Extracting Socioeconomic Patterns from the News: Modelling Text and Outlet Importance Jointly

Vasileios Lampos[1], **Daniel Preoţiuc-Pietro**[2], **Sina Samangooei**[3], **Douwe Gelling**[2], and **Trevor Cohn**[4]

[1] Department of Computer Science, University College London — v.lampos@ucl.ac.uk

[2] Department of Computer Science, The University of Sheffield — {d.preotiuc,d.gelling}@shef.ac.uk

[3] Electronics and Computer Science, University of Southampton — ss@ecs.soton.ac.uk

[4] Computing and Information Systems, The University of Melbourne — t.cohn@unimelb.edu.au

Abstract

Information from news articles can be used to study correlations between textual discourse and socioeconomic patterns. This work focuses on the task of understanding how words contained in the news as well as the news outlets themselves may relate to a set of indicators, such as economic sentiment or unemployment rates. The bilinear nature of the applied regression model facilitates learning jointly word and outlet importance, supervised by these indicators. By evaluating the predictive ability of the extracted features, we can also assess their relevance to the target socioeconomic phenomena. Therefore, our approach can be formulated as a potential NLP tool, particularly suitable to the computational social science community, as it can be used to interpret connections between vast amounts of textual content and measurable society-driven factors.

1 Introduction

Vast amounts of user-generated content on the Internet as well as digitised textual resources allow us to study text in connection to real world events across large intervals of time. Over the last decade, there has been a shift in user news consumption starting with a move from offline to online sources (Lin et al., 2005); in more recent years user-generated news have also become prominent. However, traditional news outlets continue to be a central reference point (Nah and Chung, 2012) as they still have the advantage of being professionally authored, alleviating the noisy nature of citizen journalism formats.

Here, we present a framework for analysing socioeconomic patterns in news articles. In contrast to prior approaches, which primarily focus on the textual contents, our analysis shows how Machine Learning methods can be used to gain insights into the interplay between text in news articles, the news outlets and socioeconomic indicators. Our experiments are performed on a set of EU-related news summaries spanning over 8 years, with the intention to study two basic economic factors: EU's unemployment rate and Economic Sentiment Index (ESI) (European Commision, 1997). To determine connections between the news, the outlets and the indicators of interest, we formulate our learning task as bilinear text-based regression (Lampos et al., 2013).

Approaches to learning the correlation of news, or text in general, with real world indicators have been performed in both unsupervised and supervised settings. For example, Flaounas et al. (2010) uncover interesting patterns in EU's Mediasphere, whereas Schumaker and Chen (2009) demonstrate that news articles can predict financial indicators. Conversely, Bentley et al. (2014) show that emotions in the textual content of books reflect back on inflation and unemployment rates during the 20th century. Recently, Social Media text has been intensively studied as a quicker, unobtrusive and cheaper alternative to traditional surveys. Application areas include politics (O'Connor et al., 2010), finance (Bollen and Mao, 2011), health (Lampos and Cristianini, 2012; Paul and Dredze, 2011) or psychology (De Choudhury et al., 2013; Schwartz et al., 2013).

In this paper, we apply a modified version of a bilinear regularised regression model (BEN) proposed for the task of voting intention inference from Twitter content (Lampos et al., 2013). The main characteristic of BEN is the ability of modelling word frequencies as well as individual user importance in a joint optimisation task. By applying it in the context of supervised news analysis, we are able to visualise relevant discourse to a particular socioeconomic factor, identifying relevant words together with important outlets.

Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, pages 13–17,
Baltimore, Maryland, USA, June 26, 2014.

2 Data

We compiled a data set by crawling summaries on news articles written in English language, published by the Open Europe Think Tank.[1] The press summaries are daily aggregations of news items about the EU or member countries with a focus on politics; the news outlets used to compile each summary are listed below the summary's text. The site is updated every weekday, with the major news being covered in a couple of paragraphs, and other less prevalent issues being mentioned in one paragraph to as little as one sentence. The news summaries were first published on February 2006; we collected all of them up to mid-November 2013, creating a data set with the temporal resolution of 1913 days (or 94 months).

