PolText 2016

The International Conference on the Advances in Computational Analysis of Political Text

Sponsored by the European Social Fund,
Operational Programme Efficient Human Resources 2014–2020

Proceedings of the Conference

14–16 July 2016
Dubrovnik, Croatia

Organizers

Center for Empirical Research in Political Science,
Faculty of Political Science, University of Zagreb

Text Analysis and Knowledge Engineering Lab,
Faculty of Electrical Engineering and Computing, University of Zagreb

Partners

School of Social and Political Science, University of Edinburgh

Department of Political Science, University of Oslo

Sponsors

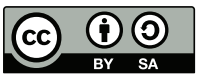
Publishers:

University of Zagreb, Faculty of Political Science
University of Zagreb, Faculty of Electrical Engineering and Computing

Editors:

Daniela Širinić, University of Zagreb
Jan Šnajder, University of Zagreb
Zoltán Fazekas, University of Oslo, Norway
Shaun Bevan, University of Edinburgh, United Kindgom

ISBN 978-953-6457-92-2 (University of Zagreb, Faculty of Political Science)
ISBN 978-953-184-220-4 (University of Zagreb, Faculty of Electrical Engineering and Computing)



Preface

The International Conference on the Advances in Computational Analysis of Political text is organized by the Center for Empirical Research in Political Science of the Faculty of Political Science, TakeLab of the Faculty of Electrical Engineering and Computing from the University of Zagreb, and two partner institutions, Universities of Edinburgh and Oslo. PolText 2016 is the first meeting in this series and will be held in Dubrovnik, Croatia on July 14–16. Technological developments, digital media, and advances in open government practices have made a vast amount of information available for social scientists. Most of this information is available as text. News portals disseminate political stories at unprecedented rates, politicians and political elites advertise their own messages through social media outlets and crowdsourcing provides new affordable and quick venues for asking citizens what they think about politics. With political texts at our fingertips, vexing research questions are emerging. Extracting, organizing, and analyzing large amounts of textual information can be quite resource-intensive with many political scientists lacking the skills necessary for dealing with such data. Fortunately, recent developments of cutting edge computational technologies such as automatic language processing, machine learning, and information extraction techniques has made research utilizing text-as-data more accessible and appealing. On the other hand, computational scholars equipped with novel technologies and linguistic solutions often have less experience with social science theories and less contextual knowledge about political data.

PolText organizers have recognized that there is a mutual benefit in connecting disparate worlds of computational text analysis and political science in analyzing political science research problems. The main aim of the conference is to facilitate this multidisciplinary cooperation. We invited contributions on computational approaches in analysing political text such as government speeches, political debates, social media, media content, party manifestos and/or legislation, but also contributions focusing on text categorization, topic modeling, information extraction, corpus analysis, sentiment analysis, stance classification and ideal point estimation, argumentation mining, political reputation analysis, techniques for multilingual text analysis and other language technologies.

We received 70 submissions, of which 21 were accepted. We thank all of the authors who submitted their work to PolText 2016. We are also grateful to our invited speakers, Stuart Soroka and Jon Oberlander, who welcomed our invitations to Dubrovnik.

This conference is sponsored by the European Social Fund – Operational Programme Efficient Human Resources 2014–2020, as an activity of the Croatian Policy Agendas Project (principal investigator: dr. Daniela Širinić) implemented by the Center for Empirical Research in Political Science of the Faculty of Political Science.

Daniela Širinić, Jan Šnajder, Zoltán Fazekas, and Shaun Bevan

Contents

Conference Program

Thursday, 14 July

19:30–21:00 Welcome reception and registration

Friday, 15 July

09:00–09:20 Welcome address

09:20–10:00 Keynote

The Potentials and Pitfalls of Automated Sentiment Analysis in Political Texts
Stuart Soroka

10:00–12:00 Session 1: Semantic Analysis
Moderator: Sebastian Popa

Semantic Annotation for the Analysis of Political Debates: A Graph-based Approach
Edoukou Philippe Armel N'Techobo, Amal Zouaq and Michel Gagnon

EUSpeech: A New Dataset of EU Elite Speeches
Gijs Schumacher, Martijn Schoonvelde, Denise Traber, Tanushree Dahiya and Erik de Vries

TopFish: Topic-Based Analysis of Political Position in US Electoral Campaigns
Federico Nanni, Cäcilia Zirn, Goran Glavaš, Jason Eichorst and Simone Paolo Ponzetto

Tracking Political Reputation with Distributional Semantic Models
Andrey Kutuzov and Alexander Bai

Discussants: Stuart Soroka and Philip Habel

12:00–12:15 Coffee break

12:15–14:00 Session 2: Mass Media Analysis
Moderator: Daniela Širinić

Topics and their Salience in the 2015 Parliamentary Election in Croatia: A Topic Model based Analysis of the Media Agenda
Damir Korenčić, Marijana Grbeša Zenzerović and Jan Šnajder

Legitimacy of New Forms of Governance in Public Discourse – An Automated Media Content Analysis Approach Driven by Techniques of Computational Linguistics
Michael Amsler and Bruno Wueest

Automated Detection of Chinese Government Astroturfers Using Network and Social Metadata
Blake Miller

Discussants: Shaun Bevan and Dan Hiaeshutter-Rice

14:00–15:00 Lunch break

15:00–16:45 Session 3: The Ins and Outs of Political Manifestos
Moderator: Zoltan Fazekas

Classifying Topics and Detecting Topic Shifts in Political Manifestos
Caecilia Zirn, Goran Glavaš, Federico Nanni, Jason Eichorst and Heiner Stuckenschmidt

How Intra-Party Disagreement Determines Issue Salience in Election Manifestos
Zachary Greene

Predicting Political Party Affiliation from Text
Felix Biessmann, Pola Lehmann, Daniel Kirsch and Sebastian Schelter

Discussants: Andrej Kutuzov and Sebastian Popa

Saturday, 16 July

09:00–09:45 Keynote

Role Theory Meets Relation Extraction: Initial Experiments on Perceptions of the EU
Jon Oberlander

09:45–11:30 Session 4: Social Media Analysis
Moderator: Shaun Bevan

Talk is Cheap: Selective Politicization of EU Dimension in the 2014 EP Elections
Sebastian Popa, Zoltan Fazekas, Hermann Schmitt, Pablo Barberá and Yannis Theocharis

The Medium is the Message: Automated Content-Analytic Techniques Across Mass Media Platforms
Dan Hiaeshutter-Rice

On Classifying Twitter Users' Policy-Relevant Community Affiliations Using DBpedia
Anjie Fang, Philip Habel, Iadh Ounis, Craig Macdonald and Xiao Yang

Discussants: Jon Oberlander and Matthew Loftis

11:30-11:45 Coffee break

11:45-13:30 Session 5: New Tools and Methods
Moderator: Goran Glavaš

The Mixture Index of Fit in Text Analysis
Juraj Medzihorsky

Collaborating with the Machines: Building Big Political Data Sets on a Budget
Matt W. Loftis and Peter B. Mortensen

Computer-Aided Dictionary Making: An Efficient Dictionary Construction Technique for Content Analysis
Kohei Watanabe

Discussants: Goran Glavaš and Zoltan Fazekas

13:30–14:30 Lunch break

14:30–16:15 Session 6: Measurement and Meaning
Moderator: Pola Lehmann
The Meaning of Democracy: Using Distributional Semantics to Account for Meaning Differences
Stefan Dahlberg , Sofia Axelsson, Markus Sahlgren and Amaru Cuba Gyllensten

Predicting Government (Non)Responsiveness to Freedom of Information Requests with Supervised Latent Dirichlet Allocation
Benjamin Bagozzi, Daniel Berliner and Zack Almquist

VisArgue – A Visual Text Analytics Framework for the Study of Deliberative Communication
Mennatallah El-Assady, Valentin Gold, Annette Hautli-Janisz, Wolfgang Jentner, Miriam Butt, Katharina Holzinger and Daniel Kleim

Discussants: Juraj Medzihorsky and Damir Korenčić

16:15–16:30 Coffee break

16:30–17:30 Session 7: Making Sense of Parliamentary Texts
Moderator: Zachary Greene

Beyond Money in Politics: Automatic Detection of Legislative Text Re-Use
Eugenia Giraudy, Matthew Burgess, Julian Katz-Samuels and Joe Walsh

Mapping the Immigration Debate in the Swedish Riksdag
Eitan Tzelgov and Petrus Sundin Olander

Discussants: Edoukou Philippe Armel N'Techobo and Randolph M. Siverson

Legitimacy of New Forms of Governance in Public Discourse - An Automated Media Content Analysis Approach Driven by Techniques of Computational Linguistics

Michael Amsler
Institute of Computational Linguistics
University of Zurich
mamsler@cl.uzh.ch

Bruno Wüest
Institute of Political Science
University of Zurich
wueest@ipz.uzh.ch

Abstract

For political scientists, it is increasingly important to explore large text collections without time-consuming human intervention. We are presenting a language technology tool kit that allows political scientists to extract information on various forms of governance from a comprehensive multilingual corpus. The tool kit allows searching for governance entities and measuring their salience, tonality and media frames. In substantial terms, our pipeline enables scholars of governance to extend their research focus to the previously neglected area of public communication.

1 Introduction

Automated approaches to analyze unstructured text data have made tremendous progress in computational linguistics in the last decades (Jurafsky and Martin, 2009). At the same time, social scientists are increasingly in need of such approaches, since the number of large, digitally available text collections is constantly growing. The obvious task then is to transfer the comprehensive computational linguistic tool set in order to meet the specific requirements of social scientific studies (Wueest et al., 2011). In this contribution, we present a pipeline of language technologies that allows the analysis of public communication in a specific yet fundamental research domain for the political sciences: democratic governance.

The denationalization and privatization of democratic governance poses formidable challenges to the traditional, territorially grounded forms of democratic authorities (Zürn, 1998). At the European and international level, new modes of governance such as supra-national and intergovernmental bodies as well as transgovernmental networks have come to supplement classic intergovernmental governance (Abbott and Snidal, 2008). At the sub-national level, regulatory agencies and public-private partnerships increasingly spread across metropolitan regions by transforming traditional regional and local state institutions (Kelleher and Lowery, 2009).

These various new forms of governance have in common that they organize political authority along functional rather than territorial lines, which also implies that they are decoupled from representative democratic control. This is why observers often declare a loss of democratic legitimacy for the political system (Follesdal and Hix, 2006; Keohane et al., 2009). However, other scholars usually point to formal accountability mechanisms such as governmental and parliamentary over-sight as well as judicial review, which can at least partly compensate a deficit in democratic legitimacy (Lodge, 2002). Other, more informal mechanisms of accountability such as media coverage, in contrast, have been either neglected or dismissed as scarcely relevant (Maggetti, 2012).

This is surprising, given that public communication plays an ever more decisive role for setting the political agenda and establishing the distribution of information on policy making in modern democratic societies (Walgrave et al., 2008; Müller, 2014; Arnold, 2004). Media coverage is assumed to hold new forms governance accountable through reputational mechanisms (Gentzkow and Shapiro, 2006). If media regularly pay critical attention to governance processes, they can encourage the formation of an informed public opinion (O'Donnell, 1998). This, in turn, mounts pressure on governance actors to explain, justify and – if necessary – correct their conduct.

In the following, we present a comprehensive corpus and language technology pipeline, which enable political scientists to assess these questions.

Proceedings of the International Conference on the Advances in Computational Analysis of Political Text (PolText 2016), pages 1–7, Dubrovnik, Croatia, 14–16 July 2016.

The paper begins by presenting our operationalization of indicators that allow the reliable measurement of governance accountability in a large-scale text analysis. Subsequently, we will describe the software pipeline and language technologies necessary to implement the operationalization, before we present a case study highlighting the feasibility of our approach.

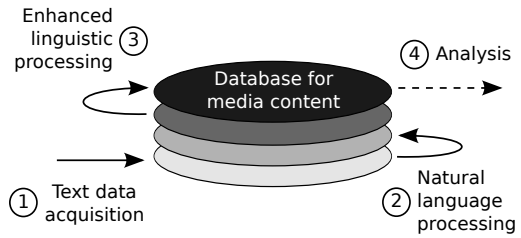
2 Measuring media coverage on governance accountability

So far, mediatized accountability mechanisms have only been dealt with in conceptual elaborations or comparative case studies that entailed manual content analyses (Maggetti, 2012; Coglianese and Howard, 1998; Gerhards and Roose, 2007). Although these contributions are theoretically insightful and empirically rich, their focus on a narrow set of actors, geographical units or media sources always faces the necessity to justify why their cases provide more than just idiosyncratic evidence. We suggest that an automated large-scale analysis helps to achieve a more broad analytical support on the question whether and how media scrutinize on the accountability of governance processes.

2.1 Sample

The anchor of the analysis is a large gazetteer of pre-defined entities related to governance (see Figure 1). These entities refer to actors (collective actors and individuals), policy fields and regulation such as treaties or directives. At the moment, a comprehensive gazetteer of entities for 3257 queries is integrated in the document retrieval. The entities cover a large variety of forms of governance: transgovernmental networks, independent as well as private regulatory authorities, metropolitan bodies, supranational parliaments and international environmental governance outcomes.

Figure 1: Stylized workflow in the language processing pipeline



In a first step, a comprehensive corpus of the following newspapers, newswires and online sources is established by retrieving all articles for the keyword gazetteer via API accesses to media content databases such as Lexis Nexis ((1) in Figure 1).

- *Quality:* Frankfurter Allgemeine, Süddeutsche Zeitung, Welt, Tageszeitung (Germany); Figaro, Le Monde (France); Neue Zürcher Zeitung, Le Temps (Switzerland); The Guardian, London Times, Independent (UK)
- *Tabloid/Freesheets:* Bild (Germany); Aujourd'hui en France, 20 minutes (France); Blick, Le Matin, 20 Minuten (Switzerland); Daily Mail, Daily Mirror, Metro (UK)
- *Magazines:* Spiegel, Stern, Zeit (Germany); Nouvel Observateur, L'Express (France); Weltwoche, Wochenzeitung, L'Hedbo (Switzerland); New Statesman, Spectator, Economist (UK)
- *Regional:* Berliner Zeitung, Stuttgarter Zeitung, Stuttgarter Nachrichten (Germany); Le Parisien, Le Progrés (France); Tagesanzeiger, Berner Zeitung (Switzerland); London Evening Standard, City A.M., Birmingham Mail, Birmingham Post (UK)
- *Online sources:* Spiegel Online (Germany), Figaro Online, Le Monde Online (France); 20 Minuten Online (Switzerland); BBC News Online (UK)
- *Newswires:* Associated Press, Agence France Presse, Deutsche Presse Agentur, BBC Monitoring, Europolitics, ENP Newswire, AWP

Since different types of media systems (Hallin and Mancini, 2004), as well as different types of media (Strömbäck and Kaid, 2008) possibly cover governance in different ways, the media sources are sampled so that there is a balanced set of outlets in our four countries (Switzerland, Germany, France, and United Kingdom). From each type of outlets, the outlet with the highest circulation (or website visits in the case of the online sources) was chosen. As far as possible, we also cover other potential variations such as different ideological leanings. In addition to these country-specific media samples, we also include a range of internationally operating newswires, which provide us with information on the general reporting on governance in disregard of specific journalistic cultures in single media outlets.

Subsequently, an additional layer of data consisting of the compressed documents along with initial meta-data (source, date-of-publication etc.) is added to the database ((2)). At a third stage, we employ a full natural language processing chain, which includes morphological analysis, tagging,

lemmatizing, and dependency parsing ((3)). Finally, a fourth layer of enhanced linguistic analysis – named entity recognition, co-reference resolution, sentiment detection, opinion mining and topic modeling – is implemented to calculate the indicators of interest we will discuss in the following ((4)).

2.2 Salience

The attention media pay to specific forms of governance is the obvious starting point of the data generation process. No media attention is the worst case in terms of question regarding the public accountability and legitimization of governance, since 'quiet politics' (Culpepper, 2010) implies low interest by the public and, correspondingly, high leverage for particular interests and dishonest conduct in governance processes. The first necessary measure therefore is salience, defined as the visibility of specific forms of governance in the media.

2.3 Tonality

A second crucial information on governance entities is the media's evaluation of these governance entities in terms of tonality. The tone of media reports on governance entities yields useful results if changes in tonality signify reactions to events on the governance processes under concern (Maggetti, 2012). For example, if a corruption scandal shakes a governance actor, we expect media reports to shift to a negative tone. This also implies that tonality has to be measured at the level of the specific entity and not at the level of text documents as a whole.

2.4 Issues

Governance entities may draw media attention for different reasons, but not all are relevant for the research objective. If a sports magazine reported on the passion of the head of the Swiss Financial Markets Supervisory Agency (Finma) for windsurfing (which arguably is true), hardly any political analyst would deem this information relevant to understand financial market regulation in Switzerland. More generally, evidence on the thematic context in which governance entities are mentioned is key to assess whether media reports on specific entities are actually covering the governance processes of interest.

2.5 Frames

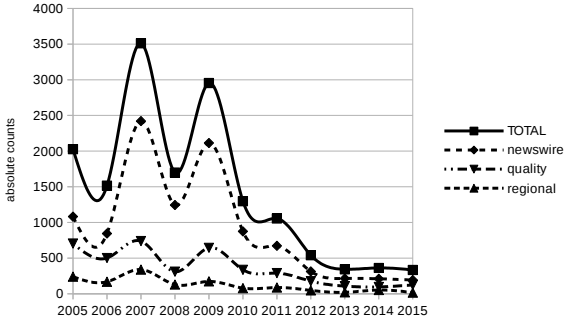
What is still missing is information on the reasons why the media report on governance entities, i.e. which interpretations and problem definitions journalists convey to the reader. To this aim, we additionally conduct a media frame analysis-(Entman et al., 2009; Goffman, 1974). In the context of this analysis, we specify frames as generic schemata of interpretation that refer to the main source of democratic legitimacy of governance entities as it is reported in the text documents. More precisely, we separate input-oriented legitimacy frames from throughput- and output-oriented ones (Easton, 1965; Schmidt, 2013). Input legitimacy is thus present if media refer to participatory aspects, civil society involvement, popular support and democratic accountability in general, or public interest representation with regards to governance processes. Throughput denotes the quality of governance processes in terms of their accountability, legality and transparency. Output legitimacy, accordingly, refers to the efficiency and effectiveness of governance.

3 The public accountability of the Kyoto Protocol

3.1 Salience

For this case study we measure salience as the occurrence of articles in the media coverage across the timeline. Although a simple measurement, the salience reveals on the one hand important insights about the presence of the respective entity and, on the other hand, offers the opportunity to closer scrutinize the content near the peaks.

Figure 2: Salience of articles referring to Kyoto Protocol (only English articles; n=15,849)



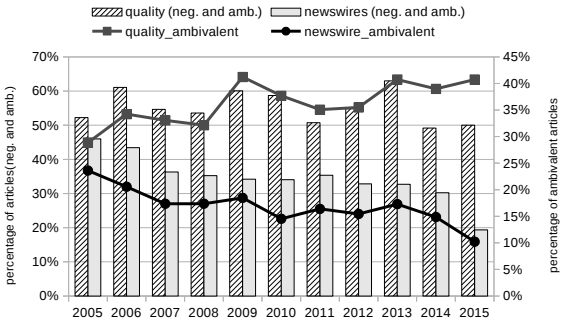
As can be seen in Figure 2, the visibility manifests itself with two clear peaks in 2007 and 2009. A closer investigation of the respective cov-

erage points towards the importance of the Fourth Assessment Report of the United Nations Intergovernmental Panel on Climate Change (IPCC) in 2007 and the 2009 United Nations Climate Change Conference in Copenhagen which triggered each an increased attendance to the subject.

3.2 Tonality

To measure tonality in the media coverage, we apply a linguistically informed sentiment analysis system, similar to (Taboada et al., 2011). The system used for this task was evaluated in another case study for the tracking of coverage tonality which yielded good results (see Wueest et al. (2014)). A more detailed description can be found in (Klenner et al., 2014). Although the tonality can be derived for singular entities in the given texts, we aggregate in this case study on the document level since the thematical focus is narrowed by the data acquisition process (i.e. the query to the media databases).

Figure 3: Comparison of negative and ambivalent tonality between media types *quality* and *newswire*



In Figure 3 we focus on the difference of tonality regarding the level of critique considering different media types: the bars show the percentage of articles of negative and ambivalent tonality (ordinate on the left-hand side). It is obvious that the coverage in quality papers is much more critical than in the newswire articles. The lines show the percentage of only the ambivalent articles (ordinate on the right-hand side) which reveals that the difference between the two media types mainly stems from the much higher percentage of ambivalent articles, that is, articles which discuss the topic under different perspectives, considering chances and risks as well as progress and failure in the implementation process.

3.3 Issues

We apply structural topic models (STM) (Roberts et al., forthcoming) to explore the thematic context in which the media writes about governance. STM is a data-driven technique, which allows us to estimate document probabilities for latent variables, called topics. STM builds on the Latent Dirichlet Allocation, a hierarchical mixed-membership model in which the document-topic and word-topic probabilities have a common prior drawn from a Dirichlet distribution (Blei et al., 2003). One of the STM's major innovations is that the prior distribution of topics (i.e. topic prevalence) can be influenced by covariates. In the following analysis, we use the newspaper names and a b-spline with 10 degrees of freedom on a monthly trend variable to control for unwanted linguistic differences across news outlets and over time. In addition, we apply a parametric evaluation of the most probable topic-word vectors in order to find the optimal number of topics. To this purpose, we use word2vec (Mikolov and Dean, 2013), which learns and aggregates term similarities through a shallow neural network process. These term similarities can then be used to compare topic coherence and exclusiveness across different topic models. For the Kyoto protocol corpus, word2vec suggests a granularity of 19 for a candidate range of 3 to 20 topics.

Figure 4: Dynamics of selected topics

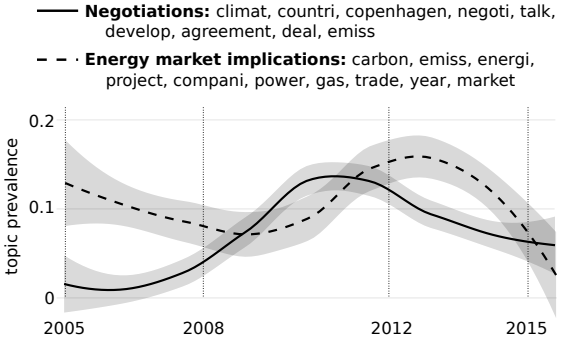


Figure 4 shows the trends in the prevalence of two especially meaningful topics over time. In addition, the list of the 10 most probable word stems for each topic is listed.

The first topic summarizes the different negotiation rounds on the Kyoto protocol, most notably the first commitment period from 2008 until 2012 with the Copenhagen summit in 2009 as key event. Reports on the different negotiations accordingly

peak in this period. The second topic, in contrast, highlights the consequences of the Kyoto protocol on the energy markets and emission trading. Quite intuitively, this topic becomes most prevalent in the aftermath the big policy decisions from 2011 on.

3.4 Frames

While we have focused on purely empirical data for the other indicators, we will report first insights from the methodological approach used for the framing measurement.

In contrast to the measurements for the other indicators which are derived generically, we rely on annotated data for the framing. More precisely, we annotate the frames using the brat annotation tool (Stenetorp et al., 2012). So far, our three annotators build a valuable training corpus of about 14,000 frames[1]. After an intensive training phase, inter-annotator agreement is constantly high (micro-averaged F1-scores for fine-grained frame categories that range between 0.66 for 23 documents during and 0.71 for 5 documents at the start of the annotation). Since the annotation task is not yet finished and the implementation of the supervised machine learning approach is still under development we report preliminary results for a baseline, using paragraph-based bag-of-words model including different settings but based on only about 15% of the frames.

First attempts have revealed that the recognition of frames is a challenging task, especially since we encounter a skewed distribution in the data (i.e. paragraphs containing frames vs. paragraphs without frames). Additionally, the distribution between the different types of frames is skewed as well (i.e. some frames occur much more than others), which then again complicates the task for a supervised learning approach. Hence we plan to implement the automated approach designed as follows: in the first stage we will apply a model that tries to detect paragraphs with mentions of democratic legitimacy (as a generic category). Second, we will then differentiate between input, output and throughput frames and apply the fine-grained frame classification in the end within this categories.

For frame detection we report an F1-score of 0.81 (micro-averaged) and 0.66 (macro-averaged) for the binary classification as a baseline. Table 1 shows precision, recall, and F1 scores for the individual categories. In the second scenario we added the annotated text passages (TP) upweighted to the bag-of-words (BoW) and word embeddings (emb.) features. Interestingly, precision was much more positively affected than recall for the frames while it was the other way around for the paragraphs not containing frames. It has to be mentioned that these first baseline results leave room for improvement, especially for recall. However, we propose a more thorough generalization based on a deeper linguistic analysis (i.e. syntactic and semantic information) for a better performance but such an approach is yet to be implemented.

Table 1: Evaluation of 10-fold cross-validation for the detection of frames in paragraphs

	Frame			**No Frame**			
	Prec.	Rec.	F1	Prec.	Rec.	F1	Acc.
BoW+emb.,	0.44	0.36	0.40	0.84	0.87	0.85	0,76
BoW+emb.+TP	0.61	0.36	0.45	0.84	0.94	0.89	0.81

In the conducted experiments the following features have proven useful for the classification task: unigrams (including lower-cased variant), bigrams, word embeddings (from GloVe (Pennington et al., 2014)), and especially the upweighted annotated text passages. Additionally, we do not include class bias.

4 Conclusion

This project starts from the assumption that the salience, tonality and issues in media reports on governance entities reveal crucial evidence on whether and how media coverage entails mechanisms of accountability. More precisely, if media adjust their attention according to events related to specific governance entities, if media react to failure with a negative tone – and to success with a positive tone – and if the media really cover the issues related to the area of responsibility of these governance entities, media coverage actually constitute an 'accountability forum' for this governance entity.

Acknowledgments

Research for this paper was funded by the NCCR Democracy at the University of Zurich.

[1] At this point we must thank Michelle Amman, Anna Sigrist and Anna-Lina Müeller for their excellent work on the manual annotation data.

References

K. W. Abbott and D. Snidal. 2008. The governance triangle: regulatory standards institutions and the shadow of the state. In Walter Mattli and Ngaire Woods, editors, *The Politics of Global Regulation*. Princeton University Press, Princeton, NJ.

R. D. Arnold. 2004. *Congress, the press, and political accountability*. Princeton University Press, Princeton, NJ.

D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning*, 3:993–1022.

C. Coglianese and M. Howard. 1998. Getting the message out: Regulatory policy and the press. *Politics*, 3:39–55.

P. D. Culpepper. 2010. *Quiet Politics and Business Power. Corporate Control in Europe and Japan*. Cambridge University Press, Cambridge, MA.

D. Easton. 1965. *A Systems Analysis of Political Life*. Wiley, New York, NY.

R. M. Entman, J. Matthes, and L. Pellicano. 2009. Framing politics in the news: Nature, sources and effects. In K. Wahl-Jorgensen and T. Hanitzsch, editors, *Handbook of Journalism Studies*. Routledge, London.

A. Follesdal and S. Hix. 2006. Why there is a democratic deficit in the EU: A response to Majone and Moravcsik. *JCMS: Journal of Common Market Studies*, 44:533–562.

M Gentzkow and J. M. Shapiro. 2006. Media bias and reputation. *Journal of Political Economy*, 114(2):280–316.

A. Offerhaus Gerhards, J. and J. Roose. 2007. Die öffentliche zuschreibung von verantwortung. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 59:105–124.

E. Goffman. 1974. *Frame Analysis: An Essay on the Organization of Experience*. Harper & Row, New York, NY.

D. C. Hallin and P. Mancini. 2004. *Comparing Media Systems: three Models of Media and Politics*. Cambridge University Press, Cambridge, MA.

D. Jurafsky and J.H. Martin. 2009. *Speech an Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall, Upper Saddle River, NJ.

C. A. Kelleher and D. Lowery. 2009. Central city size, metropolitan institutions and political participation. *British Journal of Political Science*, 39(1):59–92.

R. O. Keohane, S. Macedo, and A. Moravcsik. 2009. Democracy-enhancing multilateralism. *International Organization*, 63(1):1–31.

M. Klenner, M. Amsler, and N. Hollenstein. 2014. Verb polarity frames: a new resource and its application in target-specific polarity classification. In *Proceedings of KONVENS 2014*, pages 106–115.

M. Lodge. 2002. The wrong type of regulation? regulatory failure and the railways in Britain and Germany. *Journal of Public Policy*, 22:271–297.

M. Maggetti. 2012. The media accountability of independent regulatory agencies. *European Political Science Review*, 4(3):385–408.

K. Chen G. Corrado Mikolov, T. and J. Dean, 2013. *Efficient Estimation of Word Representations in Vector Space*. CoRR, abs/1301.3781.

L. Müller. 2014. *Patterns of Media Performance: Comparing the Contribution of Mass Media to Democratic Quality Worldwide*. Palgrave Macmillan, Houndmills, UK.

G. A. O'Donnell. 1998. Horizontal accountability in new democracies. *Journal of Democracy*, 9:112–126.

J. Pennington, R. Socher, and C. D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

M. E. Roberts, B. M. Stewart, and M. E. Airoldi. forthcoming. A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*.

V. A. Schmidt. 2013. Democracy and legitimacy in the european union revisited: Input, output and 'throughput'. *Political Studies*, 61:1467–9248.

P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*, Avignon, France, April. Association for Computational Linguistics.

J. Strömbäck and L. L. Kaid. 2008. *The Handbook of Election News Coverage Around the World*. Routledge, New York, NY.

M. Taboada, J. Brooke M., Tofiloski, K. Voll, and M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

S. Walgrave, S. Soroka, and M. Nuytemans. 2008. The mass media's political agenda-setting power: A longitudinal analysis of media, parliament, and government in Belgium (1993 to 2000). *Comparative Political Studies*, 41:814–836.

B. Wueest, S. Clematide, A. Bünzli, D. Laupper, and T. Frey. 2011. Electoral campaigns and relation mining: Extracting semantic network data from swiss newspaper articles. *Journal of Information Technology and Politics*, 8(4):444–463.

B. Wueest, G. Schneider, and M. Amsler. 2014. Measuring the public accountability of new modes of governance. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 38–43, Baltimore, MD, USA, June. Association for Computational Linguistics.

M. Zürn. 1998. *Regieren Jenseits des Nationalstaats. Globalisierung und Denationalisierung als Chance*. Suhrkamp, Frankfurt a. M., DE.

Predicting Government (Non)Responsiveness to Freedom of Information Requests with Supervised Latent Dirichlet Allocation

Benjamin E. Bagozzi
Department of Political Science
& International Relations
University of Delaware
bagozzib@udel.edu

Daniel Berliner
School of Politics
& Global Studies
Arizona State University
danberliner@gmail.com

Zack W. Almquist
Department of Sociology
& School of Statistics
University of Minnesota
almquist@umn.edu

Abstract

Understanding government responsiveness to citizen information requests is important to theories of political accountability, as well as to practitioners' abilities to monitor and improve this crucial transparency mechanism. We use supervised latent Dirichlet allocation techniques to predict the Mexican government's (non)responsiveness to *all* federal information requests filed during the period 2003-2015. After presenting our approach, we assess its value-added in both the in-sample and out-of-sample settings.

1 Introduction

Following Mexico's landmark 2002 access to information law (Berliner and Erlich, 2015), every single freedom of information request filed with federal government agencies has been made publicly available—now over one million requests in total. Understanding the Mexican government's responsiveness to these individual information requests is important for theories of government responsiveness (and its politicization), as well as for practitioners' abilities to monitor, scrutinize, and improve the quality of this critical accountability mechanism. Such laws, similar to the Freedom of Information Act in the United States, have now been adopted by over 100 countries around the world (Berliner, 2014; Berliner, 2016).

After converting the complete corpus of Mexican public information requests (2003-2015) into machine readable text, we use topic models to predict government (non)responsiveness towards Mexican information requests in both an in-sample and out-of-sample context. Specifically, we apply supervised latent Dirichlet allocation (sLDA) techniques to this text corpus, so as to evaluate the extent to which one can use the texts of individual requests to predict the (i) time until a government response and (ii) probability of a "denied request." We then evaluate the value-added of this approach against several alternatives. Finally, we assess our sLDA topics for their "politicization," and find that the topics that are most strongly associated with *nonresponsiveness* do indeed exhibit more politicization than do the topics most associated with high *responsiveness*.

2 Background

Democratic institutions are founded on the notion of responsive government, but responsiveness is usually limited and incomplete. Many scholars have studied why political actors may be more or less responsive in different circumstances — both at a macro-scale in terms of how government policies and spending respond to the preferences of the median voter (Golden and Min, 2013), and at a micro-scale in terms of individual citizen-government interactions (Lagunes, 2008; Butler and Broockman, 2011; McClendon, 2016).

Building upon the latter approach, we examine government responsiveness in one case of frequent government-citizen interaction: responses to public information requests in Mexico. To do so, we use a comprehensive dataset of over one million information requests filed with federal government agencies. These correspond to queries made by individual citizens, legal representatives, businesses, and NGOs to specific Mexican federal government agencies, and cover, for example, requests for information on government salaries, land use and zoning restrictions, or distributive programs. Due to the unique online information platform created by Mexico's 2002 *Ley Federal de Transparencia y Acceso a la Información Pública Gubernamental*, the text of each of these requests, along with associated metadeta, has been made publicly available for the years 2003-2015.

Proceedings of the International Conference on the Advances in Computational Analysis of Political Text (PolText 2016), pages 8–13, Dubrovnik, Croatia, 14–16 July 2016.

2.1 Measuring (Non)Responsiveness

Our analysis focuses upon predicting government (non)responsiveness to these information requests. We are interested both in the *timing* of response and in the *type* of response: information provided or denied. We accordingly use two separate measures to evaluate the (non)responsiveness to any given information request: (i) a binary indicator of "denied requests" (for various reasons) and (ii) information on the time-until-response.

Regarding our time-until-response measure, we create an outcome variable that corresponds to the logged number of working days (excluding weekends and official Mexican government holidays) until an information request response is provided to the requestor by the Mexican government. While the standard time limit for the Mexican government to provide a response is 20 working days, officials can request an extension of up to a maximum of 40 working days. Across our entire dataset, 66.4% of requests received responses within 20 working days while 89.3% of requests received responses within 40 working days. Our final (logged) time-until-response measure has a mean of 2.89 and range of 0.00-to-7.59.

Our binary "denied request" indicator is our primary outcome of interest, and is based upon the coding scheme developed by Fox et al. (2011), which classifies any response marked as "No es de competencia," "Inexistencia," "Reservada," "No se dará trámite," "Solicitud no corresponde al marco de la ley" and "Sin Respuesta" as a "denied request" ($= 1$), and zero otherwise. The resultant "denied request" indicator is moderately imbalanced with a sample mean of 0.23. Finally, we then also omit the final two months of information requests from our analyses below, to ensure that we do not treat any cases marked as "Sin Respuesta" as "denied" when they had simply not yet exceeded the time limits for response.

2.2 Information Request Features

We focus on the request texts themselves as our primary features of interest. These texts correspond to each requestor's own open-ended description of the specific information that they are requesting. Because public officials are the primary responders to these requests, we believe that the themes found across these requests, and their varying degrees of politicization, will help to predict government (non)responsiveness.

We thus downloaded all requests from Mexico's online information request interface. While most requestors described the nature of their requests within the designated field, a smaller subset (roughly 13%) included a portion or all of their request as an attachment. Because these attachments are relevant to our analysis, we additionally downloaded each attachment and added these into our primary request text field, along with any auxiliary request content. We then (i) removed all requests pertaining to confidential personal information and (ii) truncated all remaining requests from the thousandth string onwards.[1] This created our primary corpus of interest, which was further preprocessed using standard approaches (e.g., stemming) for the automated analysis of political texts (Bagozzi and Schrodt, 2012; Bagozzi, 2015; Berliner et al., 2016). Altogether, the above steps yielded a corpus of 1,003,756 requests.

We next appended the names of each request's designated federal government agency to our processed texts. Each information request in our sample designated a single government agency, such as the Instituto-Nacional-de-Desarrollo-Social, as the *target agency* for the information that was requested. As these agencies vary in their levels of politicization and resources, we anticipate agency-designation, like a request's textual content, to influence the degree of (non)responsiveness to a given request. Agency information was included as an additional field within the original request metadata, and encompasses roughly 300 distinct Mexican federal agencies for our sample. Further below, we evaluate the contribution of this addition to our prediction and classification tasks.

3 Supervised Latent Dirichlet Allocation

Topic models have recently been shown to be highly valid for the discovery of latent thematic content within Mexico's information request texts (Berliner et al., 2016). As such, the present paper evaluates the utility of *supervised* latent Dirichlet allocation (sLDA) models (Blei and Mcauliffe, 2008) for the prediction of government (non)responsiveness to these same request texts.

sLDA is a probabilistic topic model designed for identifying the groupings of words that are most predictive of a document-indexed response variable. sLDA estimates these groupings of

[1] Only 0.02% of our documents have more than 1,000 strings; most are attachments with extensive itemized lists.

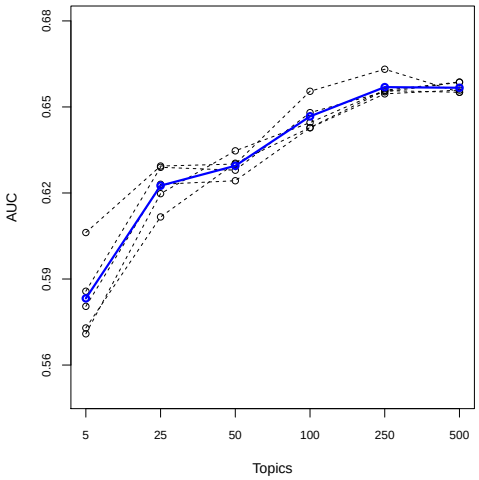
words—hereafter referred to as topics—via a three-level hierarchical model that treats each document as containing a finite mixture of underlying topics, where the topics themselves are specified as an infinite mixture over a corresponding latent set of topic probabilities. One's document-level responses are then regressed on these estimated topic frequencies so as to restrict responses to be non-exchangeable with words, while allowing for flexibility between topic frequency and response type under a generalized linear model (GLM) framework (Blei and Mcauliffe, 2008).

Under this approach, our information request texts are assumed to be mixtures of multiple *latent topics*, each with a characteristic set of words. We anticipate that a subset of these latent topics will be highly politicized, and hence expect that our modeling of all topics across all request documents will aid in the prediction of government (non)responsiveness, as measured via (i) logged time-until-response or (ii) "denied request." In each sLDA model presented below, we specify the distribution of the former response variable to be Gaussian and the latter to be logistic, and perform estimation using collapsed Gibbs sampling via the 'lda' package in R (Chang, 2015).

Researchers must assign the number of topics, k, to be estimated within sLDA. We use a five-fold cross-validation approach to identify an optimal number of topics for the task of prediction. To do so, we first draw a random sample of approximately 250,000 information requests and then randomly partition this sample into five folds of training and test data. For each set of training data, we next estimate a series of sLDA models where the number of topics, k, is sequentially set to $k = \{5, 20, 50, 100, 250, 500\}$ and where our outcome variable is assigned as the binary "denied request" measure described above. We then use each resultant sLDA model's output to initialize a validation sLDA model using each fold's corresponding test sample. With these results in hand, we calculate the area under each test sample's corresponding receiver operating characteristic curve (i.e., the AUC) for "denied requests."

