Topic-Based Agreement and Disagreement in US Electoral Manifestos

Stefano Menini[1,3], Federico Nanni[2], Simone Paolo Ponzetto[2], Sara Tonelli[1]
[1]Fondazione Bruno Kessler, Trento, Italy
[2]University of Mannheim, Germany
[3]University of Trento, Italy
{menini, satonelli}@fbk.eu
{federico, simone}@informatik.uni-mannheim.de

Abstract

We present a topic-based analysis of agreement and disagreement in political manifestos, which relies on a new method for topic detection based on key concept clustering. Our approach outperforms both standard techniques like LDA and a state-of-the-art graph-based method, and provides promising initial results for this new task in computational social science.

1 Introduction

During the last decade, the adoption of natural language processing (NLP) techniques for the study of political phenomena has gained considerable momentum (Grimmer and Stewart, 2013), arguably because of both the availability of parliamentary proceedings (van Aggelen et al., 2017), electoral manifestos (Volkens et al., 2011) and campaign debates (Woolley and Peters, 2008), and the interest of the computational social science (CSS) community in the potential of text mining methods for advancing political science research (Lazer et al., 2009).

Previous work focused on the automatic detection of sentiment expressions in political news (Young and Soroka, 2012), the identification of ideological proportions (Sim et al., 2013) and the scaling on a left-right spectrum of politicians' speeches (Slapin and Proksch, 2008). More recently, researchers looked at topic-centered approaches to provide finer-grained analyses, including segmentation methods for topic-labeled manifestos (Glavaš et al., 2016), supporting manual coders in identifying coarse-grained political topics (Zirn et al., 2016), as well as topic-based and cross-lingual political scaling (Nanni et al., 2016; Glavaš et al., 2017).

Measuring Agreement. Automatically measuring the level of agreement in political documents (Gottipati et al., 2013; Menini and Tonelli, 2016) has the potential of supporting political analyses such as the comparisons between campaign strategies (Burton et al., 2015), the study of promises kept and broken after elections (Naurin, 2011), the formation of coalitions (Debus, 2009) and the interactions between government and opposition (Hix and Noury, 2016). However, previous work relies on the availability of pre-defined topics, including supervised methods (Galley et al., 2004; Hillard et al., 2003), approaches leveraging collaboratively generated resources (Gottipati et al., 2013; Awadallah et al., 2012) or pairwise agreement detection from political debates (Menini and Tonelli, 2016).

Our Contributions. *a) New task:* Given a collection of political documents such as, e.g., electoral manifestos, we look at ways to perform an automatic, topic-based agreement-disagreement classification. *b) New approach:* We first segment the texts into coarse-grained domains. Next, coarse domains are used to extract a fine-grained list of topic-based points of view which, in turn, are used to perform classification. We achieve this by developing a novel approach for topic detection on the basis of key concept clustering techniques: this is shown to outperform not only LDA-based Topic Modeling – the *de facto* standard approach for this task in CSS (Grimmer and Stewart, 2013) – but also established unsupervised (k-means) and state-of-the-art graph-based clustering techniques. *c) Experimental study and resources:* We use manifestos from the Comparative Manifesto Project (Volkens et al., 2011). As in previous works (Zirn et al., 2016), we focus on a subset consisting of six U.S. manifestos (Republican and Democrat) from the 2004, 2008 and 2012 elections. We show

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2938–2944
Copenhagen, Denmark, September 7–11, 2017.

that our method leads to promising results when measuring the topic-based agreement between the party manifestos, thus indicating the overall feasibility of the task. Additionally, we release all code and annotations related to this paper to foster further work from the research community.

2 System Overview

We present a new system for measuring the topic-based agreement of political manifestos. Our approach consists of four main steps: i) macro-domain detection, e.g. *foreign policy, economy, welfare*, ii) key concept extraction, iii) topic detection as key concept clustering, e.g., *energy consumption, new energy solution, petroleum dependence* for the topic *green economy*, and iv) pairwise, topic-based agreement detection.

The central component of our pipeline is a new approach for fine-grained topic detection in political contents based on key concept clustering. This is because, among existing methods, supervised approaches cannot be applied here due to the scarce availability of in-domain labeled data, as well as the already remarked high complexity of the annotation process (Benoit et al., 2016). Moreover, the application of unsupervised topic detection techniques like LDA has been shown during prototyping to produce low-quality topics that are rather coarse (cf. the results in Section 3).