The text was tokenised using the NLTK library (Bird et al., 2009). News outlets with fewer than 5 mentions were removed, resulting in a total of 435 sources. Each summary contains on average 14 news items, with an average of 3 news sources per item; where multiple sources were present, the summary was assigned to all the referenced news outlets. After removing stop words, we ended up with $8,413$ unigrams and $19,045$ bigrams; their daily occurrences were normalised using the total number of news items for that day.

For the purposes of our supervised analysis, we use the response variables of ESI and unemployment rate across the EU. The monthly time series of these socioeconomic indicators were retrieved from Eurostat, EU's statistical office (see the red lines in Fig. 1a and 1b respectively). ESI is a composite indicator often seen as an early predictor for future economic developments (Gelper and Croux, 2010). It consists of five confidence indicators with different weights: industrial (40%), services (30%), consumer (20%), construction (5%) and retail trade (5%). The unemployment rate is a seasonally adjusted ratio of the non employed persons over the entire EU labour force.[2]

3 Models

A common approach to regression arises through the application of generalised linear models. These models use a feature vector input $\boldsymbol{x}$ and aim to build a linear function of $\boldsymbol{x}$ for predicting a response variable y:

$$f(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{w} + \beta \quad \text{where } \boldsymbol{x}, \boldsymbol{w} \in \mathbb{R}^m. \quad (1)$$

The objective is to find an f, which minimises a model-dependent loss function (e.g. sum squared error), optionally subject to a regularisation penalty ψ; ℓ_2-norm regularisation (ridge regression) penalises high weights (Hoerl and Kennard, 1970), while ℓ_1-norm regularisation (lasso) encourages sparse solutions (Tibshirani, 1994). Sparsity is desirable for avoiding overfitting, especially when the dimensionality m is larger than the number of training examples n (Hastie et al., 2009). Elastic Net formulates a combination of ℓ_1 and ℓ_2-norm regularisation defined by the objective:

$$\begin{aligned}\{\boldsymbol{w}^*, \beta^*\} = \underset{\boldsymbol{w},\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(\boldsymbol{x}_i^{\mathrm{T}} \cdot \boldsymbol{w} + \beta - y_i\right)^2 \\ + \psi_{\mathrm{EN}}(\boldsymbol{w}, \rho)\,,\end{aligned} \quad (2)$$

where ρ denotes the regularisation parameters (Zou and Hastie, 2005); we refer to this model as **LEN** (Linear Elastic Net) in the remainder of the script.

In the context of voting intention inference from Twitter content, Lampos et al. (2013) extended LEN to a bilinear formulation, where a set of two vector weights are learnt: one for words ($\boldsymbol{w}$) and one for users ($\boldsymbol{u}$). This was motivated by the observation that only a sparse set of users may have predictive value. The model now becomes:

$$f(X) = \boldsymbol{u}^{\mathrm{T}} X \boldsymbol{w} + \beta\,, \quad (3)$$

where X is a matrix of word $\times$ users frequencies. The bilinear optimisation objective is formulated as:

$$\begin{aligned}\{\boldsymbol{w}^*, \boldsymbol{u}^*, \beta^*\} = \underset{\boldsymbol{w},\boldsymbol{u},\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(\boldsymbol{u}^{\mathrm{T}} X_i \boldsymbol{w} + \beta - y_i\right)^2 \\ + \psi_{\mathrm{EN}}(\boldsymbol{w}, \rho_1) + \psi_{\mathrm{EN}}(\boldsymbol{u}, \rho_2)\,,\end{aligned} \quad (4)$$

where X_i is the word $\times$ user frequency matrix, and ρ_1, ρ_2 are the word and user regularisation parameters. This can be treated as a biconvex learning task and be solved by iterating over two convex processes: fixing $\boldsymbol{w}$ and learning $\boldsymbol{u}$, and vice versa (Lampos et al., 2013). Regularised regression on both user and word spaces allows for an automatic selection of the most important words and users, performing at the same time an improved noise filtering.

