Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach

Derek Greene* James P. Cross†

October 20, 2015

Abstract

This study analyzes political interactions in the European Parliament (EP) by considering how the political agenda of the plenary sessions has evolved over time and the manner in which Members of the European Parliament (MEPs) have reacted to external and internal stimuli when making Parliamentary speeches. It does so by considering the context in which speeches are made, and the content of those speeches. To detect latent themes in legislative speeches over time, speech content is analyzed using a new dynamic topic modeling method, based on two layers of matrix factorization. This method is applied to a new corpus of all English language legislative speeches in the EP plenary from the period 1999-2014. Our findings suggest that the political agenda of the EP has evolved significantly over time, is impacted upon by the committee structure of the Parliament, and reacts to exogenous events such as EU Treaty referenda and the emergence of the Euro-crisis have a significant impact on what is being discussed in Parliament.

*Insight Centre for Data Analytics & School of Computer Science & Informatics, University College Dublin, Ireland. (derek.greene@ucd.ie)

†School of Politics & International Relations, University College Dublin, Ireland. (james.cross@ucd.ie).

1 Introduction

The plenary sessions of the European Parliament (EP) are one of the most important arenas in which European representatives can air questions, express criticisms and take policy positions to influence EU politics. Indeed the plenary of the Parliament represents the closest that the European Union (EU) gets to engaging in the core democratic process of publicly-aired democratic debate. As a result, understanding how Members of the European Parliament (MEPs) express themselves in plenary, and investigating how the political discussions evolve and respond to internal and external stimuli is a fundamentally important undertaking.

In recent years, there has been a concurrent explosion of online records detailing the content of MEP speeches, and the development of data-mining techniques capable of extracting latent patterns in content across sets of these speeches. This allows us for the first time to investigate the plenary agenda of the Parliament in a holistic and rigorous manner. One approach to tracking the political attention of political figures has been to apply topic-modeling algorithms to large corpora of political texts, such as parliamentary speeches of the U.S. Senate (Quinn et al., 2010). These algorithms seek to distill the latent thematic patterns in a corpus of speeches (Blei et al., 2003), and can be used to improve the transparency of the political process by providing a macro-level overview of the activities and agendas of politicians in a time- and resource-efficient manner.

This study takes up the challenge of extracting latent thematic patterns in political speeches by developing a dynamic topic model to investigate how the plenary agenda of the EP has changed over three parliamentary terms (1999–2014). The method applies two layers of Non-negative Matrix Factorization (NMF) topic

modeling (Lee and Seung, 1999) to a corpus of 210,247 speeches from 1,735 MEPs across the 28 EU member states.

Our proposed topic modeling methodology reveals the breadth of policy areas discussed by MEPs in the Parliament, and the results presented in Section 5 indicate that the political agenda of the Parliament has evolved significantly over time. By examining a number of case studies, ranging from the Euro-crisis to EU treaty changes, we can identify the relationship between the evolution of these dynamic topics and the exogenous events driving them. By using external data sources, we can also confirm the semantic and construct validity of these topics. In order to explain some of the patterns in speech making we observe, we conclude the study with an exploration of the determinants of MEP speech-making behavior on the detected topics.[1]

2 Related Work

This section introduces two converging literatures relating to MEP behavior in the EP and the use of topic models to study political text corpora respectively.

The most prominent forms of MEP behavior that have received academic attention are plenary speeches and roll-call voting. A great deal of attention has been paid to how political institutions shape these forms of MEP behavior. For instance, the formal committee structure of the EP has been shown to provide committee members with strategic advantages due to privileged access to information, and opportunity to shape the EP's policy choices. This has led MEPs to self-select into committees dealing with salient issues with a view to influencing

[1]To provide access to the results of the project to interested parties, we make a browsable version available online: `http://erdos.ucd.ie/europarl`

outcomes within these committees (Bowler and Farrell, 1995). Within committees, holding roles such as the Chair and Rapporteur have been shown to impact upon speech making and voting behavior in the broader plenary (Hix, Simon et al., 2007).

Strict institutional rules also govern the allocation of MEP speaking time in the EP plenary (Proksch and Slapin, 2010). The total amount of speaking time for any particular issue is limited and divided between time reserved for actors with formal plenary duties such as rapporteurs, and time proportionally divided between party groups based upon their share of MEPs elected. Speaking time limits lead to competition between MEPs, and party-group leaders allocate scarce speaking time between MEPs for maximum impact (Slapin and Proksch, 2010).

Due to the limits in speaking time available, MEPs can also submit written questions/statements that are appended to the plenary records. MEPs use these written statements to interact with the Commission directly (Raunio, 1996) and provide the opportunity for 'fire-alarm oversight' of national governments guilty of failures to implement EU law (Jensen et al., 2013). MEPs have more discretion when submitting written submissions than they do with plenary speaking time.

It has been shown that the content of MEP speeches reflect latent ideological conflict between MEPs (Slapin and Proksch, 2010). Using text-analysis techniques based upon word-frequency distributions, these authors demonstrate the correspondence between the content of legislative speeches and other measures of ideological positions found in the literature based upon roll-call votes and expert surveys. To our knowledge, topic models have yet to be applied to the European Parliament plenary.

Topic models aim to discover the latent semantic structure or topics within

a text corpus, which can be derived from co-occurrences of words across documents. These models date back to the early work on latent semantic indexing by Deerwester et al. (1990), which proposed the decomposition of term-document matrices for this purpose using Singular Value Decomposition. Considerable research on topic modeling has focused on the use of probabilistic methods, where a topic is viewed as a probability distribution over words, with documents being mixtures of topics, thus permitting a topic model to be considered a generative model for documents (Steyvers and Griffiths, 2007). The most widely-applied probabilistic topic modeling approach is Latent Dirichlet Allocation (LDA) proposed by Blei et al. (2003). Following on from static LDA methods, authors have subsequently developed analogous probabilistic approaches for tracking the evolution of topics over time in a sequentially-organized corpus of documents, such as the dynamic topic model (DTM) of Blei and Lafferty (2006).

Alternative algorithms, such as Non-negative Matrix Factorization (NMF) (Lee and Seung, 1999), have also been effective in discovering the underlying topics in text corpora (Wang et al., 2012). NMF is an unsupervised approach for reducing the dimensionality of non-negative matrices, which seeks to decompose the data into factors that are constrained so as to not contain negative values. By modeling each object as the additive combination of a set of non-negative basis vectors, a readily interpretable clustering of the data can be produced without requiring further post-processing. When working with text data, these clusters can be interpreted as topics, where each document is viewed as the additive combination of several overlapping topics.

Topic-modeling methods have been adopted in the political science literature to analyze political attention. In settings where politicians have limited time-

resources to express their views (*e.g.* plenary sessions in parliaments), they must decide what topics to address. Analyzing what they choose to speak about can thus provide insight into the political priorities of the politicians under consideration. Single membership topic models, which assume each speech relates to one topic, have successfully been applied to plenary speeches made in the 105th to the 108th U.S. Senate in order to trace political attention of the Senators within this context over time (Quinn et al., 2010). This study found that a rich political agenda emerged, where topics evolved over time in response to both internal and external stimuli.