Figure 1 plots the corresponding AUCs for all k's evaluated, along with mean AUCs (the solid line), and indicates that an optimal number of topics for the task of predicting "denied requests" rests somewhere in the $k = 250$ range, since this topic number yields the highest average AUC for our cross-validation sample (i.e., 66%). We hence set $k = 250$ for all primary sLDA models below.

Figure 1: Cross-Validation Results



4 Evaluations

Having used a random sample of 25% of all request texts to identify an optimal number of topics, we next evaluate our sLDA model on our remaining (held out) texts. To do so, we first re-estimate a final ($k = 250$) sLDA model on *all* of the previously sampled 250,000 documents, separately for each outcome of interest: (i) logged time-until-response and (ii) "denied requests." We then generate in-sample and out-of-sample predictions for our two outcome variables, where for our out-of-sample predictions we use the remaining 75% of our sample data (i.e., $\approx 750,000$ request texts).

4.1 In-Sample Results

In order to assess our in-sample sLDA results for both (i) time-until response and (ii) "denied request," this subsection first discusses our topic-specific coefficient estimates, followed by an evaluation of the topics most predictive of (non)responsiveness, and then finally an assessment of in-sample classification. For both models, nearly all of our 250 topic-specific estimates are statistically significant under traditional thresholds, with the vast majority implying either an increase in responsiveness—or a slight increase nonresponsiveness—when present. However, a small number of topics exhibit very large positive effects on non-responsiveness in each model. We hence identify the two topics with the largest es-

timated effects on (i) nonresponsiveness and (ii) responsiveness from *each* sLDA response-model for further examination.

The top words associated with these 'highly predictive topics' are presented below, where we have de-stemmed all topwords, removed *target agency* names (if present), and translated each resultant word to English. The two topics that are most predictive of nonresponsiveness, Slowest$_{\#1}$ and Denied$_{\#1}$, each capture the same highly politicized theme: investigative requests pertaining to financial improprieties, accreditations, and scandals (e.g., FICREA, a collapsed credit union under fraud investigation). Notably, our sLDA estimates imply that requests associated with this topic see a 18,064% increase in the odds of a " denied request," and a 47 day increase in time-until-response. By comparison, the median increase in the odds of a "denied request," and the median increase in time-until-response—across all 250 of our topic estimates—are 110.7% and 1-day.

The second most predictive topic of a "denied request" (Denied$_{\#2}$), likewise appears to be highly politicized, with topwords associated with inquires into money-and-politics, including topwords such as "money," "where," "diputados" and "senators," and with an estimated increase in the odds of a "denied request" of 13,466%. By contrast, Slowest$_{\#2}$ instead appears to be slightly less politicized with its topwords suggesting a more general focus on government accreditation and endorsement. Nevertheless, on the whole, these four topics are far more politicized than the topwords found within Fastest$_{\#1}$, Fastest$_{\#2}$, Provided$_{\#1}$, Provided$_{\#2}$, which as can be seen below, encompass themes of politeness, benign information queries, and requests concerning commercial-product and energy-rate information.

Topics most predictive of time-until-response:

- Slowest$_{\#1}$: saving, FICREA, financial, users, CONDUSEF, bank, settlement, value, accreditation, society
- Slowest$_{\#2}$: documents, accreditation, published, any, electronic, endorses, I request, copy, contains, fact
- Fastest$_{\#1}$: do, requirements, business, can, answer, necessary, respect, you can, question, information
- Fastest$_{\#2}$: registry, brand, involved, find, commercial, I request, property, kind, medium, so

Topics most predictive of a "denied request":

- Denied$_{\#1}$: value, saving, settlement, financial, protections, any, interventions, concept, banking, society
- Denied$_{\#2}$: money, change, deputies, decommissioned, quantity, western, where, year, information, senators
- Provided$_{\#1}$: electronic, energy, CFE, municipality, consumption, rate, lighting, bills, users, latest
- Provided$_{\#2}$: IFAI, I request, information, published, cape, carry, process, following, opinion, federal

We next evaluate the in-sample classification performance of our sLDA models. In the interest of space, we focus all ensuing discussions on the binary "denied request" outcome and results. We then construct two random "coin-flip" baselines for comparison, hereafter denoted ξ, with the first generating random binary data with probability $\frac{1}{2}$, and the second generating random binary data with probability equal to the mean of our true binary response $\bar{y} = 0.23$. In this manner $\xi = \bar{y}$ provides us with a random classifier that maximizes overall accuracy, whereas $\xi = \frac{1}{2}$ provides us with a random classifier that instead favors the improved identification of cases within our less frequent outcome (i.e., nonresponsiveness).

We compare these two random classifiers against our in-sample "denied request" sLDA results with the aid of AUCs, true positive rates (TPRs), true negative rates (TNRs), F1 scores, and overall classification accuracy. Given our preference for the accurate prediction of our minority class (i.e., nonresponsiveness), we assign a cutoff of 0.25 for the calculation of our TPR, TNR, F1 score, and accuracy values.

As can be seen in Table 1, our AUC values imply that our sLDA in-sample predictions are moderately better than chance (AUC$= 66.49$)—which is a finding that is further reinforced by our sLDA model's superior F1 score and TPR values to those obtained under either $\xi = \frac{1}{2}$ or $\xi = \bar{y}$. As expected, $\xi = \bar{y}$ maximizes overall accuracy, with a value (64.34) that is superior to that of $\xi = \frac{1}{2}$ (50.06). However, the maximized accuracy obtained under $\xi = \bar{y}$ still falls slightly below that of our sLDA classifier (66.10), and comes at the cost of noticeably poorer TPR performance than either $\xi = \frac{1}{2}$ or sLDA, which as mentioned above, is valued more so than TNR in this application given our primary interest in *nonresponsiveness*.

Table 1: In-Sample Classification Statistics

	AUC	**TPR**	**TNR**	**F1score**	**Accuracy**
sLDA	66.49	52.54	70.18	41.77	66.10
$\xi = \frac{1}{2}$	50.04	50.02	50.07	31.67	50.06
$\xi = \bar{y}$	50.04	22.86	76.83	22.87	64.34

4.2 Out-of-Sample Results

We now turn to an evaluation of our sLDA model's out-of-sample classification properties. For this evaluation, we use our primary sLDA model to generate "denied request" predictions for the remaining 75% (i.e., $\approx 750,000$ documents) within our 2003-2015 request sample. Using these predictions, we then repeat the same steps as above in generating two random classifiers for comparison, $\xi = \frac{1}{2}$ and $\xi = \bar{y}$, and recalculate the previously described set of classification statistics, in Table 2.

Table 2: Out-of-Sample Classification Statistics

	AUC	TPR	TNR	F1score	Accuracy
sLDA	66.24	52.26	70.24	41.64	66.08
$\xi = \frac{1}{2}$	50.05	50.06	50.04	31.70	50.04
$\xi = \bar{y}$	50.05	22.95	77.10	23.07	64.56

Our out-of-sample results are largely consistent with our in-sample findings. As above, the sLDA model outperforms both random classifiers in AUC, TPR, F1 score, and overall accuracy, and performs second best (to $\xi = \bar{y}$) in TNR. The results reported in Table 2—across all classifiers—suggest that our out-of-sample sLDA predictions perform comparably to, albeit slightly worse than, our in-sample sLDA results. For example, our sLDA model accurately classifies 66.08% of all out-of-sample cases, whereas in the in-sample context our sLDA model's overall accuracy was 66.10%. Differences between these two sets of sLDA predictions are slightly larger when one examines AUCs (66.49 vs. 66.25), though these differences are again fairly negligible, especially relative to the effect of k on our AUCs in Figure 1.

Finally, though not reported here, we also compared these results to a "requests only" sLDA model that omits our *target agency* names as features, and found that the latter performs slightly worse than our full sLDA model. For example, the "requests only" model's out-of-sample AUC is 64.95, which is noticeably smaller than that of our primary sLDA model. Our remaining comparison metrics yielded similar conclusions: the addition of *target agency* names to our text features leads to a small but consistent improvements in accuracy.

4.3 Comparison to Alternate Approaches

We next compare our sLDA approach to three widely used alternatives: support vector machines (SVMs), logistic regression with Lasso, and random forests (RF). All three of these alternate approaches encountered computational difficulties when applied to our full training set of 250,000 documents, leading us to evaluate each of these classifiers, and sLDA, on a smaller training set ($n = 50,000$) and smaller test set ($n = 150,000$) of documents for the purposes of comparison. The results from this exercise appear in Table 3.

Table 3: Out-of-Sample Comparisons

	AUC	TPR	TNR	F1score	Accuracy
sLDA	65.84	50.28	70.99	40.71	66.21
SVM	65.27	26.54	88.53	32.21	74.23
Lasso	65.39	30.52	86.39	34.70	73.50
RF	70.23	48.52	78.82	44.29	71.83

In Table 3, sLDA performs slightly better than SVM and Lasso—but noticeably worse than RF—in terms of AUC. More generally, SVM and Lasso each appear to under-predict "denied requests," thereby ensuring that these two classifiers have higher TNR and higher overall accuracy than either sLDA or RF, albeit at the expense of worse performances on TPR and F1 score. While RF does exhibit a slightly worse TPR than sLDA, its higher F1 score, higher overall accuracy, and higher AUC suggest that RF outperforms sLDA along most dimensions of comparison, though, on the whole, both approaches (i.e., sLDA and RF) generally outperform Lasso and SVM in Table 3.

5 Conclusion

The content of Mexico's information requests, when modeled with sLDA, can help to predict government (non)responsiveness. Evidence from this exercise further suggests that politicization may increase nonresponsiveness. Future work should refine our approach so as to better accommodate (i) the imbalance in "denied request" outcomes, (ii) additional features (such as a requestor's home municipality), and (iii) the non-hierarchal structure of the Mexican information request data; while also better benchmarking our request text sLDA-classification results against alternative supervised machine learning techniques.

Acknowledgments

Almquist's research was supported in part by ARO YIP Award #W911NF-14-1-0577. He is currently a visiting scholar at the University of Washington.

References

Benjamin E. Bagozzi. 2015. The Multifaceted Nature of Global Climate Change Negotiations. *The Review of International Organizations*, 10(4): 439-464.

Benjamin E. Bagozzi and Philip A. Schrodt. 2012. The Dimensionality of Political News Reports. *Proceedings of the European Political Science Association Meetings*.

Daniel Berliner. 2014. The Political Origins of Transparency *The Journal of Politics*, 76(2): 479-491.

Daniel Berliner. 2016. Transnational advocacy and domestic law: International NGOs and the design of freedom of information laws *Review of International Organizations*, 11(1): 121-144.

Daniel Berliner and Aaron Erlich. 2015. Competing for Transparency: Political Competition and Institutional Reform in Mexican States *American Political Science Review*, 109(1): 110-128.

Daniel Berliner, Benjamin E. Bagozzi, and Brian Palmer-Rubin. 2016. What Information Do Citizens Want?: Evidence from One Million Information Requests in Mexico, Working Paper.

David M. Blei and Jon D. Mcauliffe. 2008. Supervised Topic Models *Advances in Neural Information Processing Systems*, 20:121-128.

Daniel M. Butler and David E. Broockman. 2011. Do Politicians Racially Descriminate Against Constituents? A Field Experiment on State Legislators *American Journal of Political Science*, 55(3):463-477.

Jonathan Chang. 2015. Package 'lda'. *https://cran.r-project.org/web/packages/lda/*.

Jonathan Fox, Libby Haight, and Brian Palmer-Rubin. 2011. Proporcionar transparencia ¿Hasta qué punto responde el gobierno mexicano a las solicitudes de información pública? *Gestión y Política Pública*, 20(1):3-61.

Miriam Golden and Brian Min. 2013. Distributive Politics Around the World *Annual Review of Political Science*, 16(1):73-99.

Gwyneth McClendon. 2016. Race and Responsiveness: A Field Experiment with South African Politicians *The Journal of Experimental Political Science*, 3(2).

Paul Lagunes. 2008. Irregular Transparency? An Experiment Involving Mexico's Freedom of Information Law, Working Paper.

Predicting political party affiliation from text

Felix Biessmann* **Pola Lehmann†** **Daniel Kirsch** **Sebastian Schelter ‡**

Abstract

Every day a large amount of text is produced during public discourse. Some of this text is produced by actors whose political colour is very obvious. However, though many actors cannot clearly be associated with a political party, their statements may be biased towards a specific party. Identifying such biases is crucial for political research as well as for media consumers, especially when analysing the influence of the media on political discourse and vice versa. In this study, we investigate the extent to which political party affiliation can be predicted from textual content. Results indicate that automated classification of political affiliation is possible with an accuracy better than chance, even across different text domains. We propose methods to better interpret these results, and find that features not related to political policies, such as speech sentiment, can be discriminative and thus exploited by text analysis models.

1 Introduction

Analysis and classifications of political text is and has been a very important tool to generate political science data [8]. Traditionally, experts conduct such classifications by reading and labelling the text of interest[1]. This is, however, a very time consuming task and thus sets various limits on the possible amount of data that a few experts can analyse. The growing field of automated text analysis, which allows for the analysis of much more text in less time, is therefore of great interest to political scientists. Additionally, automated text analyses allow for a more objective and replicable analysis of political text than human coders can achieve [9].

A major problem with automated text analyses is generalisation to text domains other than that on which the system has been trained [15]. While political experts can read texts from different domains and are able to detect political bias appearing in a variety of contexts and styles, machine learning algorithms are prone to poor performance generalisation across text domains if the training data is biased towards one domain only. Unfortunately, good unbiased training data is difficult to obtain. One of the best sources for automated political text analysis systems are plenary debates of the parliament: many studies are based on this type of data, as it consists of a large source of text that can be clearly associated with a party. We examine to what extent models trained on this data can generalise their predictions to other text domains, such as party manifestos and texts from social media. We discuss the effects of text length and domain shifts of text data, and investigate some possible reasons for the differences in classification performance.

We investigate the predictions of the models with three strategies: first, we test the influence of text length on the prediction accuracy. Second, we use sentiment analysis to investigate whether this aspect of language has discriminatory power. Third, univariate measures of correlation between text features and party affiliation allow us to relate the predictions to the kind of information that political experts use for interpreting texts.

In this article, section 2 gives an overview of the data acquisition and preprocessing methods, section 3 presents the model, training and evaluation procedures, in section 4 we discuss results and section 5 concludes with interpretations of the results.

*felix.biessmann@gmail.com
†pola.lehmann@wzb.eu
‡sebastian.schelter@tu-berlin.de

[1] See for example the Manifesto Project, the Comparative Agendas Project or Poltext.

Proceedings of the International Conference on the Advances in Computational Analysis of Political Text (PolText 2016), pages 14–19, Dubrovnik, Croatia, 14–16 July 2016.

2 Data Sets and Feature Extraction

We ran all experiments using publicly available data sets of German political texts, and applied standard libraries for processing the text. The following sections describe the details of data acquisition and feature extraction.

2.1 Data

Annotated political text data was obtained from three sources: a) the plenary debates held in the German parliament (*Bundestag*) b) all manifestos of parties winning seats in the election to the German parliament and c) facebook posts from all parties. The texts from plenary debates were used to train a classifier and evaluate it on this in-domain data. We employed the latter two data sources to test the generalisation performance of the classifier on out-of-domain data.

Parliament discussion data Parliament texts are annotated with the respective party label. The protocols of plenary debates are available through the website of the German Bundestag [3]; we leveraged an open source API to query the data in a cleaned and structured format [2]. Each uninterrupted part was treated as a separate speech.

Party manifesto data The party manifesto text originates from the Manifesto Corpus [12]. The data released in this project mainly comprises the complete manifestos of all parties that have won seats in a national election. Each statement or *quasi-sentence*[2] is annotated with one of 56 policy issue categories. Examples for the policy categories are *welfare state expansion, welfare state limitation, democracy, equality*; for a complete list and detailed explanations on how the annotators were instructed see [1]. Each quasi-sentence has two types of labels: the party affiliation and the manually assigned policy issue aimed at in each quasi-sentence. The length of each annotated statement in the party manifestos is rather short. The median length is 95 characters, or 12 words[3]. In order to increase the length of the texts for classification, we used the policy labels to aggregate the data into the following topics: *External Relations, Freedom and Democracy, Political System, Economy, Welfare and Quality of Life, Fabric of Society, Social Groups*. In this setting, each party had just one data point for each of the topics.

Facebook post data We crawled the facebook page of each party [4, 7, 5, 6] and extracted the post texts, excluding all comments and other information. Like the manifesto data, these texts are very short. As aggregation per topic was not possible for this data, we aggregated the texts by splitting all texts into parts of 1000 words.

2.2 Bag-of-Words Vectorisation

We tokenised all text data and transformed it into bag-of-word (BOW) vectors as implemented in scikit-learn [13]. Several options for BOW vectorisations were tried, including term-frequency-inverse-document-frequency normalisation, n-gram patterns up to size $n = 3$ and different cut-offs for discarding words which were too frequent or infrequent.

3 Classification Model and Training

We leveraged bag-of-words feature vectors to train a multinomial logistic regression model. Let $y \in \{1, 2, \ldots, K\}$ be the true party affiliation and $\mathbf{w}_1, \ldots, \mathbf{w}_K \in \mathbb{R}^d$ the weight vectors associated with the kth party. Then the party affiliation estimate is modelled as

$$p(y = k|\mathbf{x}) = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}} \text{ with } z_k = \mathbf{w}_k^\top \mathbf{x}. \quad (1)$$

3.1 Optimisation of Model Parameters

The model pipeline contained a number of hyperparameters that we optimised using gridsearch cross-validation. To this end, we split the parliament speech data into training and validation sets in a 90%/10% ratio; we trained the pipeline with each parameter setting on the training set and validated its performance on the validation set. We chose the parameters of the best performing model to train a model on the training and validation set data. None of the data in the separately held back in-domain test data nor the out-of-domain test data sets was used for this hyperparameter optimisation.

3.2 Sentiment analysis

We extracted sentiments via a publicly available key word list [14]. A sentiment vector $\mathbf{s} \in \mathbb{R}^d$ was constructed from the sentiment polarity values in the sentiment dictionary. We compute the

[2] A quasi-sentence has the length of an argument. It is never longer than one sentence.

[3] The longest statement is 522 characters (65 words) long, the 25%/50%/75% percentiles are 63/95/135 characters or 8/12/17 words, respectively.

sentiment index for attributing positive or negative sentiment to a text as the cosine similarity between BOW vectors and sentiment vector.

3.3 Interpreting bag-of-words models

Interpreting coefficients of linear models (independent of the regulariser used) implicitly assumes uncorrelated features; this assumption is violated by the text data used in this study. Thus direct interpretation of the model coefficients $\mathbf{w}_k$ is problematic, see also [17, 10]. In order to allow for better interpretation of the predictions and to assess which features are discriminative, we computed correlation coefficients between each word and the party affiliation label.

4 Results

The following section gives an overview of the results for all political bias prediction tasks. Predictions compared with the manifesto data were computed using models trained on texts from the 17th Bundestag, predictions obtained for facebook post texts were computed with models trained on the 18th Bundestag[4].

4.1 In-domain predictions

When predicting party affiliation on text data from the same domain that was used for training the model, average precision and recall values of above 0.6 are obtained. We list the evaluation results for the political party affiliation prediction on in-domain data (held-out parliamentary speech text) for the 17th Bundestag in Table 1. These results are comparable to those of [11] who report a classification accuracy of 0.61 on a five class problem predicting party affiliation in the European parliament.

4.2 Out-of-domain predictions

For out-of-domain data obtained from manifesto data, the models yield significantly lower precision and recall values between 0.3 and 0.4, see Table 2. We observe a similar effect for the facebook post data. The short texts resulted in poor prediction accuracies of 0.51 on average. Additionally, classes were highly unbalanced in this setting, since some parties have an order of magnitude more posts than others.

[4]We leveraged the speeches from the 17th legislative period for the first task as this legislature is already completed and offers more data. Results for the 18th Bundestag are similar but omitted for brevity. We employ the speeches of the 18th legislative period for the facebook posts as the posts were more recent.

Table 1: **In-domain classification performance** for data from the 17th legislative period on in-domain data. N denotes number of data points in the evaluation set.

	precision	recall	f1-score	N
cducsu	0.62	0.81	0.70	706
fdp	0.70	0.37	0.49	331
gruene	0.59	0.40	0.48	298
linke	0.71	0.61	0.65	338
spd	0.60	0.69	0.65	606
total	0.64	0.63	0.62	2279

Table 2: **Out-of-domain** classification performance (quasi-sentence level) on **manifesto data** of a classifier trained on speeches of the 17th legislative period of the Bundestag.

	prec.	recall	f1-score	N
cducsu	0.26	0.58	0.36	2030
fdp	0.38	0.28	0.33	2319
gruene	0.47	0.20	0.28	3747
linke	0.30	0.47	0.37	1701
spd	0.26	0.16	0.20	2278
total	0.35	0.31	0.30	12075

4.3 Influence of text length on accuracy

A key factor that made the prediction in the out-of-domain prediction task particularly difficult was the short length of the texts to classify, see also section 2. In order to investigate the effect of text length, we aggregated the data into longer texts, and grouped manifesto data into political topics. Table 3 shows the topic level prediction results. We obtain F1 scores of above 0.8 for all parties except for the SPD . As the facebook posts lacked topic labels, we conducted the aggregation of these texts by first concatenating all facebook posts of a party into one long text; this text was then partitioned into segments of 1000 words each. For each party 50 random segments were selected for classification. The results are shown in Table 4. Prediction accuracies comparable to the in-domain case can also be achieved for these texts. This increase is in line with previous findings on the influence of text length on political bias prediction accuracy [11].

Table 3: **Out-of-domain** classification performance (topic level) on **manifesto data**. Compared to quasi-sentence level predictions (Table 2), the predictions made on the topic level are more reliable.

	precision	recall	f1-score	N
cducsu	0.64	1.00	0.78	7
fdp	1.00	1.00	1.00	7
gruene	1.00	0.86	0.92	7
linke	1.00	1.00	1.00	7
spd	0.80	0.50	0.62	8
total	0.88	0.86	0.86	36

Table 4: **Out-of-domain** classification performance on 50 randomly selected **facebook posts** of respective party (text length: 1000 words). The average prediction performance is comparable to that on in-domain test data.

	precision	recall	f1-score	N
cducsu	0.65	1.00	0.79	50
gruene	0.67	0.12	0.20	50
linke	0.60	0.82	0.69	50
spd	1.00	0.92	0.96	50
avg / total	0.73	0.71	0.66	200

4.4 Misclassification and policy change

Automatic political text analysis requires a profound understanding of the models used. One way to better understand these models is to inspect the misclassifications of a model. A potential explanation for the misclassifications could be that parties change their policy positions over time. The confusion matrix for the 17th Bundestag in Table 5 shows that the SPD manifesto texts are often predicted as belonging to the CDU/CSU on the topic level. This was the the legislative period when the CDU under chancellor Merkel was making a strong left move with respect to socioeconomic issues.

4.5 Predicting government status

We also trained a model on government membership labels, in order to a better compare against other studies that predict party affiliation in a two party system. Table 6 shows the results for the 17th legislative period. While the in-domain prediction accuracy is close to 0.9, the out-of-domain evaluation on manifesto data drops again to a performance close to chance. This is in line with results on binary classification of political bias in the Canadian parliament [16]. The authors report classification accuracies between 0.80 and 0.87, and find a pronounced drop in performance on texts from a different domain (e.g. older texts or texts from another chamber). In our results, the aggregation into topics did not increase the accuracy in this binary setting when classifying manifesto texts. The drop in accuracy of the binary classifier on facebook data (aggregated analogous to the party affiliation case) was less pronounced: accuracies were above 0.70.

Table 5: **Topic level confusion matrices** of manifesto texts.

		Predicted				
		cducsu	fdp	gruene	linke	spd
True	cducsu	7	0	0	0	0
	fdp	0	7	0	0	0
	gruene	0	0	6	0	1
	linke	0	0	0	7	0
	spd	4	0	0	0	4

4.6 Discriminative features

Another important question when analysing automatic text classification models is whether the difference between the features of each party stems from different policies or from other aspects of the text. To address this point we analysed features that are discriminative for government membership and for parties.

Sentiment correlates with political power The drop in prediction accuracy in the government prediction task was more pronounced for manifesto texts than for facebook posts. What do facebook posts and plenary debates have in common? In contrast with the authors of manifestos, both the speakers in the parliament as well as the authors of facebook posts know which party is in government. A language feature that might capture this is sentiment. Indeed our results in Table 7 show that positive sentiment strongly correlates with government membership and the number of seats in the parliament. Previous studies also find that text features which are discriminative in a two party system are not necessarily related to policies but more to language of defence and attack [11].

Table 6: Classification accuracy on the binary prediction problem, categorising texts into government and opposition. Out-of-domain accuracy again drops close to chance performance for the manifesto data but remains higher for the facebook post texts.

	In-Domain	**Out-of-Domain**	
	Parliament	Manifestos	Facebook Posts
Accuracy	0.88	0.60	0.76

Table 7: Correlation coefficient between the average sentiment of political speeches of a party in the German Bundestag with two indicators of political power: a) membership in the government and b) the number of seats a party occupies in the parliament.

Sentiment vs.	Gov. Member	Seats
17th Bundestag	0.84	0.70
18th Bundestag	0.98	0.89

Correlations between words and parties In order to determine further discriminative features, we quantified which words were preferentially used by each party by measuring the correlation of single words with the party label. Unspecific stopwords were excluded. We find clear differences between the parties, which are in line with the parties ideologies.

Left party (linke) Frequent words include referrals to big companies (*konzerne*) and their profits (*profite*), the working class *beschaeftigte*, the social welfare program *hartz iv* as well as war (*krieg*).

Green party (gruene) Uses words related to environmental damage (*klimaschaedlichen*), exploited low wage employees (*leiharbeitskraefte*) and pensions (*garantierente*).

Social Democratic Party (SPD) Uses mostly unspecific words related to the parliament and governmental processes (*staatssekretaerin, kanzlerin, bundestagsfraktion*) and some words related to cutting of expenses (*kuerzungen*).

Christian Democratic Union/Christian Social Union (CDU/CSU) Often used words relate to a pro-economy attitude, such as competitiveness or (economic) development (*wettbewerbsfaehigkeit, entwicklung*) and words related to security (*sicherheit, stabilitaet*).

5 Conclusions and Limitations

We find that automated political bias prediction is possible with an accuracy better than chance, even beyond the training text domain. These results suggest that such systems could be helpful as assistive technology, for example for human annotators in an active learning setting.

In line with previous findings [16, 11], we find a large effect of text length and text domain on the generalisation performance of the classifier. The first effect, that longer texts are easier to classify, intuitively makes sense. Also humans are challenged when judging the political bias of shorter texts out of context [8]. However, short texts are a realistic challenge for automated political bias prediction systems: political texts from social media data and other web sources are often very short and hence difficult to analyse for both human annotators and algorithms. Both political education and science can benefit from automatic analyses of these very data streams, as these fields have a strong influence on public opinion and yet cannot be analysed by humans alone, due to the volume of data.

The second effect, i.e. the drop in generalisation performance on out-of-domain data, appears to be correlated to the first one: it can be alleviated in some cases by aggregating texts into longer segments. In the case of party affiliation prediction, the out-of-domain classification is on a par or even better than the prediction accuracy on in-domain data. However in the binary classification setting (government membership prediction), text aggregation does not help as much: aggregating manifesto data, written without the knowledge of which party would be member of the government, into longer texts does not counteract the effect of out-of-domain accuracy drop. We attribute this effect in part to the fact that sentiment appears to be a discriminative feature for government membership.

Acknowledgements

We would like to thank Friedrich Lindenberg for factoring out the https://github.com/bundestag/plpr-scraper from his Bundestag project. Michael Gaebler provided helpful feedback on an earlier version of the manuscript.

References

[1] Manifesto codebook https://manifesto-project.wzb.eu/information/documents?name=handbook_v4.

[2] https://github.com/bundestag.

[3] https://www.bundestag.de/protokolle.

[4] https://www.facebook.com/B90DieGruenen/.

[5] https://www.facebook.com/CDU/.

[6] https://www.facebook.com/linkspartei/.

[7] https://www.facebook.com/SPD/.

[8] Kenneth Benoit, Drew Conway, Benjamin E. Lauderdale, Michael Laver, and Slava Mikhaylov. Crowd-sourced text analysis: reproducible and agile production of political data. *American Political Science Rewview*, Forthcoming.

[9] Kenneth Benoit, Michael Laver, and Slava Mikhaylov. Treating words as data with error: Uncertainty in text statements of policy positions. *American Journal of Political Science*, 53:495–513, 2.

[10] Stefan Haufe, Frank Meinecke, Kai Görgen, Sven Dähne, John-Dylan Haynes, Benjamin Blankertz, and Felix Biessmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *Neuroimage*, 87:96–110, 2014.

[11] Graeme Hirst, Yaroslav Riabinin, Jory Graham, and Magali Boizot-Roche. Text to ideology or text to party status? In Isa Maks Bertie Kaal and Annemarie van Elfrinkhof, editors, *From Text to Political Positions: Text analysis across disciplines*, pages 47–70, 2014.

[12] Pola Lehmann, Theres Matthieß, Nicolas Merz, Sven Regel, and Annika Werner. *Manifesto Corpus*. WZB Berlin Social Science Center., https://manifestoproject.wzb.eu/information/documents/corpus, 2015.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[14] R. Remus, U. Quasthoff, and G. Heyer. Sentiws – a publicly available german-language resource for sentiment analysis. In *Proceedings of the 7th International Language Resources and Evaluation (LREC'10)*, 2010.

[15] Jonathan B. Slapin and Sven-Oliver Proksch. Word as data: Content analysis in legislative studies. In Shane Martin, Thomas Saalfeld, and Kaare W. Strøm, editors, *The Oxford Handbook of Legislative Studies*, pages 126–144. Oxford University Press, Oxford, 2014.

[16] Bei Yu, Stefan Kaufmann, and Daniel Diermeier. Classifying party affiliation from political speech. *Journal of Information Technology & Politics*, 5(1):33–48, 2008.

[17] Alexander Zien, Nicole Krämer, Sören Sonnenburg, and Gunnar Rätsch. The feature importance ranking measure. In *ECML/PKDD*, pages 694–709, 2009.

The Meaning of Democracy

Using Distributional Semantics to Account for Meaning Differences

Stefan Dahlberg, Sofia Axelsson University of Gothenburg

Magnus Sahlgren, Amaru Cuba Gyllensten, Gavagai, Swedish Institute of Computer Science.

Introduction

International survey research on democracy has made significant efforts to map popular support for democracy across the world. However, the concept of democracy can mean different things; it can refer to an ideal, to a political procedure, or to a set of political outcomes. When collecting survey data regarding the level of support for the concept of democracy, we do not know which of these meanings the support refers to. Perhaps differences in survey results are influenced by differences in the meaning of the concept democracy? While some scholars emphasize the procedural and institutional aspects that need to be present in a democracy, most theoretical definitions of democracy also include references to the ideals and values associated with democracy. The literature on public support for democracy has revealed significant cross-country differences in people's attitudes towards democracy. The variance is partly due to differences between high "diffuse" support for the principles of democracy, which can be found in Western, consolidated democracies, and "specific" support for the performance of democracies, which is more prevalent in new democracies (see Easton 1975; Norris 1999; Linde & Ekman 2003; Dahlberg & Holmberg 2012).

Cross-cultural survey research rests upon the assumption that if survey features are kept constant to the maximum extent, data will remain comparable across languages, cultures and countries (Diamond 2010; Lolle & Goul Andersen, 2015). Yet translating concepts across languages, cultures and political contexts is complicated by linguistic, cultural, normative or institutional discrepancies. Further, even if it is possible to unambiguously translate lexical items across languages, there may be semantic differences between various languages and various cultures in how these lexical items are used. Recognizing that language, culture and other social and political aspects affect survey results has been equated with giving up on comparative research, and consequently, the most commonly used "solution" to equivalence problems has been for researchers to simply ignore the issue of comparability across languages, cultures and countries (King et al 2004; Hoffmeyer-Zlotnik & Harkness 2005).

This paper presents our first steps towards using distributional semantics to account for semantic differences between lexical realizations of concepts across languages. Distributional semantics is a statistical approach for quantifying semantic similarities based on co-occurrence information collected from large text data (Turney & Pantel 2010). In this experiment, we have used data collected from online sources such as blogs, forums, news sites, and webpages. The reason for using such data rather than balanced corpora is that it enables us to analyze word meanings in normal, uncontrolled, unsolicited, and contemporary language use. Compared to other methodological approaches aimed at identifying and measuring cross-cultural discrepancies, this approach has the advantage of enabling us to analyze how concepts are used in their "natural habitat" (Wittgenstein 1958). Our ambition is that using distributional semantics applied to such data will enable us to uncover potential meaning differences in the use of

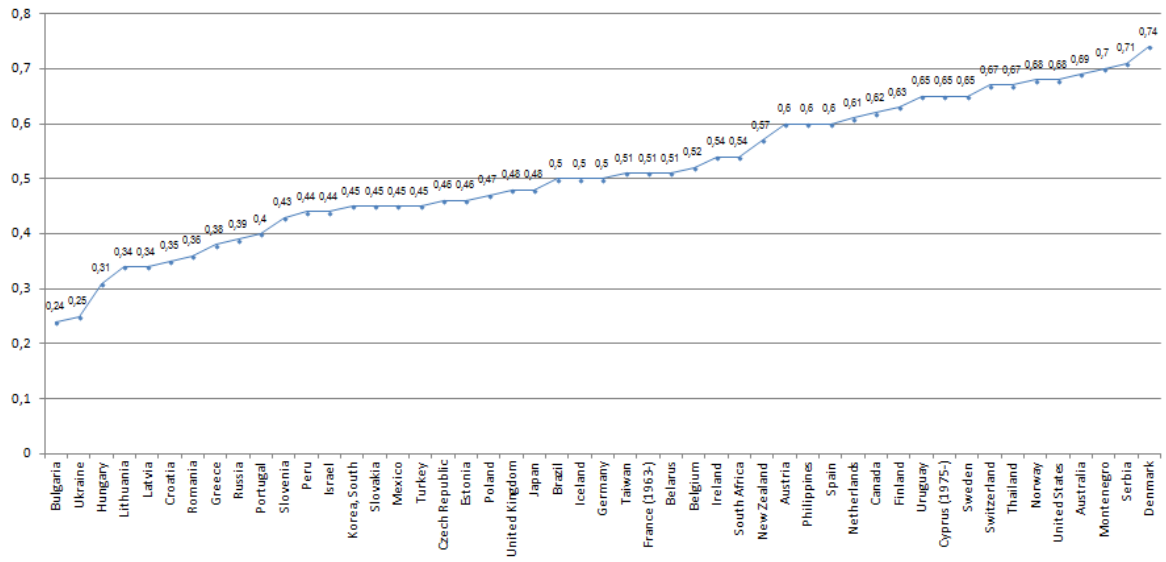
Proceedings of the International Conference on the Advances in Computational Analysis of Political Text (PolText 2016), pages 20–30, Dubrovnik, Croatia, 14–16 July 2016.

terms across languages. For this study we make use of distributional semantics through Gavagai Living Lexicon - the largest distributional thesaurus online - to obtain a word-space, which is a set of terms semantically similar to the term *democracy* across 13 European languages including Russian. In a next step we apply a manual classification of word-space terms into a set of five broad categories, pertaining to democracy at different levels of abstraction. Doing so, we take a step towards the inclusion of new variables, accounting for differences in meaning across languages, into existing survey data-sets and thereby maximizing comparability across contexts.

Satisfaction with the way democracy works

There is a rich literature on both within an between country factors that affects citizens' satisfaction with the way democracy works (for an overview, see Cutler et. al. 2013). Still there is a lot of variation left to explain. However, a crucial point regarding our attempts to gain new knowledge in this subject relates to the question about *what citizens actually are expressing their support for?*

Figure 1 Satisfaction with the way Democracy works across 49 countries



Comment: The aggregated measures of citizen's satisfaction with the way democracy works (SWoD) are based on data from two different data sources. The Comparative Studies of Electoral Systems (CSES) module 3 and 4 (2006-2016) and the European Social Survey (ESS) wave 3 (2008). In both surveys the question is reading: On the whole, how satisfied are you with the way democracy works in [country]? In contrast to the CSES questionnaire (where the response options are 1-not at all satisfied to 4-very satisfied), the ESS response options are based on an 11 point scale, stretching from 0 (extremely dissatisfied) to 10 (extremely satisfied) (for more information, see: www.europeansocialsurvey.org/data). Differences in scale and time are not optimal for comparisons. However, For 23 countries, data were overlapping between CSES and EES and the correlation between the two survey measures were r=0.81, which makes them not identical but at least very close. Based on the this correlation we have combined them into one dataset where country averages were rescaled into 0-1 with high values indicating satisfaction.

Some of the efforts to map people's conceptions of democracy across the world can be found within the literature on political support. Public support is crucial for the legitimacy of a democratic regime, yet citizens can be critical of the incumbent democratic regime or be dissatisfied with certain political institutions while still support democracy as the ideal form of government. One way to conceptualize the different levels of political support has been provided by Easton (1975). The Estonian model differentiates between "diffuse support" for the political community and for democratic principles on the one hand, and "specific support" for the regime structure and political authorities on the other. The level of specific support is contingent upon the behavior of, and outcomes delivered by, authorities in relation to citizens' expectations of authorities' performance. Diffuse support captures "attachment to the political object for its own sake" (1975:445) and is generally associated with higher levels of popular support for democracy; it is accumulated through over-time socialization that gradually transforms into generalized attitudes towards political objects. In this sense, it is also contingent upon a history of specific support, in turn generated by a regime's capacity to deliver order, protect human rights and uphold the rule of law, and generate economic development.

In *Critical Citizens*, Pippa Norris (1999 Ed.) builds upon Easton's definition and develops a five-level model for political support that includes support for the political community, regime principles, regime performance, regime institutions and political leaders. The different types of support are ordered along a continuum, ranging from diffuse support for the national community to specific support for political actors. Building upon those dimensions, the authors of *Critical Citizens* conclude that citizens in advanced industrial democratic societies are becoming increasingly sceptical towards political parties, parliaments and governments and their performance; yet popular support for democratic ideals, values and principles - part of what Easton conceived as diffuse support - remain high and widespread.

A study by Holmberg (2012) has demonstrated that public support for democracy tends to be lower in new democracies (see also Aarts & Thomassen 2008), and citizens in new democracies tend to base their evaluations of democracy more on regime performance and economic outcomes than on conceptions of abstract democratic ideals (Bratton & Mattes 2001). In another article, Dahlberg, Linde & Holmberg (2015), shows that individual level determinants of support for democracy are interacting with institutional consolidation. In more newly democratized countries, perceptions of government performance and economical outcomes are more important for expressing support for democracy; while assessments of representation and procedures are more important in established democracies.

In Search for the Meaning of Democracy

The different methods available for studying how the meaning of democracy changes with the linguistic, cultural and political context can be summarized in two different approaches: one explorative approach, which allows respondents to describe what democracy means to them, and can be carried out either through surveys by utilizing open-ended questions (Dalton, Shin & Jou 2007) or using ethnographic methods (Schaffer 2000). The other approach is to use closed-ended questions in surveys and ask respondents to rate the relative importance of different democratic properties and then deduce their understanding of democracy from these results (Bratton 2010). The different approaches have their advantages, but also limitations and caveats; survey

research using different batteries of closed-ended questions allows for global comparisons, but existing survey items suffer from validity issues as it has proved difficult to establish if democracy means the same to people across linguistically, culturally and socio-politically different societies. Ethnographic studies, in contrast, allow for "thick description" and enhance our understanding of what democracy means for people in ordinary social, cultural and political context. This method also captures both political and non-political uses of democracy, which can be used as an indicator of to what extent the concept is anchored in society. However, the ethnographic method is by default limited in its scope, which many would argue undermines cross-country comparisons. The method used in this paper combines the explorative approach of ethnographic methods with the systematic analysis used in survey research. It offers a solution both to the issue of validity and cross-cultural generalizations. Our approach enables us to analyze the use of the concept in its "natural habitat" (Wittgenstein 1958).