[1] http://www.openeurope.org.uk/Page/PressSummary/en/

[2] http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Unemployment_statistics

In our experiments, news outlets and socioeconomic indicators replace users and voting intention in the previous model formulation. To ease the interpretation of the outputs, we further impose a positivity constraint on the outlet weights $\boldsymbol{u}$, i.e. $\min(\boldsymbol{u}) \geq 0$; this makes the model more restrictive, but, in our case, did not affect the prediction performance. We refer to this model as **BEN** (Bilinear Elastic Net).

4 Experiments

Both models are applied to the news summaries data set with the aim to predict EU's ESI and rate of unemployment. The predictive capability of the derived models, assessed by their respective inference performance, is used as a metric for judging the degree of relevance between the learnt model parameters – word and outlet weights – and the response variable. A strong predictive performance increases confidence on the soundness of those parameters.

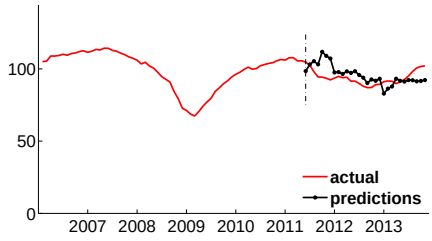
To match input with the monthly temporal resolution of the response variables, we compute the mean monthly term frequencies for each outlet. Evaluation is performed via a 10-fold validation, where each fold's training set is based on a moving window of $p = 64$ contiguous months, and the test set consists of the following $q = 3$ months; formally, the training and test sets for fold i are based on months $\{q(i-1)+1, ..., q(i-1)+p\}$ and $\{q(i-1)+p+1, ..., q(i-1)+p+q\}$ respectively. In this way, we emulate a scenario where we always train on past and predict future points.

Performance results for LEN and BEN are presented in Table 1; we show the average Root Mean Squared Error (RMSE) as well as an error rate (RMSE over $\mu(y)$) across folds to allow for a better interpretation. BEN outperforms LEN in both tasks, with a clearer improvement when predicting ESI. Predictions for all folds are depicted in Fig. 1a and 1b together with the actual values. Note that reformulating the problem into a multi-task learning scenario, where ESI and unemployment are modelled jointly did not improve inference performance.

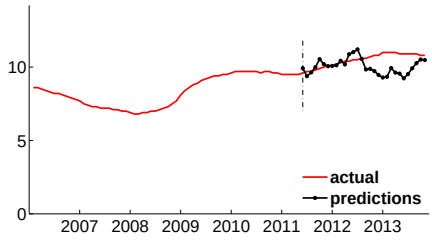
The relatively small average error rates ($< 8.8\%$) make meaningful a further analysis of the model's outputs. Due to space limitations, we choose to focus on the most recent results, depicting the models derived in the 10th fold. Following the example of Schwartz et al. (2013), we use a word cloud visualisation, where the font size is proportional to the derived weights by applying BEN, flipped terms denote negative weights and colours are determined by the frequency of use in the corpus (Fig. 2). Word clouds depict the top-60 positively and negatively weighted n-grams (120 in total) together with the top-30 outlets; bigrams are separated by '_'.

	ESI	Unemployment
LEN	9.253 (9.89%)	0.9275 (8.75%)
BEN	**8.209** (8.77%)	**0.9047** (8.52%)

Table 1: 10-fold validation average RMSEs (and error rates) for LEN and BEN on ESI and unemployment rates prediction.