Bayesian hierarchical topic models have also been used to capture the political priorities of Members of Congress as found in their official press releases (Grimmer, 2010), and structural topic models have been used to incorporate text "metadata" in the form of document-level covariates. Such covariates can include information about a document itself such as when and where it was created, alongside information about the creator of the document (Roberts et al., 2014).

In conclusion, the current literature provides some interesting insights into the factors that impact upon MEP speech-making and voting behavior, and the introduction of topic models to the study of political attention has allowed researchers to consider larger and more complete datasets of political activity across longer time periods than has previously been possible.

3 Methods

In this section we describe a two-layer strategy for applying topic modeling in a non-negative matrix factorization framework to a timestamped corpus of political

speeches. We first describe the application of NMF topic modeling to a single set of speeches from a fixed time period, and then propose a new approach for combining the outputs of topic modeling from successive time periods to detect a set of *dynamic topics* that span part or all of the duration of the corpus.

3.1 Topic Modeling Speeches

While work on topic models often involves the use of LDA, NMF can also be applied to textual data to reveal topical structures (Wang et al., 2012). The ability of NMF to account for how important a word is to a document in a collection of texts based on weighted term-frequency values is particularly useful. Specifically, applying a log-based term frequency-inverse document frequency (TF-IDF) weighting factor to the data prior to topic modeling has shown to be advantageous in producing diverse but semantically coherent topics which are less likely to be represented by the same high frequency terms. This makes NMF suitable when the task is to identify both broad, high-level groups of documents and niche topics with specialized vocabularies (O'Callaghan et al., 2015). In the context of political speech in parliaments, this is a particularly desirable attribute of the model, as it can differentiate between broad procedural topics relating to the day-to-day running of plenary and more focused discussions on specific policy issues.

3.1.1 Applying NMF

Given a corpus of n speeches, we first construct a speech-term frequency matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$, where m is the number unique terms present across all speeches (*i.e.* the corpus vocabulary). Applying NMF to $\mathbf{A}$ results in a reduced rank-k

approximation in the form of the product of two non-negative factors $\mathbf{A} \approx \mathbf{WH}$, where the objective is to minimize the reconstruction error between $\mathbf{A}$ and $\mathbf{WH}$. The rows of the factor $\mathbf{H} \in \mathbb{R}^{k \times m}$ can be interpreted as k topics, defined by non-negative weights for each of the m terms in the corpus vocabulary. Ordering each row provides a topic descriptor, in the form of a ranking of the terms relative to corresponding topic. Essentially the ordered row entries of the matrix $\mathbf{H}$ allow us to identify the most common terms characterizing each topic, thus allowing for substantive interpretation. The columns in the matrix $\mathbf{W} \in \mathbb{R}^{n \times k}$ provide membership weights for all n speeches with respect to each of the k topics. The columns in matrix $\mathbf{W}$ can be used to associate individual speeches with the topic they are related to, and when we know from meta-data what MEP makes a given speech, we can thus capture MEP contributions to a given topic.

NMF algorithms are often initialized with random factors, which can lead to unstable results where the algorithm converges to a variety of local minima of poor quality. To ensure a deterministic output, we generate initial factors using the Non-negative Double Singular Value Decomposition (NNDSVD) initialization approach (Boutsidis and Gallopoulos, 2008).

3.1.2 Parameter Selection

A key parameter selection decision in topic modeling pertains to the number of topics k. Choosing too few topics will produce results that are overly broad, while choosing too many will lead to many small, highly-similar topics. One general strategy proposed in the literature has been to compare the *topic coherence* of topic models generated for different values of k (Chang et al., 2009). A range

of such coherence measures exists in the literature, although many of these are specific to LDA. Recently, O'Callaghan et al. (2015) proposed a general measure, Topic Coherence via Word2Vec (TC-W2V), which evaluates the relatedness of a set of top terms describing a topic. This approach uses the increasingly popular *word2vec* tool (Mikolov et al., 2013) to compute a set of vector representations for all of the terms in a large corpus. By measuring the similarity between pairs of term vectors, we can assess the extent to which the two corresponding terms share a common meaning or context (*e.g.* are related to the same topic). Topics with descriptors consisting of highly-similar terms, as defined by the similarity between their vectors, should be more semantically coherent.

For the purpose of assessing the coherence of topic models, TC-W2V operates as follows. The coherence of a single topic t_h represented by its t top ranked terms is given by the mean pairwise cosine similarity between the t corresponding term vectors in the *word2vec* space:

$$\text{coh}(t_h) = \frac{1}{\binom{t}{2}} \sum_{j=2}^{t} \sum_{i=1}^{j-1} cos(wv_i, wv_j) \tag{1}$$

An overall score for the coherence of a topic model T consisting of k topics is given by the mean of the individual topic coherence scores:

$$\text{coh}(T) = \frac{1}{k} \sum_{h=1}^{k} \text{coh}(t_h) \tag{2}$$

An appropriate value for k can be identified by examining a plot of the mean TC-W2V coherence scores for a fixed range $[k_{min}, k_{max}]$ and selecting a value corresponding to the maximum coherence.

Parliamentary speeches will often be short and concise. In the case of the EP, speeches are often limited to 1-2 minutes in duration. As such, we would expect each speech to be primarily related to a single topic. This is consistent with the observations made by Quinn *et al.* (2010). Here we produce a single membership topic model (*i.e.* a disjoint clustering of individual speeches in relation to topics) by selecting the maximum membership weight for each row in the factor $\mathbf{W}$.

3.2 Dynamic Topic Modeling

3.2.1 Layer 1

When applying clustering to temporal data, authors have often proposed dividing the data into *time windows* of fixed duration (Sulo et al., 2010). Therefore, following Sulo et al. (2010), we divide the full time-stamped corpus of parliamentary speeches into τ disjoint time windows $\{W_1, \ldots, W_\tau\}$ of equal length. The rationale for the use of time windows as opposed to processing the full corpus in batch is two-fold: 1) we are interested in identifying the agenda of the parliament at individual time points as well as over all time; 2) short-lived topics, appearing only in a small number of time windows, may be obscured by only analyzing the corpus in its entirety. At each time window W_i, we apply NMF with parameter selection based on Eqn. 2 to the transcriptions of all speeches delivered during that window, yielding a *window topic model* T_i containing k_i *window topics*. This process produces a set of successive window topic models $\{T_1, \ldots, T_\tau\}$, which represents the output of the first layer in our proposed methodology.

3.2.2 Layer 2

From the window topic models we construct a new condensed representation of the original corpus, by viewing the rows of each factor $\mathbf{H}_i$ coming from each window topic model as "topic documents". Each topic document contains non-negative weights indicating the descriptive terms for that window topic. We expect that window topics from different windows that share a common theme will have similar topic documents. We then construct a condensed topic-term matrix $\mathbf{B}$ as follows:

1. Start with an empty matrix $\mathbf{B}$.
2. For each window topic model T_i:
   (a) For each window topic within T_i, select the t top ranked terms from the corresponding row vector of the associated NMF factor $\mathbf{H}$, set all weights for all other terms in that vector to 0. Add the vector as a new row in $\mathbf{B}$.
3. Once vectors from all topic models have been stacked in this way, remove any columns with only zero values (*i.e.* terms from the original corpus which did not ever appear in the t top ranked terms for any window topics).