Distributional semantics as method

Distributional semantic models collect co-occurrence statistics from text data in order to build high dimensional vector representations of terms where similarity between vectors indicates similarity of usage. The method is motivated by a structuralist meaning theory known as the distributional hypothesis, which states that words with similar meanings tend to occur in similar contexts, and that the contexts shape and define the meanings of the words (Sahlgren 2006). According to the hypothesis, if we observe two words that constantly occur in the same contexts, we are justified in assuming that they mean similar things. Distributional semantic models can thus be used to find semantically similar terms to a given target term - in effect, a distributional semantic model constitutes a statistically compiled lexicon. As an example, a distributional semantic model would likely return terms like "green", "yellow", "black", and "white" when probed with the term "red". In linguistic terms, this constitutes a *paradigm*, in which the members can often be substituted by each other in context.

We have used an online distributional semantic thesaurus - the Gavagai Living Lexicon (Sahlgren et al, 2016) - that continuously learns a distributional semantic model from online big data. To the best of our knowledge, this is the currently largest purely distributional multilingual thesaurus that updates its representations continuously with new data. As such, the resource enables us to analyze and compare the current usages of concepts in different languages. The specific distributional semantic model that is used in this study is called Random Indexing (Sahlgren 2016), a count-based method geared towards vast amounts of data that alleviates space usage problems by means of a random projection. The reason behind this is practical rather than theoretical, namely that Gavagai offers Random Indexing as a service.

Table 1 Language data

Language	Approximate amount of documented data per day
Danish (DA)	12 000
German (DE)	340 000
English (EN)	2 000 000
Spanish (ES)	320 000
Finnish (FI)	13 000
French (FR)	330 000
Hungarian (HU)	22 000
Italian (IT)	122 000
Dutch (NL)	88 000
Norwegian (NO)	12 000
Portuguese (PT)	170 000
Russian (RU	363 000
Swedish (SV)	150 000

The Gavagai Living Lexicon uses data from a range of different sources, including news and social media that are open to the public. The data is retrieved from a number of different commercial data providers, such as Trendiction[1], Twingly[2], and Gnip[3]. The data flow contains millions of documents each day; at peak periods, the flow can reach millions of documents each day, which amounts to more than a

[1] trendiction.com
[2] twingley.com
[3] gnip.com

billion terms each day. Table 1 shows the languages analyzed in this paper as well as the approximate amount of daily documents. While the Gavagai Living Lexicon currently features 20 different languages, we have chosen to include 13 on the basis of suitable data sample size. The amount of data differs considerably between languages: English is by far the largest language, followed by Russian, German and French.

For the search term democracy (taken from World Values Survey questionnaires) in indefinite form, we collected 15 semantically similar terms for each language in March 2016, by using an application, Postman[4], to call the Living Lexicon API. The strength of the relationship between the search term – democracy – its semantic neighbours is measured by a cosine value ranging from 0 to 1 where 1 is perfect semantic similarity. The 15 items collected across all 13 languages thus represent the rank order by semantic strength; the 15 most similar terms to the term democracy. Given that some languages contain larger amount of data than others, it is expected that some languages with less data - or languages newly incorporated into the Living Lexicon - will contain more noise than others.

Thematic classifications of distributional thesaurus items across languages

The theoretical attempts to portray people's conceptions of democracy, as laid out by Easton and Norris, have gained some validity in a number of empirical correlational studies (see fc. Bratton & Mattes 2001; Aarts & Thomassen 2008; Holmberg 2012; Dahlberg, Linde & Holmberg 2014).

For the empirical analysis, we have constructed a classification scheme that, drawing primarily on Norris (1999), is based on the separation between diffuse versus specific support for democracy. From this distinction follows that the separation not only is a matter of different levels of abstraction but also a difference in terms of input and output of the democratic system. If we are able to conceptualize language use for the term democracy into a smaller set of theoretically meaningful categories for different languages; we will also be able to incorporate the proportions of stances for each language within each category back to the survey-based data. These language-based variable constructs can then be used to correct for differences in meaning of the word democracy across languages.

Figure 2 Classification categories

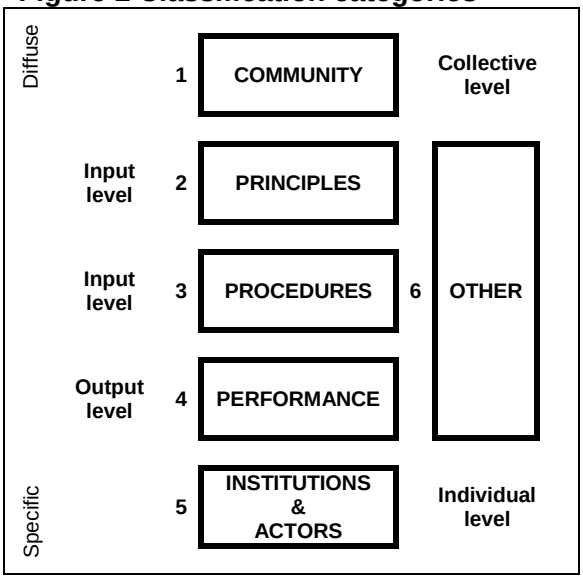


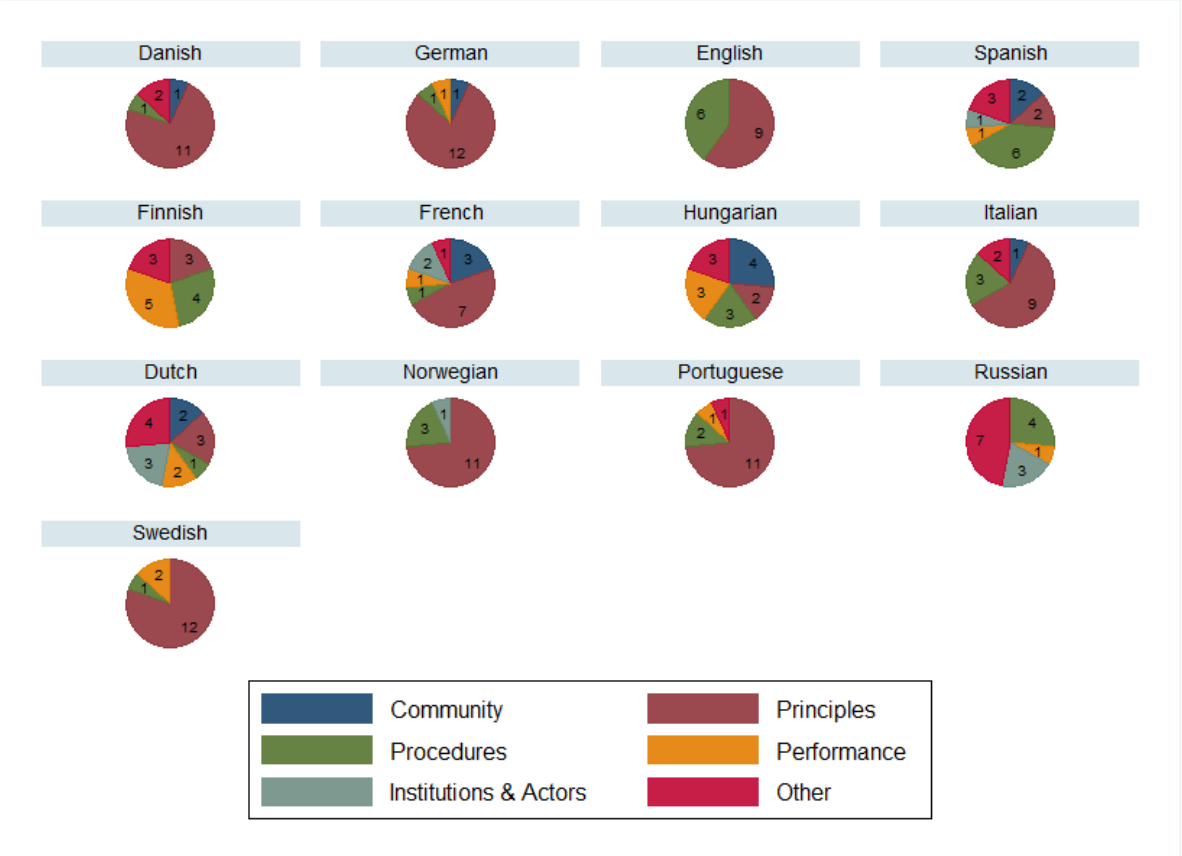
Figure 2 displays the classification scheme with six categories ranging from a more diffuse to a more specific level of abstraction. On the input side we find democracy in terms of principles – values the political system associated with – (category 2) and procedures - around which the political community or system is organized (category 3). On the output side we find democracy in terms of performance – outcomes of the political system for instance properties associated with economic development or the welfare state (category 4). Community (category 1) and institutions and actors (category 5) are neither input nor output categories; while the former denotes the political community – or the collective society in which the political system is situated – at a more abstract level, the latter refer to institutions or individual actors of the

[4] getpostman.com

political system. In addition, we have included a separate category (6) for items not corresponding to any of the previous categories, for instance items that are simply noise.

Knowing that automatic machine learned translators cannot always guarantee the interpretive sophistication required for studies of this kind, translators are employed for each language to assist with the categorisation process. In view of the language agnostic approach of this paper, this is somewhat methodologically fragile given that human translators inevitably introduce bias to the material. However, the translation was conducted in a supervised environment, where the translators were tasked with not merely providing translation suggestions of the items analyzed but of describing the items, using an official dictionary to capture the lexical meaning of the items derived from the Gavagai Living Lexicon. In addition, they were assigned to describe their views of the items in terms of "local semantics" (see Levisen 2014); how people within the specific language context generally comprehend of and make use of the items. While we know that some languages are more linguistically related than others - making direct translation of certain words easier - there might still be meaning discrepancies in the everyday word usage that would not be captured by an automatic translator. All in all, two translators were employed for each language translated so as to enhance the reliability of the translation process. Having translated all 15 items retrieved from the Gavagai Living Lexicon, classification of the items were subsequently conducted manually in a combined deductive-inductive manner using the classification scheme presented above (for an overview of the translated and classified items, see table A1).

Figure 3 Classification of terms across 13 languages



Findings

Based on the outcome from the word-spaces of our 13 languages there are several implications. First of all, the Gavagai Living Lexicon appears to perform rather well; there is some noise in the word-spaces, but not very much. Noise is here understood as random terms not related the word democracy in any sensible meaning. Secondly, using a classification scheme inspired by a combination of Easton (1975) and Norris (1999) appears to be both comprehensive as well as rather complete. For most languages not many terms went into the "other" category (see figure 2 and table A1). An exception is Russian where 47 percent of the terms did not fit in any of the categories. As for the remaining 12 languages it is evident that particularly for the north-western European languages, the term democracy is mainly spoken about in terms of principles (category 2) of democracy. English is an exception to this trend with procedures (category 3) being almost as common as principles. However, talking about democracy in terms of principles is also common in Italian and Portuguese. In Finnish, the semantically similar terms relate to several categories; nonetheless, talking about democracy in terms of performance is more common compared to the other covered languages.

Concludingly the analyses show that talking about democracy in terms of principles is, on average, the most common understanding of the democracy concept. However, we also find important variations across languages.

From the survey research on support for democracy we have, as mentioned, seen indications that performance measures are more highly correlated with support for democracy in Eastern Europe, while principles are more strongly correlated in the North-West of Europe. Our analysis on language data lends support to these patterns to some extent. We find that the term democracy is mainly spoken about in terms of principles in Western Europe. For Eastern Europe, we do not have enough languages to cover for the data; although, we find a signal in the Finnish language use that supports the performance aspect.

Notably, 13 languages is not a sufficient number to be able to make a contribution to the survey-based research field on support for democracy; we obviously need to increase the number of languages included in the distributional analysis. Another related problem is that there may not be much data available for all languages of interest.

Our current modus operandi is to collect data from established data providers, which on the one hand saves us the trouble and cost of having to crawl data ourselves, but on the other hand makes us completely dependent on the coverage and availability of data from these providers. For a lot of the languages represented in for example the World Values Survey, there is very little data available from the providers we currently use. This problem is further aggravated by our need to analyze specifically the use of the term *democracy*, which in most cases is not a very frequently occurring term. Unfortunately, this means that we need a considerable amount of data for each language to be able to perform the type of analysis suggested in this paper; scarce availability of data makes it ostensibly difficult to apply techniques that require a sound statistical foundation.

The Gavagai Living Lexicon uses a computing framework that emphasizes scalability and efficiency; hence, it is primarily design for big data environments. As we have discussed in this section, the current type of analysis only occasionally deal with such amounts of data; the typical case is rather that data is very small, which means a different type of distributional model will likely be preferable to use.

One of the main issues with the proposed approach is that we need representations of normal language use in order to uncover how the term democracy is being used by language users. In this respect, we cannot use data from editorial sources such as Wikipedia, or balanced corpora like BNC. Given that we are interested in the ways in which the concept of democracy are used in everyday language, web data seems like

the most viable option, even considering the problems with using existing data providers. More fine-tuned control of data and method is thus required for analyses of this kind. Separating data by country of origin (geo-tagging) and source (news media versus social media) is currently in the works, although issues of comparability between data sources still remain. We are also currently investigating whether other distributional semantic models are better suited for our resources and ambitions, since Random Indexing is geared towards quantities of data that are orders of magnitudes larger than what we have access to.

When these obstacles are solved, our final goal is to include the thematic categories into aggregated survey data. Ideally, our ambition is to be able to classify the meaning of democracy across as many languages as possible. The outcome from this thematic classification will then be incorporated as variables into an aggregated dataset with countries as units of analysis. Doing so, we will be able to account for the extent to which differences in support for democracy across countries are driven by differences in meaning of the targeted term democracy. For example, if we are able to find that the word democracy mainly is spoken about in terms of performance and outcome in East-European countries, then we would also expect that the levels of satisfaction will be more sensitive towards fluctuations and changes in GDP, GDP-growth, inflation or unemployment rates within these contexts; but this is of course an hypothetical empirical question that remains to be answered.

References

Aarts, Kees & Jacques Thomassen. 2008. "Satisfaction with Democracy: Do Institutions Matter?" *Electoral Studies,* 27: 5-18.

Bratton, Michael. 2009. "Democratic Attitudes and Political Participation: An Exploratory Comparison across World Regions." Paper prepared for the *Congress of the International Political Science Association*, Santiago, Chile, July.

Bratton, Michael. 2010. "Anchoring the 'D-Word' in Africa", *Journal of Democracy*, 21(4): 106-113.

Comparative Study of Electoral Systems (www.cses.org) 2015 CSES Modules 3 & 4. Full releases. June 22, 2016 version. doi: 10.7804/cses.module4.2016-06-22.

Cutler, Fred, Andrea Nuesser & Ben Nyblade 2013 "Evaluating the Quality of Democracy with Individual Level models of Satisfaction: Or, a complete model of satisfaction with democracy." Paper presented at the *ECPR General Conference*, Bordeaux, 4–7 September 2013.

Dahlberg, Stefan, Jonas Linde & Sören Holmberg. 2014. "Democratic discontent in old and new democracies - Assessing the importance of democratic input and governmental output." *Political Studies*, 63: 18-37.

Dalton, Russell J., Doh-Chull Shin & Willy Jou. 2007. "Understanding Democracy: Data from Unlikely Places." *Journal of Democracy*, 18(4): 142-256.

Diamond, Larry. 2010. "The Meanings of Democracy." *Journal of Democracy*, 21(4):102-105.

Easton, David. 1975. "A Re-Assessment of the Concept of Political Support." *British Journal of Political Science,* 5(4): 435-457.

European Social Survey, (2008). *ESS Round 4 Source Questionnaire.* London: Centre for Comparative Social Surveys, City University London.

Hoffmeyer-Zlotnok, Jürgen H. P. & Janet Harkness. (Eds.) 2005. *Methodological Aspects in Cross-National Research.* Mannheim: ZUMA.

Holmberg, Sören. 2014. "Feeling Policy Represented" in Thomassen, Jacques (Ed.),
Elections and Democracy: Representation and Accountability. Oxford: Oxford University Press.

King, Gary, Christopher J. L. Murray, Joshua A. Salomon and Ajay Tandon. 2004. "Enhancing the Validity and Cross-Cultural Comparability of Measurement in

Survey Research." *American Political Science Review,* 98(1): 191-207.

Linde, Jonas & Joakim Ekman. 2003. "Satisfaction with Democracy: A Note on a Frequently Used Indicator in Comparative Politics." *European Journal of Political Science Research,* 42(3): 391-408.

Levisen, Carsten. 2014. "The story of "Danish Happiness": Global discourse and local semantics." *International Journal of Language and Culture*, 1(2): 174-193.

Lolle, Henrik and J ørgen Goul Andersen. 2015. "Measuring Happiness and Overall Life Satisfaction: A Danish Survey Experiment on the Impact of Language and Translation Problems", *Journal of Happiness Studies* 6: 1-14.

Norris, Pippa. (Ed.) 1999. *Critical citizens: Global support for democratic governance*. Oxford: Oxford University Press.

Sahlgren, Magnus. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Stockholm: Department of Linguistics, Stockholm University.

Sahlgren, Magnus. 2008. "The Distributional Hypothesis." *Rivista di Linguistica*, 20(1): 33-53.

Sahlgren, Magnus Amaru Cuba Gyllensten, Fredrik Espinoza, Ola Hamfors, Jussi Karlgren, Fredrik Olsson, Per Persson, Akshay Viswanathan & Anders Holst. 2016. "The Gavagai Living Lexicon". *Language Resources and Evaluation Conference (LREC)*, 23-28 May 2016, Portorož (Slovenia), ELRA, 2016.

Schaffer, Frederic Charles. 2000. *Democracy in Transition: Understanding Politics in an Unfamiliar Culture*. New York: Cornell University Press.

Wittgenstein, Ludwig, 1958. *Philosophical Investigations*. Blackwell: Oxford.

Abstract

Table A1 Translated semantically similar terms across 13 European languages

Language	Word-space term (descending cosine ranking)	Classification number
	Danish	
DA	religious freedom	2
DA	tolerance	2
DA	justice, fairness	2
DA	broad-mindedness, open-mindedness	2
DA	a debate, debate arrangement	3
DA	equality, sameness	2
DA	solidarity	2
DA	blasphemy	2
DA	lads, jacks	6
DA	dignity	2
DA	brotherhood	1
DA	symbols	6
DA	equal worth, equal worthiness	2
DA	freedom	2
DA	peace	2
	German	
DE	humanity, humanness	2
DE	justice, fairness, equity	2
DE	freedom, liberty	2
DE	rule of law, constitutional legality, state of law	3
DE	mutual understanding between people/peoples	2
DE	tolerance	2
DE	human dignity	2
DE	humanity	2
DE	openness; open-mindedness	2
DE	humanity, good neighbourliness, compassion with fellow humans	2
DE	equality; equity, sameness	2
DE	welcome culture, welcoming culture	4
DE	peace	2
DE	solidarity	2
DE	neighbourhood; good neighbourliness, solidarity	1
	English	
EN	rule of law	3
EN	good governance	3
EN	human dignity	2
EN	constitutional democracy	3
EN	democratic governance	3
EN	democratic principles	2
EN	constitutionalism	3
EN	liberal democracy	3
EN	self-determination	2
EN	individual choice	2
EN	self determination	2
EN	secularism	2
EN	social justice	2
EN	freedom of speech	2
EN	freedom of expression	2
	Spanish	
ES	institutionality, institutionalism	3
ES	representative democracy	3
ES	intellectuality	2
ES	historiography	6
ES	bourgeoisie	5
ES	internal democracy	3
ES	idiosyncrasy	6
ES	diaspora	1
ES	sovereignty	2
ES	stamp	6
ES	diplomatic mission	3
ES	diplomacy	3
ES	democratic governance	3
ES	social fabric	1
ES	oil industry	4
	Finnish	
FI	system	3
FI	freedom of speech	2
FI	method, process, system, technique	3
FI	strategy, plan of action	4
FI	device, appliance, gadget	6
FI	system	3
FI	service	4
FI	equality	2
FI	intellectuality, education, civilization	2
FI	technique, technology, engineering, electronics, method	4
FI	concept, rough-copy, first draft	6
FI	operating model, concept, procedure, behaving pattern	3
FI	research group	6
FI	market economy	4
FI	public transportation	4
	French	
FR	doctrine, ideology, dogma, principles	2
FR	civilization	1
FR	political class	1
FR	cohesion, solidarity, union	2
FR	music scene	6
FR	intelligentsia	5
FR	secularity, secularism	2
FR	participatory democracy	3
FR	bourgeoisie, middle class	5
FR	utility, usefulness	2
FR	national cohesion, national unity, national solidarity	1
FR	inclusion, integration	2
FR	freedom of expression, freedom of speech	2
FR	stimulate the economy, revive the economy	4
FR	human dignity	2
	Hungarian	
HU	defense, protection, shelter	4
HU	liberalism	2
HU	democracies	3
HU	the democracy	3
HU	litigious, contentious	3
HU	palm tree	6
HU	values	2
HU	banks	4
HU	inhabitant, dweller, 'citizen'	1
HU	in the style of	6
HU	existence, being	1
HU	societies	1
HU	with authorities	6
HU	the climate	4
HU	in the societies	1
	Italian	
IT	equality	2
IT	equality	2
IT	self-determination	2

IT	laicism, secularism, secular	2
IT	sovereignty of the people	2
IT	monarchy	3
IT	representative democracy	3
IT	of the query, of the interrogation	6
IT	legality	2
IT	active citizenship	3
IT	of initiative	6
IT	equity	2
IT	legality	2
IT	brotherhood	1
IT	tolerance	2
	Dutch	
NL	society	1
NL	governments	5
NL	economies	4
NL	society	1
NL	press, media	5
NL	system, structure	3
NL	allies	5
NL	survey, questionnaire	6
NL	broadcasts	6
NL	religion	2
NL	currency	6
NL	welfare, prosperity	4
NL	expression of opinion	2
NL	equality, sameness	2
NL	vendors, sellers, salespeople,	6
	Norwegian	
NO	human worth, human value	2
NO	United Nations High Commissioner	5
NO	freedom of religion	2
NO	democracy, rule of the people	3
NO	justice, fairness	2
NO	equality	2
NO	freedom of expression	2
NO	women's rights	2
NO	tolerance	2
NO	rule of law	3
NO	gender equality	2
NO	freedom of belief	2
NO	rule of law	3
NO	human rights	2
NO	self-determination	2
	Portuguese	
PT	democratic regime	3
PT	representative democracy	3
PT	labour laws	4
PT	sovereignty	2
PT	supremacy	2
PT	human dignity	2
PT	meritocracy	2
PT	hegemony	2
PT	ring road, beltway, roundabout	6
PT	individuality	2
PT	equity	2
PT	dignity	2
PT	representativity	2
PT	status quo	2
PT	rationality	2
	Russian	
RU	paradigm	3
RU	elite	5
RU	ideology	3
RU	political elite	5
RU	diplomacy	3
RU	strategy	4
RU	intelligentsia	5
RU	gimmick, shtick, feature, zing; peculiarity	6
RU	location, site	6
RU	Scotland	6
RU	bureaucracy	3
RU	edition, release; version, explanation, theory	6
RU	product line, range; ruler; school assembly; military	6
RU	assembly; assembly	6
RU	platform; plan; platform shoes	6
RU	modification, upgrade, variation, version	6
	Swedish	
SV	religious freedom	2
SV	equality	2
SV	freedom of opinion	2
SV	freedom of the press	2
SV	social justice	2
SV	freedom of expression	2
SV	rule law, state of justice	3
SV	freedom	2
SV	tolerance	2
SV	freedom of the press	2
SV	market economy	4
SV	equal value of all, everyone's equal worth	2
SV	feminism	2
SV	gender equality	2
SV	censorship	4

VisArgue - A Visual Text Analytics Framework for the Study of Deliberative Communication

Mennatallah El-Assady[1], Valentin Gold[2], Annette Hautli-Janisz[3], Wolfgang Jentner[1], Miriam Butt[2], Katharina Holzinger[2], Daniel Keim[3]
[1]Department of Computer and Information Science
[2]Department of Politics and Public Administration
[3]Department of Linguistics
University of Konstanz, Germany
valentin.gold@uni-konstanz.de*

Abstract

For the last two decades, deliberative democracy has been intensively debated within political science and other related fields. Only recently, deliberation research has experienced a computational turn. In this paper, we present a linguistic and visual framework for the study of deliberative communication. The framework includes a range of visual analytics approaches to support research into deliberation. In particular, we propose a range of visualizations for highlighting deliberative patterns over time, speakers, and debates.

1 Introduction

For the last two decades, deliberative democracy has been intensively debated within political science and other related fields. Deliberative democracy promotes a form of democracy that is based on normative rationality and public reasoning. The ideal deliberation aims to arrive at a rationally motivated consensus instead of majoritarian decision-making (Habermas, 1981; Gutmann and Thompson, 1996). At its core, the discourse should be inclusive and based on extensive reasoning. Following Habermas, stakeholders participating in the discourse should be willing to adhere to "the unforced force of the better argument".

While the empirical turn in deliberation research (Chambers, 2003; Bächtiger and Steiner, 2005) has led to an increased understanding of deliberative decision-making, previous approaches in political sciences rely on the application of manual coding schemes determining the deliberative quality within debates (Steenbergen et al., 2003; Hangartner et al., 2007; Lord and Tamvaki, 2013). However, analyzing deliberative processes through manual coding schemes are demanding and time-consuming resulting in a limited set of debate corpora. Moreover, the coding is often subjective making it subject to critical judgments of other researchers (King, 2009; Black et al., 2010; Dacombe, 2013). As a result, manual coding poses challenges with respect to both validity and reliability.

Only recently, the computational turn in deliberation research allows to analyze large quantities of debates. Previous studies, however, focus on single (visual) elements like topic structures (Nguyen et al., 2012; Prabhakaran et al., 2014; Lin et al., 2013) or cognitive complexity to proxy for debate quality (Wyss et al., 2015) but fail to provide a coherent framework for the exploration and interpretation of deliberative communication. With the VisArgue framework, we propose a novel linguistic and visual analytics toolbox to study deliberative communication in all its diverse aspects.

VisArgue is designed on the basis of comprehensible algorithms that also allow less experienced scholars to grasp the underlying logic of the visual tools. Due to the application of many visualization approaches to the same data, different perspectives in the data are highlighted supporting a detailed analysis of the data. In other words: the VisArgue framework provides a toolbox for opening the black-box of deliberative communication.

2 VisArgue framework

The VisArgue framework is based on a collaborative research initiative involving political science, computational linguistics, and information science and visualization engineering[1]. It is designed to support scholars of deliberative communication in various ways. First, we propose a visual tool combining higher-level thematic structures with a close examination of the content (section 2.1). Second, we introduce an approach to

*Corresponding author

[1]For more information, please see http://www.visargue.uni-konstanz.de

Proceedings of the International Conference on the Advances in Computational Analysis of Political Text (PolText 2016), pages 31–36, Dubrovnik, Croatia, 14–16 July 2016.

analyze speaker behavior patterns over topic and time (section 2.2). These two visual approaches mainly support the exploration of yet unknown texts and can be applied independently of the language. Finally, based on the theoretical foundations of deliberative communication, the VisArgue framework proposes a range of visualizations explicitly focusing on deliberative communication. These visualizations range from a rather simple statistical toolkit (section 2.3) to a visual analytics approach combining close and distant reading for the exploration of deliberative patterns (section 2.4). So far, only German communication data can be processed within these visualizations.

The framework is implemented using a client-server architecture. Users can access the tools using their internet browsers which makes installing extra software unnecessary. The web-client works independently of the user's operating system. The software architecture is based on a Java back-end and a JavaScript front-end. The processed data is saved in a database (MongoDB) and is then loaded into the user's cache – making cached data accessible to the visualizations without the need to process it multiple times. To tackle privacy issues, users have to use authentication to access the web-client. This ensures only authorized access to the data of each user.

In the following sections, we will provide an overview on some of the visual analytics tools. We will briefly describe the rationale and give examples of these visualizations. In order to provide a coherent picture, we rely on data on the arbitration on Stuttgart 21 (henceforth: S21). S21 is a railway and urban development project in Southern Germany. To reconcile conflicts between proponents and opponents, an arbitration procedure was established to discuss the facts of the project. The arbitration lasted for 9 sessions. Overall, this results in a corpus of around 9.100 turns with almost 70 speakers.

2.1 Lexical Episode Plots

The Lexical Episode Plots (Gold et al., 2015b) combine the logic of what Digital Humanities scholars call "distant reading" with the logic of "close reading". Primarily, the visual tool is used to explore yet unknown texts. In general, it can not only be applied to communication data, but also to any other (sequential) text data type. The contribution of this visual analytics approach is twofold: First, a novel text mining method to identify thematic clusters within a text is introduced. Second, these clusters are presented in an interactive visualization enabling an exploratory data analysis.

With respect to the applied algorithm identifying the clusters, we rely on a comprehensive method enabling less experienced users to grasp the mathematical foundations of the algorithm. The basic idea is based on the concept of lexical chaining (Morris and Hirst, 1991). Hereby, we attempt at extracting word-sequences that appear more densely than expected within a text segment given their count in the whole word sequence of the text. Hence, each extracted cluster represents a span of text in which the frequency of a specific term is significantly higher than its average in the document. The clusters are not only based on unigrams, but also on higher-order n-grams, i.e. two or more words that form an entity term (like "computational social science"). Additionally, based on a likelihood ratio test, for each term cluster, we compute its level of significance.

In a second step, the lexical episodes are visualized. The visual design follows the mantra: overview first, zoom and filter, detail on demand (Shneiderman, 1996). In general, each episode is visualized as a vertical bar to the left of the text. The bars span from the first to the last occurrence of the term within a cluster segment. Each bar is assigned a different color – bars that include the same term are assigned the same color. Scholars can visually explore the episode clusters, interactively. First, episodes can be filtered based on the level of significance. By interactively changing the significance level, users can control the number of episodes displayed in the visualization. Second, they can zoom in and out to switch between a distant and close reading of the textual data. Finally, by clicking on an episode bar, the terms are highlighted within the text representation.

Figure 1 shows the visualization of the Lexical Episode Plots. The visualization reveals the sequential structure of the arbitration on S21 and highlights the most important thematic clusters. For instance, in the first session, the members of the arbitration committee discussed the transport of goods (*Güterverkehr*), the switches (*Weichen*), and the emergency concept (*Notfallkonzept*). Moreover, the visualization reveals that Ms. Starke (*Frau_Starke*) was the most referred person in the beginning of the arbitration.

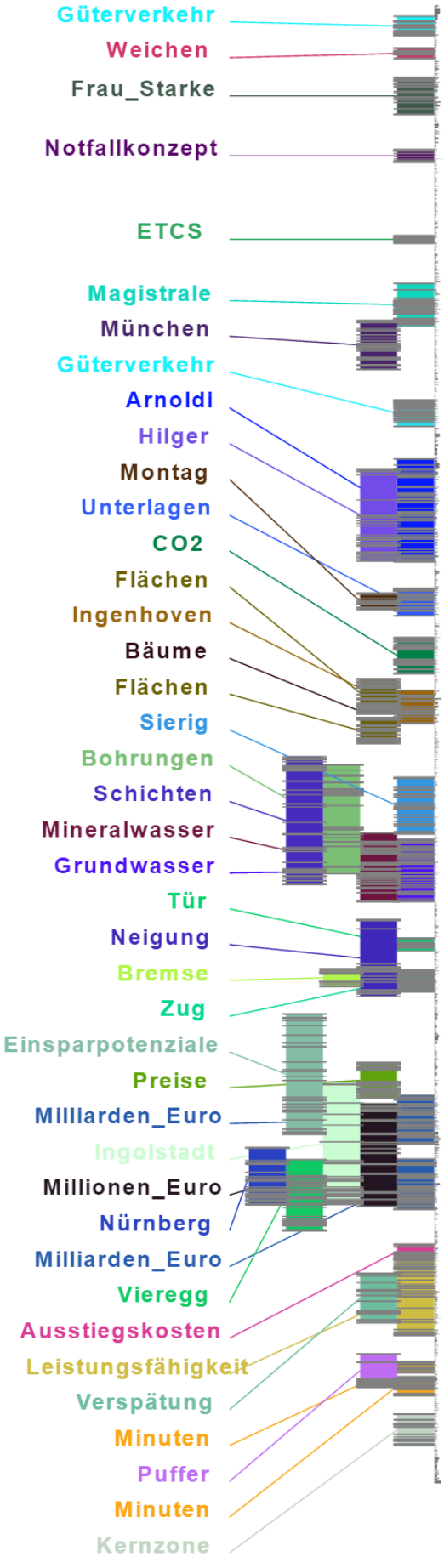


Figure 1: Lexical Episode Plots for S21

2.2 ConToVi

ConToVi (El-Assady et al., 2016), the Conversation Topic Visualization, was introduced to analyze speaker behavior patterns. ConToVi tracks the movement of speakers across the thematic landscape of a conversation. It is designed to explore the dynamics of conversations over time, highlighting speaker interactions and behavior patterns. Hence, compared to the Lexical Episode Plots, it adds a new dynamic layer to the analysis.

To uncover the topics in a given text, we utilize a hierarchical topic modeling algorithm that is developed to cope with the sequential structure of conversations (El-Assady, 2015). This algorithm was designed to specifically address the challenges with transcribed spoken data – namely more noisy data containing non-standard lexical items and syntactic patterns. Using the results of the topic modeling algorithm span a floor for the representation of speaker dynamics. In Figure 2, the movement of speakers in the topic space is shown. The topics are represented on the circular plot. Topics that are addressed more often are visualized by larger segments on the circular plot. With 16 topics shown, the movements and interactions of speakers over time can be visually tracked turn by turn. For instance, while in the previous turn the yellow speaker has addressed the topic on the left side, in this turn, the speaker moves to a different topic on the upper right side. Similarly, before the yellow speaker changed his or her topic, the light green speaker moved from a topic on the right side to the topic depicted at the bottom of the circular plot.

Beside demonstrating dynamics of speakers over time, ConToVi allows retracting the speakers' paths through the topic space. Since one of the main theoretical assumptions of deliberative communication requires speakers to listen and respond to each other, we assume deliberative debates to be characterized by overlapping paths. This is illustrated in Figure 3 for one session of the arbitration. The moderator of the debate moves back and forth addressing most topics in this session. In general, the moderator also addresses topics not related to the moderation of the debate but actively intervenes in the substantive issues of the debate. Speaker A and B are both less involved in the debate with Speaker A showing a tendency to the upper left topics – however, to some degree, the paths overlap.

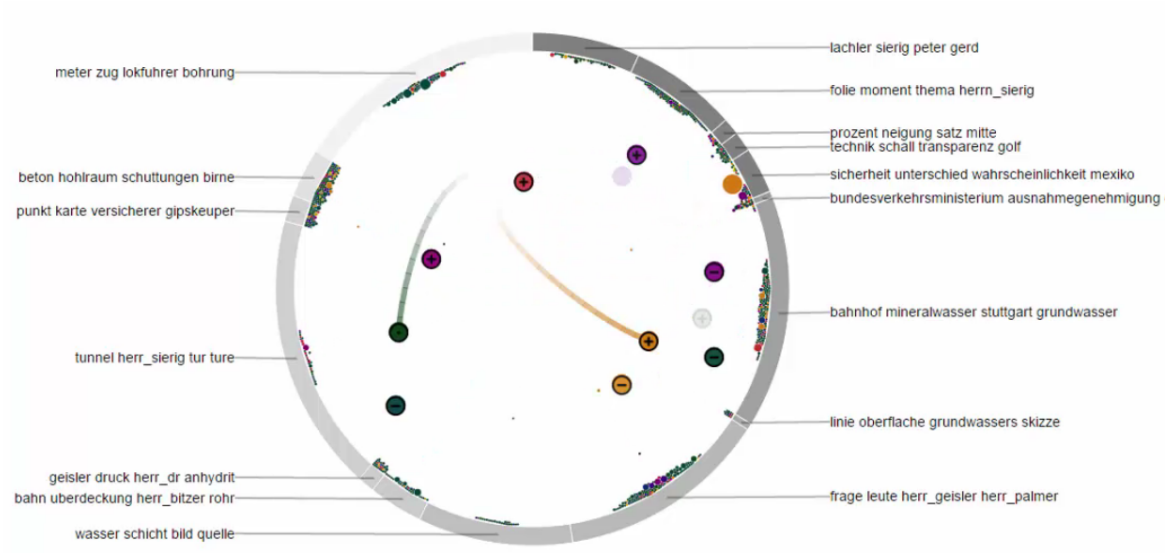


Figure 2: ConToVi Visualization

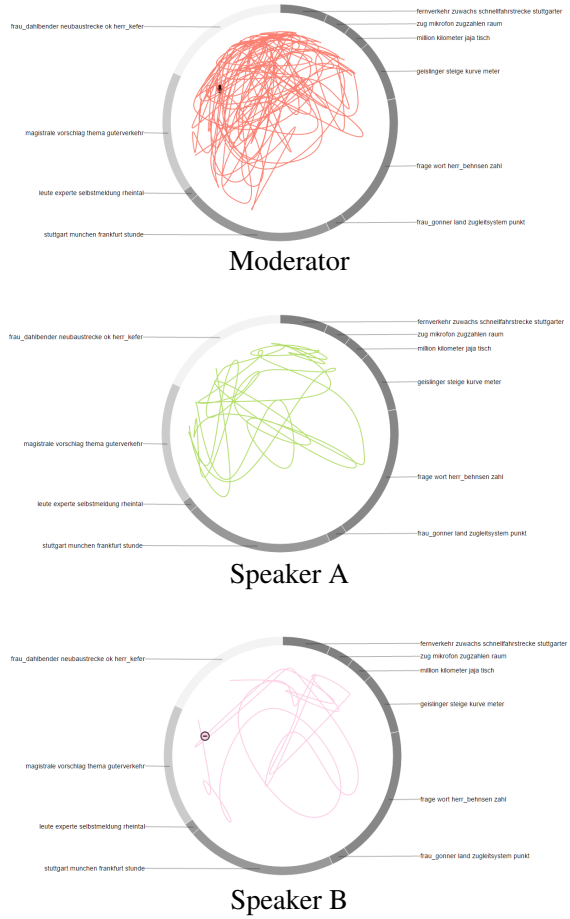


Figure 3: Speaker Paths

2.3 Deliberation Statistics

To arrive at a visual representation of deliberative communication, deliberation needs to be measured. As part of the VisArgue project, we propose a computational linguistic parsing system annotating the degree of deliberation for four dimensions: participation, respect, justification, and accommodation (Gold et al., 2015a; Gold and Holzinger, 2015). These four dimensions result from the application of natural language processing tools, unsupervised content extractions, dictionary applications, and statistical analyses. The four dimensions are further subdivided in different subdimensions belonging to similar theoretical concepts. For instance, within the broad dimension of justification, we determine the type and degree of reason-giving, the certainty with which information are exchanged, and the reference to norms. In total, the computational linguistic pipeline results in 53 individual measures of deliberative communication.

In order to support the analysis of deliberative communication, the VisArgue framework offers the possibility to quickly access descriptive statistics with respect to the 53 measures. In Figure 4, we demonstrate the general visual rationale for generating the statistics. Based on the type of measure, scholars can drag and drop the measures from the left side panel to the right panel. Besides specifying the x- and y-axis according to the scholars needs, they are provided the opportunity to name the visualization. After all is set, by clicking on the button, the visualization is created.