Similar to LDA-based approaches, we view each topic as a cluster of words or phrases. However, given that we are in a domain with topics built around rather specific lexical cues, we do not rely on the entire vocabulary of the documents. Instead, we build clusters that are made up of semantically similar key concepts extracted from the documents themselves, including both single and multiwords (i.e. keywords and keyphrases). In the next paragraph we present an overview of each component of our system.

1) Domain Detection. We are given as input sentences from a political manifesto. The first step of our work is to classify them into the seven macro-domains defined by the Manifesto Project, namely *external relations*, *freedom and democracy*, *political system*, *economy*, *welfare and quality of life*, *fabric of society*, *social groups*. To achieve this goal, we use ClassyMan, a system developed in a previous work (Zirn et al., 2016), which predicts the domains and domain shifts between pairs of adjacent sentences.

2) Key concept Extraction. Next, for each domain we process each sentence using Keyphrase Digger (KD) (Moretti et al., 2015). KD is a rule-based (hence domain-agnostic) system for key concept extraction that combines statistical measures with linguistic information, and which has shown competitive performance on the SemEval 2010 benchmark (Kim et al., 2010). We set the tool to extract lemmatized key concepts up to three tokens. For each key concept, we compute its tf-idf, considering each domain as a different document. The result is a list of key concepts for each domain, with a score representing their relevance to the domain.

3) Key concept Clustering. Starting from the flat lists of key concepts extracted by KD, we adopt a recursive procedure to merge them into meaningful clusters. First, we build a distributional semantic vector for each key concept by averaging the embeddings of each word in the key concept (we use the GloVe embeddings from Pennington et al. (2014) with 50 dimensions, pre-trained on Wikipedia). Next, we build a semantic graph representation where *a)* each node consists of a key concept, *b)* the weight of each edge is the cosine distance between their respective embedding vectors and *c)* edges are directed, pointing to the node of the key concept with the higher tf-idf. For ties, we create multiple edges. To direct the nodes we adopt tf-idf, since we want to weigh the key concepts according to the relevance for the macro-domain we are processing. This allows us to obtain well-defined groups within the domains.

To reduce the number of weak edges, we set a cosine similarity threshold of 0.8,[1] and we set an edge between two multi-word keyphrases if they have at least one word in common (e.g. *ethnic minority*, *black minority*).

Finally, we obtain clusters of semantically related key concepts from the graph as follows: *a)* we extract all groups of key concepts with an edge directed to the same node and create a first set of clusters. Then, *b)* clusters sharing at least 50% of the key concepts are merged. Next, *c)* the clusters purity is improved by removing the less relevant key concepts. These are identified as those key concepts whose cosine distance is more than 1.5 times the standard deviation from the centroid of

[1] We evaluated the clustering output with different thresholds ranging from 0.6 to 0.9. The value of 0.8 is the one leading to the best accuracy. In Table 1 and 2 we report the results using this threshold.

the cluster. At the end of this process, we obtain for each domain a set of clusters or *topics*, made of semantically related key concepts. The number of clusters is determined dynamically during the process and does not need to be defined a priori.

4) Statement Extraction. We use the clusters of key concepts to identify pairs of statements, related to the same topic, from the Republican (R) and Democratic (D) manifestos. For each topic, we first collect the statements from D and R manifestos having among their key concepts one of the key concepts defining the topic and then we pair groups of three statements from D with groups of three statements from R. We use groups of three statements because it allows us *i)* to obtain a sufficient number of pairs to perform automatic classification, and *ii)* to improve the quality of the manual annotations. We noticed during an initial evaluation that annotators focus easier on groups of 3 sentences rather than larger groups and that, on the other hand, using less than 3 sentences decreases the chances to obtain at least two statements in agreement/disagreement within a pair.

5) Agreement Classification. The last step is the automatic classification of agreement and disagreement between Republicans and Democrats. To classify pairs of statements, we rely on a supervised machine learning approach with the set of features used in Menini and Tonelli (2016), in which a similar task is addressed. The classification relies on features related to surface information such as lexical overlap and negation, to the semantic content of the statements (e.g. sentiment) and to their relation (e.g. entailment).