The matrix $\mathbf{B}$ has size $n' \times m'$, where $n' = \sum_{i=1}^{\tau} k_i$ is the total number of "topic documents" and $m' << m$ is the subset of terms remaining after Step 3. The use of only the top t terms in each topic document allows us to implicitly incorporate feature selection into the process. The result is that we include those terms that were highly descriptive in each time window, while excluding those terms that never featured prominently in any window topic. This reduces the computational cost for the second factorization procedure described below.

Having constructed $\mathbf{B}$, we now apply a second layer of NMF topic modeling

to the matrix to identify k' *dynamic topics* that potentially span multiple time windows. The process is the same as that outlined previously in Section 3.1. Here the TC-W2V coherence measure is used to detect number of dynamic topics k'. The resulting factors can be interpreted as follows: the top ranked terms in each row of $\mathbf{H}$ provide a description of the dynamic topics; the values in the columns of $\mathbf{W}$ indicate to what extent each window topic is related to each dynamic topic.

We track the evolution of these topics over time as follows. Firstly, we assign each window topic to the dynamic topic for which it has the maximum weight, based on the values in each row in the factor $\mathbf{W}$. We define the temporal *frequency* of a dynamic topic as the number of distinct time windows in which that dynamic topic appears. The set of all speeches related to this dynamic topic across the entire corpus corresponds to the union of the speeches assigned to the individual time window topics, which are in turn assigned to the dynamic topic.

The resulting outputs of the two-layer topic modeling process are 1) A set of τ window topic models, each containing k_i *window topics*. These are described using their top t terms and the set of all associated speeches; 2) A set of k' *dynamic topics*, each with an associated set of window topics. These are described using their top t terms and set of all associated speeches; and 3) A ranking of every MEPs *contributions* relative to all window and dynamic topics in the corpus.

Table 1 shows a partial example of a dynamic topic. We observe that, for the four window topics, there is a common theme pertaining to climate change. The evolution of the climate change topic can be seen in the emergence of the terms 'Copenhagen', 'conference' and 'summit' in 2009-Q4 and 2010-Q1, at exactly the time when the Copenhagen climate change summit was underway. Detecting the evolution of topics in this manner is one of the advantages of taking a dy-

Rank	2008-Q4	2009-Q1	2009-Q4	2010-Q1
1	energy	climate	climate	climate
2	climate	change	change	copenhagen
3	emission	future	copenhagen	change
4	package	emission	developing	summit
5	change	integrated	emission	emission
6	renewable	water	conference	international
7	target	policy	summit	mexico
8	industry	target	agreement	conference
9	carbon	industrial	global	global
10	gas	global	energy	world

Table 1: Example of 4 window topics, described by lists of top 10 terms, which have been grouped together in a single dynamic topic related to climate change.

namic approach. While the variation across the term lists reflects the evolution of this dynamic topic over the time period (2008-Q4 to 2010-Q1), the considerable number of terms shared between the lists underlines its semantic validity.

4 Data

During August 2014 we retrieved all plenary speeches available on Europarl, the official website of the European Parliament, corresponding to parliamentary activities of MEPs during the 5th – 7th terms of the EP.[2] This resulted in 269,696 unique speeches in 24 languages. While we considered the use of either multilingual topic modeling or automated translation of documents, issues with the accuracy and reliability of both strategies lead us to focus on English language speeches in plenary – either from native speakers or translated – which make up the majority of the speeches available on Europarl. A corpus of 210,247 English language speeches was identified in total, representing 77.95% of the original col-

[2] `http://europarl.europa.eu`

lection. In terms of coverage of speeches from MEPs from the member states, this ranges from 100% for the United Kingdom, through 87% for Germany, down to 66.2% for Romania. However, the most recent state to accede to the EU, Croatia, represents an outlier in the sense that only 2.6% of speeches were available in English at the time of retrieval due to EP speech translation issues.

We subsequently divide the corpus into 60 quarterly time windows, from 1999-Q3 to 2014-Q2. We select a quarter as the time window duration to allow for the identification of granular topics, while also ensuring there exists a sufficient number of speeches in each time window to perform topic modeling. Initial experiments performed on shorter durations with small numbers of speeches per window often yielded results with a smaller number of coherent topics. In addition, a quarterly time window is appropriate in order to avoid empty time windows occurring due to the summer recess of the EP.

For each time window W_i we construct a speech-term matrix $\mathbf{A}_i$ as follows:

1. Select all speech transcriptions from window W_i, and remove all header and footer lines.
2. Find all unigram tokens in each speech, through standard case conversion, tokenization, and lemmitization.
3. Remove short tokens with < 3 characters, and tokens corresponding to generic stop words (*e.g.* "are", "the"), parliamentary-specific stop words (*e.g.* "adjourn", "comment"), and names of politicians.
4. Remove tokens occurring in < 5 speeches.
5. Construct $\mathbf{A}_t$, based on the remaining tokens. Apply standard TF-IDF term weighting and document length normalization.

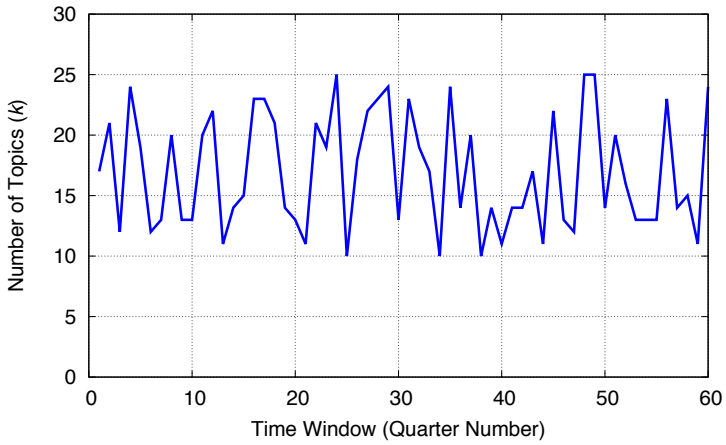
The resulting time window data sets range in size from 679 speeches in 2004-Q3 to 9,151 speeches in 2011-Q4, with an average of 4,811 terms per data set.

5 Experimental Results

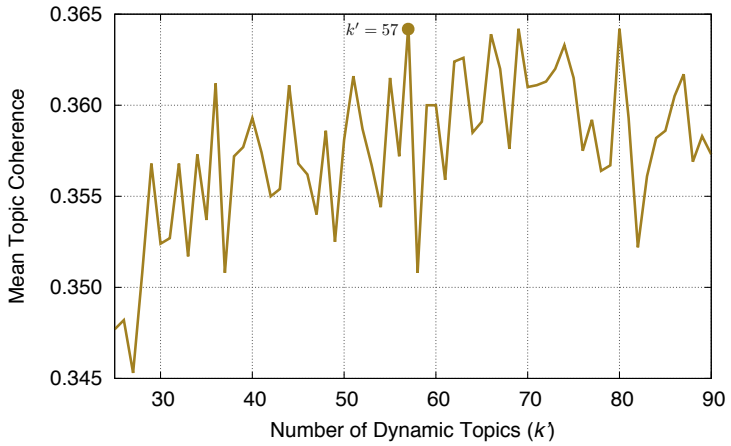
5.1 Experimental Setup

After pre-processing the data, the first task was to identify k, the number of topics in each window. To do this, we applied NMF with parameter selection as described in Section 3.1. Given the relatively specialized vocabulary used in EP debates, when building the *word2vec* space for parameter selection, as our background corpus we used the complete set of English language speeches. We used the same *word2vec* settings and number of top terms per topic ($t = 10$) as described in O'Callaghan et al. (2015). At each time window, we generated window topic models containing $k \in [10, 25]$ topics, and then selected the value k that produced the highest mean TC-W2V coherence score (Eqn. 2). The illustration of the number of topics per window in Fig. 1a shows that there is considerable variation in the number of topics detected for each window, which does not correlate with the number of speeches per quarter (Pearson correlation 0.006). This suggests our results are not driven by the volume of speeches, but rather variation in topics being discussed across different windows.