One of these visualizations is shown in Figure 5. It depicts the degree of reason-giving for

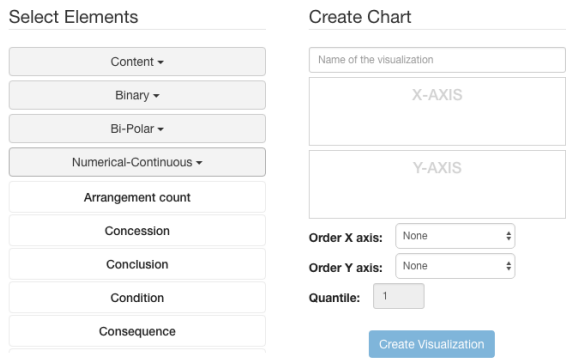


Figure 4: Statistics Visualization

each speaker in one of the sessions on S21, in relation to the mean level of reason-giving in this session. The green bars to the right indicate more reason-giving than on average, the red bars to the left less reason-giving, respectively. In general, we also provide the possibility to aggregate the statistics with regard to some metadata of the speakers, e.g. the position towards the project.

Figure 5: Degree of Reason-Giving per Speaker

2.4 Lexical Units

In order to explore and interpret the various measures of deliberative communication, we propose Lexical Units Visualization that is based on the annotation system but allows a distant reading of all annotations. Similar to the Lexical Episode Plots, the visualization combines the logic of close and distant reading and can be used to interactively explore the discourse.

For instance, in Figure 6, we demonstrate the visual approach for five deliberative annotations in one of the sessions on S21. The five annotations are visualized next to each other enabling a distant comparison of textual features. Again, similar to the Lexical Episode Plots, the text of the debate is shown in black and each segment is colored with its respective annotations. Each segment represents an Elementary Discourse Unit (EDU). Based on Marcu (2000), we assume the text between two punctuation marks to belong to the same event (Polanyi et al., 2004) and, hence, to be collocated in one EDU. The first bar in Figure 6 visualizes argumentation (red), the second bar conventional implicatures (blue), the third bar event modality (purple), the fourth bar information certainty (gree), and finally, the last bar emotions (yellow). The figure reveals overlapping segments of deliberative annotations and by providing zoom functionality, close reading can provide more insights into the debate and the reasons for these overlapping segments of deliberative behavior.

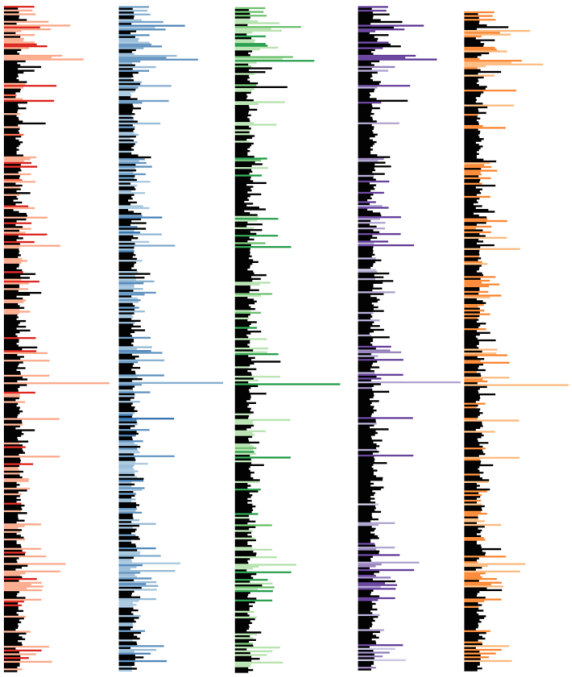


Figure 6: Lexical Units Visualization

3 Conclusion

In this paper, we introduce the VisArgue framework, a set of interactive visualization approaches to explore and interpret deliberative communication. These visual analytics tools are based on the result of a natural language processing pipeline combining various measurement approaches. We conclude that the turn in deliberation research towards computational analysis is the next step for analyzing large quantities of communication data.

References

André Bächtiger and Jürg Steiner. 2005. Introduction. *Acta Politica*, 40:153–168.

Laura W. Black, Stephanie Burkhalter, John Gastil, and Jennifer Stromer-Galley. 2010. Methods for Analyzing and Measuring Group Deliberation. In Erik P. Bucy and R. Lance Holbert, editors, *Sourcebook of Political Communication Research: Methods, Measures, and Analytic Techniques*, chapter 17, pages 323–345. Routledge, New York, NY.

Simone Chambers. 2003. Deliberative democracy theory. *Annual Review of Political Science*, 6(1):307–326.

Rod Dacombe. 2013. Thinking about the quality of deliberative politics: a critical look at the discourse quality index. Paper presented at the SSPP Annual Research Conference 2013, June 14, King's College London.

Mennatallah El-Assady, Valentin Gold, Carmela Acevedo, Christopher Collins, and Daniel Keim. 2016. ConToVi: Multi-Party Conversation Exploration using Topic-Space Views. *Computer Graphics Forum*, 35(3):431–440.

Mennatallah El-Assady. 2015. Incremental Hierarchical Topic Modeling for Multi-Party Conversation Analysis. Master's thesis, University of Konstanz.

Valentin Gold and Katharina Holzinger. 2015. An Automated Text-Analysis Approach to Measuring the Quality of Deliberative Communication. Paper prepared for presentation at the 2015 Annual Meeting of the American Political Science Association (APSA), San Francisco, USA.

Valentin Gold, Mennatallah El-Assady, Tina Bögel, Christian Rohrdantz, Miriam Butt, Katharina Holzinger, and Daniel Keim. 2015a. Visual Linguistic Analysis of Political Discussions: Measuring Deliberative Quality. *Digital Scholarship in the Humanities*. First published online: 10 September 2015.

Valentin Gold, Christian Rohrdantz, and Mennatallah El-Assady. 2015b. Exploratory Text Analysis using Lexical Episode Plots. In E. Bertini, J. Kennedy, and E. Puppo, editors, *Eurographics Conference on Visualization (EuroVis) - Short Papers*. The Eurographics Association.

Amy Gutmann and Dennis F. Thompson. 1996. *Democracy and Disagreement. Why moral conflict cannot be avoided in politics, and what should be done about it*. Harvard University Press, Cambridge, MA.

Jürgen Habermas. 1981. *Theorie des kommunikativen Handelns*. Suhrkamp, Frankfurt am Main.

Dominik Hangartner, André Bächtiger, Rita Grünenfelder, and Marco R. Steenbergen. 2007. Mixing habermas with bayes: Methodological and theoretical advances in the study of deliberation. *Swiss Political Science Review*, 13(4):607 – 644.

Martin King. 2009. A critical assessment of Steenbergen et al's Discourse Quality Index. Roundhouse Vol 1 Issue 1.

Ching-Sheng Lin, Samira Shaikh, Jennifer Stromer-Galley, Jennifer Crowley, Tomek Strzalkowski, and Veena Ravishankar. 2013. Topical positioning: A new method for predicting opinion changes in conversation. In *Proceedings of the Workshop on Language in Social Media (LASM 2013)*, page 41, Atlanta, GA.

Christopher Lord and Dionysia Tamvaki. 2013. The politics of justification? Applying the 'Discourse Quality Index' to the study of the European Parliament. *European Political Science Review*, 5:27–54, 3.

Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge, MA.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational linguistics*, 17(1):21–48.

Viet-An Nguyen, Jordan Boyd-Graber, and Philip Resnik. 2012. SITS: A Hierarchical Nonparametric Model using Speaker Identity for Topic Segmentation in Multiparty Conversations. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics.

Livia Polanyi, Chris Culy, Martin van den Berg, Gian Lorenzo Thione, and David Ahn. 2004. Sentential structure and discourse parsing. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 80–87.

Vinodkumar Prabhakaran, Ashima Arora, and Owen Rambow. 2014. Staying on topic: An indicator of power in political debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, October. Association for Computational Linguistics*.

Ben Shneiderman. 1996. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages*, pages 336–343, Washington, WA. IEEE Computer Society Press.

Marco R. Steenbergen, André Bächtiger, Markus Spörndli, and Jürg Steiner. 2003. Measuring political deliberation: A discourse quality index. *Comparative European Politics*, 1(1):21–48.

Dominik Wyss, Simon Beste, and André Bächtiger. 2015. A Decline in the Quality of Debate? The Evolution of Cognitive Complexity in Swiss Parliamentary Debates on Immigration (1968–2014). *Swiss Political Science Review*, 21(4):636–653.

Intra-Party Disagreement at Party National Congresses and Issue Salience in Election Manifestos: Evidence from Germany

Zachary Greene
University of Strathclyde
16 Richmond St.
Glasgow, G1 1XQ
zacgreene@gmail.com

Abstract

Parties emphasize issues to attract votes and mobilize intra-party groups. The platform creation process creates opportunities for intra-party groups to contribute to its content. In the face of internal disagreement and diverse topical interests, however, manifesto writers must find the most acceptable compromise between issue emphasis and avoidance. I propose that divided parties discuss issues in greater detail. Members from a range of backgrounds see the manifesto as a venue to detail and extend the details of internal compromise. This perspective leads to the prediction that divided parties emphasize their policy goals in greater detail. I evaluate this perspective using evidence of internal disagreement and priorities from a scaling model of speeches at parties national meetings and a structural topic model of election manifestos in Germany.

1 Introduction

Party leaders prefer to only address issues that increase electoral support and ignore topics that highlight internal conflict. Yet, they also face competing demands from within the party. If party leaders use manifestos to detail policy compromises between groups that internally disagree, manifestos will also include issues that appeal beyond electoral motivations. Do parties artfully avoid conflict or use manifestos to negotiate internal coalitions?

I propose that parties' campaigns address issues to maintain diverse group support while also addressing topics popular to public opinion. I hypothesize that leaders use election programs to build support among internal groups. Well informed, active and influential intra-party groups demand more detailed proposals that outline the compromises the leader will take upon entering office. The outcome of these negotiations would increase attention to these issues as intra-party groups seek to limit the leader's future policy-making activities. Intra-party division, therefore, fuels attention on issues in the party's platform. I contrast this approach with a hypothesis following from an issue competition perspective that would predict a strategy of issue avoidance.

I examine these competing propositions using data from parties' national congresses and election manifestos for parties in Germany. These parties allow members to give speeches at these meetings before holding elections to select the party's platform. Like studies of intra-party heterogeneity (Debus and Bräuninger, 2008) (Greene and O'Brien, 2016), I use automated content analysis of speeches at national congresses to develop measures of intra-party division and issue diversity. I examine the hypotheses by combining estimates from an unsupervised scaling model (e.g. Wordfish) (Slapin and Proksch, 2008), with a Structural Topic Model (Roberts et al., 2013). I first scale the level of disagreement at party national meetings using Wordfish and combine these estimates with the relative attention to issues at these meetings using a Structural Topic Model. I then use these estimates as structural covariates to predict the relative salience of issues in parties' manifestos. The results indicate that party manifestos often address issues their leaders would rather avoid for electoral reasons.

Proceedings of the International Conference on the Advances in Computational Analysis of Political Text (PolText 2016), pages 37–42, Dubrovnik, Croatia, 14–16 July 2016.

2 Empirical Approach

I predict that intra-party disagreement either 1) decreases the attention to issues, or 2) increases attention to issues in party platforms. To adjudicate these hypotheses, I collected the texts of manifestos for parties in Germany from 1990 to 2015 (Volkens et al., 2011). These parties produce manifestos following discussion at national meetings. These meetings offer the opportunity for members from diverse backgrounds to express goals, run for internal positions and eventually select future le(Ceron, 2012; Greene and Haber, 2016). These parties often demonstrate high levels of parliamentary discipline and their leaders dominate the legislative process. If party leaders truly dominate the manifesto writing process, as many have argued, then intra-party disagreement should be pushed under the rug to avoid electoral defeat.

In the first stage, I estimate the relative disagreement and attention to issues expressed at parties' national meetings. I begin by scaling the relative positions of speakers to find the variance of positions expressed, which I label intra-party disagreement. I then estimate the topics expressed in speeches at the national meetings using a STM with a variable to measure differences across years. Based on the distribution of words allocated to each topic, I measure the expected proportion of words in each manifesto based one the primary substantive topics expressed by the model as predicted by structural covariates for each year. By aggregating these results to measure the level of overall disagreement and effective expected number of issues (ENI), these results provide the structural components in a model predicting attention to the issues identified by a STM of parties' manifestos.

2.1 Issue Salience

I predict attention to issues in parties manifestos as identified by a Structural Topic Model (STM). Like a correlated topic model, this approach allows me to estimate the attention to issues in parties' manifestos, but also allows me to include substantive covariates (metadata) that predict the proportion of words on each topic. This approach has been used in a variety of political applications such as the framing and content of news reports, survey responses (Roberts et al., 2014), twitter feeds and even religious statements (Lucas et al., 2015; Genovese, 2015).

I collected parties' manifestos from the Comparative Manifesto Project (Volkens et al., 2011). I then processed the texts by converting them into raw text files that could be read into the stm package in R developed by Roberts et al. (2014). I converted all documents from .pdf files to .txt. I then removed all punctuation and numbers, and converted all words to lower case stems using the Snowball stemmer for German. I further removed words that only appear in one percent of documents. Roberts et al. (2014) outline the generative process for estimating an STM model with k topics (fixed in advance by the researcher) for each document (for additional information, see Roberts et al. 2014). The STM applies a data generating process to each of the documents and then sorts through the data to discover the most likely values for each of the parameters. The model starts at the document level, before estimating the topic and the topic/word distributions based on the observed word frequencies in each document. Each document contains a mixture over topics to allow them to contain multiple topics. Following Roberts et al. (2014), the document's attention to a topic is estimated from a logistic-normal generalized linear model with covariates for each document. Following Roberts et al. (2014), I include a number of structural variables that allow me to predict topic prevalence within each of the manifestos based on their disagreement and ENI.

2.2 Salience and Disagreement at Party Congresses

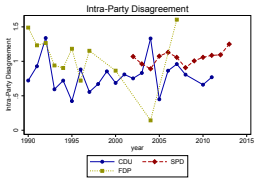
While scholars have long sought to study intra-party cohesion in parliamentary settings, innovations in text analysis allow researchers to study disagreement between members of the same party through speeches (Bäck et al., 2014) or the motions voted on in party meetings (Ceron, 2012; Ceron, 2013; Ceron, 2014). Like Greene and Haber (2016), I gather speeches from party national meetings to measure the disagreement within each congress.[1]

[1]Party congress transcripts are well adapted for studying the effect of intra-party dynamics on electoral manifestos. Parties hold congresses either annually or at least in the year prior to an election to decide on the party's electoral strategy, policy goals and leadership. Party leaders likely dominate the

I use the transcripts from these meetings to measure intra-party disagreement using Wordfish to scale the most important underlying dimension of expressed conflict. Wordfish is an unsupervised scaling model that estimates party positions from the frequency of words used in a set of documents by distributing the primary parameters (the mean and standard deviation) according to a Poisson distribution.The model predicts the count of each word in each speech. The model further includes fixed effects for the speaker , word, and a word specific weight that captures the importance of word and estimates the speaker's position.[2]

To estimate disagreement, I model the positions of all speeches currently available and in a usable format for the CDU (1990-2011), the FDP (1990-2009) and the SPD (2001-2013) to estimate intra-party disagreement. I then estimate the standard deviation of the positions at each party congress.[3] I present an overview of these estimates for the CDU, the FDP and the SPD in Figure 1. As past analyses note, the CDU faced somewhat higher levels of disagreement in the early 1990s as in response to the challenges of controlling an unpopular government, but became more focused in the opposition (1997-2005).

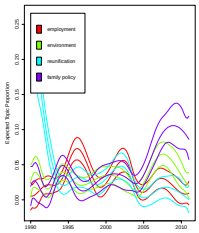
Figure 1: Intra-Party Disagreement



I also use speeches to estimate the distribution of topics as a baseline for the manifesto analysis. I follow the same STM procedure that I undertake for parties' election manifestos, but also include splines for the year in which the party congress was held. This allows me to then estimate the expected proportion of words on each issue by year using the point estimate for each topic and year.

I present the predicted expected proportion of words for selected topics from year splines in Figure 2 based on a 25 topic model. I predict four topics that I label as employment, the environment, reunification, and family politics based on the words that are the most exclusive and frequent for the topic (FREX).[4]. As Figure 2 shows, the amount of attention to this topic spiked early in the 1990s to reflect the fall of communism and the individuals seek to shelter from the former soviet bloc. [5] Finally, topic 18 includes a large number of words related to children and families.[6] This topic became increasingly important within Germany in the late 2000s and to the CDU in particular, as the expected topic proportion indicates in Figure 2.

Figure 2: CDU Party Congress Topics



I replicate this process for each party to estimate the relative attention to issues in parties' con-

meetings' agenda, but members often select between manifestos and even leadership candidates at these meetings. In many cases, party members that seek to express their support for issues face few limitations in speaking at these meetings (Kernell, 2015b; Kernell, 2015a). Policy motivated party members can speak at these meetings to draw the leadership's attention toward their preferred goals or express support for a set of issues.

[2]Like Wordscores (Laver et al., 2003), Wordfish has been used to study expressed preferences in a range of settings such as party election manifestos, speeches in parliament, Twitter posts, motions in party meeetings and speeches at party congresses (Proksch and Slapin, 2009; Ceron, 2012; Bäck et al., 2014; Greene and Haber, 2016).

[3]In the analysis of CDU party congresses, I anchor the estimates using the positions of Helmut Kohl in 1994 and Angela Merkel in 2010 to reflect the party's increasingly centrist views.

[4]Topic 2 stands out as related to employment and social help (arbeit, arbeitsmarkt, steueurreform, arbeitslos, sozialhilf). Topic 15 includes terms related to the environment and globalization (okolog, markt, globalisier, umwelt, energi). Topic 5 also includes some points related to the environment (umwelpolitik, umweltsminist, klimaschutz) more closely as it relates to climate change and government reforms

[5]Topics 6, 16 and 19 are closely related as the most exclusive and common terms for the categories refer to asyllum seakers, and immigrants in the first, the fall of the DDR and East Germany (Ost, DDR West, dresdn) in the second and to the politics of assylum and refugees (fluchtlingskonvention, asylbewerb, herkunftsland, drangend) in the third.

[6]such as familienpolitic, elt, kind, kindergeld, betreuungsgeld

gresses. I then use the predicted proportion of words from each party congress on each topic as the issue level components to measure the ENI at the party congress following Greene (Greene, 2015; Greene and O'Brien, 2016). I use the measures of intra-party disagreement, and ENI from the party congress prior to the election as the primary independent variables. These variables predict the prevalence of words within each manifesto topic.

3 Analysis

I predict the attention to topics in manifestos by estimating a structural topic model. The sample for the main analysis is limited to 17 manifestos in Germany from 1990-2012. Below, I present the results of the structural topic model for a 14 topic model.[7] I first review the distribution of words across documents and the content of each topic. I then predict the effect of the covariates on the expected distribution of words in each topic based on each parties' level of internal disagreement. The competing hypotheses suggest that parties will either increase attention to issues in the face of increased disagreement or they will avoid them.

Figure 3: Manifesto Topic Prevalence

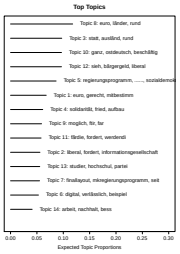
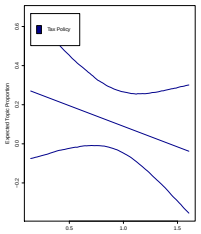


Figure 3 shows the proportion of words along with those that score highly as both frequent and exclusive (FREX). The most prevalent topics represent issues traditionally important in German politics such as the environment and economy, family politics and childcare, and Eurozone labour market policies. The least common set of words are those in Topic 14, that generally refers to general statements on German labor policies. [8] I focus the analysis on the effect of the structural covariates to consider the effect of disagreement.

I predict that electorally motivated, but divided parties' manifestos will avoid addressing issues in detail. However, if party leaders are constrained by their organizations or require their support, parties' manifestos will actually include greater detail. The inclusion of the structural covariate for disagreement allows me to examine the effect of disagreement on topic prevalence.

Figure 4: Expected Proportion Tax Policy



I present the expected prevalence of words on those topics for which the structural covariates have a clear effect. While intra-party disagreement has a neutral effect on most issues, disagreement has a negative effect in the most prevalent topic, which I label as tax policy.[9] I show the predicted change in the expected proportion of a manifesto's focus on words related to tax policies in Figure 5 with 95% confidence intervals. As the first hypothesis predicts, as disagreement increases, the expected proportion of words focused on tax policies decreases. While the 95% confidence intervals hug the zero line at the lower bound through the range of the figure, 90% confidence intervals are slightly above zero or values lower than the mean level of disagreement (.91). These results support the perspective that divided parties avoid or de-emphasize the most important issues.

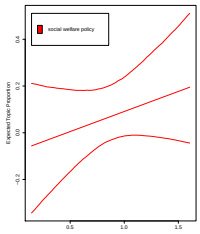
However, the effect of disagreement reverses

[7] Models using 10, 12, and 16 categories produce topic estimates largely similar substantive topics.

[8] While some of the FREX terms presented in the middle categories appear uninteresting in Figure 3, a deeper look into these lists reveal that the topics are substantively useful. The topics reveal some degree of semantic coherence and exclusivity. Although some topics are less coherent, the combination of European issues with the national ones suggests that these categories likely reflect the reality that many issues are balanced between the national and European level.

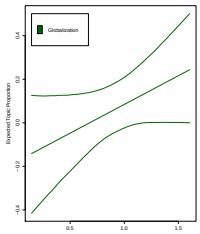
[9] The most frequent words that also are the most exclusive (FREX) for topic 8 are words such as kindergrundfreibetrag (child tax allowance), burgerversicher (general insurance), gesundheitsprami (health care premiums).

Figure 5: Expected Proportion Social Welfare



on other policies. I show the effects of disagreement on the expected proportion of a document on social welfare policy[10] and on issues related to globalization[11] in Figures 5 and 6 with 95% confidence intervals. In both cases, at low levels of disagreement, the topics are hardly discussed. At higher levels, disagreement is associated with a greater proportion of words.[12] For these issues, the results suggest evidence consistent with the second hypothesis; parties include greater discussion in the face of increased disagreement.

Figure 6: Expected Proportion Globalization



ENI performs largely as expected.[13] These results imply that as parties discuss more issues in their party congresses, they emphasize economic development in greater detail, but address less attention to globalization, civil society groups, and education in their manifestos.

4 Discussion

This paper seeks to disentangle the influence of disagreement on the issues parties include their election manifestos. This study forwards evidence from party national congresses on the influence of intra-party heterogeneity on their election manifestos. By measuring intra-party disagreement and issue diversity from unsupervised content analysis of party congresses in Germany, I show evidence that intra-party disagreement influences the attention parties give issues in their manifestos.

The results from a structural topic model are mixed. Disagreement is associated with decreased attention to the most important topic.[14] But, disagreement is also associated with increased attention on topics that could be linked to the second most important dimension of conflict, European politics, post-materialism and quality of life issues. These results indicate that competing perspectives on issue disagreement likely result from the multi-dimensionality of European politics.

[10]The social welfare topic includes words that score highly on the FREX measure such as regierungsprogramm, sozialdemokrati, mindestlohn, and brgerinnenprojekt.

[11]The globalization topic includes references to europea, liberal, okologi, FTIR, FILR, and soil.

[12]These positive relationships persist using alternate modeling choices such as changing the number of topics and additional structural components, although the effects only weakly become significant above approximately mean levels of disagreement (.91) at the 90% level.

[13]It is positively associated with Topic 3. This topic includes words associated with the DDR, foreign states, the economy, new, creation, and state., and the tax policy topic. It is negatively associated with the globalization policy topic, Topic 12 (FREX words: civil society, and a basic income for citizens), and Topic 13 (FREX words: education and school)

[14]The fuller content of the tax policy topic closely relates to the issues important to the historical left-right division within Germany (Lipset and Rokkan, 1967)

References

Hanna Bäck, Marc Debus, and Jochen Müller. 2014. Who takes the parliamentary floor? the role of gender in speech-making in the swedish riksdag. page 1065912914525861.

Andrea Ceron. 2012. Bounded oligarchy: How and when factions constrain leaders in party position-taking. 31(4):689–701.

Andrea Ceron. 2013. Brave rebels stay home: Assessing the effect of intra-party ideological heterogeneity and party whip on roll-call votes.

Andrea Ceron. 2014. Inter-factional conflicts and government formation do party leaders sort out ideological heterogeneity? page 1354068814563974.

Marc Debus and Thomas Bräuninger. 2008. Intra-party factions and coalition bargaining in germany. In *Intra-Party Politics and Coalition Governance*, pages 121–145.

Federica Genovese. 2015. Politics ex cathedra: Religious authority and the pope in modern international relations. *Research and Politics*, 2(4):2053168015612808.

Zachary Greene and Matthias Haber. 2016. Leadership competition and disagreement at party national congresses. 46(3):611–632.

Zachary Greene and Diana Z. O'Brien. 2016. Diverse parties, diverse agendas? female politicians and the parliamentary party's role in platformformation.

Zachary Greene. 2015. Competing on the issues how experience in government and economic conditions influence the scope of parties' policy messages. page 1354068814567026.

Georgia Kernell. 2015a. Party nomination rules and campaign participation. page 0010414015574876.

Georgia Kernell. 2015b. Strategic party heterogeneity. page 0951629814568401.

Michael Laver, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data. 97(2):311–331.

Seymour M. Lipset and Stein Rokkan. 1967. Cleavage structures, party systems, and voter alignments: an introduction.

Christopher Lucas, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer, and Dustin Tingley. 2015. Computer-assisted text analysis for comparative politics. page mpu019.

Sven-Oliver Proksch and Jonathan B. Slapin. 2009. How to avoid pitfalls in statistical analysis of political texts: The case of germany. 18(3):323–344.

Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Edoardo M. Airoldi, and others. 2013. The structural topic model and applied social science. In *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*.

Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley. 2014. stm: R package for structural topic models. 1:12.

Jonathan B. Slapin and Sven-Oliver Proksch. 2008. A scaling model for estimating time-series party positions from texts. 52(3):705–722.

Andrea Volkens, Onawa Lacewell, Pola Lehmann, Sven Regel, Henrike Schultze, and Annika Werner. 2011. The manifesto data collection.

The Medium is the Message: Automated Content-Analytic Techniques Across Mass Media Platforms

Dan Hiaeshutter-Rice
Communication Studies
University of Michigan
Ann Arbor, MI
dhrice@umich.edu

Abstract

Recent technological change has meant that political communication increasingly occurs across multiple mediums – including newspapers, television, and radio, but also a wide range of social media. The differences that exist in message content, language, and audience – not to mention the varying affordances of different platforms – greatly impact what users choose to communicate, how they communicate it, and how it understood by others. In short: the nature of political campaigns, and the way in citizens learn about and understand politics, is shifting. This working paper is the first step towards developing an empirical analysis to study this shift. Preliminary analyses of content from Facebook, Twitter, debates, and announcement speeches over the past 12 months of the 2016 US primary season make clear that (a) there is a considerable portion of campaign content in social media, and (b) that content is both similar in general themes and different in emotional affect from what we see in more traditional campaign communications.

1 Introduction

The dramatic increase in the quantity and type of media platforms available has significantly increased the visibility of political candidates. Consider the 2016 presidential race, for example: for most voters, Senator Bernie Sanders' Twitter account or Senator Ted Cruz's Instagram account are just a few seconds away. Instead of waiting for a daily newspaper or a radio address to hear the thoughts of a candidate, one can simply pull their phone out of their pocket and access the Facebook posts of Governor Jeb Bush. This expansion of campaigning venues means there is far more information about candidates and their campaigns than there was before.

This new media landscape matters. The differences that exist in message content, language, and audience, as well as varying affordances of these different platforms, greatly impacts what users (such as presidential candidates) choose to communicate, how they communicate it, and how it is understood by others. It is thus essential for political communication scholarship to consider relationships between the *content* of the message (tone, words, etc.); the *target* of the message (the intended direct and indirect audiences); and *mediation* of the message (the technical/technological constraints involved in, as well as the social norms attached to, the media platform). This paper addresses the first of these.

2 Background

This project seeks to explore how changing the communication stream through which campaign information is broadcast changes factors citizens consider when forming their ideological positions. A first step is the application of standard (dictionary-based) language processing tools to explore differences in language across both traditional and "new" communications mediums. It will capitalize on the tremendous amount of content being generated by the 2016 American presidential primary races. Part of the aim here is logistical: this project will develop new ways of structuring and managing corpora from multiple platforms – in this case, (1) scraped candidate

Proceedings of the International Conference on the Advances in Computational Analysis of Political Text (PolText 2016), pages 43–47, Dubrovnik, Croatia, 14–16 July 2016.

Facebook posts from the Facebook API, (2) scraped debate transcripts, (3) candidate speech transcripts, and (4) candidate Twitter data through the Twitter API.[1] Part of the aim is methodological: new analyses require that we reconsider whether the same dictionaries and/or algorithms can be effectively deployed across multiple mediums, for instance. Discussion will consider relationships between the content of the message (tone, words, etc.); the target of the message (the intended direct and indirect audiences); and mediation of the message (the technical/technological constraints involved in, as well as the social norms attached to, the media platform). And extracted differences across corpora will offer an early sense for how mediums matter for political campaigns in 2016.

This paper builds towards a large project that will evaluate content of political campaigns across mediums. That project is motivated by Marshall McLuhan's "the medium is the message" argument, which claims that mediums (such as television, newspapers, Twitter, etc.) are *themselves* messages, entirely independent of the content that they present. (McLuhan, 1964). McLuhan uses the example of a heinous crime covered by a television newscast. He argues that the important thing to understand is how the presence of the television inside the home changes public opinion about the crime in ways that were not possible before the advent of in-home TV sets. In essence, television brought the crime into the home by altering the way the users experienced the content of the story. For political campaigning, this paper begins to explore the idea that mediums fundamentally alter the content of the messages transmitted through them. This matters for how we think about the kinds of political information that is available to citizens.

From McLuhan's seminal work, we can turn to the initial question posed in this paper. How do mediums change the content of the message? To attempt to answer these questions, this project uses data from the 2016 presidential election across a variety of mediums.

3 Measuring Text

There are a variety of approaches used to understand political text (See Grimmer and Stewart, 2013 for discussion). This paper starts with dictionary based processing. Dictionary processing can help us get at some of the ways in which mediums change political communication, although clearly not all of them. However, as this is a relatively new theoretical approach, a basic descriptive approach can provide the crucial first insights.

Dictionary processing takes advantage of user-defined dictionaries to categorize and classify text. For example, Young and Soroka (2012) use this method to contribute to a large body of work that argues that language processing through established dictionaries can provide insights into the tone/sentiment of the message. Automated text processing is also applied in Murthy and Petto (2015) who find that print media and Twitter systematically differ in the sentiment (positive/negative) attached to candidates; Evans et al. (2014) discover that Democrats, women, and incumbents Tweet differently than their opposites; and Golbeck et al. (2010) who revealed that the content of politician's Twitter accounts does not provide new information or insights into government. But automated text processing can only get us so far as it inherently relies on sample selection and the dictionary used. As such, decisions about what words to put in the dictionary and their relationship to other words matter greatly for the findings of the project.

4 Data

The first step this project makes in evaluating mediated communication is by looking at the actual words that are used in each medium. To restate the argument presented above, we should not expect the same message to exist across mediums. Perhaps the easiest way to think about mediated content is basic word counts. While admittedly a simplistic analysis, we can use actual examples to highlight how this matters. A candidate who says "We need to address the immigration issue" is saying something different than a candidate who says "We need to address the immigration issue but not with amnesty but start with taking control of our own borders." Undoubtedly some of the mediums are going to produce fewer words per post/tweet/statement/etc. In and of itself, this fact actually suggests that the theoretical argument presented here has weight. If some platforms constrain content in specific ways, it is inevitable that content will have to differ across mediums.

[1] This includes announcement speeches from all candidates, all twenty debate transcripts (8 Democratic and 12 Republican), 12,026 Facebook posts and 29,479 Tweets.

Figure 1: Word Share Percentage by Candidate and Medium

Candidate	Total Words	Announce	Debate	Facebook	Twitter	Candidate % of Total Words
Bush	93777	2.21	21.38	24.34	52.07	8.27
Carson	177519	2.10	8.59	63.02	26.29	15.65
Clinton	162436	2.70	29.33	29.05	38.92	14.32
Cruz	158034	1.42	20.61	39.02	38.95	13.94
Kasich	128184	3.84	21.87	29.74	44.55	11.30
Rubio	113114	1.68	34.09	22.61	41.62	9.97
Sanders	164507	1.97	26.28	30.12	41.62	14.51
Trump	136406	4.17	29.80	22.71	43.31	12.03
Medium % of Total	NA	2.51	23.99	32.58	40.92	NA

Figure 1 presents some basic descriptive information on the corpora used in this paper. It demonstrates that, among the mediums in this paper, the majority of words come from online social media platforms. Some candidates, such as Ben Carson, are especially prolific on mediums like Facebook where he posted over 111,000 words. There are mediums that candidates use that are not represented here, television ads and other speeches prominent among them. Yet given the average person speaks at about 150 words per minute (Hulme et al., 1984), to equal the words on Facebook, Dr. Carson would have had to produce 740 minutes of television content. This means 740 individually different television ads (Kirmani and Wright, 1989). Then general takeaway, then, is that online platforms allow for more words to be produced than traditional venues and that campaigns seem to be producing far more online content than spoken words.

Figure 2: Most Common Words

Announcement	Debate	Facebook	Twitter
People	People	Will	Will
Will	Will	People	Today
America	Going	New	Great
Know	Know	Today	New
Can	Think	President	President
Country	Country	America	Now
One	Need	Great	Can
Going	Can	Can	People
President	Well	Hillary	Live
Need	Get	Join	Watch

We can also look at the most common words that are used across mediums. This will give us a sense of the general ideas and themes that exist in the platforms. For the most part, Figure 2 suggests that there is not a lot of variation within the most common words being used. This should not be that much of a surprise as some words are just more common in the English language than others. The two major differences that stand out are the presence of "Hillary" in Facebook posts as well as "live" and "watch" in Twitter. Secretary Clinton is actually largely responsible for using her own name, using it almost 1000 times in her posts.[2] The Twitter result can be explained by the propensity for candidates to put links to their news conferences, debates, YouTube clips, and ads in their posts.

5 Emotion and Mediums

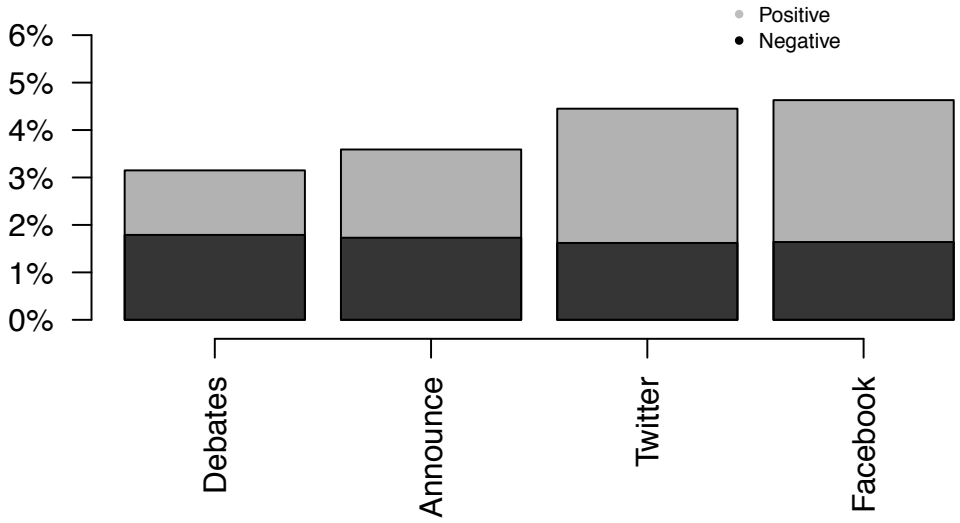
While the previous investigations do provide some basic understanding of the different mediums, a more complex analysis is needed to prove that mediation changes messages. Thus, going beyond word counts we can think about how the words that are used change based on the words that surround them. This is an important step in considering how a medium can fundamentally change a message. If candidates are talking about the same thing but applying different affect to their words as mediums change, then the message is necessarily different.

Applying the Linguistic Inquiry and Word Count's (LIWC) positive and negative emotion dictionaries (Pennebaker, et al. 2001)[3] provides an initial test of this theory. It is here that we start to see how mediums can change the message of politicians. Figure 3 highlights the key differences between positive and negative emotion words across mediums.

[2] As a point of comparison, Bernie Sanders uses his name 439 times in his posts and Donald Trump is at, somewhat unexpectedly, only 530 self-references.

[3] There are certainly other software packages available that do similar work, such as the General Inquirer (Stone et al., 1966) and SentiStrength (Thewall et al., 2010) among others.

Figure 3: Percentage of Emotion Words By Medium



Emotion word dictionaries do not capture the emphasis or tone in which they are delivered during an in-person medium, such as a debate or announcement speech.

These findings show that negative emotion words are a relatively stable percentage of words that are used in each medium (between 1.62% and 1.82%). However, live mediums (debates and announcement speeches) contain significantly fewer positive emotion words than online mediums. This finding appears to run counter to conventional wisdom on how debates play out. It is not that debates consist of candidates throwing around emotionally charged language and offering combative responses. Instead, relative to the other mediums presented here, we see a comparatively sedate and toned down corpus. Debates are just as negative as other platforms, but they are significantly less positive.

Results also find that the most emotionally charged commentary (positive and negative) is in Facebook and Twitter. Additionally, the data presented above was not "pre-whitened" in any way. This inevitably creates noise in the online mediums. Because hashtags and internet slang were left in as is and emojis were removed, the findings presented understate the emotion proportion of words used. This finding reflects scholarship that finds high levels of emotional content in social media platforms (Bollen et al., 2011; Woolley et al., 2010) as well as different sentiment in online mediums that we might expect (Groshek & Al-Rawi, 2013). Combined with the common word findings, the communication produced by the 2016 presidential candidates may be similar in the general theme, but there is variation in the affect attached to the messages. This matters not only for understanding how mediums change messages but also in the broader question of how medium specific messages could affect individual level preferences and ideologies.

6 Discussion and Future Work

As stated above, the goal is to extract meaningful differences in the produced politician communication. Therefore, the next steps for this project include expanding the corpora available as well as more advanced methodologies. As the 2016 presidential election cycle continues, data will be added as they are generated. Currently, additional Facebook data is being collected for more candidates as well as more speech transcripts and campaign coverage by major newspapers. In addition, available campaign television ads will also be transcribed. These additional data will provide more insight into the mediums in question.

Some of the key considerations for understanding how mediums alter content is to look at components of the text such as: specificity, inclusive/exclusive group language, and sentence complexity. As such, beyond the methods applied in this paper, the project could take advantage of standard supervised machine learning topic modeling as well as N-gram analyses as well as other tools that can help tease out the important implications of mediated communication.

All told, this initial inquiry suggests that mediums do matter for political communication. While it does not appear that the mediums systematically change the most common words,

findings here do suggest that at the very least the affect of statements differ across mediums. As such, political communications in the mediums presented here seem to contain the same general ideas while the words that surround those ideas have differing levels of affect. This matters as citizens react to emotion in political campaigning (Brader, 2004 & 2005). It is, thus, possible to claim that a citizen who is only exposed to one of these mediums is getting a very different set of messages than another citizen who only sees a different medium.

References

Brader, Ted (2005). Striking a responsive chord: How political ads motivate and persuade voters by appealing to emotions. *American Journal of Political Science*, *49*(2), 388-405.