3 Evaluation

3.1 Topic Extraction

Having a set of manifestos annotated with coarse-grained domains – using *ClassyMan*, which achieved a micro F1-Score of 0.78 across the seven macro-topics in a 10-fold cross validation setting – the central step of our pipeline is to detect clusters of key concepts representing fine-grained topics in each macro domain. To do that, we adopt the method described above, that we call here *Key Concept Clusters*. We examine its performance in comparison with two types of baselines.

LDA Baselines. We first employ vanilla LDA, a common approach for topic detection in CSS (Grimmer and Stewart, 2013), relying on the assumption that tokens often co-occurring together in a corpus belong to the same topic. For this task, we use the Mallet topic model package.[2] Given the fact that our method for key concept clustering identifies on average 30 topics per domain, we create a corpus for each domain with all its sentences and we run LDA with 10,000 iterations to obtain 30 topics. We test LDA by considering all the tokens in the corpus (*Vanilla LDA*) and only the extracted key concepts (*Key concept LDA*).

Clustering Baselines. The second type of baseline adopts the same representation of key concepts used in our approach, i.e., we represent candidate phrases by averaging the embeddings of their constituent words. We test two different clustering approaches to group them into topics: the first uses *K-means* (with 30 clusters). The second (*Graph-based*) builds a fully-connected semantic relatedness graph by measuring the cosine similarity between all pairs of key concepts: topic clustering is then obtained by finding all maximal cliques in the graph using the Bron-Kerbosch algorithm.

Evaluation. In order to assess the overall quality of the topics produced by each approach, we adopt the word-intrusion post-hoc evaluation method (Chang et al., 2009) using the platform presented in Lauscher et al. (2016). For each approach, we randomly pick 100 topics and for each topic we keep two sets of key concepts, respectively the four and eight top-relevant elements of the cluster.[3] Then, we add to these four/eight words a new word from another topic (i.e. the intruder), and we shuffle the obtained five/nine words. Finally, we ask three political science experts to identify the intruder. The more the topics are coherent, the easier the intruder is detected. While this type of post-hoc evaluation is extremely time-consuming – no less than 45 minutes of work for annotator for each produced ranking, thus hindering the experimental assessment of, for instance, the role of different numbers of topics for each baseline – it is necessary given the already remarked limits of existing gold standards manually-created for the task (Mikhaylov et al., 2012; King et al., 2017).

Results. As shown in Table 1, our system outperforms the other methods with an accuracy of 0.86 in the word-intrusion task with four key con-

[2] http://mallet.cs.umass.edu/

[3] For LDA, Mallet provides already ranked results. For the other approaches, the most relevant key concepts are the key concepts closest to the centroid of the cluster.

Method	Acc.@4	Acc.@8
Vanilla-LDA	0.22	0.35
Key concept-LDA	0.29	0.36
Graph-based Clusters	0.46	0.44
k-means Clusters	0.72	0.67
Key concept Clusters	0.86	0.67

Table 1: Topics evaluation: accuracy in word intrusion task. The table reports the accuracy values on the first 4 and 8 key concepts in the clusters.

Method	Kappa@4	Kappa@8
Vanilla LDA	0.32	0.46
Key concept LDA	0.50	0.40
Graph-based Clusters	0.39	0.32
k-means Clusters	0.65	0.61
Key concept Clusters	0.79	0.62

Table 2: Inter-annotator agreement (IAA) evaluation (Fleiss' kappa) in the word intrusion task. The table reports the IIA on the first 4 and 8 key concepts in the clusters.

cepts in each cluster, while it decreases to 0.67 if we extend the evaluation to include eight key concepts. Besides, inter-annotator agreement (Fleiss' kappa), reported in Table 2, varies a lot across the different methods. In particular, the agreement in the intrusion task with four key concepts is higher for clusters generated with our method (0.79), while it is very low using LDA (0.32). This confirms the findings by Chang et al. (2009) that LDA topics are often difficult to interpret.

If we extend the evaluation to the first eight elements of each cluster, we notice that the difference between the agreement with our pipeline (0.62) and LDA (0.46) decreases. This shows that, with *key concept clusters*, increasing the number of key concepts in a topic affects their interpretation, although there is still an improvement with respect to the other approaches.

