirsiavash, Deva Ramanan, and Charless Fowlkes. 2009. Bilinear classifiers for visual recognition. In *Advances in Neural Information Processing Systems*, volume 22, pages 1482–1490.

Daniel Preoţiuc-Pietro, Sina Samangooei, Trevor Cohn, Nicholas Gibbins, and Mahesan Niranjan. 2012. Trendminer: An Architecture for Real Time Analysis of Social Media Text. In *Sixth International AAAI Conference on Weblogs and Social Media*, pages 38–42. AAAI Press, July.

Ignacio Quesada and Ignacio E Grossmann. 1995. A global optimization algorithm for linear fractional and bilinear programs. *Journal of Global Optimization*, 6(1):39–76, January.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World Wide Web (WWW)*, pages 851–860. ACM.

Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society - Series B (Methodological)*, 58(1):267–288.

Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. 2010. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pages 178–185. AAAI.

Ming Yuan and Yi Lin. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society - Series B: Statistical Methodology*, 68(1):49–67.

Peng Zhao and Bin Yu. 2006. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7(11):2541–2563.

Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, April.