Brader, Ted (2006). *Campaigning for hearts and minds: How emotional appeals in political ads work*. University of Chicago Press.

Bollen, Johan, Mao, Huina & Pepe, Alberto (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *ICWSM*, *11*, 450-453.

Evans, Heather. K., Cordova, Victoria & Sipole, Savannah. (2014). Twitter style: An analysis of how house candidates used twitter in their 2012 campaigns. *PS: Political Science & Politics*, *47*(02), 454-462.

Grimmer, Justin & Stewart, Brandon. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, mps028.

Groshek, Jacob & Al-Rawi, Ahmed (2013). Public sentiment and critical framing in social media content during the 2012 US presidential campaign. *Social Science* Computer *Review*, 0894439313490401.

Hulme, Charles, Thomson, Neil, Muir, Claire & Lawrence, Amanda (1984). Speech rate and the development of short-term memory span. *Journal of experimental child psychology*, *38*(2), 241-253.

Kirmani, Amna & Wright, Peter (1989). Money talks: Perceived advertising expense and expected product quality. *Journal of Consumer Research*, *16*(3), 344-353.

McLuhan, Marshall. (1994). *Understanding media: The extensions of man*. MIT press.

Murthy, Dhiraj & Petto, Laura R. (2015). Comparing Print *Coverage* and Tweets in Elections A Case Study of the 2011–2012 US Republican Primaries. *Social Science Computer Review*, *33*(3), 298-314.

Pak, Alexander & Paroubek, Patrick (2010, May). Twitter *as* a Corpus for Sentiment Analysis and Opinion Mining. In *LREc* (Vol. 10, pp. 1320-1326).

Pennebaker, James, Francis, Martha & Booth, Roger. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, *71*, 2001.

Stone, Philip. J., Dunphy, Dexter. C., & Smith, Marshall. S. (1966). The General Inquirer: A Computer Approach to Content Analysis.

Thelwall, Mike, Buckley, Kevan, Paltoglou, Georgio., Cai, Di, & Kappas, Arvid. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, *61*(12), 2544-2558.

Woolley, Julai. K., Limperos, Anthony. M., & Oliver, Mary Beth (2010). The 2008 presidential election, 2.0: A content analysis of user-generated political Facebook groups. *Mass Communication and Society*, *13*(5), 631-652.

Young, Lori & Soroka, Stuart (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication*, *29*(2), 205-231.

Topics and their Salience in the 2015 Parliamentary Election in Croatia: A Topic Model based Analysis of the Media Agenda

Damir Korenčić[1] **Marijana Grbeša-Zenzerović**[2] **Jan Šnajder**[3]
[1]Deparment Electronics, Ruđer Bošković Institute, Croatia
[2]Faculty of Political Science, University of Zagreb, Croatia
[3]Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia
damir.korencic@irb.hr grbesa@fpzg.hr jan.snajder@fer.hr

Abstract

There is a growing interest in automated content analysis for agenda-setting studies. While topic models have shown to be useful for this purpose, they are generally troubled with low topic quality and coverage. To alleviate this, Korenčić et al. (2015) proposed a semi-supervised topic modeling methodology. The aim of this work is to gain a better understanding of their methodology, by conducting a preliminary study of the media agenda during the 2015 parliamentary election in Croatia. Our goal is to analyze the topics and their salience during the official election campaign and the lively post-election negotiation period. We report on the methodological insights gained from this study and a preliminary analysis of the media agenda during the election period.

1 Introduction

Agenda-setting has been one of the most influential media effects theories for decades. Its underlying idea is that the media have the capacity to shape the public's perception of importance of particular issues (McCombs and Shaw, 1972; Scheufele, 2000). This effect is highly relevant during an election period, as it is widely acknowledged that the salience of media issues may influence voters' choices.

Agenda-setting studies often rely on quantitative content analyses of newspaper texts. In such studies, the media agenda is *measured* in terms of the salience of the issues: the newspaper documents are coded for issues, and the salience of each issue is taken to correspond to its frequency across the corpus. Recently, there has been a growing interest in the use of automated content analysis leveraging natural language processing techniques; in particular, *topic models* (Blei et al., 2003) have gained wide popularity. Existing studies from the domains of computer and political science demonstrate the usefulness of topic models for agenda analysis (Grimmer, 2010; Quinn et al., 2010; Kim et al., 2014). However, they also identify the problems related to topical coverage and quality (Chuang et al., 2013; Chuang et al., 2015), which may seriously hamper the validity of an agenda-setting study.

Recently, Korenčić et al. (2015) proposed a methodology based on topic modeling that mitigates the above deficiencies by using a semi-supervised, human-in-the-loop acquisition of topics, aiming for high-quality topics that better correspond to media issues. In a nutshell, the methodology consists of two steps: an agenda discovery step, in which topics are induced automatically and revised manually, and an agenda measuring step, in which the articles are tagged with topics. Korenčić et. al demonstrate that the approach can outperform a supervised classifier, while it additionally facilitates the discovery of topics. However, there is a number of non-trivial design choices associated with using their methodology, including technical (e.g., model parameters) and conceptual (e.g., the granularity and choice of topics) ones.

The aim of this paper is to put to practice the methodology of Korenčić et al., and gain an understanding of its advantages and potential caveats. To this end, we conduct a preliminary study on data collected during one of the most interesting periods in modern Croatian political history: the 2015 parliamentary election and the lively post-election period of negotiations on government formation. This paper makes two contributions: we report on (1) the methodological insights gained by applying this methodology and (2) a preliminary analysis of the media agenda during Croatian 2015 parliamentary election.

2 Corpus

We collected the data for our study from seven leading Croatian news sites: *Večernji list*, *Jutarnji*

Proceedings of the International Conference on the Advances in Computational Analysis of Political Text (PolText 2016), pages 48–54, Dubrovnik, Croatia, 14–16 July 2016.

list, *Slobodna Dalmacija*, *Glas Slavonije*, *T-portal*, *Novi List*, and *RTL Televizija*. We first selected the news feeds that correspond to domestic and regional news, and then collected the articles published during the official election campaign (from October 21st to November 6th, 2015), as well as in the period between the election day and the constitution of the Parliament (from November 8th to December 28th, 2015). We next removed very short texts (those with less than 40 alphanumeric tokens) and non-texts (error messages, subscription previews, photo galleries, etc.). Finally, we performed deduplication by using word-level edit distance to form groups of almost identical texts and keeping from each group only one text per news outlet. After filtering and deduplication, the final corpus consists of 15,394 news articles.

3 Agenda Discovery

The first step in the methodology of Korenčić et al. (2015) is *agenda discovery*. This is an exploratory step, whose purpose is to chart a wide range of *topics* present in the media. We use the term "topic" instead of "issue" to avoid misunderstandings that may stem from a narrow understanding of the term "issue", which is commonly employed in agenda-setting studies to denote policy issues such as health, defense, economy etc., as opposed to less substantive campaign contents; cf. (de Vreese, 2004; Zeh and Hopmann, 2013). In contrast, the term "topic" refers here to a broader range of different contents, varying from "issues" to more vague contents (such as intra-party conflicts and similar). Furthermore, we use the term *semantic topic*[1] to refer to topics as perceived by humans, including issues, processes, events, and entities. A semantic topic stands in contrast to a topic induced automatically by a topic model, to which we will refer as a *model topic*.

Ideally, model topics will correspond to semantic topics; in reality, however, model topics can contain noise or correspond to more than one semantic topic (Chuang et al., 2013). The objective of the agenda discovery step, then, is to detect the semantic topics and map them to model topics. To this end, we rely on human inspection of topics obtained by using several different models, each run on the same data. Namely, studies have shown that topics of a single model may not cover all semantic topics (Chuang et al., 2015). By analyzing the topics several models, we can compensate for the incomplete coverage of the individual models.

3.1 Topic models

To discover the semantic topics, we use the LDA topic model (Blei et al., 2003), available as part of the Gensim package (Řehůřek and Sojka, 2010). Text preprocessing consists of stop-word and non-word removal, and stemming using a Croatian stemmer of Ljubešić et al. (2007). Models are trained using a fast online learning algorithm (Hoffman et al., 2010). We set model hyperparameter $\alpha = 50/T$ (where T is the chosen number of topics), while we set $\beta = 0.01$ (Griffiths and Steyvers, 2004). Model learning parameters are set to $S = 1000, \tau_0 = 1.0, \kappa = 0.5$, as proposed by Korenčić et al. (2015). As input for the annotation process, we constructed three LDA models: two models with $T = 50$ topics (using different random seeds) and one model with $T = 100$ topics.

3.2 Semantic topic discovery

After obtaining the 200 topics from the three models, we presented the topics to seven human annotators: two authors and five master students of journalism. The annotators were instructed to perform a three-step annotation as follows. First, they were asked to deduce the meaning of the model topic by inspecting the list of words with high probability within the topic, and the list of news articles ranked by proportions of the topic within the article. After the first step, a number of semantic topics relating to the model topic were detected. In the second step, annotators consulted a shared list of extracted semantic topics to check whether the topics they detected have not already been detected by other annotators, and, if this was not the case, to add the topics to the list. Finally, the annotators used tags to link the model topics with the semantic topics, and vice versa, and also provided a short textual description for each model topic.

The actual annotation round was preceded by a training session, in which the annotation procedure was explained and demonstrated, followed by a test round and a discussion. Model topic inspection was performed with a GUI application deployed on a server and accessed via remote desktop clients. The annotators used the application to browse the topics and inspect the lists of words and news articles. Each annotator processed about 30 topics. The assignment balanced the topics across the three

[1] Korenčić et al. (2015) used the word "theme" for the same concept.

models. On average, an annotator spent 10 minutes on a single topic (min. 5.5 and max. 16.8). The total annotation effort was 33 person-hours.

3.3 Semantic topics revision

The procedure outlined above yielded 106 semantic topics. However, a closer inspection revealed errors in the annotations: some semantic topics were repeated, some were named ambiguously, while in some cases the link between the semantic topic and the underlying model topics was questionable. We speculate that the annotation quality could be ameliorated by investing more time in annotator training and by enforcing a more strict annotation procedure. We also observed that some topics, such as *weather reports* and *traffic disruptions*, while annotated correctly, are ultimately irrelevant for agenda analysis. For these reasons, we decided to carry out one additional revision round.

Another, less surprising finding was that the obtained topics are not mutually exclusive – rather, the topics are of different levels of abstractness and constitute a hierarchy. While inspecting the discovered topics and relations among them, we found it convenient to manually organize the topics into a taxonomy.[2] For instance, we put the semantic topic *election polls* under *election forecasts*, which, together with *election results*, we put under *electoral process*. We found that such a taxonomy was very useful for identifying and scoping the topics of interest. More concretely, we could use the taxonomy to chose a suitable level of topic granularity.

Topics were revised and organized in a taxonomy jointly by all three authors, which took about three hours. After the second round, a list of 71 semantic topics remained, organized in a taxonomy with the following 21 top-level categories: *prosecutions of public figures*, *post-election negotiations*, *foreign policy*, *terrorism and refugee crisis*, *Catholic Church*, *institution of the president*, *armed forces / Croatian army*, *electoral process*, *ecology*, *energetics*, *education*, *tourism*, *decentralization and reform of local and regional government*, *health care*, *media and journalists*, *trade unions and workers' rights*, *economy*, *intra-party conflicts*, *agriculture*, *brain drain and demography*, and *independent events*. The last category mostly pertains to specific events that occurred during the election campaign, but which do not fit well in any other category.

[2]We note that there exist models specifically designed for the extraction of topic hierarchies; e.g. (Griffiths et al., 2004).

4 Agenda Measuring

The detected semantic topics provide the analyst with a general overview of the media agenda. The next step is to measure the salience of the detected topics. For this preliminary study, we decided to focus on topics from two top-level categories: *electoral process* and *post-election negotiations*.

4.1 Defining custom semantic topics

We began our analysis with the inspection of the semantic topics in the two selected categories, using the same method as for the agenda discovery step. The inspection revealed that some topics overlap, while others seemed to be missing relevant content. We therefore decided to introduce new topics that better capture the issues of interest. We dub these topics *custom semantic topics*, as they are not the output of the topic discovery process, but were later constructed specifically for the purpose of agenda analysis. We defined six custom semantic topics of interest by combining existing semantic topics; an exception is the *party negotiations* topic, whose content we split into custom topics *negotiations* and *negotiations–substance*. The complete list of custom topics and semantic topics belonging to two selected categories is shown in Table 1.

What has become obvious at this point is the need for text exploration tools that would complement and improve exploration based on the inspection of topic models (browsing topic-related words and articles). We envisage that such tools would enable keyword-based text retrieval for a deeper exploration of semantic topics, text similarity-based search for tracing rare issues, ability to seed entirely new topics, as well as the interactive modification of topic models, perhaps along the lines of (Hu et al., 2014). We consider this an interesting challenge for future work.

4.2 Measuring topic salience

After defining the custom topics, we proceed to measure their salience by counting the news articles in our corpus that deal with this topic. We do this by tagging each document with custom semantic topics. This process is essentially used as a proxy for human coding of the articles with topics.

To perform the tagging, Korenčić et al. (2015) propose to build a new topic model specifically customized towards the topics of interest. Namely, when using a non-customized model, there is no guarantee that model topics produced by the

Top-level categories	Custom semantic topics	Semantic topics	Description
Electoral process	**election mathematics**	election forecasts, election polls, election results	pre- and post-election polls, speculation and statistics, forecasts, turnout, results, parliament combinatorics
	election procedures and regulation	election procedures and regulation, voting outside the place of residence, election rules and DIP, irregularities	election calendar, candidacy, monitoring, Ivan Turudić, electoral commission, candidates' debates, ethical commission, voting rules, irregularities
	election program and campaign	economic election program, media coverage of elections	communication and bickering of parties and politicians, election programs and campaigns
Post-election negotiations	**negotiations**	party negotiations, split within Most	negotiations and position taking, accusations and bickering, split within Most
	negotiations–substance	party negotiations	reform of local government, exclusive economic zone, economic and fiscal measures
	appointment of the PM designate and constitution of the Parliament	appointment of the mandate, presidential consultations, constituting the parliament	legal procedure and political process of the PM candidate appointment and constituting the Parliament

Table 1: List of semantic topics and derived custom semantic topics for the selected top-level categories

stochastic inference procedure will match any of the semantic topics of interest. Even if they would match to a certain extent, one would still have to manually inspect them and map to semantic topics.

Model customization is achieved by constructing, for each semantic topic, a list of *seed words* – words highly indicative for that topic. Once we have such lists, the model is built with probabilistic priors set to enforce topics that assign high probability to seed words. The underlying idea is that models built in this way will produce topics that correspond to custom semantic topics we are interested in. We follow the procedure of Korenčić et al. (2015) for obtaining the seed words: for each custom semantic topic, we inspect the list of highly probable words for that topics, and for each such word, we inspect a list of news articles estimated as related to it by a word-article association measure. If the majority of articles indeed deal with the considered topic, we add the word to the seed words list for that topic. Table 2 shows the seed words for the considered custom topics.

Custom semantic topics	Seed words
election mathematics	mandate, result, poll, win, vote, voter, exit, preferential, turnout, advantage, constituency
election procedures and regulation	committee, DIP, donation, report, spend, donate, promotion, GONG, financing, law, electoral silence, violation, campaign, debate, complaint, observer
election program and campaign	economic, program, VAT, promise, electoral, termination, Prnjavor, demographic, irrigation, debt
negotiations	Petrov, negotiation, Božo, Prgomet, meeting, non-party, independent, Petrina, Drago, key, Grmoja, tripartite, reply, support, forming, pressure
negotiations–substance	reform, local, self-governance, belt, devalvation, inflation, rationalization, model, termination, Lovrinović
appointment of the PM designate and constitution of the Parliament	PM-designate, signature, consultations, forming, Pantovčak, round, session, Reiner, constitutive, convocation, elected

Table 2: Seed words for the chosen topics

For document tagging, we follow the procedure outlined by Korenčić et al. (2015): for each news article, using the customized model, we first infer the topic probability distribution, and then tag the article with the semantic topic corresponding to its most probable model topic.

An important insight we gained in this step is that some semantic topics are difficult to detect using topic models. For rare issues, custom topic model seeded with issue-specific words will produce a topic that includes other similar topical content, ultimately decreasing the tagging precision. We believe that a better alternative to detecting such topics is to describe them with a set of discriminative keyphrases, similarly to traditional dictionary approach to coding (Krippendorff, 2012). A case in point are the *Ljubljana Bank* and *voting outside the place of residence* topics, for which tagging based on a boolean keyword-based query fared much better than tagging using a customized topic model.

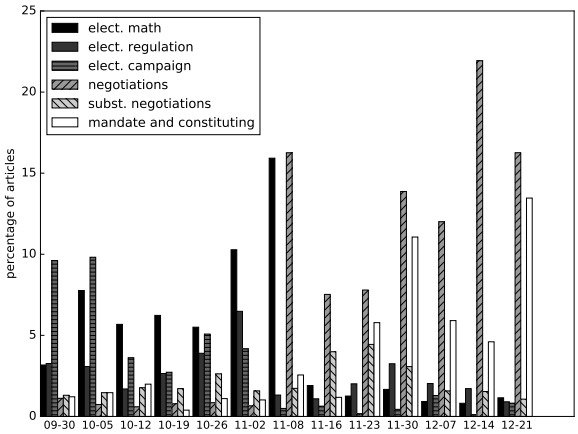


Figure 1: Electoral process and negotiations

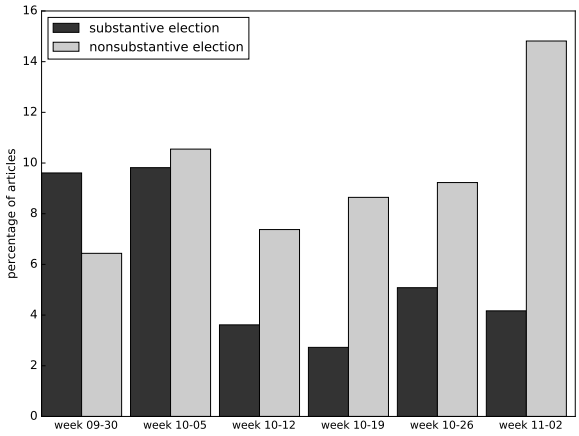


Figure 2: Substantive vs. non-substantive political topics in the pre-election period

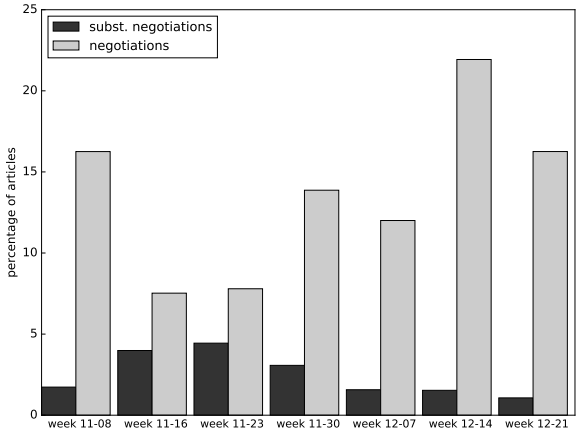


Figure 3: Substantive vs. non-substantive negotiation topics in the post-election period

5 Preliminary Analysis

In this section we present the results of using our model to conduct a preliminary analysis of the media agenda during Croatian 2015 parliamentary election.

5.1 Topic-event correlation

We note that validity is an important consideration when applying automated content analysis for political science research (Grimmer and Stewart, 2013; Lacy et al., 2015; Zamith and Lewis, 2015). While a thorough investigation of validity of our approach does not fit the scope of this paper, we ran a sanity check by analyzing how the inferred salience of topics correlates with real-life events.

In Fig. 1 we show the frequency of the articles across the six topics we considered. We find that the salience of semantic topics (defined as the number of articles tagged with the semantic topic) is very well correlated with real-life events. This correlation confirms the predictive validity (Grimmer and Stewart, 2013) of the model. Concretely, the *election mathematics* topic (including contents such as poll results, prediction of winners and losers, election results, etc.) was very salient in the week preceding the election day, and rocketed on the election day (November 8th).

Further evidence in support of the validity can be found by considering the events that took place after the election day. As none of the parties won the majority necessary to form the Government, both major parties – Social Democratic Party (SDP) and Croatian Democratic Union (HDZ) – tried to win over the newly established party of Most (The Bridge), which won a significant number of seats. Negotiations between the parties got excessive media coverage, which is successfully registered by our topic model. Furthermore, the week-to-week salience of the *negotiations* topic, also captured by our model, corresponds to the real-life events that triggered the visibility of this topic in the public discourse (bargains and disputes between the parties, search for the PM designate who would prompt Most to support one of the major parties, etc.). The same goes for the topic *appointment of the PM designate and constitution of the Parliament*, which was the most prominent in the days preceding the finally arranged constitution of the Parliament and formation of the Government.

5.2 "Game-schema" coverage

After the sanity check, we turned to a more insightful analysis from a political communication perspective. Building on the acknowledged distinction between substantive and less substantive elec-

tion coverage (cf., for instance, Zeh and Hopmann (2013)), we divided the semantic topics into "substantive" and "non-substantive" ones. For the pre-election period, we categorize *election mathematics* and *election procedures and regulation* as non-substantive topics, and *election program and campaign* as a substantive topic. For the post-election period, we differentiated between no-substance negotiation topics (such as conflicts between parties, bargains, etc.) and substantive negotiations that evolved around certain policy issues (cf. Table 1). Figures 2 and 3 show the week-to-week salience of these topics in the pre-election and post-election period, respectively.

The interesting finding is the clear dominance of the "non-substantive" content over the "substantive" content during the pre-election period. This primarily refers to the dominance of articles that focused on "horse-race" issues (e.g., opinion polling, who's ahead and who's behind, prediction of results) and the campaign hoopla, as opposed to articles that covered election programs and similar. Expectedly, the gap between substantive and non-substantive content was widening as the election was approaching. Interestingly, Figure 3 shows that this discrepancy was even stronger in the post-election period, suggesting that during the negotiation process media were more interested in parties' political bargain than in substantial content of negotiations. Whether this is due to media's interest in hoopla or due to the fact that politicians did not put substantial issues on the table is of course not the focus of this study. Overall, the analysis reveals the dominance of the "game schema" over more issue-centered information in the media coverage of elections, already witnessed in a number of countries (Patterson, 1993; Strömbäck and Dimitrova, 2006; Zeh and Hopmann, 2013).

6 Conclusion

We used a semi-supervised topic modeling methodology of Korenčić et al. (2015) to carry out a preliminary study of the media agenda during the 2015 parliamentary election in Croatia. The methodology consists of agenda discovery, in which model topics are manually mapped to semantic topics, and agenda measuring, in which news articles are tagged with topics. The primary purpose of our study was to gain a better understanding of the entire modeling process. In the agenda discovery step, the main methodological insights we gained is the need for a stricter annotation procedure and the importance of constructing a taxonomy of topics. In the agenda measuring step, we found the need for exploratory tools that would complement and improve the inspection of topic models and learned that some topics might be detected more precisely using keyphrases rather than topic model coding.

In a preliminary analysis of the media agenda, we were able to confirm the predictive validity of our model. Furthermore, we demonstrated the applicability of topic modeling by investigating the presence of substantive vs. non-substantive contents in the media coverage of the election. The results corroborate the common established assumption about the rise of game-oriented coverage at the expense of issue-related contents. It should be noted, however, that the findings presented here are just a few of the results that were obtained using topic modeling analysis, as a more detailed report would exceed the scope of this paper.

For future work, we plan to devise a stricter annotation procedure based on cross-checking, and test it using topic quality and coverage as the criteria. We also intend to experiment with text exploration tools complementary to topic models. Future validation should include more rigid quantitative validation measures and comparisons with the findings obtained by human coding. Finally, future research of media election coverage should be more focused in scope and include only articles specifically pertaining to election. This would weed out redundant content and may yield even more insightful results.

Acknowledgments

The first author has been supported by the Croatian Science Foundation project number 9623 and the The Croatian Policy Agendas Project[3] funded by the European Social Fund. We thank the students for their help in annotating the data: Frane Basioli, Doris Berečić, Janja Jagić, Mila Kovačević, and Karla Milevoj. We also thank the TakeLab[4] team for their technical assistance.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Jason Chuang, Sonal Gupta, Christopher Manning, and

[3] www.cepis.hr
[4] takelab.fer.hr

Jeffrey Heer. 2013. Topic model diagnostics: Assessing domain relevance via topical alignment. In *Proceedings of ICML*, pages 612–620. JMLR Workshop and Conference Proceedings.

Jason Chuang, Margaret E. Roberts, Brandon M. Stewart, Rebecca Weiss, Dustin Tingley, Justin Grimmer, and Jeffrey Heer. 2015. TopicCheck: Interactive alignment for assessing topic model stability. In *Proceedings of NAACL-HLT*, pages 175–184. ACL.

Semetko Holli A de Vreese, Claes H. 2004. *Political campaigning in referendums: Framing the referendum issue*. Routledge.

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235.

Thomas L. Griffiths, Michael I. Jordan, Joshua B. Tenenbaum, and David M. Blei. 2004. Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems 16*, pages 17–24. MIT Press.

Justin Grimmer and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, pages 1–31.

Justin Grimmer. 2010. A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1–35.

Matthew Hoffman, Francis R. Bach, and David M. Blei. 2010. Online learning for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems 23*, pages 856–864. Curran Associates, Inc.

Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. 2014. Interactive topic modeling. *Machine Learning*, 95(3):423–469.

Yeooul Kim, Suin Kim, Alejandro Jaimes, and Alice Oh. 2014. A computational analysis of agenda setting. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, pages 323–324. ACM.

Damir Korenčić, Strahil Ristov, and Jan Šnajder. 2015. Getting the agenda right: Measuring media agenda using topic models. In *Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications*, pages 61–66. ACM.

Klaus Krippendorff. 2012. *Content Analysis: An Introduction to its Methodology*. Sage.

Stephen Lacy, Brendan R Watson, Daniel Riffe, and Jennette Lovejoy. 2015. Issues and best practices in content analysis. *Journalism & Mass Communication Quarterly*.

Nikola Ljubešić, Damir Boras, and Ozren Kubelka. 2007. Retrieving information in Croatian: Building a simple and efficient rule-based stemmer. *Digital Information and Heritage*, pages 313–320.

Maxwell E McCombs and Donald L Shaw. 1972. The agenda-setting function of mass media. *Public Opinion Quarterly*, 36(2):176–187.

Thomas E Patterson. 1993. Out of order: How the decline of the political parties and the growing power of the news media undermine the american way of electing presidents. *New York: Alfred Knopf*.

Kevin M Quinn, Burt L Monroe, Michael Colaresi, Michael H Crespin, and Dragomir R Radev. 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209–228.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. University of Malta.

Dietram A. Scheufele. 2000. Agenda-setting, priming, and framing revisited: Another look at cognitive effects of political communication. *Mass Communication and Society*, 3(2–3):297–316.

Jesper Strömbäck and Daniela V Dimitrova. 2006. Political and media systems matter: A comparison of election news coverage in Sweden and the United States. *The Harvard International Journal of Press/Politics*, 11(4):131–147.

Rodrigo Zamith and Seth C Lewis. 2015. Content analysis and the algorithmic coder: What computational social science means for traditional modes of media analysis. *The ANNALS of the American Academy of Political and Social Science*, 659(1):307–318.

Reimar Zeh and David Nicolas Hopmann. 2013. Indicating mediatization? Two decades of election campaign television coverage. *European Journal of Communication*, 28(3):225–240.

Tracking Political Reputation with Distributional Semantic Models

Andrei Kutuzov
University of Oslo
Postboks 1080 Blindern 0316
Oslo, Norway
andreku@ifi.uio.no

Aleksander Bai
Oslo and Akershus
University College of Applied Sciences
Institute of Information Technology
aleksander.bai@hioa.no

Abstract

In this paper, an unsupervised approach to tracing sentiment dynamics for named entities is presented. It is based on training successive word embedding models, updating them with new textual data (for example, daily news flow). Then, for the named entities under analysis, semantic distances towards evaluative adjectives are measured in the trained models, providing data about typical features associated with the given entity. Examples of applying this approach to English and Russian material are described.

1 Introduction

Sentiment analysis is a well-established field within natural language processing (Pang and Lee, 2008). It is aimed at automatically detecting the tonality of people's opinion about some entity. One of frequent use cases for sentiment analysis is tracking how the reputation of a person (for example, a politician) changes over time. The presented paper is dedicated to this particular application.

We propose to extract sentiment dynamics from mass media texts, for example news pieces. For this purpose we employ distributional semantics and machine learning, particularly neural embedding models, which have gained considerable attention in the recent years.

We present a comparatively fast and simple method of calculating how the public opinion trends change for the particular named entities, using sets of evaluative words as 'anchors'. Essentially, our idea is that one can describe sentiment towards a word or a multi-word entity in a distributional model as a set of distances between this word's semantic representation and the semantic representations of a limited number of evaluative words (mostly adjectives). A naive example would be the entity '*Adolf Hitler*', with the semantic representation in the model much closer to the representation of '*bad*' than to '*good*', and the entity '*Albert Einstein*' closed to the representation '*smart*' than to '*dumb*'. Such representations can then be compared across models trained on different time periods to trace sentiment dynamics. We have tested and verified the method on both English and Russian text corpora.

2 Related Work

The conceptual basis for neural embeddings is distributional semantics with its main idea: meaning is a sum of contexts (Firth, 1957). It means that it is possible to represent words or multi-word entities (including named entities, such as persons) with vectors of their contexts. Co-occurrence data is typically mined from large text corpora, supposedly representing language as a whole.

Prediction-based distributional models (the most famous tool in the field is arguably *word2vec*[1], implementing *Continuous Skipgram* and *Continuous Bag-of-Words* learning algorithms) attempt to learn optimal lexical vectors (embeddings) by predicting words based on their contexts (Mikolov et al., 2013). Their objective function causes words that occur in similar contexts to learn similar embeddings during the training process, which allows finding the 'nearest associates' for any given word, as well as distances between pairs of words.

Such models have been shown to outperform more traditional count-based models (Baroni et al., 2014), and are increasingly widespread in natural language processing applications requiring semantic representations. In this research, we augment the concept of neural embeddings with the idea of successively updating distributional models with new textual data (see Section 4).

[1] https://code.google.com/archive/p/word2vec/

Proceedings of the International Conference on the Advances in Computational Analysis of Political Text (PolText 2016), pages 55–60, Dubrovnik, Croatia, 14–16 July 2016.

Note that word embeddings are now sometimes also used as a source of data for compiling polarity lexicons for sentiment analysis techniques; see, for example, (Pablos et al., 2016) and (Castellucci et al., 2016). However, we use polarity lexicons available from previous work to select our evaluatDive words, so we avoid influencing the evaluation words by the corpa itself.

3 Description of corpora and polarity lexicons

To experiment with tracing sentiment changes, we use lemmatized corpora of English and Russian news text, where each word is annotated with its part of speech (PoS) to help disambiguation for distributional models.

The English corpus consists of *The Signal Media Dataset*[2], which contains 265,512 blog articles and 734,488 news articles from September 2015. The size of the corpus (after lemmatizing and removing stop words) is 238,822,447 words.

The second corpus is the collection of news articles in Russian, also published in September 2015. It contains about 500,000 texts extracted from about 1000 Russian-language news sites. The size of the corpus (after lemmatizing and removing stop-words) is 59,167,835 words.

As for polarity lexicons, as already said, we employed ready-made vocabularies from previous work. For the English part, we used the *AFINN-111* lexicon (Nielsen, 2011), which has 2477 words and phrases separated into positive, negative and ambiguous subsets. Each word is rated between minus five (negative) and plus five (positive) and has been labeled manually by (Nielsen, 2011). We removed the words that had a frequency of less than 50 instances per million word tokens, and removed words that were not adjectives. There are for instance many verbs in the AFINN-111 that have a positive or negative polarity, but which are not suited to describe a politician's reputation (like *cutting* or *wow*). In addition we also removed words with a small absolute polarity score (between +2 and -1) before we ended up with 20 positive words and 16 negative words. Below are some examples of the words from the positive and negative lexicon:

1. *like* (positive)
2. *good* (positive)
3. *great* (positive)
4. *important* (positive)
5. *kind* (positive)
6. *hard* (negative)
7. *bad* (negative)
8. *difficult* (negative)
9. *wrong* (negative)
10. *limited* (negative)
11. ...

For Russian we used *RuSentiLex* vocabulary (Loukachevitch and Levchik, 2016). From there, we extracted all adjectives annotated as positive or negative. Then we removed those that are either not present in the Russian corpus, or their frequency there is less than 50 instances per million word tokens. The reason for this is that we want to operate with the words for which our distributional modes have enough co-occurrence data to train meaningful word embeddings. This left us with lexicons of several dozens adjectives for each polarity. We manually filtered out words which were clearly not fit for describing a politician's reputation, so the final negative lexicon contained 12 words and the positive lexicon contained 25 words. Below is the sample of words from the resulting lexicon:

1. добрый *kind* (positive)
2. достойный *worthy* (positive)
3. красивый *beautiful* (positive)
4. любимый *lovely* (positive)
5. мирный *peaceful* (positive)
6. независимый *independent* (positive)
7. плохой *bad* (negative)
8. преступный *criminal* (negative)
9. слабый *weak* (negative)
10. террористический *terrorist* (negative)
11. ...

[2] http://research.signalmedia.co/newsir16/signal-dataset.html

4 Updating models

Prediction-based distributional models can be updated with new co-occurrence data in a straightforward way. Note that this is usually not the case with count models which demand computationally expensive calculations (for example, very large matrix factorization in the SVD algorithm) each time new texts are added.

This simplifies our aim of tracking reputation changes over time. Given a permanent stream of media texts mentioning politicians, it is possible to constantly update or re-train our base model (trained on some large 'reference' corpus) with new textual data, thus introducing temporal dimension into word vectors. New events in the world cause shifts of associations in the minds of text producers (journalists, bloggers, etc). These shifts are reflected in changing frequencies of typical words co-occurring with this or that named entity. As the model is being updated, these new contexts cause word vectors to "drift" and adapt to new training data, as described in, for example, (Kulkarni et al., 2015). This, in turn, results in changing word positions in vector space, related to other words.

As already said, we need a 'reference' or 'baseline' model which aims to mimic some background knowledge, before the model is exposed to daily updates. For English, we used the British National Corpus[3] (about 50 million words) to train this reference model, while for Russian it was the corpus of news articles published in the months preceding September 2015, precisely June, July and August (taken from the same source as the September articles). This corpus contains about 250 million words. Similar reference models were used in previous research (Kutuzov and Kuzmenko, 2016), and showed acceptable performance in detecting semantic shifts over time. In fact, one can use any suitable corpus for the purposes of training the reference model, like Wikipedia or other freely available large text collections.

Then, *Continuous Bag-of-Words* (Mikolov et al., 2013) models were trained for both corpora, using negative sampling with 10 samples, vector size 300, symmetric window size 20 and 5 iterations. Words with frequency less than 10 were ignored during training. For the training itself and other operations with the models, we employed the *Gensim* library (Řehůřek and Sojka, 2010).

After that, we successively updated these models with texts released in the several-days-long time periods belonging to September 2015. Granularity of 2 or 3 days was chosen in order to enlarge the amount of data fed to models: for example, some one-day Russian corpora corresponding to weekends contained only several thousand words. For this reason, we additionally tried to include week-ends in 3-days periods, to make news stream more evenly distributed. As a result, average time period size in tokens was 18,774,000 for English data and 5,332,000 for Russian data.

We once again emphasize that our models were not re-trained from scratch with new texts added from new corpora. Instead, we continued training the same baseline model, gradually updating word vectors with new contexts. All interim states were saved as separate models, and in the end we had several models for each language, reflecting successive time periods

5 Reputation as geometrical location in the embedding space

We extracted named entities for the English model using NLTK[4] PoS tagger and named entities classifier (Bird et al., 2009). Then, we subjected the top 50 entities by frequency to manual selection. After removing irrelevant entities, we ended up with 39 entities that represent either politicians (like *Donald Trump* and *David Cameron*) or entities important for political news (like *Afghanistan* and *Syria*).

For Russian, we experimented with a set of politicians' names extracted from the annotated dataset provided by the organizers of *FactRuEval-2016* shared task (Starostin et al., 2016). As it consists mainly of contemporary news texts, the most frequent person names there were politicians, and we had only to filter out a few erratic annotations. In the end, we've got a list of 50 most frequent names, starting with Vladimir Putin, Petr Poroshenko and Barack Obama. Note that for multi-word entities to be properly handled by our models, we pre-processed the training corpora and joined parts of these compounds into one token (for example, *'barack_NOUN obama_NOUN'* became *'barack::obama_NE'*, where NE is '*named entity*').

[3] http://www.natcorp.ox.ac.uk/

[4] http://www.nltk.org

Figure 1: Expression of different sentiment features (evaluative adjectives) for David Cameron in the English model.

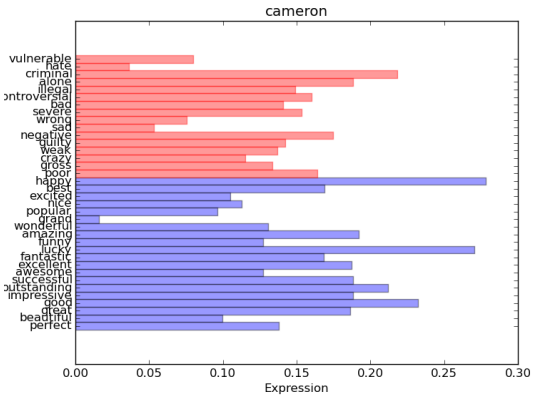
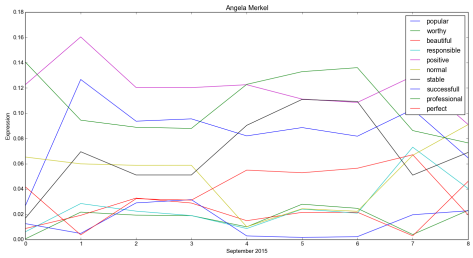


Figure 2: Temporal dynamics for expression of different sentiment features (evaluative adjectives) for Angela Merkel in the Russian models



We then applied simple sentiment analysis techniques to trace how sentiment of named entities denoting politicians change over time period represented by the models described in the previous section. In the space of a distributional model, it means that the vectors of these named entities 'move' closer to or farther from a set of evaluation vectors from our polarity lexicons. For example, for a given period (and a given sequence of temporal models), a given person can be described as getting more and more positive ('*good*') coverage in the media, but at the same time becoming more and more associated with something to laugh at, etc.

In this way, one can track not only changes of sentiment itself, but also the details or dynamics of sentiment changes. It is done through analysis of the drift in the 'nearest associates' set: we can see what evaluative adjectives moved closer to the named entity we are interested in. This makes our approach even more reliable when there is the need to know the cause of sentiment dynamics.

We work with 3 'abstraction levels', on which a user can overview sentiment trends for a particular politician or for a list of politicians:

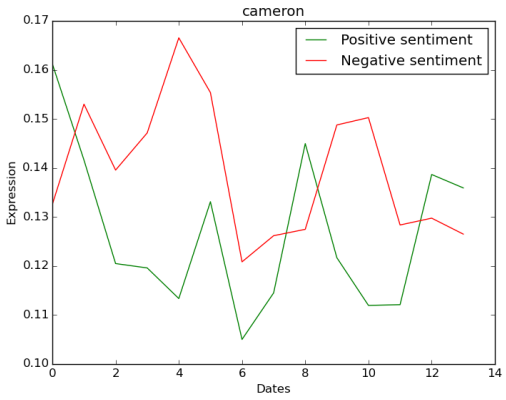
1. **Detailed**: the degrees of expression for all positive and negative features are given. This data can be visualized in two ways:
   (a) for one particular time period (see Figure 1); Here an analyst can check which positive and negative features a particular person was associated with for a given period. Each feature's value in itself can be either positive or negative, and the stronger a feature is associated with the person, then higher the expression is. Here positive features are shown in blue bars, while negative features are shown in red bars. It means it is easy to see what aspects of the sentiment that requires attention.
   (b) for several periods, showing dynamics for all the features; this type of abstraction is show in Figure 2. It is highly detailed and can be used to estimate the temporal tendencies for the distinct features over time.

