Tweeting Europe: A text-analytic approach to unveiling the content of political actors' Twitter activities in the European Parliament

Mark Belford* Derek Greene† James P. Cross‡

June 17, 2016

Abstract

Twitter is an important platform for communication and is frequently used by Members of the European Parliament (MEPs) to campaign and engage in discussion with constituents and colleagues in the parliament. Examining the issues that MEPs talk about on Twitter can thus inform us about their political priorities. Topic modelling aims to summarise a corpus of documents by capturing the underlying hidden structure of the data and presenting the user with an overview of the key subjects and themes discussed in the corpus, known as topics. This paper aims to quantify and explore the content that MEPs pay attention to on Twitter by applying a new ensemble approach for topic modelling which involves applying two layers of Non-Negative Matrix Factorisation (NMF). The resulting set of issues paid attention to by MEPs are explained by considering the effects of events, issue characteristics, and MEP characteristics.

*Insight Centre for Data Analytics, University College Dublin, Ireland. (mark.belford@insight-centre.org)

†Insight Centre for Data Analytics, University College Dublin, Ireland. (derek.greene@ucd.ie)

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

1 Introduction

Agenda dynamics in political systems have long been a topic of interest to political scientists (Baumgartner and Jones, 1991; Jones and Baumgartner, 2005; Baumgartner and Jones, 2002). Tracking how political actors' attention to policy issues evolves over time, and exploring the factors that determine the evolution of political agendas can help us understand how political systems engage with and respond to new information and policy challenges. In this study, we take up the challenge of capturing and explaining agenda dynamics in the European Parliament (EP), by examining the issues Members of the European Parliament (MEPs) communicate about on Twitter. Twitter has become an important arena where the political elite can engage with and communicate the current policy agenda directly to the public. Due to the public nature of these discussions, we can use Twitter data to uncover and explore the issues to which political actors primarily devote their online attention, and how this attention evolves over time.

In order to examine the nature and content of MEPs activity on Twitter, we introduce and apply a novel topic-modeling approach to a corpus of over 1.28 million tweets posted by 584 MEP accounts from the 8th EP, between July 2014 and April 2016.[1] Topic modeling is an unsupervised machine learning approach that seeks to uncover the latent structure of a corpus of documents (Blei et al., 2003). The output of a topic model is a summary of the corpus content in the form of a set of topics, in which each topic is represented by a ranked list of the top terms that describe it. Analysing these topics allows one to track MEP

[1] In this study we focus on the set of MEPs from Anglophone countries (UK, Ireland, Malta) and the set of Tweets emanating from these MEPs in English, but our future work will expand the analysis to all MEPs.

attention regarding different issues and the way it evolves over time. We expect the major drivers of MEP topic attention to be related to expected and unexpected events and characteristics of the MEPs themselves.

Popular approaches for topic modeling have involved the application of probabilistic algorithms (Blei et al., 2003; Steyvers and Griffiths, 2007), and also, more recently, matrix factorisation algorithms (Wang et al., 2012). In both cases, these algorithms generally include stochastic elements in their initialisation, which can affect the final ordering of the topics and the rankings of the terms that describe those topics. This is problematic when seeking to capture MEP attention, as the set of topics and ranked terms describing them can change based on parameter choices. Such issues represent a fundamental "instability" in these algorithms – different runs of the same algorithm on the same data can produce different outcomes. Most authors do not address this issue and instead simply utilise a single random initialisation and present the results of the topic model as being definitive. Another challenge in topic modeling is the identification of coherent topics on short texts, such as tweets (Aiello et al., 2013). The noisy and sparse nature of this data makes this more difficult when compared to working on longer, cleaner texts such as political speeches or news articles.

Here we consider the idea of *ensemble* machine learning techniques, the rationale for which is that the combined judgement of a group of algorithms will often be superior to that of an individual (Breiman, 1996). Such techniques have been well-established for both supervised classification tasks (Opitz and Shavlik, 1996) and also for unsupervised cluster analysis tasks (Strehl and Ghosh, 2002b). In the case of the latter, the goal is to produce a "better" clustering of the data, which also avoids the issue of instability. The application of unsupervised ensembles

generally involves two distinct stages: 1) the generation of a collection of different clusterings of the data; 2) the integration of these clusterings to yield a single more accurate, informative clustering of the data. A variety of different strategies for both generation and integration have been proposed in the literature (Ghaemi et al., 2009).

In this paper we propose an ensemble algorithm for topic modelling, based on the generating and integration of the results generated from multiple runs of Non-negative Matrix Factorization (NMF) (Lee and Seung, 1999) on samples of a corpus of short texts. The integration aspect of the algorithm builds on previous work involving the combination of topics from different time periods with NMF (Greene and Cross, 2015). We apply this approach to tweets in our corpus which were posted by a subset of MEPs from Anglophone member states during 2014–2016.[2] This analysis reveals a diverse set of topics being discussed by MEPs, ranging from discussions on internal EP activities, reactions to exogenous events, through to canvassing for the referendum on EU membership (Brexit). Our proposed algorithm allows us to robustly identify these topics, and chart their evolution across the time period under examination.

The rest of the paper is structured as follows. In Section 2 we review the existing literature in the areas of political text analysis, topic modeling, and ensemble methods. In Section 3 we discuss the rationale behind our analysis and the determinants of MEP attention to different issues, and then propose a suitable methodology in Section 4. Then Section 5 summarises our data collection and preparation, while the findings of our analysis are presented and explained in

[2]We later plan to analyse the whole set of MEPs on Twitter, but do not do so now due to the difficulties associated with multi-lingual topic modelling.