Final Tuning. We next tune clustering to classify fine-grained topics as in agreement or disagreement. Tuning is performed as to maximize clustering accuracy while obtaining a sufficient number of topics shared by both Democrats and Republicans. Since a cosine similarity threshold of 0.8 in the clustering process leads to clusters that are too specific, often addressed only by one of the two parties, we reduce the threshold to 0.7, so that the topics are likely to be covered by both manifestos. In addition, we want to compare the agreement focusing on small clusters, composed by a maximum of 10 key concepts. To obtain them, we iterate the clustering process over the key concepts of larger clusters, progressively increasing the cosine similarity threshold until there are no groups larger than 10 key concepts. We reach this goal with a threshold of 0.85. Using these settings, the accuracy (Acc. @4) of the clusters decreases to 0.74, but we obtain clusters that allow us to extract a total of 351 pairs covering 87 fine-grained topics. Table 3 shows some of the clusters extracted.

3.2 Agreement Classification

Data Annotation. The statements in the pairs have been annotated by three scholars of political science in terms of agreement, disagreement or none of the two. The annotation results in 158 pairs in disagreement, 135 in agreement and 58 neither in agreement nor in disagreement, with an inter-annotator agreement (IAA) of 0.64 (Fleiss' Kappa). Note that only in three cases the annotators claimed that the meaning of a sentence pair did not match with the topic detected with our approach. This additional finding highlights again the quality of our method for topic detection based on key concept clustering.

Agreement Classification. Agreement classification is carried out using Support Vector Machine (SVM) tested in two configurations. In the first setting, we train and test the classifier with 10-fold cross validation over the manually annotated pairs from the political manifestos. In the second configuration, we explore instead a cross-domain approach: we train the SVM on the 1960 Elections dataset from Menini and Tonelli (2016) and use all the pairs in our gold standard of political manifestos as test set. This experiment is aimed at assessing the impact of training on comparable data are from the same domain (i.e., transcript of political speeches vs. manifestos).

The results of both configurations are shown in Table 4, where they are compared to a random baseline. The results show that the set of features used suits our task, classifying the data with an accuracy comparable to the performance of human annotators, if we consider IAA as an upper bound for the task. We achieve nevertheless results that are in a lower range than Menini and Tonelli (2016), thus suggesting that agreement and disagreement is harder to detect in political mani-

Macro-domain: External Relations
japan, korea, missile, north_korea, south_korea, weapon_north_korea
extremism, renounce_terrorism, nuclear_terrorism, proliferation, security, terrorism
Macro-domain: Freedom and Democracy
culture, freedom, ideology, religion, society, tolerance, tradition
democracy, discrimination, first_amendment_rights, freedom, issue, law, rights_of_citizenship
Macro-domain: Political System
budget, budget_act, cost, cut, deficit, shortfall, tax
congressional_republican, election, republican, republican_platform, romney, vote
Macro-domain: Economy
alternative_fuel, electricity, fuel, gas, transportation_fuel
bailout, credit, loan, mortgage, payment, savings
Macro-domain: Welfare and Quality of Life
ailment, chronic, disease, health, illness, obesity, treatment_of_disease
global_energy_forum, industry, new_energy_solution, new_global_energy, solar_energy_generation
Macro-domain: Fabric of Society
crime, criminal, high-profile_criminal_conviction, prosecution
religious_freedom, religious, religious_discrimination
Macro-domain: Social Groups
agricultural agricultural_america, agricultural_production, agriculture, farm, rural, rural_america
hispanic, latino, latino_population

Table 3: Examples of key concept clusters extracted for each macro-domain.

Classification	Accuracy
Random Baseline	0.54
10-fold cross validation	0.66
1960 Elections training	0.61

Table 4: Results on agreement classification.

festos than in speeches. Finally the accuracy of the classifier in the cross-domain setting is lower than the one obtained with in-domain cross-validation, but still comparable with that of human annotators.

4 Conclusion

In this paper, we presented a system for supporting automatic topic-bases analyses of agreement and disagreement in political manifestos. This approach goes beyond established approaches for the task, which are either too coarse-grained or rely intensively on manual annotations.

Our method can provide insights into agreement and disagreement between parties, covering several topics of internal and foreign policy. By examining the results, we find an overall cross-party agreement of 46% regarding the discussed issues. However, this agreement varies substantially if we consider the different macro-domains. For example, while we notice a strong disagreement over the domain *political system*, especially for what concerns the responsibilities of previous administrations, other domains, such as *external relations*, present a more balanced ratio of agreement and disagreement between Republicans and Democrats. The possibility of measuring agreement at a finer level (topics) that is offered by our approach, shows, for example, that between 2004 and 2012 two opposite positions have been defined regarding the Middle East. On the contrary, there has been a general agreement on the role of the U.S. concerning the relations with Europe.