2. **Coarse**: the averaged degrees of positive and negative features are given (2 numbers for each time period), see Figure 3; this level of analysis abstracts from the particular features and instead provides dynamics of positive and negative attitude towards a named entity (how strongly it is associated with positive and negative adjectives in general).

3. **General opinion**: only a total score is given that reflects the dynamics of general attitude towards the named entity (calculated as difference between averaged positive and negative degrees). An example of visualization for this level of abstraction is given in Figure 4. On this plot, among other trends, one can see how the Ukrainian politician Yulia Timoshenko received more and more negative attitude in Russian mass media (was associated with negative features) in September 2015.

By using the different abstraction levels (detailed, coarse, general) it is feasible to fine-tuned analysis of the sentiment data, when one can 'zoom in' or 'zoom out' depending on the particular research aim.

Figure 3: Coarse positive and negative sentiment dynamics for David Cameron in the English models (September 2015)



For instance, as shown in Figure 1 it's obvious that certain adjectives are more strongly associated with David Cameron than others. He is generally considered a smiling guy which is reflected by a strong associationed with the feature *happy*. There is also a strong association with the negative feature *criminal*, which might be explained by the accusation of a criminal offense by a former friend of Cameron. This attracted much negative publicity in September 2015 for David Cameron, and is clearly picked up by the sentiment analysis.

By *zooming out* to a more coarse overview, this negative publicity for David Cameron can also be illustrated in Figure 3 where there is a spike in positive sentiment around 16-17 September 2015, before a drop in positive sentiment a week later. The first positive shift is around the time David Cameron answered questions from the public[5], and this was considered a positive event for Cameron that might explain the shift in positive sentiment. However, shortly after this successfull event there was an accusation by a former friend about David Cameron having performed an criminal offense[6]. It is apparent from the coarse overview in Figure 3 that this lead to a drop in positive sentiment and a rise in negative sentiment.

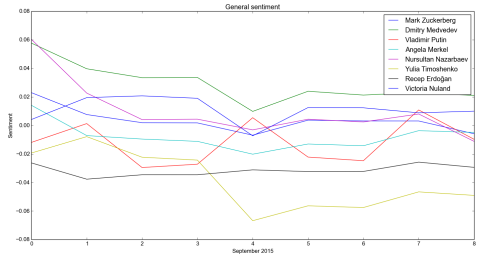
6 Conclusion

We have presented our data and methods, providing examples drawn from experiments on large corpora of English and Russian news texts. Our 'embedding sentiment' approach is completely unsupervised, requiring only substantial amounts of relevant texts and the lists of named entities that are of interest. The ready-made lists of evaluative adjectives (polarity lexicons) are available for most languages and can be easily adapted depending on what features are under analysis. Our outlined methods provide large visualization potential as well, and it is feasible to see sentiment trends on either a detailed, coarse or general level.

At the same time, there is still room for improvement, especially concerning proper algorithm evaluation and comparison with other established sentiment analysis techniques.

[5]http://www.bbc.com/news/uk-politics-34264683
[6]http://www.bbc.com/news/uk-politics-34312744

Figure 4: General sentiment dynamics for a sample of named entities in the Russian models (September 2015)



References

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 238–247.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*. O'Reilly Media, Inc.

Giuseppe Castellucci, Danilo Croce, and Roberto Basili. 2016. A language independent method for generating large scale polarity lexicons. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

John Firth. 1957. *A synopsis of linguistic theory, 1930-1955*. Blackwell.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant de-

tection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635. International World Wide Web Conferences Steering Committee.

Andrey Kutuzov and Elizaveta Kuzmenko. 2016. Cross-lingual trends detection for named entities in news texts with dynamic neural embedding models. In *Proceedings of the First International Workshop on Recent Trends in News Information Retrieval co-located with 38th European Conference on Information Retrieval (ECIR 2016)*, pages 27–32.

Natalia Loukachevitch and A. Levchik. 2016. Creating Russian sentiment lexicon. In *Proceedings of OSTIS-2016 conference*, pages 377–382.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Finn Årup Nielsen. 2011. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Aitor García Pablos, Montse Cuadros, and German Rigau. 2016. A comparison of domain-based word polarity estimation using different word embeddings. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA.

A. S. Starostin, V. V. Bocharov, S. V. Alexeeva, A. A. Bodrova, A. S. Chuchunkov, S. S. Dzhumaev, I. V. Efimenko, D. V. Granovsky, V. F. Khoroshevsky, I. V. Krylova, M. A. Nikolaeva, I. M. Smurov, and S. Y. Toldova. 2016. Factrueval 2016: Evaluation of named entity recognition and fact extraction systems for russian. In *Computational Linguistics and Intellectual Technologies: papers from the Annual conference 'Dialogue'*, pages 702–721.

TopFish: Topic-Based Analysis of Political Position in US Electoral Campaigns

Federico Nanni[1], Cäcilia Zirn[1], Goran Glavaš[1,3], Jason Eichorst[2], Simone Paolo Ponzetto[1]
[1] Data and Web Science Group, University of Mannheim
B6 26, DE-68161 Mannheim, Germany
[2] Collaborative Research Center SFB 884, University of Mannheim
L13 15-17, DE-68161 Mannheim, Germany
[3] Text Analysis and Knowledge Engineering Lab, University of Zagreb
Unska 3, HR-10000 Zagreb, Croatia
{federico,caecilia,goran,simone}@informatik.uni-mannheim.de
eichorst@uni-mannheim.de

Abstract

In this paper we present TopFish, a multi-level computational method that integrates topic detection and political scaling and shows its applicability for a temporal aspect analysis of political campaigns (pre-primary elections, primary elections, and general elections). It enables researchers to perform a range of multidimensional empirical analyses, ultimately allowing them to better understand how candidates position themselves during elections, with respect to a specific topic. The approach has been employed and tested on speeches from the 2008, 2012, and the (ongoing) 2016 US presidential campaigns.

1 Introduction

The competition for votes in US elections provides an opportunity for candidates to communicate their positions. Evidence suggests that campaign statements are designed to inform voters of the types of policy a candidate will pursue in legislative (Ringquist and Dasse, 2004) and executive offices (Marschall and McKee, 2002).
Converging on a position, however, is a complicated process. Candidates must not only satisfy the interests of voters in the general election, but also win in primary elections where party identification is shared among candidates and support is ultimately won from informal organizations within the party (Masket, 2009).

Adequately capturing this process, namely the development of candidates' positions and reputations in campaigns is a challenging empirical problem that relies on processing large amounts of political texts. Significant advancements in quantitative methods from the field of natural language processing (NLP) have enabled coarse-grained analyses of texts produced in presidential campaigns (Medzihorsky et al., 2014; Sim et al., 2013; Gross et al., 2013). However, in all of these works positions are analysed based on the content of the whole documents. Put differenly, there is still an empirical gap with respect to fine-grained analysis of politicians positions towards particular topics and how these topically-bounded positions change over time.

In this paper, we present *TopFish*, a computational method that (1) identifies parts of public campaign speeches that correspond to topics of interest and (2) determines candidates positions specifically towards each of these topics. TopFish combines a topical classifier following the idea of our previous work on party manifesto classification (Zirn et al., 2016) and the Wordfish tool (Slapin and Proksch, 2008), which is commonly used for quantitatively estimating candidate positions in political science analyses (Grimmer and Stewart, 2013).

In order to show why there is the need for a more fine-grained position analysis on topic level, we apply TopFish to speeches delivered in presidential election campaigns. In a qualitative analysis, we discuss how candidates' positions do not only vary with respect to topics, but how they also change in different phases of an election campaign. In other words, we show how some topic-based positions of some candidates change from pre-primaries, over primaries, to general election.

The approach we present is weakly-supervised because it depends on an appropriate topic-labeled dataset, yet it does not require any manual annotations for positions themselves. Therefore, it can be easily applied to other types of political texts

Proceedings of the International Conference on the Advances in Computational Analysis of Political Text (PolText 2016), pages 61–67, Dubrovnik, Croatia, 14–16 July 2016.

such as online discussions or debate transcripts.

2 Related Work

During the last decade, there has been a consistent growth in application of natural language processing (NLP) methods in political science research (Grimmer and Stewart, 2013). Here we cover the most relevant lines of work.

Topic detection in political text. The detection of topics in political documents has been performed adopting unsupervised techniques such as latent semantic analyses (LSA) (Hofmann, 1999) and latent dirichlet allocations (LDA) (Blei et al., 2003) as well as supervised adaptations like Supervised LDA (sLDA) (Mcauliffe and Blei, 2008) and labeled LDA (lLDA) (Ramage et al., 2009). For example, (Quinn et al., 2010) present a method that estimates a hierarchical structure of topics in political discussions, while Balasubramanyan et al. (2012) describe an adaptation of sLDA for studying the topic-based polarization of debates in the US and Gottipati et al. (2013) explore the potential of Debatepedia for determining political topics and positions. Zirn and Stuckenschmidt (2014) propose a method for analyzing and comparing documents according to a set of predefined topics based on lLDA, while Nanni and Fabo (2016) combine entity linking (Rao et al., 2013) and labeled LDA in order to overcome the most common limitation of unsupervised topic modeling techniques, namely the interpretability of the results.

Fully supervised approaches for topic detection have been also performed (see for example Hillard et al. (2008)). However, as these solutions rely on expert knowledge for establishing in advance a set of relevant topics and on annotating a large set of training data, they generally are more time-consuming to build. In contrast, we show that for our approach a small set of annotated data is enough, and we explore the use of external annotated training sources.

Political position scaling. While there has been a long term interest in modelling ideological beliefs using automated systems (see for example Abelson and Carroll (1965)), only in recent years we have seen a growth of advanced computational techniques for performing the task. In 2003, Laver, Benoit and Garry presented Wordscores (Laver et al., 2003), a supervised approach that relies on a set of pre-defined reference texts to determine the position of political documents in space. Inspired by it, in 2008 Slapin and Proksch developed Wordfish (Slapin and Proksch, 2008), a completely unsupervised solution for scaling documents on a single dimension.

The techniques presented above analyse coarse-grained political positions on document level and do not fully exploit the potential of topic-based political scaling.

Text-based analyses of political campaigns. In the last decade, computer-based analysis of political campaigns has attracted the attention of journalists (Silver, 2012) and academics (Foot et al., 2003). Scharl and Weichselbraun (2008) studyied trends in political media coverage before and after the 2004 U.S. presidential election applying NLP methods. Recently, Prabhakaran et al. (2014) studied the topic dynamics of interactions during the 2012 Republican presidential primary debates. Transcriptions of speeches have been employed by Gross et al. (2013) adopting the method presented in (Sim et al., 2013) to study the US 2008 and 2012 campaigns and in particular to test the Etch-a-Sketch hypothesis[1]. We will address the same hypothesis in our qualitiative evaluation part in subsection 4.2.

3 Topic Detection and Scaling

In this section, we describe in detail the two steps of TopFish, which consists of identifying the topics in the speeches and separately scaling the topic-specific positions based on parts of text belonging to a particular topic of interest.

3.1 Identification of topics in speeches

In the first step, our goal is to identify the topics that are discussed in the collected candidate speeches. We decide to use the classification scheme developed by the Comparative Manifesto Project (Volkens et al., 2011), which distinguishes between seven topical domains: *External Relations*, *Freedom and Democracy*, *Political System*, *Economy*, *Welfare and Quality of Life*, *Fabric of Society* and *Social Groups*. We assume that those domains, which are used to capture all topics tackled in party election programs, also correspond to major coarse-grained topics of interest in electoral speeches.

[1] From Mitt Romney's own words: "I think you hit a reset button for the fall campaign [i.e., the general election]. Everything changes. Its almost like an Etch-a-Sketch. You can kind of shake it up and we start all over again."

In order to determine the topics addressed in a political speech, we follow the idea of a classification approach we introduced in Zirn et al. (2016). This classifier, initially designed to annotate topics in political manifestos, extends a local supervised topic classifier with predictions from topic-shift classifiers and topic distribution knowledge in a global optimization framework. The global optimization step, however, is helpful when applied to the manifestos, as they cover many different topics (potentially all seven) and require classification on sentence level. For the speeches, however, we choose to classify text at paragraph level because whole paragraphs most often belong to the same topic because politicians tend to express their arguments coherently. Moreover, as each speech generally focuses on a few specific topics (for example *External Relations* and *Economy*), and does not cover the entire spectrum of topics, we decided that the optimization step used for manifesto classification would be superfluous in this setting. We thus only apply on speeches a local supervised topic classifier, trained on manifestos, that combines lexical with semantic textual similarity features (Zirn et al., 2016).

We train this local classifier on two different datasets and compare their performance on a gold standard of speeches labeled on paragraph level.

Training set: manifestos. We train the classifier on party manifesto programs labeled on sentence level. A sub-part of the training set was annotated manually by human experts, the rest was labeled automatically with the method presented in (Zirn et al., 2016). The advantage of such a domain transfer approach is the fact that we need no manual topic annotations on speeches. The downside is, however, that the language of manifestos might differ from the language used in speeches. In the next section, we quantify the drop in performance due to the domain change.

Training set: annotated speeches. We manually annotated a small part of the presidential election campaign speeches on paragraph level with their categories. We train the above described system on this data and, in the next section, report the results. We explore whether investing human resources for annotating speeches pays off with more accurate classification results.

3.2 Position analysis

In order to determine the positions of politicians based on their speeches on a left-right spectrum, we adopt Wordfish (Slapin and Proksch, 2008), which is widely adopted for such tasks in political science research (Grimmer and Stewart, 2013). This method is designed to take documents as input and estimates their positions on a one-dimensional scale. Our goal is to determine fine-grained positions towards the topics contained in the speeches instead of the overall position of the whole speech. We therefore apply the classifier described in this section to identify the topics within a speech and divide a speech into subdocuments containing the text for a single topic only. Finally, we apply Wordfish to the subdocuments.

4 Evaluation

We first quantitatively assess the correctness of the topic classification on a small manually-labeled evaluation dataset of speeches. Then, in order to assess the quality of our fine-grained political scaling approach, we apply it to speeches of thee presidential election campaigns and do a qualitative analysis of the results.

Gold Standard Annotation We asked two scholars of political science to annotate a subset of 10 speeches from the US presidential election campaigns of 2008, 2012 and 2016. The set comprises samples of seven candidates. Our annotators labeled each of the 779 selected paragraphs one of the 7 topical classes listed in subsection 3.1. The inter-annotator agreement across the seven topical classes is $\kappa = 0.55$, which is only moderate and thus confirms the difficulty of the task.

4.1 Evaluation of Topic Classification

We compare three different settings to classify the topics in the speeches.

Baseline. As a baseline, we apply a Support Vector Machine (SVM) using a simple bag-of-words features on the gold standard performing 10-fold cross validation.

ClassySpeech. We apply the classifier described in 3.1 to our gold standard and perform 10-fold cross validation. We refer to this model as *ClassySpeech* in the following.

ClassyMan. We train the classifier described in 3.1 in a semi-supervised fashion on a set of party

Model	F_1
ClassyMan	36.2
Standard SVM	71.2
ClassySpeech	78.6

Table 1: Topic classification performance, micro F1-score, 10-fold CV (in %)

manifestos. We first train the local topic classifier model on six manually sentence-level labeled manifestos and then use the globally-optimized classifier (Zirn et al., 2016) to label the collection of 466 unlabeled manifestos from the Comparative Manifesto Project (Volkens et al., 2011). We then re-trained the local topic classifier on this set of 466 automatically labeled manifestos. In this setting we did not need to topically label any speeches. We apply the classifier trained on the manifestos resulting (from now on referred to as *ClassyMan*) to our gold standard set of speeches.

The results of the three models are shown in table 1. As it is evident from Table 1, the baseline performs quite well with an F_1-score of around 71% , re-confirming the already well-known efficiency of the simple bag-of-words-based supervised topic classification models. The drop in performance caused by the domain adaptation (i.e., the low performance of the model trained on manifestos) indicates that, even if the topics discussed in electoral manifestos and in political campaigns are the same, the language in which they are convened seems to be significantly different. Finally, the best performance is achieved by the ClassySpeech model, the local topic-classifier trained on a small set of manually labeled speeches. The fact that the ClassySpeech model drastically outperforms the ClassyMan model shows that having little of in-domain annotations (i.e., annotated speeches) matters more than having a lot of annotations on out-of-domain texts (i.e., manifestos).

4.2 Qualitative Analysis of Topic-Specific Positions

Election campaigns are a long and complex process that represents the essence of contemporary democracies. In the United States, the practice of selecting candidates for the presidential elections spans more than a year, being a major focus of American and international media. More specifically, in our work we identify three major phases in the presidential race: a) the pre-primaries, when politicians announce their candidacy for president and begin to establish their positions; b) the primaries: when candidates sharpen their profile in order to win the support of the party; and c) the presidential elections: when party nominees have to satisfy the interests of a spectrum of voters as large as possible.

Dataset preparation. After collecting speeches made by the most prominent Republican and Democrat candidates of the last three general elections (2008, 2012, 2016), we divided them in three temporal groups, namely: before primaries (i.e. before the 1st of January of the election year), primaries (between January and June of the election year) and elections (after June of the election year). Using the ClassySpeech model, we topically annotated all of the collected political speeches at paragraph level. Next, we grouped together all paragraph from the same topic and the same period (e.g. all text from all Barack Obama's primary campaign speeches labeled with topic *External Relations*).

Analysis. In the third step of the analysis we ran Wordfish on the collection of temporally and topically divided speeches. In order to understand the usefulness of our fine-grained analysis (i.e., the combination of the two dimensions – time and topic), we compared the its qualitative results with two different more coarse-grained studies. In the first study, we ran Wordfish on the entire speech collection of each candidate (i.e., without any temporal and topical slicing). In the second study we considered only the temporal dimension, i.e., we excluded the topical division.

As shown in Fig. 1, the two coarse-grained analyses do not add any new knowledge, by re-confirming already well known facts, such as the global position of candidates over the political spectrum and a common trend in political campaigns, namely the convergence to the center of the selected party candidates after the primary race (see McCain in particular).[2] In contrast, the fine-grained temporally and topically sliced analysis proposed in our study enables to dig deeper into the candidate's process of converging on a specific position[3]. As a matter of fact, it presents

[2]To know more about the Etch-a-Sketch Hypothesis and how to automatically detect it, see Gross et al. (2013)

[3]Other analyses can be found at:
`https://federiconanni.com/topfish`

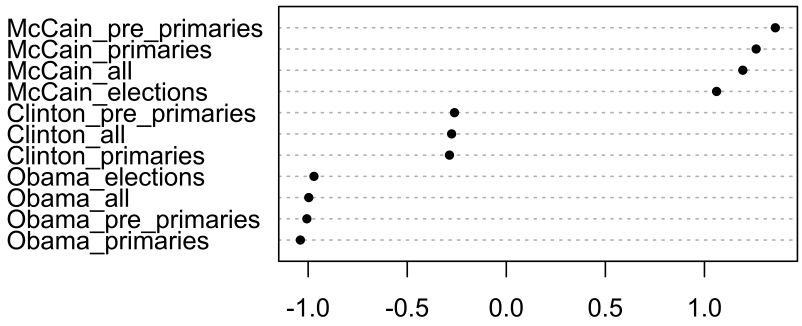


Figure 1: Coarse-grained comparative analyses, using Wordfish.

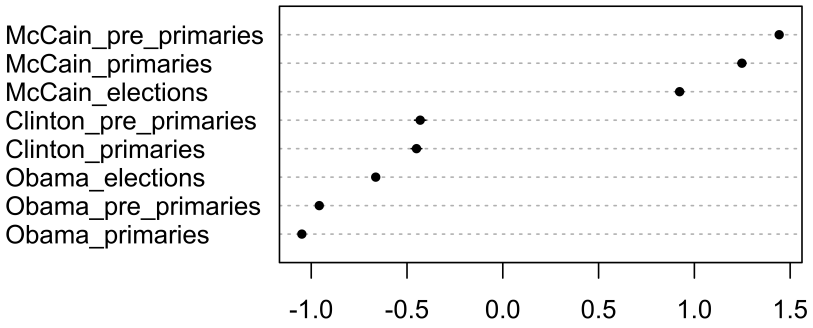


Figure 2: Wordfish position estimates regarding the topic *External Relations*.

a more clear understanding on how candidates have been positioning themselves regarding different relevant political issues, such as *External Relations* (see Fig. 2) and *Welfare and Quality of Life* (see Fig. 3). Additionally, it highlights interesting variations on the established idea of positioning during political campaigns (e.g. the shift to-the-left of Barack Obama presented in Fig. 3) which are completely ignored by a coarse-grained overview on the race.

5 Conclusion

In this paper we presented TopFish, a multilevel computational approach that combines topic detection and political scaling with temporal aspects of political campaigns (pre-primary election, primary election, and general election). We show how this solution enables researchers to perform a range of multidimensional empirical analyses, ultimately allowing them to understand how candidates position themselves during the entire campaign race. The topic-detection method here adopted has been tested against two other solutions, showing its robustness. Additionally, the presented approach has been employed and tested on speeches from the 2008, 2012 and the ongoing 2016 US presidential campaigns, showing its usefulness for examining in a more fine-grained fashion how candidates determine their political space.

Acknowledgments

The authors thank the DFG for Funding under the SFB 884 Political Economy of Reforms C4 project.

References

Robert P Abelson and J Douglas Carroll. 1965. Computer simulation of individual belief systems. *The American Behavioral Scientist (pre-1986)*, 8(9):0_24.

Ramnath Balasubramanyan, William W Cohen, Douglas Pierce, and David P Redlawsk. 2012. Modeling polarizing topics: When do different political communities respond differently to the same news? In *ICWSM*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

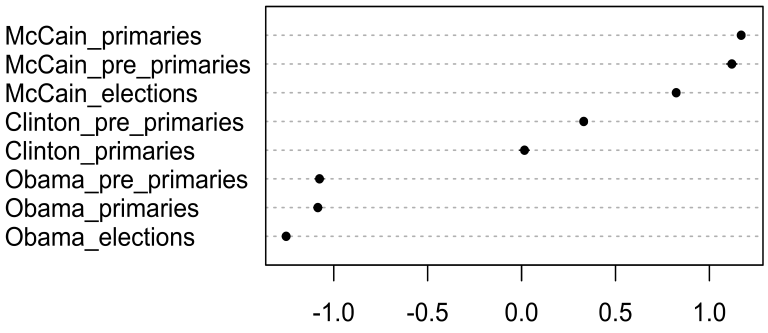


Figure 3: Wordfish position estimates regarding the topic *Welfare and Quality of Life*.

Kirsten Foot, Steven M Schneider, Meghan Dougherty, Michael Xenos, and Elena Larsen. 2003. Analyzing linking practices: Candidate sites in the 2002 us electoral web sphere. *Journal of Computer-Mediated Communication*, 8(4):0–0.

Swapna Gottipati, Minghui Qiu, Yanchuan Sim, Jing Jiang, and Noah A. Smith. 2013. Learning topics and positions from Debatepedia. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1858–1868, Seattle, Washington, USA, October. Association for Computational Linguistics.

Justin Grimmer and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, page mps028.

Justin Gross, Brice Acree, Yanchuan Sim, and Noah A Smith. 2013. Testing the etch-a-sketch hypothesis: A computational analysis of mitt romney's ideological makeover during the 2012 primary vs. general elections. In *APSA 2013 Annual Meeting Paper*.

Dustin Hillard, Stephen Purpura, and John Wilkerson. 2008. Computer-assisted topic classification for mixed-methods social science research. *Journal of Information Technology & Politics*, 4(4):31–46.

Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc.

Michael Laver, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(02):311–331.

Melissa J Marschall and Robert J McKee. 2002. From campaign promises to presidential policy: Education reform in the 2000 election. *Educational Policy*, 16(1):96–117.

Seth Masket. 2009. *No middle ground: How informal party organizations control nominations and polarize legislatures*. University of Michigan Press.

Jon D Mcauliffe and David M Blei. 2008. Supervised topic models. In *Advances in neural information processing systems*, pages 121–128.

Juraj Medzihorsky, Levente Littvay, and Erin K Jenne. 2014. Has the tea party era radicalized the republican party? evidence from text analysis of the 2008 and 2012 republican primary debates. *PS: Political Science & Politics*, 47(04):806–812.

Federico Nanni and Pablo Ruiz Fabo. 2016. Entities as topic labels: Improving topic interpretability and evaluability combining entity linking and labeled lda. *To appear in the proceedings of Digital Humanities 2016*.

Vinodkumar Prabhakaran, Ashima Arora, and Owen Rambow. 2014. Staying on topic: An indicator of power in political debates. In *EMNLP*, pages 1481–1486.

Kevin M Quinn, Burt L Monroe, Michael Colaresi, Michael H Crespin, and Dragomir R Radev. 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209–228.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. 2009. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*, pages 248–256. Association for Computational Linguistics.

Delip Rao, Paul McNamee, and Mark Dredze. 2013. Entity linking: Finding extracted entities in a knowledge base. In *Multi-source, Multilingual Information Extraction and Summarization*, pages 93–115. Springer.

Evan J Ringquist and Carl Dasse. 2004. Lies, damned lies, and campaign promises? environmental legislation in the 105th congress. *Social Science Quarterly*, 85(2):400–419.

Arno Scharl and Albert Weichselbraun. 2008. An automated approach to investigating the online media coverage of us presidential elections. *Journal of Information Technology & Politics*, 5(1):121–132.

Nate Silver. 2012. *The signal and the noise: Why so many predictions fail-but some don't*. Penguin.

Yanchuan Sim, Brice DL Acree, Justin H Gross, and Noah A Smith. 2013. Measuring ideological proportions in political speeches. *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jonathan B Slapin and Sven-Oliver Proksch. 2008. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3):705–722.

Andrea Volkens, Onawa Lacewell, Pola Lehmann, Sven Regel, Henrike Schultze, and Annika Werner. 2011. *The Manifesto Data Collection*. Manifesto Project (MRG/CMP/MARPOR), Wissenschaftszentrum Berlin für Sozialforschung (WZB).

Cäcilia Zirn and Heiner Stuckenschmidt. 2014. Multidimensional topic analysis in political texts. *Data & Knowledge Engineering*, 90:38–53.

Cäcilia Zirn, Goran Glavaš, Federico Nanni, Jason Eichorst, and Heiner Stuckenschmidt. 2016. Classifying topics and detecting topic shifts in political manifestos. PolText.

Semantic Annotation for the Analysis of Political Debates: A Graph-based Approach

Philippe N'techobo
Ecole Polytechnique de Montréal
ntechobo.edoukou-philippe-armel@polymtl.ca

Amal Zouaq
University of Ottawa
azouaq@uottawa.ca

Michel Gagnon
Ecole Polytechnique de Montréal
michel.gagnon@polymtl.ca

Abstract

In this paper, we propose a method to automatically extract a summarized view based on graphs that represents the topics discussed during political debates and the relations between these topics. To this end, we use semantic annotators based on Linked Data to extract the topics. We also propose an open relation extraction approach based on shallow parsing and disambiguate the extracted relations using VerbNet, a semantic lexicon. Then, we exploit DBpedia, a cross-domain knowledge base, to enrich the textual extractions with available predicates between these topics. Finally, we construct an abstract graph representation of the debates.

1 Introduction

Many governments have been uploading data to help the public better understand their activities, leading to political debates corpora. These corpora are usually very large, cover a variety of topics and are not always organized in a linear fashion. Thus, a high level representation of the most important topics and relations between them provides a basis for exploring such corpora, and offers efficient information access mechanisms.
In this paper, we propose an approach for the extraction of a graph-based high-level representation of Canadian parliamentary debates. This graph represents a summary of a particular debate date, subject or Member of Parliament intervention. This high-level representation can be used for semantic search, question answering or for abstractive summarization. In this graph, vertices represent the discussed topics and the edges represent semantic relations between those topics. One novelty of our approach is that the discussed topics are not only represented as plain keywords as in (Lin et al., 2015), but also as semantic annotations based on the Linked Open Data (LOD) cloud. The LOD (also referred as the Semantic Web) is a paradigm for publishing structured data on the Web in which the information is not described in document silos, but instead constitute a global interconnected space. Semantic annotations based on the LOD have flourished with the appearance of numerous services such as AlchemyAPI[1], DBpedia Spotlight[2] and Open Calais[3]. Using these annotations, we can benefit from the knowledge available on the Linked Data cloud to model the semantics of documents in general and political debates in particular. In our work, we rely on the cross-domain knowledge base DBpedia (Lehmann et al, 2015) which represents Wikipedia content. The exploitation of DBpedia for the automatic understanding of textual content has grown considerably in recent works (Lehmann et al, 2015), but to our knowledge, DBpedia has not been used for the analysis of political corpora. Moreover, the cross-domain nature of DBpedia makes it a suitable base for handling the variety of the discussed topics in parliamentary debates.
Relations between the obtained annotations are identified using an open relation extraction approach. Several kinds of relations are extracted. First we identify textual relations in the debates using shallow syntactic patterns defined manually. Second, we query DBpedia with SPARQL to retrieve available predicates between the identified topics, thus enriching the political debates with linked data knowledge. Finally, our approach

[1] http://www.alchemyapi.com/
[2] https://github.com/dbpedia-spotlight/dbpedia-spotlight/
[3] http://http://www.opencalais.com/

Proceedings of the International Conference on the Advances in Computational Analysis of Political Text (PolText 2016), pages 68–74, Dubrovnik, Croatia, 14–16 July 2016.

relies on VerbNet to identify high-level relations based on the textual relations.

The remainder of the paper is structured as follows. In section 2, we present a literature review on graph-based document representations, relation extraction and the analysis of political debates in general. In section 3, we present our methodology for graph-based extraction from political debates. Finally, section 4 presents the results of our experiments on a subset of the Hansard, the Canadian parliamentary debates.

2 Related work

Graph-based representation of textual content. T-VSM (Becker et al., 2003) (Term - Vector Space Model) is a popular model for the representation of textual content. This bag-of- word model ignores the order and links between terms in textual documents (Jin et al., 2003). Theoretical representations based on graphs have been proposed to address these limitations. Jin et al. (2003) propose three types of graphs to represent a document and capture the relations between terms. One of the proposed approaches models co-occurrence relations between terms. In this representation, the edges indicate the number of times terms appear together in the same group (sentence, paragraph, etc.). Graph models are also used in several works to perform keyword retrieval (Mihalcea et al, 2003; Abilhoa et al., 2014) automatic summarization (Erkan et al., 2004; Ganesan et al., 2010), or conceptual representation of text (Augenstein et al., 2012; Hensman et al. 2004). While some works use graphs as intermediate data structure to perform their tasks (Erkan et al., 2004; Ganesan et al., 2010), other studies present their final output in a graph-based representation (Augenstein et al., 2012; Hensman et al. 2004).

Relation extraction. A relation extraction phase is necessary to build our conceptual graph. The existing techniques for relation extraction can be classified as supervised (Kambhatla et al., 2004), semi-supervised (Brin et al, 1998) and unsupervised (Riloff et al. 1999). Supervised methods use semantic and syntactic features to decide if two entities are related. Because finding the optimal subset of important features is difficult, kernel methods have been designed to explore fully and implicitly the representation of textual input in a higher level dimensional space (Bach et al., 2007). When there are not enough examples for training, semi-supervised methods (Brin et al. 1998) can be used to automatically infer rules or extraction patterns for relations. Unsupervised approaches (Riloff et al., 1999), are based on rules generally defined manually. Among unsupervised approaches, one paradigm that has been adopted by the research community is the Open Relation Extraction (ORE), where the set of relations to be extracted is not defined a priori. In our work, we implemented an ORE approach based on shallow syntactic analysis patterns, following the principles defined in (Banko et al., 2007), which state that most of English relations can be extracted with few grammatical rules. One downside of ORE is that the extracted relations, being arbitrary, are difficult to be reused and interpreted by other systems. We propose a solution to this problem by linking the extracted textual relations to the VerbNet knowledge base (Schuler et al., 2007). We then obtain high-level relations whose semantics is clearly defined in VerbNet.

Political Debates Analysis. Most research on political corpora concerns topic extraction (van Wees et al., 2011) or the prediction of the winners of an election or a debate (Kaplan et al., 2012). Generally, the purpose of topic extraction is the discovery of patterns in government operations (van Wees et al, 2011) or the identification of politicians interests (Gurciullo et al., 2015). These approaches are based on social network analysis techniques (Derényi et al., 2005), traditional information retrieval metrics, such as TF-IDF, or on methods that are used to identify the temporal evolution of the discussed topics, such as Dynamic Modelling Topic (Blei et al., 2006) or Non-negative Matrix Factorization (NMF) (Paatero et al., 1994). Other works have been done on linking political documents to the Semantic Web. In general, the aim is to convert existing political unstructured or semi-structured documents to RDF (Marx et al., 2010) and visualize them (Nigel Shadbolt et al., 2012).

3 Methodology

3.1 Corpus

Our dataset is the *Canadian Hansard* corpus, a collection of Canadian parliamentary debates. It contains debates from 1994 to Today. The debates are organized in dates and topics. Each debate date contains several *orders of business*. The orders of business are divided into *subjects of business*, which contain transcriptions of interventions made by Members of Parliament.

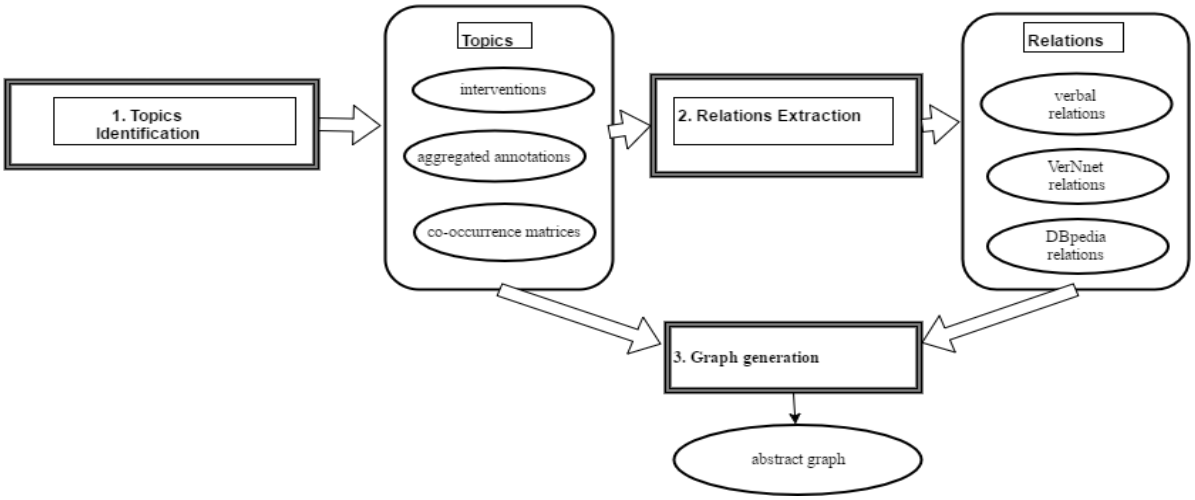


Figure 1. General Architecture

In our experimentation dataset composed of 1217 debate dates, there is on average 245 words per intervention and 254 interventions per debate date.

3.2 Architecture

As described in Figure 1, our architecture is composed of two main modules. The first one, *Topics Identification*, extracts the discussed topics as semantic annotations and assigns a relevance score to the annotations. A co-occurrence matrix is computed to allow the second module, *Relation Extraction*, to identify semantic relations between these annotations. The topics and relations are then combined to build the abstract graph.
Several aspects make this task difficult. First, we must filter the most important annotations with adequate metrics. Next, we need to identify relations in free text. Finally, the semantics of the relations must be revealed through a disambiguation phase.

3.3 Topics Identification

The objective of this phase is to identify the most important topics discussed in the debates. For this, we rely on AlchemyAPI, a semantic annotator based on Linked data, which is reported as being among the best current annotators by several evaluation studies (Gagnon et al., 2013; Jean-Louis et al., 2014)
Semantic annotation consists of linking sequences of words in a document to an existing concept defined in a knowledge base. We say that such a sequence of words is a *mention* of the concept to which it is linked. This process is generally done in two stages, named *spotting* (identifies the mentions) and *disambiguation* (ensures that each mention refers to the correct knowledge base concept).

Spotting Enhancement. Although AlchemyAPI is considered as one of the best annotators, we have identified some deficiencies in its spotting process. In particular, the returned annotations are sometimes partial and do not always identify the correct surface forms. For example, in the sentence *The Canada Transportation Act review couldn't be more timely*, AlchemyAPI returns the annotations *Canada Transportation* and *Act review* separately. The correct annotation in this case would be *Canada Transportation Act review*. In order to correct the annotations returned by AlchemyAPI, we implemented a spotting enhancement phase, where annotations are corrected by extending them to the longest noun phrase around them. This process is based on patterns built manually using Parts of Speech (POS) tagging[4] obtained with TreeTagger (Schmid et al., 1995). Our patterns are defined as regular expressions and are summarized in Table 1.

ID	POS Patterns	Examples
P1	CD? (JJ|VVN)? **N** N+	$life_N \rightarrow life_N\ imprisonment_N$
P2	CD? (JJ|VVN)? N* IN (JJ|DT|DT JJ|)? (CD|N) N*	$life_N \rightarrow sentence_N\ of_{IN}\ life_N\ imprisonment_N$
P3	N IN PP$ JJ? (CD|N) N*	$speech_N \rightarrow end_N\ of_{IN}\ my_{PP\$}\ speech_N$
P4	N POS N	$safety_N \rightarrow Canadian_N\ 's_{POS}\ safety_N$

Table 1. Spotting enhancement patterns.

Annotation aggregation. Given that our goal is to provide an abstract representation and a summarized view of the corpus as a whole, we group

[4] https://courses.washington.edu/hypertxt/csar-v02/penntable.html

the annotations that share the same lemma leading to aggregated annotations. For instance, *child* is grouped with *children*.

Annotation relevance. Finally, we associate relevance scores to the aggregated annotations. We consider three metrics: term frequency (TF), the relevance score returned by AlchemyAPI and TF-IDF.

At this stage, the annotations constitute the vertices of our abstract graph. The next stage is to extract relations between pairs of semantic annotations from the sentences where they co-occur. To this end, we build a co-occurrence (by sentence) matrix that is used for the identification of the most relevant textual relations and the construction of a *labelled graph.*

3.4 Relations extraction

If two annotations co-occur frequently, our hypothesis is that there is some relationship between them. In each sentence where two annotations co-occur, we attempt to extract a relation in the form *subject - relation - object*, where *subject* and *object* are semantic annotations and *relation* is a verbal expression between these two annotations.

To extract the relations, we used morpho-syntactic patterns, which are much less expensive than other techniques (such as dependency trees) and would therefore be suitable for processing a large corpus. Using POS tags, we designed relation extraction patterns manually as regular expressions. Table 2 summarizes the implemented patterns, and for each pattern an example is shown. Note that the extracted relation is indicated in bold. In the table, ANN represents a semantic annotation, NOT a generalized negation form, BE and HAVE the auxiliaries *be* and *have*, and VERB a generic verb. The other symbols are TreeTagger's POS tags and regular expression symbols.