The process above yielded 1,017 window topics across the 60 time windows. We subsequently applied dynamic topic modeling as described in Section 3.2. For the number of terms t representing each window topic, we experimented with values from 10 to the entire number of terms present in a time window. However,



(a) Number of window topics identified per time window, from 1999-Q3 (#1) to 2014-Q2 (#60).



(b) Plot of mean TC-W2V topic coherence scores for different values for the number dynamic topics of k', across a candidate range $[25, 90]$.

Figure 1: Identifying optimal number of topics using topic coherence

values $t > 20$ did not result in significantly different dynamic topics. Therefore, to minimize the dimensionality of the data, we selected $t = 20$. This yielded a matrix of 1,017 window topics represented by 2,710 distinct terms.

Our next task was to identify a value for the parameter k for the dynamic part of the model, i.e.- the number of dynamic topics in the corpus. To do this, we calculated TC-W2V coherence scores for a set of topic models with a range

$k' \in [25, 90]$ and then compared these coherence scores to identify the appropriate parameter value. The resulting plot (see Fig. 1b) indicated a maximal value at $k' = 57$, although a number of close peaks exist in the range [62,80]. When we manually inspected the results of the most coherent topic models for these values of k', they were highly similar in terms of the topics detected, with minor variations corresponding to merges or splits of strongly-related topics.

5.2 Dynamic Topic Validation

5.2.1 Summary

The 57 topics identified in our experiments are diverse in terms of their thematic content and temporal signatures. Table 2 lists the top 20 dynamic topics in the data, ranked by their TC-W2V topic coherence scores. We report the temporal frequency of the topics, together with a manually-assigned descriptor for discussion purposes.[3] The frequency of dynamic topics ranged from 11, which appeared in < 10 time windows, to a broad 'Plenary administration' topic which appeared in 57 out of 60 windows. The top 10 terms reported for each topic in Table 2 are highly coherent upon inspection, and clearly distinguish between different topics.

In general, we observed two distinct categories of dynamic topics. The first reflects the day-to-day politics of EU in terms of legislating and debating issues related to the core EU competencies (*e.g.* 'Energy', 'Agriculture'), while the other reflects unanticipated exogenous shocks and MEPs reactions to these events (*e.g.* Euro-crisis, September 11th attacks). These two categories exhibit differing temporal signatures. For instance, we see a considerable difference between the

[3]Full details of all window topics and dynamic topics are available at REDACTED

Topic	Short Label	Top 10 Terms	Coh.	Freq.
13	Transport	transport, railway, rail, passenger, road, network, freight, system, train, infrastructure	0.54	19
42	The Balkans	kosovo, serbia, balkan, resolution, bosnia, albania, iceland, herzegovina, macedonia, process	0.50	12
33	Air transport	air, passenger, transport, aviation, airport, traffic, airline, flight, sky, single	0.48	10
29	Adjusting to globalisation	fund, globalisation, egf, worker, adjustment, mobilisation, european, redundant, application, eur	0.47	15
6	Energy	energy, gas, renewable, efficiency, supply, source, electricity, market, target, project	0.47	36
39	Education & culture	programme, education, culture, language, cultural, youth, sport, learning, young, training	0.43	21
8	Fisheries	fishery, fishing, fish, stock, fisherman, fleet, sea, common, policy, measure	0.43	34
2	Human rights	rights, human, fundamental, freedom, democracy, law, charter, resolution, union, violation	0.43	52
45	Maritime issues	port, sea, maritime, safety, ship, accident, oil, vessel, transport, inspection	0.43	10
21	Healthcare	health, patient, environment, safety, public, care, healthcare, action, disease, mental	0.42	18
26	Child protection	child, internet, pornography, sexual, school, exploitation, young, victim, education, crime	0.42	14
56	Road safety	road, safety, vehicle, transport, system, driver, accident, motor, noise, ecall	0.41	12
16	Research	research, programme, innovation, framework, funding, industry, technology, development, cell, institute	0.41	15
15	Turkish accession	turkey, turkish, accession, progress, cyprus, negotiation, union, membership, croatia, macedonia	0.41	20
35	Tax	tax, vat, taxation, rate, system, fraud, states, evasion, car, transaction	0.41	11
32	Trade - WTO & aid	trade, wto, world, development, developing, international, negotiation, aid, free, relation	0.39	19
47	Product labelling & regulation	product, medicinal, medicine, tobacco, labelling, safety, consumer, regulation, organic, advertising	0.39	11
11	Trade - Trade partnerships	agreement, partnership, morocco, trade, negotiation, data, cooperation, association, korea, fishery	0.39	18
49	Regional funds	policy, region, cohesion, development, regional, strategy, structural, fund, economic, area	0.39	22
17	CFSP	security, policy, defence, common, foreign, military, nato, immigration, aspect, european	0.39	19

Table 2: List of top 20 dynamic topics, ranked by their TC-W2V topic coherence. For each dynamic topic, we report a manually-assigned short label, the top 10 terms, coherence, and frequency (*i.e.* number of windows in which it appeared).

broad topic on fisheries policy (Fig. 4c), when compared to the two topics arising from the events during the financial crisis and subsequent Euro-crisis as shown in Figure 4a. This distinction between dynamic topic types reflects two different forms of political process in the Parliament.

5.2.2 Intra-Topic Validity

To examine the intra-topic semantic validity of these dynamic topics, we examined the distribution of TC-W2V coherence values for all dynamic topics, when evaluated in the *word2vec* space built from the complete speech corpus. These coherence values correspond to the mean of the pairwise cosine similarities between the top-10 terms for each topic in the *word2vec* space (see Eqn. 1). As evidenced by the coherence values reported in Table 2, the most coherent topics often correspond to core EU competencies. Unsurprisingly, broad administrative topics prove to be least coherent (*e.g.* 'Commission questions', 'Council Presidency', 'Plenary administration'). Overall the mean topic coherence score of 0.36 is considerably higher than the lower bound for TC-W2V (*i.e.* minimum value = -1), suggesting a high level of semantic validity.