In the future, we hope that the pipeline presented in this paper will support political science researchers in studying topics such as party polarization through the analysis and comparison of electoral manifestos, parliamentary proceedings and campaign speeches. On the computational side, we will to extend our approach to cross-lingual data, in order to enable computer-assisted political analysis across different languages.

Downloads. The code for topic detection as key concept clustering process is available at https://dh.fbk.eu/technologies/keyphrase-clustering.

References

Astrid van Aggelen, Laura Hollink, Max Kemman, Martijn Kleppe, and Henri Beunders. 2017. The debates of the european parliament as linked open data. *Semantic Web*, 8(2):271–281.

Rawia Awadallah, Maya Ramanath, and Gerhard Weikum. 2012. Polaricq: Polarity classification of political quotations. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 1945–1949.

Kenneth Benoit, Drew Conway, Benjamin E Lauderdale, Michael Laver, and Slava Mikhaylov. 2016. Crowd-sourced text analysis: reproducible and agile production of political data. *American Political Science Review*, 110(02):278–295.

Michael John Burton, William J. Miller, and Daniel M. Shea. 2015. *Campaign Craft: The Strategies, Tactics, and Art of Political Campaign Management: The Strategies, Tactics, and Art of Political Campaign Management*. New York: Praeger.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296.

Marc Debus. 2009. Pre-electoral commitments and government formation. *Public Choice*, 138(1-2):45.

Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg. 2004. Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 669–676.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2016. Unsupervised text segmentation using semantic relatedness graphs. In *Proceedings of the 5th Joint Conference on Lexical and Computational Semantics*, pages 125–130.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2017. Unsupervised cross-lingual scaling of political texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 688–693.

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.

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.

Dustin Hillard, Mari Ostendorf, and Elizabeth Shriberg. 2003. Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (Short Papers)*, pages 34–36.

Simon Hix and Abdul Noury. 2016. Government-opposition or left-right? the institutional determinants of voting in legislatures. *Political Science Research and Methods*, 4(02):249–273.

Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2010. Semeval-2010 task 5 : Automatic keyphrase extraction from scientific articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 21–26.

Gary King, Patrick Lam, and Margaret Roberts. 2017. Computer-assisted keyword and document set discovery from unstructured text. *American Journal of Political Science*.

Anne Lauscher, Federico Nanni, Pablo Ruiz Fabo, and Simone Paolo Ponzetto. 2016. Entities as topic labels: combining entity linking and labeled lda to improve topic interpretability and evaluability. *IJCol: Italian journal of computational linguistics*, 2(2):67–88.

David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. 2009. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721.

Stefano Menini and Sara Tonelli. 2016. Agreement and disagreement: Comparison of points of view in the political domain. In *Proceedings of the 26th International Conference on Computational Linguistics*, pages 2461–2470.

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.

Giovanni Moretti, Rachele Sprugnoli, and Sara Tonelli. 2015. Digging in the dirt: Extracting keyphrases from texts with KD. In *Proceedings of the 2nd Italian Conference on Computational Linguistics*.

Federico Nanni, Cäcilia Zirn, Goran Glavaš, Jason Eichorst, and Simone Paolo Ponzetto. 2016. TopFish: Topic-based analysis of political position in US electoral campaigns. In *Proceedings of the International Conference on the Advances in Computational Analysis of Political Text*, pages 61–67.

Elin Naurin. 2011. *Election Promises, Party Behaviour and Voter Perceptions*. Palgrave Macmillan.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–43.

Yanchuan Sim, Brice Acree, Justin H Gross, and Noah A. Smith. 2013. Measuring ideological proportions in political speeches. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 91–101.

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), Berlin: Wissenschaftszentrum Berlin für Sozialforschung (WZB)*.

John T Woolley and Gerhard Peters. 2008. The American presidency project.

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

Cäcilia Zirn, Goran Glavaš, Federico Nanni, Jason Eichorts, and Heiner Stuckenschmidt. 2016. Classifying topics and detecting topic shifts in political manifestos. In *Proceedings of the International Conference on the Advances in Computational Analysis of Political Text*, pages 88–93.