ID	POS Pattern	Example
PR1	ANN [^ANN]* **(BE|HAVE) NOT (TO|IN)?** (CD|DT|JJR|JJS|PP$)? ANN	the *[liberal party]*$_{ANN}$ **was**$_{BE}$ **not**$_{NOT}$ my$_{PP\$}$ *[affiliation]*$_{ANN}$
PR2	ANN [^ANN]* **AUX? NOT? VERB RP? (TO|IN)?** (CD|DT|JJR|JJS)? ANN	the *[liberal party]* $_{ANN}$ **will**$_{AUX}$ **provide**$_{VERB}$ one$_{CT}$ *[formal apology]*$_{ANN}$
PR3	ANN [^ANN]* **AUX? NOT? VERB RP? TO VERB RP? (TO|IN)?** (CD|DT|JJR|JJS)? ANN	this *[government]*$_{ANN}$ **can**$_{AUX}$ **continue**$_{VERB}$ **to**$_{TO}$ **protect**$_{VERB}$ *[Canadians]*$_{ANN}$
PR4	ANN [^ANN]* **AUX? NOT? VERB RP? [^(ANN|VERB)]* TO VERB RP?** (CD|DT|JJR|JJS|PP$)? ANN	*[bill c-50]*$_{ANN}$ **proposes**$_{VERB}$ **important**$_{JJ}$ **reforms**$_{N}$ **to**$_{IN}$ *[Canada 's election act]*$_{ANN}$
PR5	ANN **of VERB** ANN	*[idea]*$_{ANN}$ **of preventing**$_{VERB}$ *[crime]*$_{ANN}$

Table 2. Relation extraction patterns

In addition to extracting verbal relations, one of our goals is to disambiguate textual relations with high-level relations. In fact, as we reported before, one limitation of ORE approaches is that the extracted relations are difficult to be reused, given that the set of relations which can be extracted is not known a priori. To solve this problem, we rely on VerbNet. To abstract our relations, we link their core verb to a VerbNet class. We define a core verb of a verbal relationship as the main verb of the relation, the one that defines the action. For example, in the relation *[criminal]* $_{ANN}$ *[can be released on]*$_{REL}$ *[word]*$_{ANN}$ the core verb *release* will be mapped to its corresponding VerbNet class *free-80-1*.

However, an English verb may belong to more than one VerbNet class. A disambiguation process is therefore required. We solve this issue by integrating in our system the disambiguation system ClearWSD[5].For example, in the relation *[Conservatives] [take on] [Bill 51]*, the core verb *take* has six candidate classes: *bring-11.3*, *characterize-29.2*, *convert-26.6.2*, *cost-54.2*, *fit-54.3*, *hire-13.5.3*, *performance-26.7.2*, *require-103* and *steal-10.5*. ClearWSD returns in this case *steal-10.5* as the correct VerbNet class.

One advantage of using semantic annotation is the possibility of enriching our debate corpora with DBpedia knowledge. The final step in our approach is to enrich the obtained graph with relations found in DBpedia. These relations are extracted based on SPARQL queries that tests the existence of predicates for each annotation pair.

4 Evaluation

Our evaluation corpus consists of fifteen subjects of business chosen randomly. For each of them, we constructed the corresponding labeled graph. In our evaluation, we separately examine each module of our system. First we evaluate the different metrics associated with the annotations. These metrics are assessed for their ability to filter the most important annotations of the corpus. We

[5] https://github.com/jgung/ClearWSD

also evaluate the precision of the spotting enhancement and the relation extraction phases (precision is the ratio of correct items among the total number of items returned by the system). We conclude with a qualitative evaluation of the abstract graph. To measure the accuracy of our system, we built, for each module, a gold standard using six human raters. We asked each rater to evaluate five different subjects of business for each module. We therefore obtained, for each subject of business, two evaluations. We measured the agreement between raters using the Gwet AC1 metric (Viswanathan et al., 2011).

To evaluate the relevance metrics, we extracted, for each of them, the top-10 and top-50 annotations. We presented these annotations to the raters and asked them if each evaluation was relevant or not. We then used these results to compute the precision for each metric. The results are presented in Table 3. In this evaluation, we found that TF led generally to the best performances.

	P_{TF}	$P_{TF\text{-}IDF}$	$P_{Alchemy}$	AC1
Top-50	72.8	71.2	64.33	0.47
Top-10	76.4	72	67.66	

Table 3. Precisions of the metrics used

The evaluation of the spotting enhancement process led to good results, with an overall precision score of 83.7%. The evaluators' agreement is 0.8 in this case, which is almost perfect. Most errors in this module are due to the limitations of shallow syntactic patterns. Some errors could be avoided with more complex tools such as syntactic dependency trees.

We also assessed the relation extraction module on the following aspects: the well-formedness of the extracted relations, their relevance and, finally, the correctness of the associated VerbNet class. The results are shown in Table 4. We obtained 77.9% and 72% for the well-formedness and the relevance, respectively.

By analyzing the raters' evaluations, we found that 56% of the extracted relations have at least one VerbNet correct candidate class (which adequately expresses the semantics of the extracted relation), and among these, 84% are correctly disambiguated. Thus, in total, 47% of the evaluated relations are correctly disambiguated. One can observe that topics identification and relation extraction perform well individually.

Aspect evaluated	Precision	AC1
Well-formedness	77.9	0.65
Correctness	72	0.48
At least one correct VerbNet class	56.4	0.67
Disambiguation	47	0.66

Table 4. Evaluation of the extracted relations

Finally, we evaluated the obtained graphs qualitatively (using a likert scale 0, 0.5, 1 where 1 represents yes, 0 no and 0.5 a "somewhat" answer) by asking the raters if the vertices (representing the discussed topics), the edges (representing the relations) and the graph as a whole were globally relevant and matching the topics discussed in the debates. For the graph, we also asked the raters whether the graph was summarizing accurately the ideas discussed in each selected subject. We obtained an average precision of 83% for the vertices, 78% for the relations and 73% for the full graph.

5 Conclusion

In this work, we presented an approach for corpus abstraction that we applied to political debates. Our approach is based on a graph constructed with semantic annotations and relations between these annotations extracted from texts and DBpedia. By performing a manual evaluation of the generated graphs, we concluded that it was possible to generate a graph summarizing a political corpus using semantic annotations and shallow syntactic patterns for relation extraction.

Although our results are satisfactory, several improvements are possible. The first one would be to evaluate our system on a larger corpus, consisting of at least a full debate date involving several subjects and interventions from Members of Parliament. This type of evaluation is expensive because the evaluation is done manually.

Similarly, we could benefit from richer knowledge bases to enhance the obtained graph with available predicates. In our experiments, we noticed that there are very few semantic relations between DBpedia concepts. Although DBpedia is considered by many as a cross-domain knowledge base (Mendes et al, 2012), it does not describe sufficiently the relations between these concepts (Font et al., 2015). To remedy this problem, our model could be enriched by including other LOD knowledge bases such as Yago (Hoffart et al., 2013) and Wikidata (Vrandečić et al., 2014).

Finally, while the graph-based representation is our current final output, one possible application would be to connect a text generator to the obtained graph to produce abstracts in textual form, allowing the comparison of our system with other summarization systems that produce textual output.

References

Abilhoa, W. D., & de Castro, L. N. (2014). A keyword extraction method from twitter messages represented as graphs. Applied Mathematics and Computation, 240, 308-325.

Augenstein, I., Padó, S., & Rudolph, S. (2012). Lodifier: Generating linked data from unstructured text. In The Semantic Web: Research and Applications (pp. 210-224). Springer Berlin Heidelberg.

Bach, N., & Badaskar, S. (2007). A review of relation extraction. Literature review for Language and Statistics II.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., & Etzioni, O. (2007). Open information extraction

Becker, J., & Kuropka, D. (2003, July). Topic-based vector space model. In Proceedings of the 6th international conference on business information systems (pp. 7-12).

Blei, D. M., & Lafferty, J. D. (2006, June). Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning (pp. 113-12c0). ACM.

Brin, S. (1998). Extracting patterns and relations from the world wide web. In The World Wide Web and Databases (pp. 172-183). Springer Berlin Heidelberg.

Derényi, I., Palla, G., & Vicsek, T. (2005). Clique percolation in random networks. Physical review letters, 94(16).

Font, L., Zouaq, A., & Gagnon, M. (2015, November). Assessing the Quality of Domain Concepts Descriptions in DBpedia. In 2015 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS) (pp. 254-261). IEEE.

Gagnon, M., Zouaq, A., & Jean-Louis, L. (2013, May). Can we use linked data semantic annotators for the extraction of domain-relevant expressions?. In Proceedings of the 22nd international conference on World Wide Web companion (pp. 1239-1246). International World Wide Web Conferences Steering Committee.

Gurciullo, S., Smallegan, M., Pereda, M., Battiston, F., Patania, A., Poledna, S., Hedblom, D., Oztan, B., T., Herzog. A., John. P. & Mikhaylov, S. (2015). Complex Politics: A Quantitative Semantic and Topological Analysis of UK House of Commons Debates. arXiv preprint arXiv:1510.03797.

Hensman, S. (2004, May). Construction of conceptual graph representation of texts. In Proceedings of the Student Research Workshop at HLT-NAACL 2004 (pp. 49-54). Association for Computational Linguistics.

Hoffart, J., Suchanek, F. M., Berberich, K., & Weikum, G. (2013). YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Artificial Intelligence, 194, 28-61.

Jean-Louis, L., Zouaq, A., Gagnon, M., & Ensan, F. (2014). An assessment of online semantic annotators for the keyword extraction task. In PRICAI 2014: Trends in Artificial Intelligence (pp. 548-560). Springer International Publishing.

Jin, W., & Srihari, R. K. (2007, March). Graph-based text representation and knowledge discovery. In Proceedings of the 2007 ACM symposium on Applied computing (pp. 807-811). ACM.

Kambhatla, N. (2004, July). Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In Proceedings of the ACL 2004 on Interactive poster and demonstration sessions (p. 22). Association for Computational Linguistics.

Kaplan, I., & Rosenberg, A. (2012, December). Analysis of speech transcripts to predict winners of US Presidential and Vice-Presidential debates. In Spoken Language Technology Workshop (SLT), 2012 IEEE (pp. 449-454). IEEE.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Kleef, P., Auer, S. & Bizer, C. (2015). DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web, 6(2), 167-195.

Lin, F. R., Chou, S. Y., Liao, D., & Hao, D. (2015, January). Automatic Content Analysis of Legislative Documents by Text Mining Techniques. In System Sciences (HICSS), 2015 48th Hawaii International Conference on (pp. 2199-2208). IEEE.

Marx, M., & Aders, N. (2010). From documents to data: linked data at the dutch parliament.

Mendes, P. N., Jakob, M., & Bizer, C. (2012, May). DBpedia: A Multilingual Cross-domain Knowledge Base. In LREC (pp. 1813-1817).

Mihalcea, R., & Tarau, P. (2004, July). TextRank: Bringing order into texts. Association for Computational Linguistics.

Nigel Shadbolt, Kieron O'Hara, Tim Berners-Lee, Nicholas Gibbins, Hugh Glaser, Wendy Hall, et al. Linked open government data: Lessons from data. gov. uk. IEEE Intelligent Systems, 27(3):16–24, 2012.

Riloff, E., & Jones, R. (1999, July). Learning dictionaries for information extraction by multi-level bootstrapping. In AAAI/IAAI (pp. 474-479).

Schmid, H. (1995). Treetagger| a language independent part-of-speech tagger. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 43, 28.

van Wees, J., Marx, M., & van Doornik, J. (2011). Applying Social Network Analysis to Parliamentary Proceedings.

Vrandečić D., & Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. Communications of the ACM, 57(10), 78-85.

Viswanathan M, Berkman ND. Development of the RTI Item Bank on Risk of Bias and Precision of Observational Studies [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2011 Sep. Appendix A, AC1 Statistic. Available from: http://www.ncbi.nlm.nih.gov/books/NBK82266/

EUSpeech: a New Dataset of EU Elite Speeches

Gijs Schumacher
University of Amsterdam
g.schumacher@uva.nl

Martijn Schoonvelde
Vrije Universiteit, Amsterdam
h.j.m.schoonvelde@vu.nl

Denise Traber
University of Zurich
traber@ipz.uzh.ch

Tanushree Dahiya
University of Amsterdam
tanushree.dahiya@student.uva.nl

Erik de Vries
University of Amsterdam
erik@devries.pm

Abstract

This paper presents EUSpeech, a new dataset of 18,403 speeches from EU leaders (i.e., heads of government in 10 member states, EU commissioners, party leaders in the European Parliament, and ECB and IMF leaders) from 2007 to 2015. These speeches vary in sentiment, topics and ideology, allowing for fine-grained, over-time comparison of representation in the EU.

1 Introduction

This paper presents EUSpeech, a new dataset of 18,403 speeches from EU leaders (i.e., heads of government in 10 member states, EU commission members, party leaders in the European Parliament, and ECB and IMF leaders) from 2007 to 2015 (Schumacher et al., 2016).[1] These speeches vary in sentiment, topics and ideology, allowing for fine-grained, over-time comparison of representation in the EU. This paper illustrates the possibilities of working with EUSpeech for scholars interested in elite-mass interactions in the EU. To this end, the next section first introduces EUSpeech. We then present a Wordfish scaling analysis, identifying a clear anti-Europe vs pro-Europe dimension in EP speeches (Slapin and Proksch, 2008; Proksch and Slapin, 2009). Furthermore, we use sentiment analysis to show that speech sentiment responds to objective economic and political factors (Young and Soroka, 2012).

2 EUSpeech

EUSpeech consists of all publicly available speeches from the main European institutions plus the IMF and the speeches of prime ministers—or president in the case of France—of 10 EU countries for the period after 1 January 2007.[2] Most countries and institutions have a dedicated website that stores information on the decisions, background, media appearances and speeches of members of government. In most cases websites clearly demarcated speeches from other types of oral communication such as interviews or debates.[3]

Table 1 gives an overview of the institutions and countries in our dataset and the websites we collected speeches from.[4] In most cases we used the official government websites.[5] Interestingly, most official government websites delete the speeches of outgoing premiers or presidents, leaving us with

[1] This dataset is available on Harvard Dataverse: https://dataverse.harvard.edu/dataverse/euspeech.

[2] These countries are Czech Republic, France, Germany, Greece, Netherlands, Italy, Spain, United Kingdom, Poland and Portugal.

[3] We did not collect these other types of oral communication because they depend on third parties.

[4] EUSpeech also includes the Python scripts we used to scrape the speech texts and metadata.

[5] For France we found a non-governmental website that had collected all the speeches from the relevant Presidents.

Proceedings of the International Conference on the Advances in Computational Analysis of Political Text (PolText 2016), pages 75–80, Dubrovnik, Croatia, 14–16 July 2016.

	Total	English	Speakers	Source	Wayback machine	Time period
Institution						
IMF	509	509	-	imf.org	No	01/2007 - 11/2015
European Council	236	220	2	consilium.europe.eu	No	11/2009 - 09/2015
European Commission	6140	5991	-	europa.eu	No	01/2007 - 11/2015
European Central Bank	1008	990	-	ecb.europa.eu	No	01/2007 - 11/2015
European Parliament	3665	2698	26	europarl.europa.eu	No	01/2007 - 11/2015
ALDE	48	43	1	alde.eu	No	10/2010 - 11/2014
ECR	56	55	1	ecrgroup.eu	No	07/2009 - 10/2015
Country						
Czech Republic	273	39	4	vlada.cz	Yes	06/2009 - 11/2015
France	1451	0	3	vie-publique.fr	No	01/2007 - 10/2015
Germany	580	1	1	bundeskanzlerin.de	No	10/2008 - 11/2015
Greece	484	94	4	primeminister.gov.gr	Yes	10/2009 - 11/2015
Netherlands	392	132	2	rijksoverheid.nl	No	02/2007 - 11/2015
Italy	867	63	5	governo.it	Yes	01/2008 - 9/2015
Poland	4	0	3	premier.gov.pl	No	11/2011 - 11/2015
Portugal	139	6	3	portugal.gov.pt	Yes	10/2009 - 12-2015
Spain	1764	768	2	lamoncloa.gob.es	No	01/2007 - 11/2015
United Kingdom	787	787	3	gov.uk nationalarchives.gov.uk	Yes	03/2007 - 11/2015

Table 1: **Number of speeches per country, language and institution**

only the speeches of the incumbent premier or president. To solve this problem we used the Wayback Machine, allowing us to travel back to the governments' website prior to the change of government.[6] This way we were able to retrieve most speeches, although some missing speeches were unavoidable.[7] We did not collect speeches by interim prime ministers.

Table 1 also gives an overview of the number of speeches, the number of speakers and the period for which the speeches were collected for each country and institution. There is variation between countries on all of these criteria. Clearly, some countries had more changes in leadership than others.[8] Some countries have more speeches than others for at least two reasons: larger countries tend to have more speeches than smaller ones, and some countries are simply more diligent than others in keeping track of these speeches.[9]

All speeches were scraped using Python.[10] For each country, institution, and language, we saved the text of all speeches, as well as metadata like date, speaker, title and speech length in a single csv file. We also cleaned the scraped speeches, discarding sentence structure and interpunction, leaving us with term-document matrixes. This allows us to extract comparable measures of position (scaling models) and sentiment (sentiment models).[11]

In the next two sections, we illustrate how the EUSpeech data can be used for fine-grained, over-time analysis of representation in the EU, using sentiment analysis and scaling models.

3 Sentiment Analysis

3.1 Method

Sentiment analysis uses a dictionary that indicates whether words have positive or negative sentiment. We combined two dictionaries containing positive and negative sentiment scores of English words for in total 5875 words (Wilson et al., 2005; Mohammad and Turney, 2010). These

[6] `https://archive.org/web/`

[7] The Wayback Machine makes occasional snapshots of websites. In some cases there are a few months between the last snapshot and the change of government, thus leading to some gaps in the data.

[8] For some countries we were unable to find speeches from 2007 or 2008. These were probably never published online or are hiding in the dark corners of the internet.

[9] What is important here is whether the selection of speeches on the website is a random selection of speeches or whether specific speeches have been taken out. If a speech was important in signifying a certain position or sentiment of a leader it is unlikely to have been taken out. It is more likely that irrelevant speeches at say the opening of a rather irrelevant event run the risk of not being put online. Some important speeches, however, may not have appeared because the leader took an unpopular position that was later retracted. Unfortunately, this remains speculation.

[10] The cleaning scripts are available to users of EUSpeech as well and can be adjusted to suit their research goals.

[11] In results not presented in this paper, we also apply topic models (Grimmer, 2010), complexity analysis (Kincaid et al., 1975) and noun usage (Cichocka et al., 2016) to the speeches.

dictionaries assign identical sentiment scores to words that appear in both but combined they contain more words than they do separately. We first matched the words in the dictionaries with those in the term-document matrices. Then we calculated positive and negative sentiment scores for each speech by counting the number of positive or negative words and dividing by the total number of words. We do this for all 12,297 English-language speeches, and 6,106 speeches which were translated in English using *Google Translate*.

3.2 Results

Figure 1 reports results from the sentiment analysis. Figure 1a displays mean sentiment per quarter over all speeches. We draw two conclusions from figure 1a: (1) speeches contain almost 4 times more positive sentiment than negative sentiment; (2) positive sentiment drops dramatically after 2014. Figure 1b displays sentiment (positive and negative) by institute. Levels of sentiment differ per institute, but overall positive sentiment is present more than negative sentiment. On negative sentiment (top panel) Greece and the European Parliament score highest, and the European Council, Italy and European Commission score lowest. On positive sentiment (bottom panel) the institutions (EC, ECB, IMF, EU Council and EP) and Greece score lowest. It appears that, on average, the European institutions (plus IMF) communicate with less sentiment than the prime ministers. The prevalence of high negative sentiment and low positive sentiment for the case of Greece may reflect the disastrous economic developments there.

Figure 1c presents positive and negative sentiment for a selection of (better-known) speakers. Except for one speaker (Marcel de Graaff, co-president of Europe of Nations and Freedom), all speakers use on average more positive than negative sentiment. On average, the radical speakers in this sample (Tsipras, Farage, Bisky and De Graaff) deliver speeches with relatively more negative, and less positive sentiment than the other leaders. Speakers often seen as relatively technocratic politician types (e.g. Monti, Van Rompuy and Prodi) deliver speeches with relatively little (positive and negative) sentiment.

Finally, Figure 1d presents results from four different regression analyses of positive or negative sentiment in prime minister speeches and institutions speeches on quarterly GDP growth of the Eurozone and quarterly GDP growth of the respective country, and a political crisis variable as measured by the number EU council meetings in each time period.[12]. Figure 1d shows that negative sentiment in speeches from European institutes (plus IMF) shrinks with economic growth and the number of EU Council meetings. In other words, the better the economy or the more political crisis, the less negative sentiment in speeches. Our analysis of negative sentiment in prime minister speeches is similar in the sense that country GDP growth reduces negative sentiment. However, eurozone growth increases negative sentiment. This means that negative sentiment is especially high if the eurozone is growing, but the economy of the prime minister's country is shrinking. If both country and the eurozone economies are growing, these two effects should cancel each other out. Eurozone growth and the number of EU council meetings stimulate positive sentiment in the speeches by the European institutes (plus IMF). Hence, our results in the analysis of positive sentiment are the exact reverse of the results of negative sentiment. This is not the case for the PM speeches. Here we find no effect of growth. This suggests a loss aversion mechanism: more negative sentiment in response to economic decline, but no changes to positive sentiment.

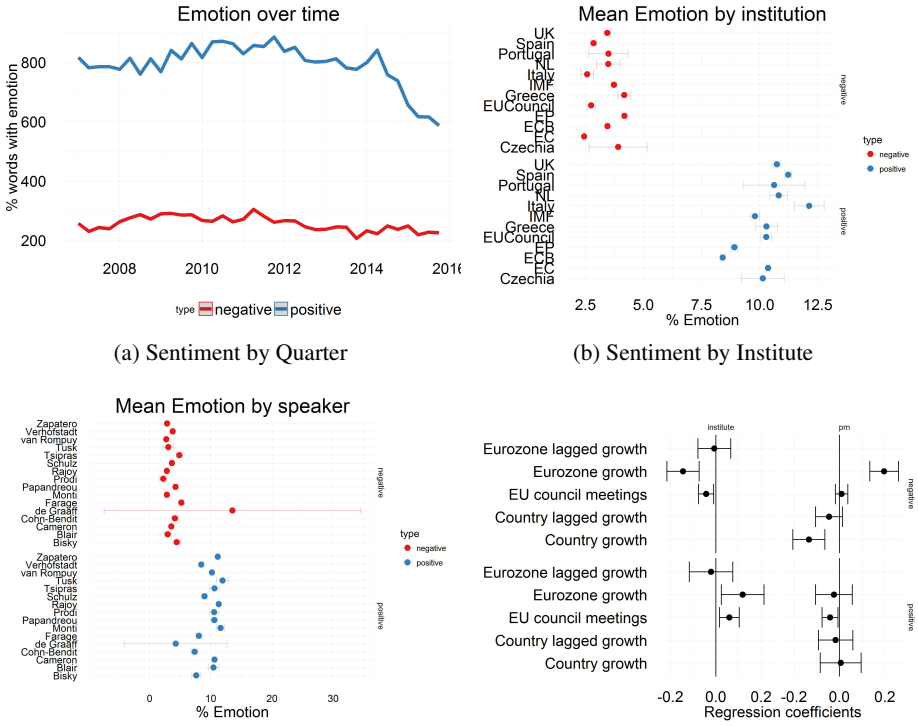
4 Scaling Models

4.1 Method

As a second illustration of the EUSpeech data we scale the European Parliament speeches for each EP group leader using *Wordfish* (Slapin and Proksch, 2008; Proksch and Slapin, 2009). *Wordfish* extracts substantively relevant quantities in an unsupervised manner, scaling these speeches on a latent dimension (Slapin and Proksch, 2008). Using select anchors (words and documents) we can retrieve the meaning of the latent dimension that *Wordfish* produces.[13] This approach relies on the assumption that the content of the political texts is predominantly ideological, and therefore informative of the policy position expressed by each actor (Grimmer and Stewart, 2013). For this analy-

[12] We also control for whether the text is translated or originally English (not presented)

[13] Exactly because we do not know *a priori* what the dominant ideological dimension is in the European Union we decided on using *Wordfish* rather than *Wordscores* which assumes we know the latent dimension.



(a) Sentiment by Quarter

(b) Sentiment by Institute

(c) Sentiment by Speaker

(d) Regression analysis Sentiment

Figure 1: **Sentiment Analysis**

sis we translated the roughly 1000 non-English EP speeches using *Google Translate*.

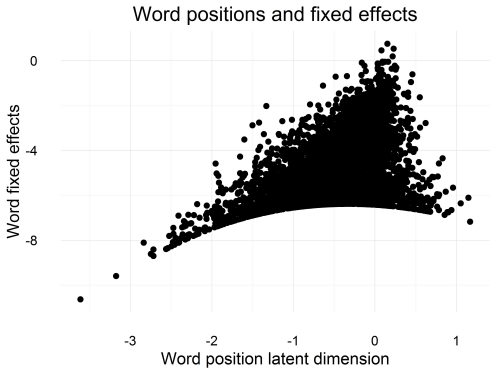
Results

The upper left panel in figure 2 demonstrates the placement of words along the single, latent continuum that wordfish estimates on the x-axis, and the words' fixed effects on the y-axis. This figure is usually referred to as an Eiffel Tower Plot. In the middle of the x-axis there are words that occur a lot, but do not distinguish positions. On the extremes of the x-axis we find words that occur less often, but are strong indicators for distance between documents. To make sense of both dimensions figure 2a lists some of the high-scoring (high betas) words on both ends of the dimension. Figure 2b shows word placement (a dot) on the ideological dimension (x-axis) and the word fixed effect (y-axis). The latter indicates how often the word occurs. Words high on the y-axis occur often and therefore do not distinguish well between documents. A word like "house" is such a word. Other words do distinguish well between documents, because some politicians use them and others do not. Some words do not occur that often (score low on y-axis) but are only used in some documents and not in others (extreme score on x-axis). We look to these words to identify the dimension that is estimated by the wordfish procedure. On the left-hand side of figure 2b we find negative word stems such as "abolit", "abort", "undemocrat" and "totalitarian". On the right-hand side we find stems such as "colegisl", "communitarian" and "Eurobond". On the basis of this we propose to identify the latent dimension as an anti-Europe vs pro-Europe dimension. Admittedly, we present here the words that make the most sense to make this case. On both ends of the dimension we also find words that are not easily placeable on any dimension. It is likely that splitting up (parts of) speeches according to topic will increase the clarity of the estimated dimension.

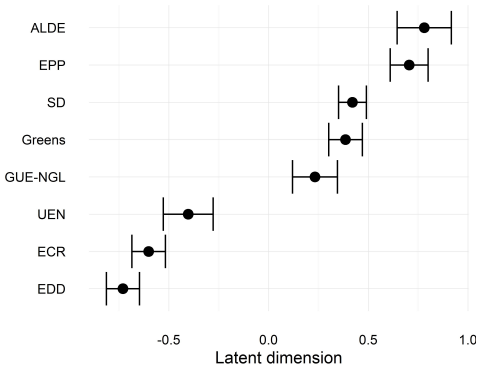
The wordfish analysis also estimates positions of the speeches on the latent dimension. For each party we calculate the mean of these positions and a 95% confidence interval (see figure 2c). The anti-European parties EDD, ECR and UEN cluster on the left of our dimension. The pro-European, mainstream parties EPP, ALDE and SD cluster on the right. To further validate our findings we compare our wordfish party estimates to the Euro-manifesto 2009 estimates of party positions on the

Anti EU words	Pro EU words
abolit	bargain
abort	bolder
atroc	colegisl
bureaucrat	communautair
disproportion	eurobond
fraud	eurogroup
monstros	formalis
resent	interinstitut
totalitarian	method
undemocrat	recapital

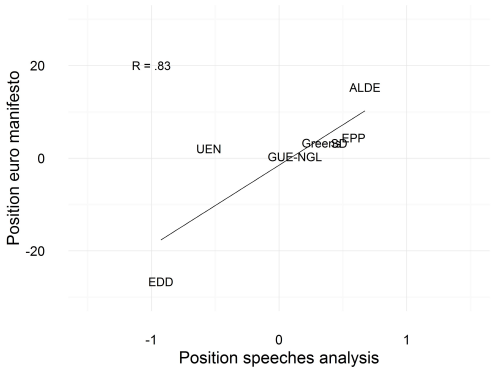
(a) Pro and Anti EU words



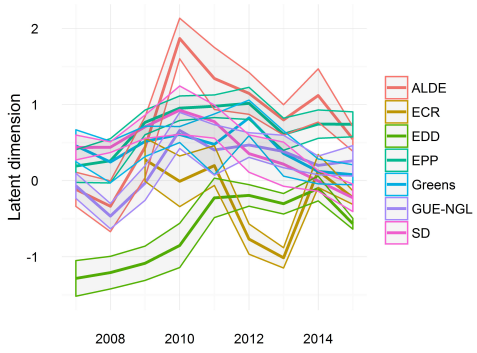
(b) Word Positions and Fixed Effects



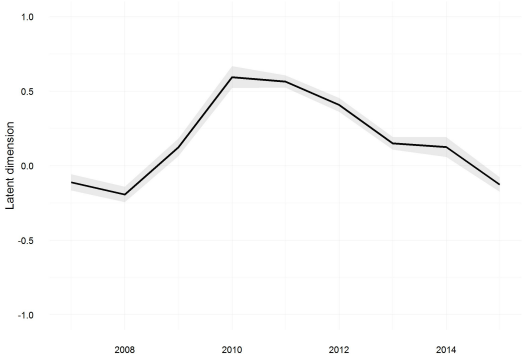
(c) Position by EP party



(d) Comparison to EU manifesto data



(e) Party position over time



(f) Position over time

Figure 2: **Ideological scaling**

anti-European vs pro-European scale. Figure 2d presents this data. The correlation between the two is .83. Hence, it is quite clear that our model measures a pro versus anti-European ideological dimension.

The last two plots display time trends. Figure 2e shows the mean position of parties over time with 90% confidence intervals. As is clear, there is quite some overlap between parties. The EDD is consistently the most anti-European party. The ECR fluctuates a bit more. The ENL is also anti-European, but has been omitted from this plot, since was only recently founded. ALDE and EDD are the most pro-European parties, but especially ALDE was more in the middle of the ideological scale until 2009. We ran a regression model to explain these party position changes. One, very strong predictor of party position change is party leadership change. The shift by ALDE coincides with the transition from Graham Watson to Guy Verhofstadt as party leader. Party leader changes also explain the ECR shifts. Interestingly, appoint-

ing a leader from the UK brings about a shift towards a more anti-European position. The somewhat dramatic changes to occur due to a leadership change, also suggests that leaders in European parties do not really take the middle ground of their party MEPs position. Otherwise, the party position would be more stable over time.

The final question is: what is the time trend? For this purpose we took the (unweighted) average per year of the party positions. Figure 2f displays this time trend. Initially, we see a shift towards a more pro-European position. This is primarily caused by the appointment of Verhofstadt as the ALDE leader, and by the moderation of the EDD. But after 2011 there is towards the middle of the ideological scale, towards a more euro-skeptical position. Here it is primarily ALDE and SD that moderated their pro-European position and the emergence of the ENL that shifts the mean. But also the Greens, EDD and ECR shift to a more anti-EU position.

5 Conclusion

In this paper we presented EUSpeech, a new dataset of 18,403 speeches of EU leaders, containing variation in sentiment and ideology, allowing for fine-grained analysis of representation the European Union. In analyses not presented here we also find interesting and predictable variation in speech topics, speech complexity and speech word usage. With these findings in mind, we think that EUSpeech will be a valuable resource for scholars interested in elite-mass interactions in the European Union.

6 Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 649281, EUENGAGE.

References

Aleksandra Cichocka, Michal Bilewicz, John T Jost, Natasza Marroush, and Marta Witkowska. 2016. On the grammar of politics—or why conservatives prefer nouns. *Political Psychology*, xx(xx):xx–xx.

Justin Grimmer and B. M. Stewart. 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3):267–297.

Justin Grimmer. 2010. A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1–35.

J.P. Kincaid, R.P. Fishburne, R.L. Rogers, and B.S. Chissom. 1975. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Memphis, Tennessee: Naval Air Station.

Saif M. Mohammad and Peter D. Turney. 2010. Emotions evoked by common words and phrases: using mechanical turk to create an emotion lexicon. *CAAGET '10 Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34.

S.-O. Proksch and J B Slapin. 2009. How to Avoid Pitfalls in Statistical Analysis of Political Texts: The Case of Germany. *German Politics*, 18(3):323–344.

Gijs Schumacher, Martijn Schoonvelde, Tanushree Dahiya, and Erik De Vries. 2016. Euspeech. dx.doi.org/10.7910/DVN/XPCVEI, Harvard Dataverse, V1.

J B Slapin and S.-O. Proksch. 2008. A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science*, 52(3):705–722.

T Wilson, J Wiebe, and P Hoffman. 2005. Recognizing contextual polarity in phrase level sentiment analysis. *Acl*, 7(5):12–21.

Lori Young and Stuart Soroka. 2012. Affective news: The automated coding of sentiment in political texts. *Political Communication*, 29(2):205–231.

Computer-aided dictionary making: An efficient dictionary construction technique for content analysis

Kohei Watanabe
London School of Economics
and Political Science
k.watanabe1@lse.ac.uk

Abstract

Dictionary-based content analysis has long been popular in social sciences, but manual construction of dictionaries is costly. Latent Semantic Scaling (LSS) is a computer-aided technique for dictionary construction, with which users can produce valid content analysis dictionaries with either 10 to 20 exemplary 'seed words' or about 10 manually scored documents. In this paper, political science examples show that the accuracy of computerized content analysis with LSS dictionaries is comparable to manually compiled dictionaries. R implementation of LSS is also publicly available.

1 Introduction

Use of keyword dictionaries, such as the General Inquirer Dictionary (Stone et al., 1966), LIWC (Francis and Pennebaker, 1993), the Regressive Imagery Dictionary (Martindale, 1975) and DICTION (North et al., 1984) has long been a popular approach to computerized content analysis. The technological simplicity of dictionary-based content makes its use intuitive for non-expert users and it is portable across different platforms.

In the dictionary-based approach, accuracy in computerized content analysis is achieved by careful choice of entry words. A good political science example is the policy position dictionary compiled by Laver and Garry (2000). Despite the fact that the dictionary was created in the 1990s with words chosen by the authors from British party manifestos, it was able to accurately locate the economic policy positions of the Conservatives (Con), the Liberal Democrats (LD) and Labour (Lab) in the 2000s (Figure 1). The correlation between machine and expert scores was as high as r=0.843.

However, valid content analysis dictionaries are only available for a very limited range of topics or types of documents. If existing content analysis dictionaries are utilized for an analysis of documents distinct from these, it raises concerns regarding the validity of results (Grimmer and Stewart, 2013). For example, the Lexicoder Sentiment Dictionary (LSD) successfully measured positive-negative tones in newspaper coverage to predict the outcome of the 2006 Canadian federal election (Young and Soroka, 2012), but it was not able to analyze British political parties' sentiments toward immigration policy as expressed in their 2010 manifestos (Figure 2). The correlation between crowd-sourced coders (Amazon MT) and LSD is only r=0.102.

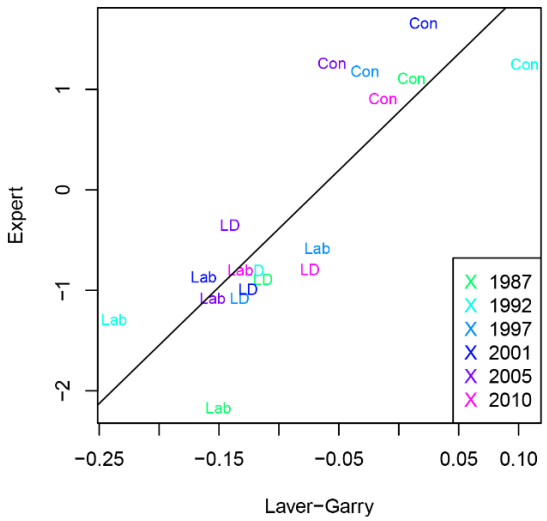


Figure 1: Economic policy position in 1987-2010 UK manifestos by Laver-Garry dictionary.

The inability of LSD to analyze sentiment toward immigration policy is due to the difference in vocabulary between newspaper articles on

Proceedings of the International Conference on the Advances in Computational Analysis of Political Text (PolText 2016), pages 81–87, Dubrovnik, Croatia, 14–16 July 2016.

general elections in Canada and political pamphlets on immigration policy in Britain. When existing dictionaries appear unsuitable for an analysis of documents of interest, a new dictionary has to be created, but it usually requires a much time and labor, undermining the very benefit of computerized content analysis.

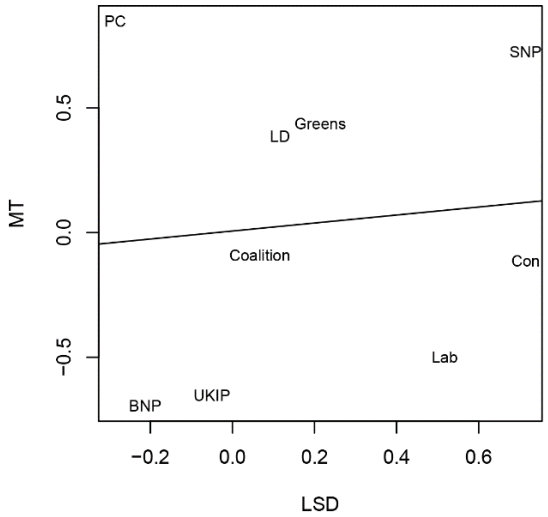


Figure 2: Immigration sentiment in 2010 UK manifestos by LSD

2 Computer-aided dictionary construction

LSS assists construction of subject-specific content analysis dictionaries based on statistical analyses of large corpora. It can be used either as (1) a lexicon-expansion technique or (2) a supervised document scaling technique. Its dictionary construction is based on the following four steps.[1]

2.1 Corpus preprocessing

LSS utilizes subject-specific large corpora to statistically estimate semantic values of words. The minimum size of a corpus for LSS is around 10 million words. In the corpus, documents have to be unitized into sentences, and all the proper nouns and function words should be removed before processing.

2.2 Word selection

LSS selects words that frequently occur with target words, aiming to collect modifiers of the target words, such as 'economy' or 'immigration'. Word selection is performed by collocation analysis of the corpus, and words that appear statistically significantly ($p<0.001$) more frequently than expected enter the dictionary. Collocation is defined as occurrence within 10-word windows from target words and measured by likelihood ration statistic (Hoey, 2012).

2.3 Word scoring

Entry words are scored by cosine similarities to pre-defined 'seed words'. For example, English positive and negative seed words are {good, nice, excellent, positive, fortunate, correct, superior} and {bad, nasty, poor, negative, unfortunate, wrong, inferior} (Turney and Littman, 2003).