5.2.3 Inter-Topic Validity

To assess the inter-topic semantic validity of the results, we examine the extent to which any meaningful higher-level grouping exists among the 57 dynamic topics. To do this we apply average linkage agglomerative clustering to the topics. Using the approach described in Greene et al. (2008), we re-cluster the row vectors from the second-layer NMF factor $\mathbf{H}$ using normalized Pearson correlation as a similarity metric. Here the vectors correspond the weights of each dynamic topic with respect to the 2,710 terms noted above. The dendrogram for the hierarchical clustering is shown in Fig. 2. Following the interpretation provided in Quinn et al. (2010), the lower the height at which any two topics are connected in the dendrogram, the more similar their term usage patterns in EP sessions.

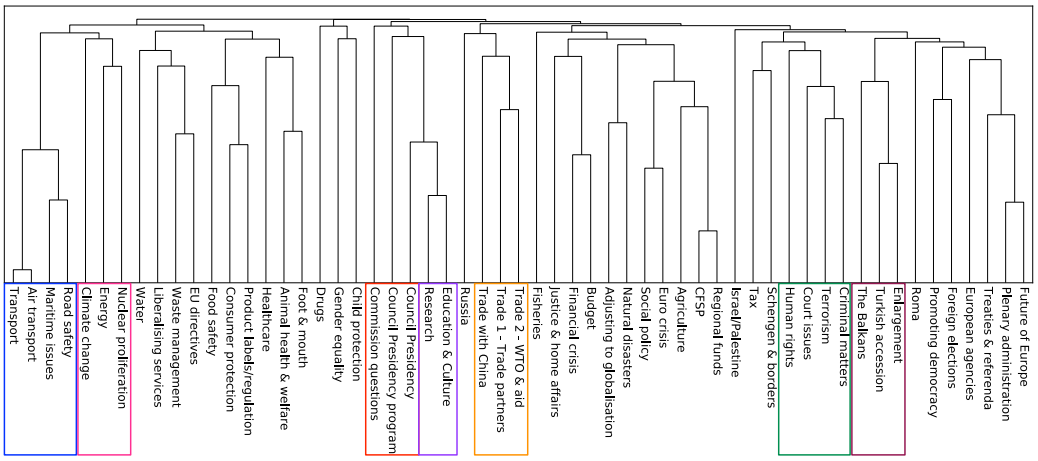


Figure 2: Dendrogram for average linkage hierarchical agglomerative clustering of 57 dynamic topics.

We observe a number of higher-level groupings of interest, which are highlighted in Fig. 2. These include groups related to transport, energy concerns, institutional interactions, education and research, trade relations, and EU enlargement. The presence of these higher-level associations between topics provide semantic validity for the results presented, where topics that one might expect to be related are found to be correlated with respect to rows in their NMF factor $\mathbf{H}$ (*i.e.* similar terms appear in the set of topic descriptors (words) that define them as topics).

5.2.4 External Validation

To assess the extent to which the dynamic topics identified correspond to EU policy areas, and thus provide evidence of construct validity, we compare the 57 dynamic topics to an existing taxonomy of subjects used by Europarl to classify legislative procedures. The taxonomy retrieved from the EP website has several different levels, ranging from broad top-level subjects (*e.g.* '3 Community policies'), to highly-specific low-level subjects (*e.g.* '3.10.06.05 Textile plants, cot-

Subject	Matched Topic: Top 10 Terms	Sim.
1.10 Fundamental Rights In The Union	rights, human, fundamental, freedom, democracy, law, charter, resolution, union, violation	0.66
4.40 Education, Vocational Training & Youth	programme, education, culture, language, cultural, youth, sport, learning, young, training	0.63
5.20 Monetary Union	euro, economic, growth, stability, pact, bank, policy, monetary, economy, ecb	0.62
4.70 Regional Policy	policy, region, cohesion, development, regional, strategy, structural, fund, economic, area	0.62
3.50 Research & Technological Development	research, programme, innovation, framework, funding, industry, technology, development, cell, institute	0.57
3.60 Energy Policy	energy, gas, renewable, efficiency, supply, source, electricity, market, target, project	0.53
6.10 Common Foreign & Security Policy	security, policy, defence, common, foreign, military, nato, immigration, aspect, european	0.52
3.20 Transport Policy in General	transport, railway, rail, passenger, road, network, freight, system, train, infrastructure	0.51
4.60 Consumers' Protection in General	product, medicinal, medicine, tobacco, labelling, safety, consumer, regulation, organic, advertising	0.50
3.70 Environmental Policy	waste, recycling, directive, packaging, management, environment, electronic, fuel, environmental, radioactive	0.50

Table 3: Top 10 legislative procedure subjects with corresponding matching dynamic topics, ranked by cosine similarity of the match.

ton'). We compare our results to the second level of the taxonomy, containing 48 subjects (*e.g.* '3.10 Agricultural policy and economies', '3.20 Transport policy in general'). For each subject code, we create a "subject document" consisting of the description of the subject and all lower-level subjects within that branch of the taxonomy. We then identify the most similar dynamic topic by comparing the top 10 terms for that topic with subject documents, based on cosine similarity.

Table 3 shows the best matching subjects and topics identified using this approach. To give a couple of examples, the topic hand-coded as relating to 'Tax' from our topic model was correctly matched with the Europarl subject code '2.70 Taxation' broadly defined at level-2 of the taxonomy, and with '2.70.01 Direct taxation' and '2.70.02 Indirect taxation' defined separately at level-3 of the taxonomy. When looking at the topic manually labeled as relating to 'Drugs', cosine similarity matches this with the level-2 subject '4.20 Public health', which has a level-3 sub-category relating to '4.20.04 Pharmaceutical products and industry'.

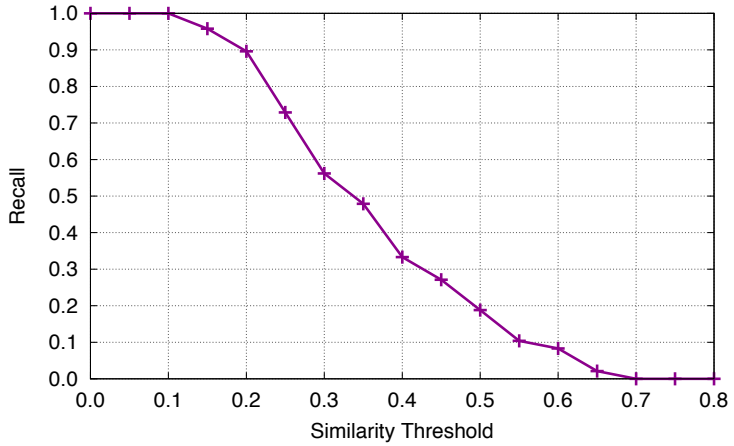


Figure 3: Recall plot for EP taxonomy subjects relative to dynamic topics, for increasing thresholds for cosine similarity.

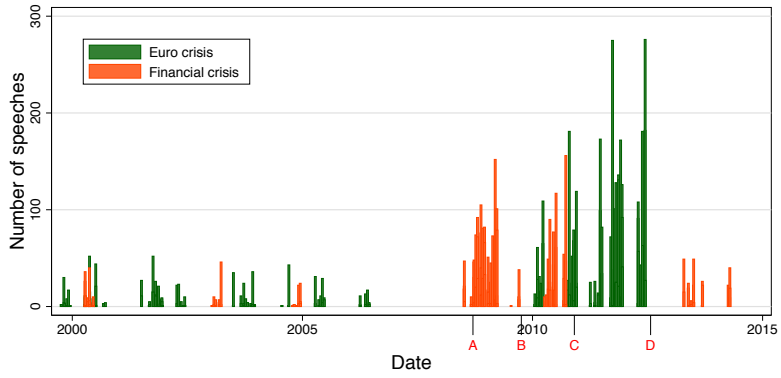
When taken in the context of the matches shown in Table 3, this indicates that our dynamic topics provide good coverage of the policy areas that might be expected to feature during EP debates, and thus increases our confidence in the construct validity of the model.