(a) ESI

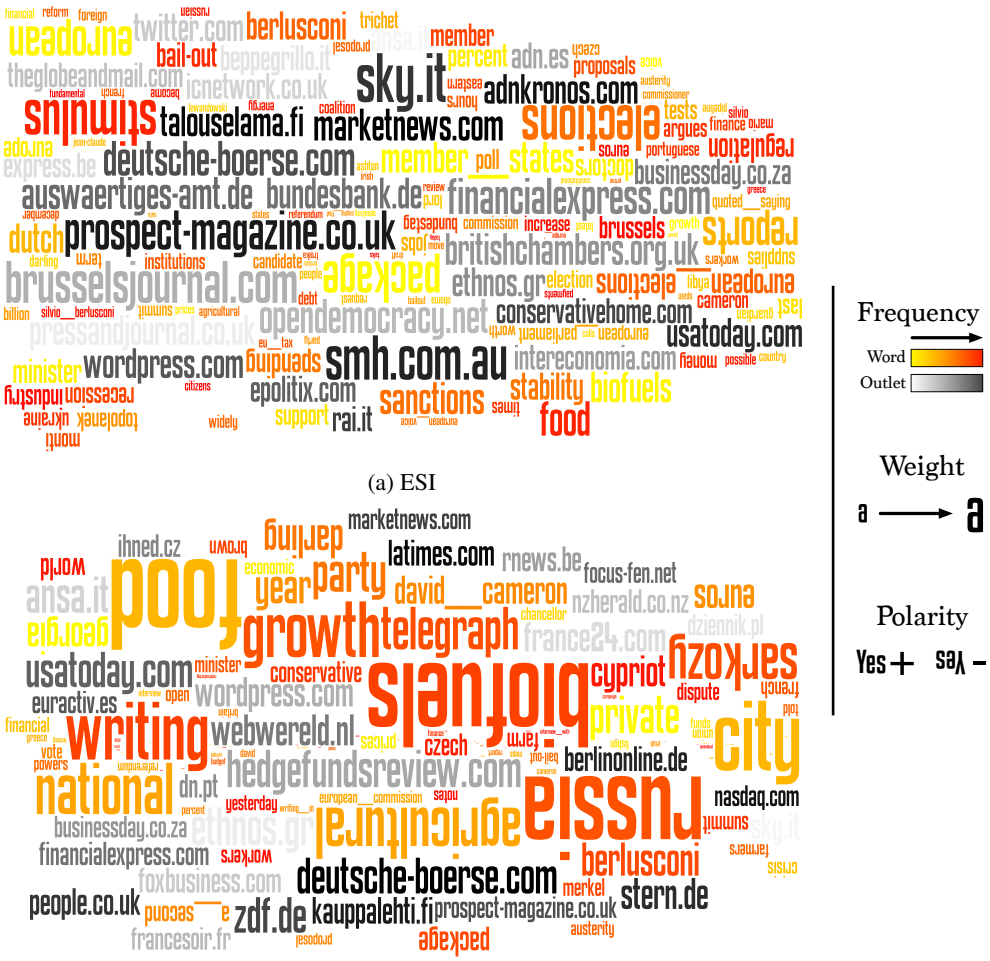


(b) Unemployment

Figure 1: Time series of ESI and unemployment together with BEN predictions (smoothed using a 3-point moving average).

5 Discussion and Future Work

Our visualisations (Fig. 2) present various interesting insights into the news and socioeconomic features being explored, serving as a demonstration of the potential power of the proposed modelling. Firstly, we notice that in the word cloud, the size of a feature (BEN's weight) is not tightly connected with its colour (frequency in the corpus). Also, the word clouds suggest that mostly different terms and outlets are selected for the two indicators. For example, '*sky.it*' is predominant for ESI but not for unemployment, while the opposite is true for '*hedgefundsreview.com*'. Some of the words selected for ESI reflect economical issues, such as '*stimulus*' and '*spending*', whereas key politicians



(a) ESI

(b) Unemployment

Figure 2: Word clouds for words and outlets visualising the outputs of BEN.

like '*david__cameron*' and '*berlusconi*', are major participants in the word cloud for unemployment. In addition, the visualisations show a strong negative relationship between unemployment and the terms '*food*', '*russia*' and '*agriculture*', but no such relationship with respect to ESI. The disparity of these selections is evidence for our framework's capability to highlight features of lesser or greater importance to a given socioeconomic time series. The exact interpretation of the selected words and outlets is, perhaps, context-dependent and beyond the scope of this work.