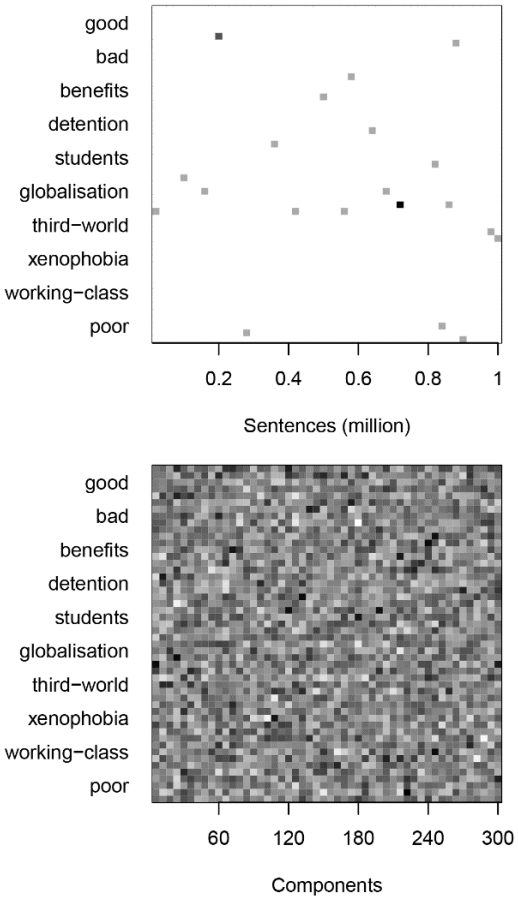


Figure 3: Notional illustration of dimension reduction by Singular Value Decomposition.

Yet, in LSS, cosine similarities are not calculated in the raw term-sentence matrix, but in a reduced term-sentence matrix utilizing Singular Value Decomposition (SVD) to decompose a large sparse matrix into a smaller dense matrix (Figure 3), a technique known as Latent Semantic Analysis (Deerwester et al., 1990). When X denotes the term-sentence matrix, SVD decomposes it into three matrices, U, D and V.

[1] Available in R at https://github.com/koheiw/LSS.

$$X \approx \hat{X} = UDV' \quad (1.1)$$

$$\hat{S} = UD \quad (1.2)$$

With the matrix $\hat{S}$, LSS estimates the sentiment of words by their cosine similarity to the seed words: The sentiment score v_i for a word w_i is a total cosine similarity to seed words weighted by seed scores p_j, which were simply +1 for the positive seed words and −1 for the negative seed words. Here $\cos(w_i, s_j)$ denotes cosine similarity between two row vectors corresponding to entry word w_i and seed word s_j in the matrix $\hat{S}$.

$$v_i = \sum_{j}^{n} \cos(w_i, s_j) \cdot p_j \quad (1.3)$$

2.4 Document scoring

Once scores are assigned to entry words, dictionary construction is completed and dictionaries are ready for content analyzing documents. In content analysis, users can either apply LSS dictionaries as (1) words with continuous scores, or (2) words in two discrete categories.

With continuous scores, document scores are weighted means of word scores, as in Wordscore (Laver et al., 2003): when entry words $w_{i...I}$ occur in a document a total of m times, and v_i is the word score and f_i is the frequency count of an entry word w_i, its document score d is computed thus:

$$d = \frac{1}{m} \sum_{i}^{I} v_i \cdot f_i \quad (1.4)$$

An LSS dictionary can also be transformed into two sets of words by splitting words by the median score, making its structure identical to traditional content analysis dictionaries. In this case, the document score d is the difference between the normalized frequency of the two sets of words:

$$d = \frac{n_{\text{upper}} - n_{\text{lower}}}{l} \quad (1.5)$$

Where n_{upper} and n_{lower} are numbers of words belonging to the upper and lower half of the dictionary, and l is the total number of words in the document.

3 Example: Immigration sentiment dictionary

With the general English positive-negative seed words, I constructed an immigration sentiment dictionary using LSS without any manual intervention. The corpus for dictionary was British newspaper articles between 2009 and 2010, which contains 15,343 stories or 11.6 million words. Target words were defined by glob patterns, "immingra*" and "migra*".

This immigration sentiment dictionary is comprised of 1,000 words. The most positive and negative words are presented in Table 1. While many of the positive words relate to legal and economic aspects of migration (litigants, detention, benefits, scroungers), negative words mainly concern the social classes and origins of migrants (poor, working-class, frontier, eastern). There are words related to animal migration (conservationist) or migraine (epilepsy, headache), but these words do no harm in analyzing political documents.

Rank	Entry Word	Score
1	issues	0.615
2	policies	0.601
3	ensure	0.585
4	benefits	0.444
5	litigants	0.430
6	huge	0.430
7	detention	0.410
8	wobbling	0.401
9	impromptu	0.396
10	documents	0.390
11	handed	0.374
12	conservationist	0.351
13	joint	0.339
14	restrictive	0.333
15	students	0.326
16	reduced	0.323
17	lounge	0.322
18	bring	0.321
19	major	0.319
20	scroungers	0.313
981	warned	-0.451
982	failure	-0.454
983	areas	-0.457
984	stemming	-0.466
985	makeup	-0.476
986	epilepsy	-0.478
987	countries	-0.506
988	exposed	-0.510
989	eastern	-0.510
990	intentioned	-0.516

991	benefited	-0.527
992	poorer	-0.539
993	frontier	-0.549
994	white	-0.559
995	headache	-0.613
996	negative	-0.646
997	tide	-0.683
998	xenophobia	-0.761
999	working-class	-0.778
1000	poor	-1.302

Table 1: Most positive and most negative entry words for an immigration sentiment dictionary.

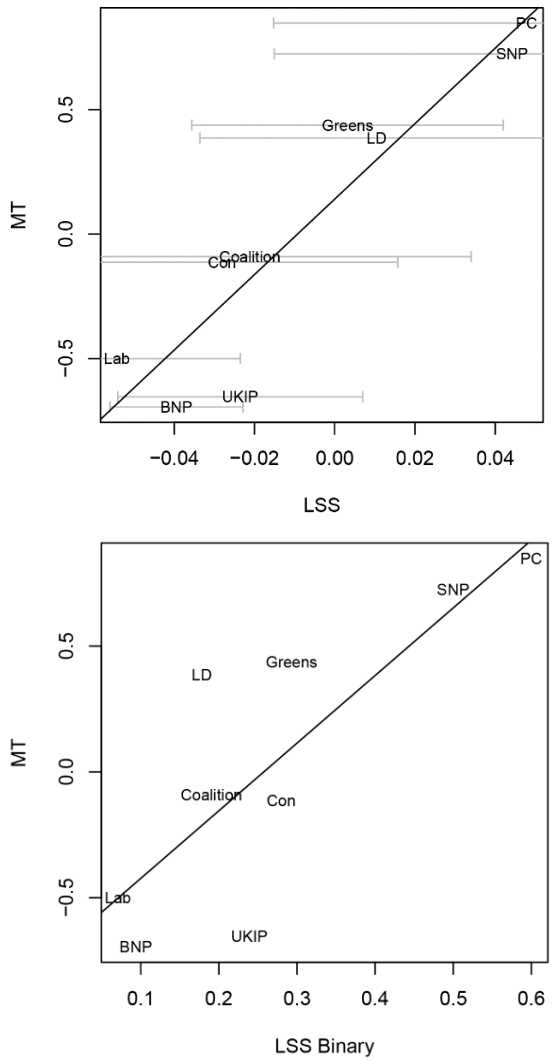


Figure 4: Immigration sentiment in 2010 UK manifestos by LSS.

I applied the immigration sentiment dictionary to the sections on immigration policy in the 2010 UK manifestos. As shown in the first plot in Figure 4, 95% confidence intervals were large due to only brief mentions of immigration in the party manifestos, but point estimation was very accurate (r=0.925). Even when the dictionary was dichotomized by the median score (second plot in Figure 4), it still achieved high correlation with manual scores (r=0.808).

4 Automated seed word selection

Users of LSS can construct their own set of seed words, but selection of seed words for complex dimensions is usually challenging. Therefore, when valid seed words are absent, users can employ a machine learning algorithm to select seed words from a corpus with manually scored documents.

In this automated seed selection, the system tests the suitability of candidates for seed words individually against manually scored documents to obtain polarity of the seed words; then seed candidates are paired with other seed words with opposite polarities to construct a seed set made up of around 10 pairs.

To test the suitability of each seed candidate, the system has to create a large number of tentative dictionaries, but it can be completed very quickly by initial calculation of pair-wise cosine similarities between all of these seed candidates. The system calculates pair-wise cosine similarities in an SVD-reduced matrix $\hat{S}$ (Equation 1.2) that is created from a corpus. The cosine similarities for all pairs are stored in a symmetric matrix D, which has K columns and rows corresponding to the seed candidates $c_{k...K}$. Given the similarity matrix D, a temporary dictionary for a seed word c_k is a kth row or column vector of the matrix D.

$$d_k = D_{\cdot k} = D_{k\cdot} \qquad (2.1)$$

First, the system creates K temporary dictionaries in this way, and applies them to the training set (Equation 1.4) to obtain correlation coefficients r_k between scores computed by the temporaries d_k and scores manually assigned. These correlation coefficients allow the system to identify the importance and polarity of the seed candidates. The importance of seed candidates is measured by the sizes of the correlation coefficients; the polarity of seed candidates is given by the signs of the correlation coefficients. The system selects only 50 seed candidates with the largest absolute correlation coefficient from both sides of polarity, and assigns seed scores p_k in the following manner:

$$p_k = \begin{cases} +1, & r_k > 0 \\ -1, & r_k < 0 \end{cases} \quad (2.2)$$

Then, seed words are given adjusted scores to make scoring of documents more consistent when they are combined into a single seed set. An adjusted seed score $\acute{p}_k$ is a seed score weighted by the inverse of average squared similarity to other seed candidates in the matrix D (Equation 2.1):

$$\acute{p}_k = p_k \cdot \frac{1}{\sum D_{\cdot k}^2 \cdot \frac{1}{K}} \quad (2.3)$$

Second, with the one hundred seed candidates polarities, the system constructs pairs of seed words $\{c_k, c_l\}$, searching for partner c_l for c_k such that (1) the partner has opposite polarity $p_l \neq p_k$, (2) the dictionary $d_{\{k,l\}}$ yields a higher correlation coefficient than the separate dictionaries $r_{\{k,l\}} > r_k$ and $r_{\{k,l\}} > r_l$, and (3) the correlation becomes the strongest with the partner $r_{\{k,l\}} \geq r_{\{k,\bar{k}\}}$. Starting from the seed candidate with the largest absolute correlation coefficient $|r_k|$, all other seed candidates enter this stepwise paring process. This process continues until at least five pairs have been found; new pairs decrease the overall correlation. The process takes only around 30 seconds on a laptop computer.

In the above process, the system can easily construct a dictionary with a large number of entry words with any set of seed words. Scores assigned to entry words $v_{k\ldots K}$ are calculated simply by taking inner products of the weighted seed scores and a subset of the similarity matrix $\acute{D}$ that only has columns corresponding to the seed words:

$$v_k = \acute{D} \cdot \acute{p}_k \quad (2.4)$$

5 Example: Economic policy position dictionary

As an example of this automated seed word selection, I created an economic policy position dictionary with a corpus of UK economic news stories published prior to elections in 1987, 1992 and 1997, which contains 45 million words in 63,759 news articles. Target words were defined by a glob pattern "economy*". The training set for machine learning was party manifestos from the three pre-millennium elections (9 documents).

In this instance, seed words were selected from words relevant to economy (the same criteria as entry words selection). From the economy-related words, a supervised learning algorithm identified pairs of seed words, producing dictionaries that replicate manual scoring. Through forward step-wise selection, LSS discovered 10 pairs of seed words and assigned weighted seed scores to them (Table 2).

Step	Seed Word	Seed Score
1	rate	478.58
8	run	347.15
10	bottom	191.53
6	miracles	148.99
7	treasury	140.51
9	remain	121.42
3	mpg	110.77
5	tight	107.67
2	acceleration	102.93
4	backdrop	102.84
10	improve	-99.97
8	provide	-109.08
3	generate	-112.00
4	unbalance	-118.95
1	damage	-130.90
7	harm	-131.83
5	based	-137.36
2	appraisal	-148.59
9	disruption	-189.70
6	general	-408.29

Table 2: Seed words for economic policy position dictionary.

The economic policy position dictionary created with the seed words accurately scored not only the British election manifestos from 1987-1997 but also those from 2001-2005, showing its out-of-sample validity (first plot in Figure 5): its errors were smaller than in the Laver-Garry dictionary (Figure 1) particularity in the extreme ranges, although the 2010 manifestos were not very accurately located. Even when the dictionary was dichotomized, the result remained very similar (second plot in Figure 5).

I also applied a Bayesian model, Wordscore, to the same training set to obtain a benchmark for the supervised LSS. The result in Figure 6 clearly shows Wordscore's inability to accurately score the 2000s manifestos by a model created from the 1987-1997 manifestos. This highlights

the advantage of corpus-based dictionary construction by supervised LSS. That is, since words in the LSS dictionary were scored based on statistical analysis of the large corpus instead of the small training documents, it is unaffected by noise in the training set. The clearest indication of the absence of overfitting is the reasonably large confidence intervals for manifestos from 1987-1997 in Figure 5.

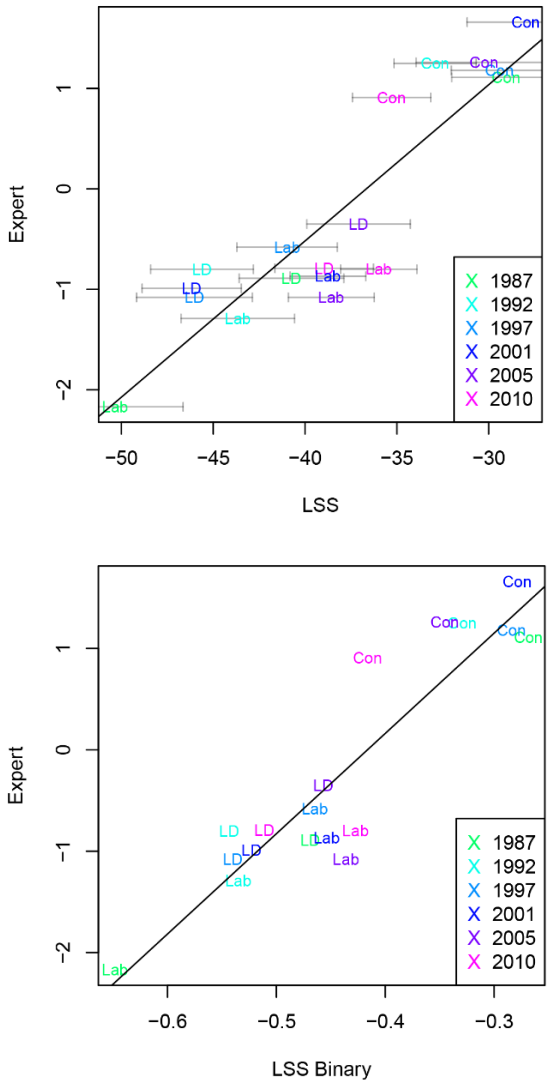


Figure 5: Economic policy position in 1987-2010 UK manifestos by LSS.

Finally, LSS was still not able to score the 2010 manifestos as accurately as the Laver and Garry dictionary, presumably because of the structural break in language of economic policy after the 2008 economic crisis. However, its accuracy can be improved by including economic news articles from 2001, 2005 and 2010 to the corpus. A new dictionary constructed with the extended corpus accurately scored manifestos in the 2000s, better distinguishing the Conservative from the Liberal Democrats and Labour in the 2010 manifestos (Figure 7).

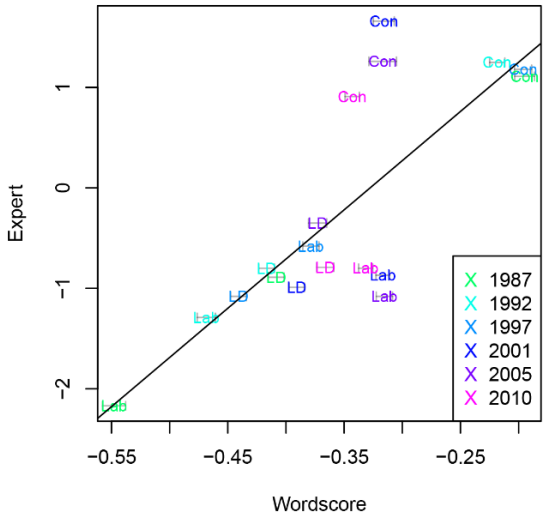


Figure 6: Economic policy position in 1987-2010 UK manifestos by Wordscore.

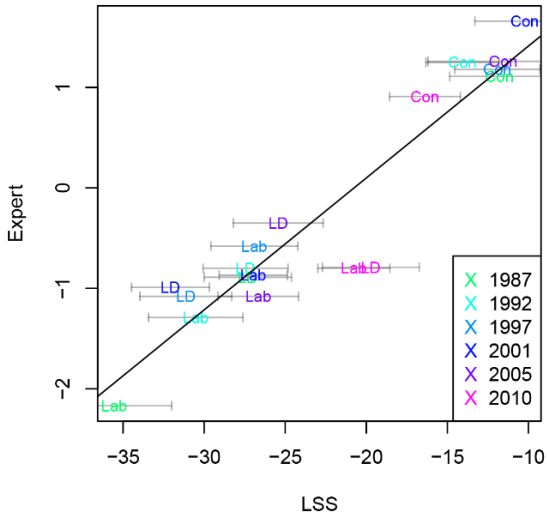


Figure 7: Economic policy position in 1987-2010 UK manifestos by LSS with extended corpus.

6 Conclusion

As evidenced in the examples, LSS dramatically reduces human involvement in dictionary construction: In the lexicon expansion, only 14 manually chosen seed words were required to create a subject-specific sentiment dictionary. In supervised machine learning, only 9 manually scored documents were sufficient for automatically discovering seed words. Further, the accu-

racy of content analysis using the dictionaries produced by LSS is comparable to manually compiled dictionaries.

LSS also has an advantage over other supervised techniques that rely on parameter estimation of small training data. By statistically analyzing large corpora, LSS discovers more general semantic values of words, achieving a greater degree of external validity. As a result, LSS dictionaries content analyze unseen documents more accurately than other models.

Reference

Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407.

Martha E. Francis and James W Pennebaker. 1993. LIWC: Linguistic Inquiry and Word Count. Technical report, Southern Methodist University, Dallas, Texas.

Justin Grimmer and Brandon M. Stewart. 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political. *Political Analysis*.

Jesse Hoey. 2012. The Two-Way Likelihood Ratio (G) Test and Comparison to Two-Way Chi Squared Test. *arXiv:1206.4881 [stat]*, June.

Michael Laver, Kenneth Benoit, and John Garry. 2003. Extracting Policy Positions from Political Texts Using Words as Data. *American Political Science Review*, 97(2):311–331.

Michael Laver and John Garry. 2000. Estimating Policy Positions from Political Texts. *American Journal of Political Science*, 44(3):619, July.

Colin Martindale. 1975. *Romantic progression : the psychology of literary history*. Hemisphere Publishing ; New York ; London, Washington, DC.

Robert North, Richard Lagerstrom, and William Mitchell. 1984. DICTION Computer Program: Version 1. Technical report, July.

Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. The M. I. T. Press.

Peter D. Turney and Michael L. Littman. 2003. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Trans. Inf. Syst.*, 21(4):315–346, October.

Lori Young and Stuart Soroka. 2012. Affective News: The Automated Coding of Sentiment in Political Texts. *Political Communication*, 29(2):205–231.

Classifying Topics and Detecting Topic Shifts in Political Manifestos

Cäcilia Zirn[1], Goran Glavaš[1,3], Federico Nanni[1], Jason Eichorst[2], Heiner Stuckenschmidt[1]
[1] Data and Web Science Group, University of Mannheim
B6 26, DE-68161 Mannheim, Germany
[2] Collaborative Research Center SFB 884, University of Mannheim
L13 15-17, DE-68161 Mannheim, Germany
[3] Text Analysis and Knowledge Engineering Lab, University of Zagreb
Unska 3, HR-10000 Zagreb, Croatia
{caecilia,goran,federico,heiner}@informatik.uni-mannheim.de
eichorst@uni-mannheim.de

Abstract

General political topics, like social security and foreign affairs, recur in electoral manifestos across countries. The Comparative Manifesto Project collects and manually codes manifestos of political parties from all around the world, detecting political topics at sentence level. Since manual coding is time-consuming and allows for annotation inconsistencies, in this work we present an automated approach to topical coding of political manifestos. We first train three independent sentence-level classifiers – one for detecting the topic and two for detecting topic shifts – and then globally optimize their predictions using a Markov Logic network. Experimental results show that the proposed global model achieves high classification performance and significantly outperforms the local sentence-level topic classifier.

1 Introduction

The Comparative Manifesto Project (CMP), initiated by Volkens et al. (2011), collects party election programs (so-called manifestos) from elections in many countries around the world. The goal of the project is to provide a large data collection to support political studies on electoral processes. A sub-part of the manifestos has been manually topically coded by political scientists. Each manifesto sentence has been labeled with one of over fifty political topics, divided into 7 coarse-grained domains.[1] While manual annotations are very useful for political analyses, they come with two major drawbacks. First, it is very time-consuming and labor-intensive to manually annotate each sentence with the correct category from a complex annotation scheme. Secondly, coders' preferences towards particular categories might cause annotation inconsistencies and disallow for comparability between manifestos annotated by different coders Mikhaylov et al. (2012).

Automated topic classification of political manifestos does not only save human resources, but it additionally provides comparable and reproducible annotations. Thus, in this work we develop a supervised framework for classifying the broad domain of sentences in political manifestos, with the specific goal of assisting human coders. Our pipeline consists of three different classifiers predicting the domains and domain shifts between pairs of adjacent sentences. They rely on a variety of features including bags-of-words and semantic textual similarity (STS) (Agirre et al., 2012; Šarić et al., 2012). In the second step, we exploit the global context of the manifestos and combine the sentence-level predictions of these three local classifiers in a global Markov Logic-based optimization setting (Richardson and Domingos, 2006), where we introduce additional global information as constraints on the prior distribution of topics, topic shifts and sequences of topics.

We evaluate each of the local classifiers and show that the introduction of global information is justified by the fact that the globally-optimized Markov Logic classifier significantly outperforms the local topic classifier and reaches the satisfactory performance of almost 80% F_1 score.

2 Related Work

The body of work on automated analysis of political texts is substantial (Grimmer and Stewart, 2013). Approaches to classification of political

[1] https://manifestoproject.wzb.eu/coding_schemes/mp_v5

Proceedings of the International Conference on the Advances in Computational Analysis of Political Text (PolText 2016), pages 88–93, Dubrovnik, Croatia, 14–16 July 2016.

texts can be roughly divided into two major groups – dictionary based methods (Kellstedt, 2000; Young and Soroka, 2012) and methods that employ supervised classification models (Purpura and Hillard, 2006; Stewart and Zhukov, 2009; Verberne et al., 2014; Karan et al., 2016). The idea behind all dictionary-based methods is similar – they first identify words that distinguish categories and then measure the occurrence frequencies of those words in texts, regardless of whether the task is recognition of racial policies from media sources (Kellstedt, 2000) or detection of affects and sentiment in political texts (Young and Soroka, 2012).

The counting principle of the dictionary-based approaches might be suitable when classifying larger fragments of text such as paragraphs or whole documents. However, all dictionaries are of limited coverage and are thus unable to capture less obvious indicator terms. This is even more emphasized when classifying short texts (e.g., sentences) as it is unlikely that many dictionary words will appear in such a short text. Along with the fact that sets of indicator words need to be compiled manually, this is why the research focus shifted to supervised classification models. Stewart and Zhukov (2009) label 8000 Russian military statements and train an ensemble of classifiers to predict whether the statements originate from activists or conservatives. Purpura and Hillard (2006) propose a two-level hierarchical classification of US legislative documents using support vector machines and standard TF-IDF weighted bag-of-words features. Karan et al. (2016) propose a very similar approach for classifying Croatian legal documents, using only document titles as input. Considering that titles are significantly shorter pieces of text, they combined traditional bag-of-word features with semantic vector representations (i.e., word embeddings) to avoid the sparseness issues.

Classification of short texts has been shown to be more challenging than document level classification. Short texts contain less words and thus require an additional semantic information, as opposed to only lexical (i.e., symbolic) input. Phan et al. (2008) build a framework for classifying short and sparse text and web snippets. They use external databases, such as MedLine, as the source of semantic knowledge that reveals hidden topics. Similarly, (Hu et al., 2009) exploit world knowledge to cluster short text snippets. The snippets do not provide enough vocabulary overlap when using only bag-of-words representations. Therefore, the authors enrich the text with internal semantics, i.e. deep understanding of the text, and external semantics from resources like Wikipedia and WordNet. The lack of appropriate knowledge bases for the political domain, however, make such approaches not applicable in our case. Instead, besides lexical features, we rely on word embeddings – general vector representations that capture well semantics of words – to topically classify manifesto sentences.

Hachey and Grover (2004) classify the rhetorical status of a sentence for text summarization. Besides lexical features, they add information such as the position of a sentence in the document and named entities. They then apply sequence labeling to predict rhetorical roles for a sequence of sentences in a document. Similarly, in this work we combine various sources of information for local sentence topic classification. We then include these classifiers in a sequence labeling model for identifying globally optimal topic sequences of a given manifesto. We decide to employ Markov logic network as a sequence labeling model because it has been already successfully applied to numerous sequence labeling tasks in natural language processing (Poon, 2010; Che and Liu, 2010; UzZaman et al., 2012; Zirn et al., 2011).

3 Topic Classification of Political Manifestos

Our goal is to support human annotators to assign manifesto sentences to political categories. The CMP distinguishes between over 50 fine-grained political categories that are grouped into seven topical areas: *External Relations*, *Freedom and Democracy*, *Political System*, *Economy*, *Welfare and Quality of Life*, *Fabric of Society* and *Social Groups*.

We first build a local sentence-level classifier that predicts one of the seven topics based on the information extracted from the sentence. Next, we employ two topic-shift classifiers that predict whether two adjacent sentences are on the same topic or not. Finally, we add information on distributions of topics and topic sequences on top of the predictions and combine all components in a global Markov Logic framework, which determines the optimal topical classification for all sentences of a manifesto.

3.1 Local Topic Classifier

The local sentence-level topic classifier makes predictions taking into account only the information from the sentence itself. To this end, a linear SVM classifier with the following set of lexical and numerical features was employed:

1. The bag-of-words term-vector of the sentence;
2. The topic of the preceding sentence;
3. The semantic similarity between the current and preceding sentence, which is computed by greedily aligning most similar words from the two sentences. Let P be the set of greedily aligned pairs (w_1, w_2) of words (where w_1 is from the first sentences, and w_2 is from the second sentence). The raw semantic similarity between the sentences is then given as:

$$sim(s_1, s_2) = \sum_{(w_1,w_2)\in P} \cos(v_{w_1}, v_{w_2})$$

   where v_w is the semantic embedding vector of the word w. We used the pretrained set of 200-dimensional GloVe embeddings[2] (Pennington et al., 2014) to compute the raw semantic similarity score. Because the similarity given by the above-mentioned formula depends on the length of the sentences, we normalized the score by the length of the sentences
4. For each topic class we also computed a numeric feature indicating the level of relative relevance of the sentence words for that class. We computed the relative frequencies of lemmas in sentences belonging to each of the topic classes on the train set. For example, if the word *"social"* appeared n times in all sentences of the train set labeled with the topical class *"Social Fabric"* and these sentences together contain N words, then $\frac{n}{N}$ is the relative relevance of the word *"social"* for the *"social security and welfare"* topic. Let $rr(w, c)$ be the relative relevance of the word w for the topical class t. The relevance score of the sentences s for the class t is then computed as follows:

$$rs(s, c) = \frac{1}{|s|} \sum_{w\in s} rr(w, c)$$

   where $|s|$ is the total number of words in the sentence s. For each sentence, one relevance score (i.e., one feature) is computed for each of the topical classes.

3.1.1 Topic-Shift Classifiers

We employ binary classifiers that predict whether two given adjacent sentences are on the same topic or not. We used the following set of features for the detection of local topic shifts:

1. Bag-of-words term-vector of the first sentence (f^1);
2. Bag-of-words term-vector of the second sentence (f^2);
3. Length (in no. words) of the first sentence (f^3);
4. Length (in no. words) of the second sentence (f^4);
5. Semantic similarity between the two sentences (f^5, cf. Section 3.1);
6. Ngram overlap between the two sentences (f^6) – the number of shared content words, normalized by the length of the sentences.

Considering the large size of the feature space due to the lexical BoW features f^1 and f^2, we first attempted to feed all features to a single linear SVM classifier. However, we observed that the numerical features (f^3–f^6) yield no improvements in classification performance over using only BoW vectors (f^1–f^2). We then fed only the numerical features to the SVM classifier with a non-linear RBF kernel and obtained similar cross-validation performance on the train set as when using the linear SVM classifier with only the bag-of-words features. Considering that the two classifiers – (1) the linear SVM using the bag-of-words features and (2) the RBF SVM with four numeric features – address the same task with completely disjoint sets of features, we decided to incorporate local predictions of both classifiers into the global optimization framework.

3.2 Topic Distribution Information

In addition to the information we gain from the sentence content, we make use of knowledge about the distribution and sequences of topics in manifestos. One salient observation is that topics are usually

[2] http://nlp.stanford.edu/data/glove.6B.zip

tackled in several consecutive sentences, so successive sentences tend to share the same label. If we take this observation a step further, we can measure the probability of topic transitions (e.g., conditional probability of a sentence of topic *Economy* following a sentence of topic *Social fabric*). We estimate these conditional probabilities on the train set. This does not only help us to decide whether two consecutive sentences share the same label, but gives us an estimate for probable sequences of topics.

3.3 Global Optimization

Markov logic (Richardson and Domingos, 2006) can be interpreted as a template language combining first-order logic with maximum entropy models. The user can specify types of data and encode prior knowledge about the information used in the classification scenario, and it searches the most probable world given the evidence.

A Markov network $\mathcal{M}$ is an undirected graph whose nodes represent a set of random variables $\mathbf{X} = \{X_1, ..., X_n\}$ and whose edges model direct probabilistic interactions between adjacent nodes. More formally, a distribution P is a log-linear model over a Markov network $\mathcal{M}$ if it is associated with:

- a set of features $\{f_1(D_1), ..., f_k(D_k)\}$, where each D_i is a clique in $\mathcal{M}$ and each f_i is a function from D_i to $\mathbb{R}$,
- a set of real-valued weights $w_1, ..., w_k$, such that

$$P(\mathbf{X} = \mathbf{x}) = \frac{1}{Z} \exp\left(\sum_{i=1}^{k} w_i f_i(D_i)\right),$$

where Z is a normalization constant.

A Markov logic network is a set of pairs (F_i, w_i) where each F_i is a first-order formula and each w_i a real-valued weight associated with F_i. With a finite set of constants C it defines a log-linear model over possible worlds $\{\mathbf{x}\}$ where each variable X_j corresponds to a ground atom and feature f_i is the number of true groundings (instantiations) of F_i with respect to C in possible world $\mathbf{x}$. Possible worlds are truth assignments to all ground atoms with respect to the set of constants C. We explicitly distinguish between weighted formulas and *deterministic* formulas, that is, formulas that always have to hold.

Given a set of first-order formulas and a set of ground atoms, we wish to find the formulas maximum a posteriori (MAP) weights, that is, the weights that maximize the log-likelihood of the hidden variables given the evidence.

3.3.1 Model

We model each sentence of the manifesto as a constant $s \in S$. In the same manner, topics 1-7 are represented as constants. First, we specify that each sentence s is mapped to exactly one topic t as a deterministic formula:

$$\forall s, t : |t| map(s, t) = 1$$

As we intend to predict the correct mappings, $map(s, t)$ is our hidden predicate. We introduce the predicate $next(s1, s2)$ stating that sentence $s1$ is followed by $s2$ to model the sequences of sentences in a manifesto. This allows us to encode our observation that subsequent sentences share the same topic:

$$\forall s, c : next(s_1, s_2) \wedge map(s_1, t) \Rightarrow map(s_2, t)$$

In contrast to the first formula, this one can be violated with a certain penalty, thus the formula is given a weight. Estimations about the transition between two particular topics are modelled alike by replacing t by particular variables t_1, t_2.

The predictions from the local sentence classifiers are modeled with the predicate $localConf(s, t, conf)$, where $conf$ represents the confidence for sentence s to be mapped to a particular topic t. We use this confidence as the weight for the corresponding formula:

$$\forall s, t : localConf(s, t, conf) \wedge map(s, t)$$

Each of the sentence-pair classifiers is modeled (separately) via a predicate called $flip$.

$$\forall s, t : shift(s_1, s_2, conf) \wedge map(s_1, t) \Rightarrow \neg map(s_2, t)$$

It expresses the confidence of a sentence pair belonging to two different topics: the label of the first sentence is "flipped" if the formula is true, i.e. if the confidence by the classifier (included as the weight for the formula) is high enough.

4 Experiments

In our experiments we used six U.S. manifestos (Republican and Democrat manifestos from 2004, 2008, and 2012 elections). In all experiments, we perform folded cross-validation and report the micro-averaged results over folds.

Topic	P	R	F_1
External Rel.	83.7	86.6	85.1
Freedom & Dem.	68.0	59.9	63.7
Pol. system	69.7	65.7	67.6
Economy	73.9	77.4	75.6
Welfare & QoL	72.8	72.8	72.8
Fabric of Soc.	74.8	76.0	75.4
Soc. Groups	71.2	67.9	69.5
Micro-avg.	74.9	74.9	74.9

Table 1: Local topic classification, 10-fold CV (%)

Model	P	R	F_1
Linear, bow feat.	56.6	54.6	55.6
RBF, num. feat.	98.5	27.4	42.9

Table 2: Topic-shift classification, 10-fold CV (%)

Topic Classification Table 1 shows the results of the local topic classifier obtained via the 10-fold CV. The classification performance is best for *External relations* (more easily recognizable due to re-occurring country names) and worst for *Freedom and democracy* (as lexical clues typical for this class tend to frequently appear in sentences of other topics as well).

Topic Shift Classification The performance of the two topic-shift classifiers is given in Table 2. These results indicate that detecting topic shifts is a more difficult task that predicting the topics of individual sentences. This is expected, as correctly identifying the topic shift logically amounts to correctly predicting topics for two consecutive sentences.

Global Classification The predictions of local classifiers are combined with the topic distribution information in a Markov Logic Network (MLN). We use RockIt (Noessner et al., 2013) as the MLN engine.

To evaluate the impact of each component, we start the experiments with a reduced set of formulas and incrementally add more constraints. As a baseline, we simply use the predictions by the local classfier (setting L). In the second setting, we encode rules for transitions (setting T) between particular topics. This is directly compared to a simpler setting S where we just assign consecutive sentences the same label instead of adding an own transition rule for every possible sequence of topics. The results of these combinations applied to the validation set are shown in the first part of the Table 3. Adding the information about consecutive sentences and transitions improves over the local classifier performance for 4 points, reaching 78.3%.

Setting	MaP	MaR	MaF_1	miF_1
L	73.5	72.3	72.8	74.9
L,T	80.7	73.1	75.2	78.3
L,S	78.3	74.5	75.9	78.3
L,S,P_{bow}	74.2	73.0	73.6	75.6
L,S,P_{num}	78.6	76.7	**77.5**	**79.3**
L,S,P_{bow},P_{num}	74.4	73.2	73.7	75.8

Table 3: Global classification (validation-set): MaP/MaR/MaF_1 = Macro precision/recall/F_1-measure; miF_1 = micro F_1-measure

As precision and recall are more balanced for setting S and it needs significantly less rules, we prefer it over setting T for the following experiments. We now employ the predictions of the topic-shift classifiers: P_{BOW} are the predictions of the linear SVM model with BOW features and P_{num} denotes the predictions of the non-linear SVM using numerical features. We first test each one seperately, then both together (setting L + N). The lower part of table 3 shows the results. The best performance of 79.3% F_1 score is obtained for the model using predictions P_{BoW}. The combination of both sentence pair classifiers drops performance, which is not surprising due to the performance of classifier P_{num}.

5 Conclusion

We presented an approach for sentence-level topical classification of party manifesto, which can be used to assist human coders in the CMP project and will allow for better reproducibility and comparability of the manually coded manifestos and will speed up the annotation process. We intend to conduct future experiments that evaluate the benefits of the application to the coding process. Furthermore, we showed that the addition of contextual and structural information about the documents improves the topical classification performance. Our approach could benefit from a cross-lingual information, i.e., from exploiting topical sequences common across different countries and languages.

Acknowledgments

The authors thank the DFG for Funding under the SFB 884 Political Economy of Reforms C4 project. Furthermore we thank our master student Xiaochen Zhao for assisting the literature research.

References

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1+2*, pages 385–393. Association for Computational Linguistics.

Wanxiang Che and Ting Liu. 2010. Jointly modeling wsd and srl with markov logic. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 161–169. Association for Computational Linguistics.

Justin Grimmer and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297.

Ben Hachey and Claire Grover. 2004. Sentence classification experiments for legal text summarisation. In *Proc. 17th Annual Conference on Legal Knowledge and Information Systems (Jurix-2004)*, pages 29–38.

Xia Hu, Nan Sun, Chao Zhang, and Tat-Seng Chua. 2009. Exploiting internal and external semantics for the clustering of short texts using world knowledge. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 919–928. ACM.

Mladen Karan, Daniela Širinić, Jan Šnajder, and Goran Glavaš. 2016. Analysis of policy agendas: Lessons learned from automatic topic classification of croatian political texts. In *Proceedings of the Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) at ACL 2016*, page in press.

Paul M Kellstedt. 2000. Media framing and the dynamics of racial policy preferences. *American Journal of Political Science*, pages 245–260.

Slava Mikhaylov, Michael Laver, and Kenneth R Benoit. 2012. Coder reliability and misclassification in the human coding of party manifestos. *Political Analysis*, 20(1):78–91.

Jan Noessner, Mathias Niepert, and Heiner Stuckenschmidt. 2013. Rockit: Exploiting parallelism and symmetry for MAP inference in statistical relational models. *CoRR*, abs/1304.4379.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–43.

Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi. 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on World Wide Web*, pages 91–100. ACM.

Hoifung Poon. 2010. Markov logic in natural language processing: Theory, algorithms, and applications. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, page 3. Citeseer.

Stephen Purpura and Dustin Hillard. 2006. Automated classification of congressional legislation. In *Proceedings of the 2006 international conference on Digital government research*, pages 219–225. Digital Government Society of North America.

Matthew Richardson and Pedro Domingos. 2006. Markov logic networks. *Mach. Learn.*, 62(1-2):107–136.

Frane Šarić, Goran Glavaš, Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. 2012. Takelab: Systems for measuring semantic text similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1+2*, pages 441–448. Association for Computational Linguistics.

Brandon M Stewart and Yuri M Zhukov. 2009. Use of force and civil–military relations in russia: an automated content analysis. *Small Wars & Insurgencies*, 20(2):319–343.

Naushad UzZaman, Hector Llorens, James Allen, Leon Derczynski, Marc Verhagen, and James Pustejovsky. 2012. Tempeval-3: Evaluating events, time expressions, and temporal relations. *arXiv preprint arXiv:1206.5333*.

Suzan Verberne, Eva Dhondt, Antal van den Bosch, and Maarten Marx. 2014. Automatic thematic classification of election manifestos. *Information Processing & Management*, 50(4):554–567.

Andrea Volkens, Onawa Lacewell, Pola Lehmann, Sven Regel, Henrike Schultze, and Annika Werner. 2011. *The Manifesto Data Collection*. Manifesto Project (MRG/CMP/MARPOR), Wissenschaftszentrum Berlin für Sozialforschung (WZB).

Lori Young and Stuart Soroka. 2012. Affective news: The automated coding of sentiment in political texts. *Political Communication*, 29(2):205–231.

Cäcilia Zirn, Mathias Niepert, Heiner Stuckenschmidt, and Michael Strube. 2011. Fine-grained sentiment analysis with structural features. In *IJCNLP*, pages 336–344.

Author Index