5.3 Case Studies

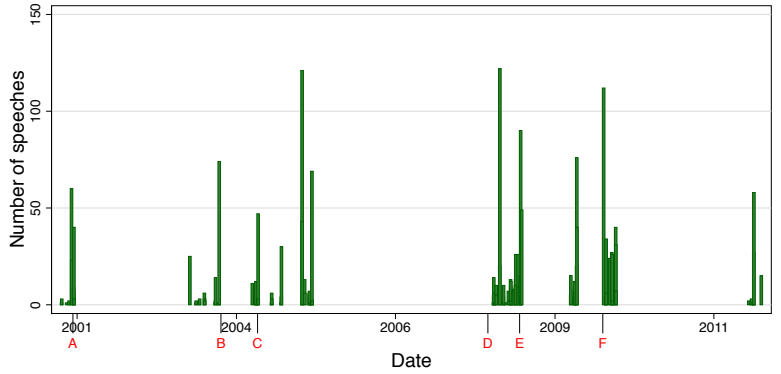
In order to further investigate the construct validity of our topics, we focus on three case studies to demonstrate how our topic modeling strategy captures variation in MEP topic attention over time, and how this attention reacts to external stimuli.

5.3.1 Financial/Euro-crisis

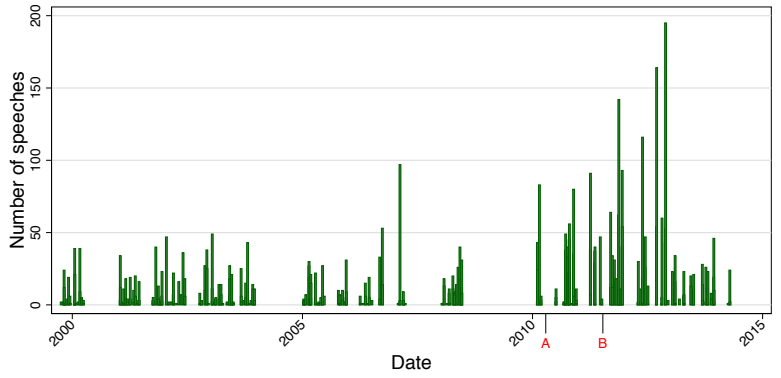
Our first case study relates to MEP attention as illustrated by two topics covering the financial/Euro-crisis, and is illustrated in Fig. 4a. This is an interesting case study, as the initial financial crisis peaked in 2008, and the Euro-crisis that followed has gone through a number of phases with major events in 2009, 2010



(a) "Financial & Euro crises" dynamic topics



(b) "'Treaty changes & referenda" dynamic topic



(c) "Fisheries" dynamic topic

Figure 4: Time plots for three sample dynamic topics across all time windows, from 1999-Q3 (time window #1) to 2014-Q2 (time window #60).

and 2012. As such, these events can be thought of as exogenous shocks that only garner MEP attention after they occur, and their exogenous nature provides a way to externally validity the dynamic topic modeling approach in use here. Fig. 4a demonstrates a number of distinct peaks in MEP speech making on both the financial crisis topic (in orange) and the Euro-crisis topic (in green). Attention to the financial crisis starts to rise in 2008-Q3 and initially peaks in 2008-Q4 (point A in Fig. 4a). This peak in activity corresponds to the date when the Lehman Brothers investment bank collapsed (15/9/2008). The other peaks in activity in Fig. 4a correspond to important events in the Euro-crisis. Point B corresponds to the revelations about under-reporting of Greek debt following the Greek parliamentary election in October 2010, Point C to the Irish bailout (November 2010), and Point D to Mario Draghi's statement that the ECB was "ready to do whatever it takes to preserve the euro" (July 2012). Draghi's statement temporarily at least reassured markets and put a lid on the Eurozone crisis. This goes some way to explaining why fewer speeches relating to this topic are observed after Point D.

5.3.2 Treaty Reform

Our second case study relates to EU treaty reforms. This topic is of interest, because one would expect a large amount in variation in MEP attention to the topic over time, as Treaty revision and reform and the referenda that accompany them are rare event and should only garner MEP attention when such events occur. Fig. 4b shows MEP attention to the treaty change and referenda topic between 2000 and 2014 in terms of the number of speeches associated with this topic. Three distinct treaties were discussed and debated over this period. The first was

the Nice treaty, which was agreed upon in 2001 and put to the vote in a referendum in Ireland in June 2001. The 'No' vote in that resulted from this referendum accounts for Point A in Fig. 4b. The next set of treaty related events to occur were the negotiations and failed ratification of the Constitutional Treaty between 2003 and 2005. This process accounts for Point B in Fig. 4b, which corresponds to the Intergovernmental Conference negotiating the treaty text starting in October 2003. In the end the Constitutional Treaty was rejected by the French and Dutch in referenda in May/June 2005. Point C indicates the date of the signing of the Enlargement treaty in May 2004. The Lisbon treaty was negotiated to replace the failed Constitutional treaty, and we observe a significant peak in MEP speeches directly relating to the Lisbon treaty when it was signed (Point D), and when the first Irish referendum failed to ratify the treaty in June 2008 (Point E). A similar peak in MEP speeches relating to treaty reform corresponds to the second Irish referendum that eventually approved the Lisbon treaty in October 2009 (Point F).

5.3.3 Fisheries Policy

Our third and final case study relates to fisheries policy. Fisheries is an interesting policy area for the dynamic topic modeling approach to detect, because it is more associated with the day-to-day functioning of the EU as a regulator of the fisheries industry, when compared to more headline making policies and events like the economic crisis and treaty changes. Fig. 4c demonstrates the prevalence of the fisheries topic over time. As can be seen, MEPs are seen to pay a reasonably stable level of attention to fisheries in terms of the numbers of speeches being made between 2000 and 2010. This trend is interrupted in 2010, when an increase in

MEP attention to the fisheries topic is observed. This can be explained by the fact that in 2009 the European Commission launched a public consultation on reforming EU fisheries policy, the results of which were presented to the Parliament and Council in April 2010. The launch of this working document corresponds to a increase in the number of MEP speeches related to the fisheries topic as detected by the dynamic topic model (Point A). The peak in MEP speech making relating to this topic (Point B) corresponds with Commissioner Maria Damanaki introducing a set of legislative proposals designed to reform the common fisheries policy in a speech to the European Parliament in July 2011.

In general, the fact that the variation over time that we observe in MEP attention to these case study topics appears to be driven by exogenous events provides a form of construct validity for our topic modeling approach.