In this paper, we presented a framework for performing a supervised analysis on news. An important factor for this process is that the bilinear nature of the learning function allows for a joint selection of important words and news outlets. Prediction performance is used as a reference point for determining whether the extracted outputs (i.e. the model's parameters) encapsulate relevant information regarding to the given indicator. Experiments were conducted on a set of EU-related news summaries and the supervising socioeconomic factors were the EU-wide ESI and unemployment. BEN outperformed the linear alternative (LEN), producing error rates below 8.8%.

The performance of our framework motivates several extensions to be explored in future work. Firstly, the incorporation of additional textual features may improve predictive capability and allow for richer interpretations of the term weights. For example, we could extend our term vocabulary using n-grams with $n > 2$, POS tags of words and entities (people, companies, places, etc.). Furthermore, multi-task learning approaches as well as models which incorporate the regularised learning of weights for different countries might give us further insights into the relationship between news, geographic location and socioeconomic indicators. Most importantly, we plan to gain a better understanding of the outputs by conducting a thorough analysis in collaboration with domain experts.

Acknowledgements

VL acknowledges the support from the EPSRC IRC project EP/K031953/1. DPP, SS, DG and TC were supported by EU-FP7-ICT project n.287863 ("TrendMiner").

References

R. Alexander Bentley, Alberto Acerbi, Paul Ormerod, and Vasileios Lampos. 2014. Books average previous decade of economic misery. *PLoS ONE*, 9(1).

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Johan Bollen and Huina Mao. 2011. Twitter mood as a stock market predictor. *IEEE Computer*, 44(10):91–94.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Social media as a measurement tool of depression in populations. In *Proceedings of ACM WebSci'13*, pages 47–56.

European Commision. 1997. *The joint harmonised EU programme of business and consumer surveys*. European economy: Reports and studies.

Ilias Flaounas, Marco Turchi, Omar Ali, Nick Fyson, Tijl De Bie, Nick Mosdell, Justin Lewis, and Nello Cristianini. 2010. The Structure of the EU Mediasphere. *PLoS ONE*, 5(12), 12.

Sarah Gelper and Christophe Croux. 2010. On the construction of the European Economic Sentiment Indicator. *Oxford Bulletin of Economics and Statistics*, 72(1):47–62.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Arthur E. Hoerl and Robert W. Kennard. 1970. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67.

Vasileios Lampos and Nello Cristianini. 2012. Nowcasting events from the Social Web with statistical learning. *ACM TIST*, 3(4):72:1–72:22.

Vasileios Lampos, Daniel Preoţiuc-Pietro, and Trevor Cohn. 2013. A user-centric model of voting intention from Social Media. In *Proceedings of ACL'13*, pages 993–1003.

Carolyn Lin, Michael B. Salwen, Bruce Garrison, and Paul D. Driscoll. 2005. Online news as a functional substitute for offline news. *Online news and the public*, pages 237–255.

Seungahn Nah and Deborah S. Chung. 2012. When citizens meet both professional and citizen journalists: Social trust, media credibility, and perceived journalistic roles among online community news readers. *Journalism*, 13(6):714–730.

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 AAAI ICWSM'10*, pages 122–129.

Michael J. Paul and Mark Dredze. 2011. You Are What You Tweet: Analyzing Twitter for Public Health. In *Proceedings of AAAI ICWSM'11*, pages 265–272.

Robert P. Schumaker and Hsinchun Chen. 2009. Textual analysis of stock market prediction using breaking financial news: the AZFin text system. *ACM TOIS*, 27(2):12:1–12:19.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE*, 8(9).

Robert Tibshirani. 1994. Regression shrinkage and selection via the lasso. *JRSS: Series B*, 58:267–288.

Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *JRSS: Series B*, 67(2):301–320.