5.4 Explaining MEP Speech Counts

We now focus our attention on the 7th European Parliament that sat between 2009 and 2014. We focus on this term, as a set of interesting covariates are available at the MEP level that can help us explain MEP contributions to a given topic. The dependent variable we seek to explain is the observed variation in the number of speeches each MEP makes on each of our identified dynamic topics. We employ a count-model framework suitable for analyzing count data (Cameron and Trivedi, 2013). The first issue to note with the count variable under consideration is that there is a large number of zeros observed. This is due to the fact that, for many topics, a considerable number of MEPs are recorded as making no speeches. This is likely due to the process through which the topic model generates our de-

pendent variable. As described in Section 3.1, we apply a single membership topic modeling approach where each speech is associated with one topic. This assumption is generally unproblematic, given the short amount of time allowed for speeches and the concentrated nature of the messages MEPs seek to communicate in them. However, any speeches that might contain multiple topics are only counted towards a single topic in the model. The result is that, in some cases, the "true" number of topics addressed by MEPs is under-represented and an inflated zero count is observed. In order to account for the inflated zero count, we model MEP speech-making as a two-stage process using a zero-inflated negative binomial regression model. A zero-inflated negative binomial model includes a Logit regression component to capture the binary process determining whether or not a MEP speaks on a topic, and a negative-binomial regression component that seeks to capture the count process determining the number of speeches made, given that a MEP has chosen to speak on a topic.

In order to explain the variation observed in our dependent variable, we include variables relating to MEP's ideology, voting behavior, and the institutional structures in which they find themselves embedded within. We account for the left-right ideological position of a MEP's national party (as a proxy for MEP ideology) using data from Scully et al. (2012). Following Proksch and Slapin (2014), we also include a measure of how often MEPs vote against their party group in favor of their national party and vice versa. The idea behind including these variables is that MEPs rebelling against one party affiliation in favor of another will either try to explain such behavior in their speeches thus increasing the count, or hide their behavior by making no speeches, thus decreasing the count. These data were taken from an updated version of the Hix et al. (2006) dataset provided by

those authors. In order to capture an MEP's committee positions we include dummies for committee membership, chairs, and Rapporteurs in committees that are directly related to a given topic. Committees were manually matched with topics to achieve this. We control for whether or not an MEP serves in the Parliamentary leadership. Controls are also included for the total number of speeches made by an MEP and the percentage of MEP speeches that are available in English as these are liable to affect the observed MEP speech count. Finally, we also include dummy variables to control for an MEP's country of origin, EP party-group membership, and the topic on which they are speaking. All institutional and control variables were scraped from the legislative observatory of the European Parliament.

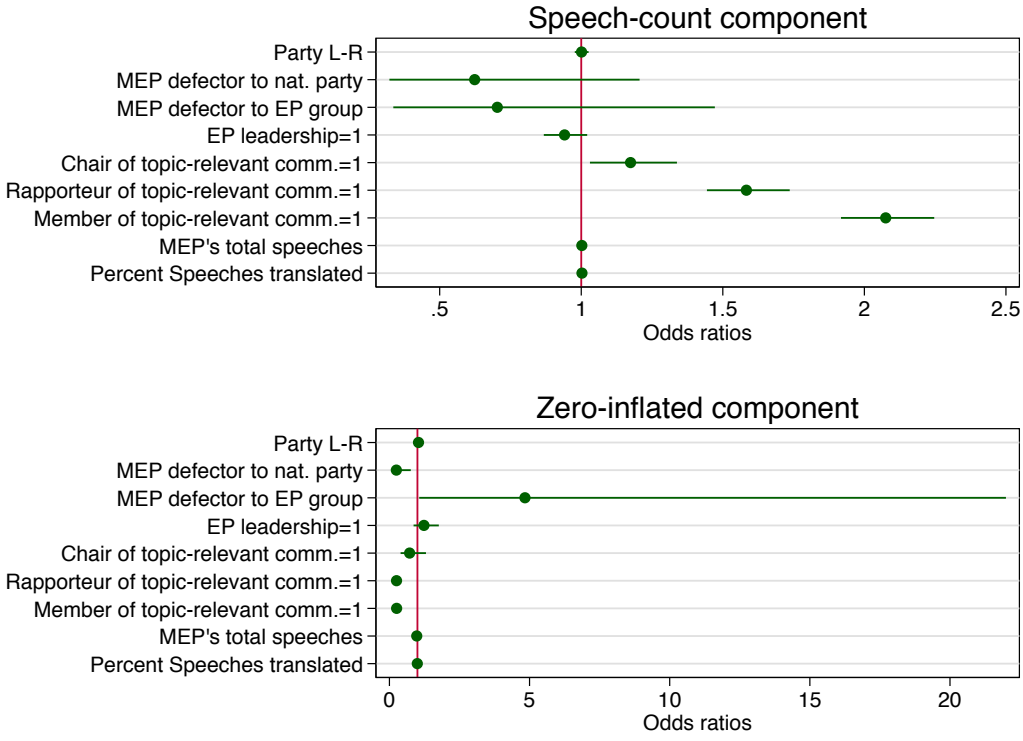


Figure 5: Plot of coefficients for regression model.

The regression presented in Fig. 5 provides further validation for the results of

our topic modeling approach. The coefficients of the model have been exponentiated so as to represent odds ratios and aid interpretation. For the Logit component of the model accounting for zero inflation, an odds ratio above 1 implies that an increase in that covariate leads to an increase in the odds that no speech is made, while an odds ratio below 1 implies an increase in that variable leads to a decrease in the odds of a speech being made. For the count component of the model an odds ratio above 1 implies a positive relationship between the predictor and outcome variable, while an odds ratio below 1 implies a negative relationship between the predictor and outcome variable.

We begin with the zero-inflated component of the model at the bottom of Fig. 5. The model suggests that a MEP's national party ideology impacts upon whether or not they make speeches on a given topic, with more right-wing MEPs tending to make no topic speeches more often than left-wing MEPs. Furthermore, MEPs defecting to national parties tend to make speeches more often than those not defecting, while the opposite is true for MEPs defecting to EP groups from national parties. This is in line with the findings of Proksch and Slapin (2014) who demonstrate that MEPs who rebel against their European party groups tend to make more speeches explaining why they do so, while those rebelling against their national party tend to make less speeches advertising their defection from the national party majority.

Of the institutional related variables, holding a leadership position or a chair of a topic relevant committee has no significant relationship to MEP speech making, while being a member of a committee relevant to a topic, or holding a Rapporteurship for such a topic-relevant committee significantly impact upon whether or not MEPs make a speech on that topic. The odds that an MEP makes no speeches on

a given topic decrease by a factor of 0.255 if an MEP is a Rapporteur of a topic-relevant Committee and decrease by a factor of 0.259 if that MEP is a member of a topic-relevant committee.

Moving to the speech-count component of the model at the top of Fig. 5, the results further reinforce our expectations that MEP positions within the Parliamentary committee system impact upon how much attention they pay to a particular topic. When an MEP holds a committee chair, Rapporteurship, or committee membership relevant to a particular topic, the odds that said MEP will make a speech on that topic increase by a factor of 1.173, 1.582, and 2.077 respectively. These results reinforce the idea that the committee system fundamentally shapes speech-making activities in the broader plenary of the EP.

In contrast, the results relating to MEPs rebelling against either national or EP party group affiliations differ significantly from the results of the zero-inflated component of the model and those presented by Proksch and Slapin (2014). Neither of these variables has a significant effect on the number of speeches made by an MEP on a topic once they have decided to speak on that topic, and the sign of the 'MEP defecting to national party' variable is reversed. These results can be explained by the fact that voting cohesion in the EP is generally quite high, and while party groups will allow rebels to explain their vote, this is a relatively rare event. In contrast, allocating speaking time to MEPs with committee roles is much more common, hence why these variables affect both the initial decision to speak on a topic and the number of speeches made on that topic. Essentially, MEPs with important committee positions will be asked to speak on more than one occasion when the topic being discussed is relevant to their committee role.

6 Conclusions

In this paper, we have proposed a new two-layer matrix factorization methodology for identifying topics in large political speech corpora over time, designed to identify both niche topics related to events at a particular point in time and broad, long-running topics. We applied this method to a new corpus of all $\approx 210k$ English language plenary speeches from the European Parliament during a 15-year period. In terms of providing substantive insight into the political processes of the European Parliament, the topic-modeling method has allowed us to unveil the political agenda in the Parliament, and the manner in which this has evolved over the time period under consideration. By considering three distinct case studies, we have demonstrated the distinctions that can be drawn between the day-to-day political work of the Parliament in policy areas such as fisheries on the one hand, and the manner in which exogenous events such as economic crises and failed treaty referenda can give rise to new topics of discussion between MEPs on the other. Once the Parliamentary agenda was extracted from the corpus of speeches, we explored the determinants of MEP attention to particular topics in the 7th sitting of the Parliament. We demonstrated how MEP ideology and voting behavior affect whether or not they choose to contribute to a topic, and once such a decision has been made, we demonstrated how the committee structure of the Parliament structures MEP contributions on a given topic.

The initial insights provide by the dynamic topic modeling approach presented here demonstrate how these methods can uncover latent dynamics in MEP speech-making activities and thus provide new insights into how the EU functions as a political system. Much remains to be explored in terms of the patterns in polit-

ical attention that emerge from the topic modeling approach. For instance, one would expect that political attention might well translate into influence over policy outcomes decided upon in the Parliament. Tracing influence to date has been difficult, as a macro-level picture of where and on what topics MEP attention lays has been unavailable. Linking political attention to political outcomes would help to unveil who gets what and when in European politics, which is a central concern for a political system often criticized for lacking democratic legitimacy.

References

Blei, D. M. and Lafferty, J. D. (2006) 'Dynamic topic models', in 'Proc. 23rd International Conference on Machine Learning', pp. 113–120.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003) 'Latent Dirichlet Allocation'. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022.

Boutsidis, C. and Gallopoulos, E. (2008) 'SVD based initialization: A head start for non-negative matrix factorization'. *Pattern Recognition*.

Bowler, S. and Farrell, D. M. (1995) 'The organizing of the European Parliament: Committees, specialization and co-ordination'. *British Journal of Political Science*, Vol. 25, No. 02, pp. 219–243.

Cameron, A. C. and Trivedi, P. K. (2013) *Regression analysis of count data*, Vol. 53 (Cambridge university press).

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C. and Blei, D. M. (2009) 'Reading Tea Leaves: How Humans Interpret Topic Models', in 'NIPS', pp. 288–296.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W. and Harshman, R. A. (1990) 'Indexing by Latent Semantic Analysis'. *Journal of the American Society of Information Science*, Vol. 41, No. 6, pp. 391–407.

Greene, D., Cagney, G., Krogan, N. and Cunningham, P. (2008) 'Ensemble Non-negative Matrix Factorization Methods for Clustering Protein-Protein Interactions'. *Bioinformatics*, Vol. 24, No. 15, pp. 1722–1728.

Greene, D., O'Callaghan, D. and Cunningham, P. (2014) 'How Many Topics? Stability Analysis for Topic Models', in 'Proc. European Conference on Machine Learning (ECML'14)', (Springer), pp. 498–513.

Grimmer, J. (2010) 'A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases'. *Political Analysis*, Vol. 18, No. 1, pp. 1–35.

Hix, S., Noury, A. and Roland, G. (2006) 'Dimensions of politics in the European Parliament'. *American Journal of Political Science*, Vol. 50, No. 2, pp. 494–520.

Hix, Simon, Noury, A. and Roland, G. (2007) *Democratic politics in the European Parliament* (Cambridge University Press).

Jensen, C. B., Proksch, S.-O. and Slapin, J. B. (2013) 'Parliamentary Questions, Oversight, and National Opposition Status in the European Parliament'. *Legislative Studies Quarterly*, Vol. 38, No. 2, pp. 259–282.

Lee, D. D. and Seung, H. S. (1999) 'Learning the parts of objects by non-negative matrix factorization'. *Nature*, Vol. 401, pp. 788–91.

Lin, C. (2007) 'Projected gradient methods for non-negative matrix factorization'. *Neural Computation*, Vol. 19, No. 10, pp. 2756–2779.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013) 'Efficient Estimation of Word Representations in Vector Space'. *CoRR*, Vol. abs/1301.3781.

O'Callaghan, D., Greene, D., Carthy, J. and Cunningham, P. (2015) 'An Analysis of the Coherence of Descriptors in Topic Modeling'. *Expert Systems with Applications (ESWA)*.

Proksch, S.-O. and Slapin, J. B. (2010) 'Position taking in European Parliament speeches'. *British Journal of Political Science*, Vol. 40, No. 03, pp. 587–611.

Proksch, S.-O. and Slapin, J. B. (2014) *The politics of parliamentary debate: Parties, rebels and representation* (Cambridge University Press).

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H. and Radev, D. R. (2010) 'How to analyze political attention with minimal assumptions and costs'. *American Journal of Political Science*, Vol. 54, No. 1, pp. 209–228.

Raunio, T. (1996) 'Parliamentary questions in the European Parliament: Representation, information and control'. *The Journal of Legislative Studies*, Vol. 2, No. 4, pp. 356–382.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B. and Rand, D. G. (2014) 'Structural Topic Models for Open-Ended Survey Responses'. *American Journal of Political Science*, Vol. 58, No. 4, pp. 1064–1082.

Scully, R., Hix, S. and Farrell, D. M. (2012) 'National or European Parliamentarians? Evidence from a New Survey of the Members of the European Parliament'. *JCMS: Journal of Common Market Studies*, Vol. 50, No. 4, pp. 670–683.

Slapin, J. B. and Proksch, S. O. (2010) 'Look who's talking: Parliamentary debate in the European Union'. *European Union Politics*, Vol. 11, No. 3, pp. 333–357.

Steyvers, M. and Griffiths, T. (2007) *Latent Semantic Analysis: A Road to Meaning* (Laurence Erlbaum), chap. Probabilistic topic models.

Sulo, R., Berger-Wolf, T. and Grossman, R. (2010) 'Meaningful selection of temporal resolution for dynamic networks', in 'Proc. 8th Workshop on Mining and Learning with Graphs', (ACM), pp. 127–136.

Wang, Q., Cao, Z., Xu, J. and Li, H. (2012) 'Group matrix factorization for scalable topic modeling', in 'Proc. 35th SIGIR Conf. on Research and Development in Information Retrieval', (ACM), pp. 375–384.