Section 6.

2 Related Work

2.1 Policy Agendas and Political Attention

Major efforts to track and explain policy agendas have developed in recent years. Beginning in the early 1990s, the Policy Agendas Project (PAP) and the Comparative Agendas Project (CAP) have tracked policy agendas across different political systems, including the EU. The major claim in both of these projects is that the variation in the attention that political figures pay to different issues across time can be described by a *punctuated equilibrium* dynamic, whereby issue attention is stable for long periods of time, but these periods are punctuated by short bursts of increased attention (Baumgartner and Jones, 1993). The sudden punctuations in political attention have been explained by factors including the bounded rationality of the political figures involved (Jones, 1994), (re-)framing of policy choices (Jones and Baumgartner, 2005), and the influence of exogenous shocks on political priorities (Jones and Baumgartner, 2012; John and Bevan, 2012), all of which lead to abrupt spikes in issue attention. Despite some conceptual and measurement issues (Dowding et al., 2015), evidence for the existence of this type of agenda dynamic is found across a multitude of political systems (Baumgartner et al., 2009).

In the EU context, and building upon the techniques developed by the PAP/CAP to capture the aforementioned punctuated-equilibrium dynamic, most academic attention has focused on the evolving policy agenda of the European Council

(Alexandrova et al., 2012, 2013). Similar to what has been found in other contexts, a punctuated equilibrium dynamic appears to be in play in the European Council, with long periods of agenda stability interrupted with sharp spikes in issue attention. Institutional, contextual and issue-specific factors are found to explain these punctuations. The agenda of the EP plenary has been examined by Greene and Cross (2015), using a dynamic topic model technique related to the one proposed in this study. They find 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. They also demonstrate the usefulness of topic-modeling techniques for uncovering latent patterns in political texts. To date, the policy agendas of other EU institutions have been neglected due to the challenges associated with capturing the diverse, diffuse, and multifaceted nature of the policy agendas found in institutions like the Commission and Council of Ministers.

2.2 Twitter Use in Politics

In recent years, political figures and the political institutions of the EU have adopted Twitter as a communication tool *en masse*. They have been found to utilise Twitter as a campaign tool (Gibson, 2015; Strandberg, 2013; Jungherr, 2014a; Obholzer and Daniel, 2016), as a means to increase their exposure and profile (Vergeer et al., 2013; Theocharis et al., 2015), and as a means of getting insight into public opinion (Anstead and O'Loughlin, 2015). A comprehensive review of the use of Twitter in politics was provided by Jungherr (2014b). Here we focus on literature directly relevant to the current study.

An innovative set of studies has utilised the structure of Twitter networks as a source of data that can tell us about the latent ideological positions of political elites, media sources, and the general public (Barberá, 2015; Conover et al., 2011; King et al., 2011). Measures of ideology generated using this approach have been shown to replicate more conventional measures of ideology, thus validating Twitter networks as a source of substantive information about political processes. Ecker (2015) suggests that some caution should be used when extracting individual-level positional data from Twitter networks, based on the connection between political representatives position in an online social network like Twitter and their individual voting records. This is reasonable advice given our current levels of understanding about what twitter data represents, and the fast-evolving nature of the platform as a political communication tool.

In the context of EU politics, there has been an explosion of the use of Twitter as a political communication tool across all EU institutions.[3] Some academic attention has been paid to electioneering on Twitter in the EU context. Lorenzo-Rodríguez and Madariaga (2015) demonstrate that the degree to which candidates adopt social media as a form of campaigning is related to the profile of their party, incumbency, rates of internet use in their home country, and ballot paper positioning. In an in-depth study of the use of Twitter in the 2014 EP elections, Nulty et al. (2015) examine questions relating to the adoption and use of social media by candidates. The volume and content of Twitter activity are examined over the course of the campaign. Patterns in these aspects of social media use are explained with reference to explanatory variables including the gender, incumbency status,

[3] See `http://europa.eu/contact/social-networks/index_en.htm` for an up to date list of EU actors and institutions on Twitter and other online platforms

ideology, and pro-Europeanness of candidates. The dynamics of Twitter use over the course of the campaign are also investigated, with increasing activity on the social network being observed as the campaign progressed and the election approached. Finally, the content of Tweets was also shown to evolve over the course of the campaign as demonstrated by the evolving use of Twitter hashtags. Hashtag use was shown to be related to the emergence of *Spitzenkandidaten*, with different terms co-occurring with references to each individual candidate (Schulz, Verhofstadt, and Junker). Hashtag use and the positive/negative sentiment expressed in Tweets was shown to vary depending on the country of origin of EP candidates.

Studying the same election period, Barberá et al. (2015) propose a new method for measuring the ideological positions of individual MEPs and party groups in the Parliament based on the structure of the Twitter network. Their results demonstrate that two major dimensions structure the Twitter networks that actors/parties find themselves situated within: the traditional left-right dimension; and a second dimension relating to how Europhobe/Europhile a given actor/party is.

Less attention has been paid to the use of online communication tools in the EP in non-election periods. Larsson (2015) finds that MEPs tend to use Twitter less regularly outside election time, suggesting that the 'permanence' in online campaigning is relatively low. While it is certainly the case that on average Twitter use by MEPs is lower outside election time, focusing on the quantity of tweets alone tells us nothing about the content of tweets, the networks through which they propagate, and the manner in which MEPs use the medium as a way to communicate the internal politics of the EP to a wider audience. The idea that Twitter can give us an insight into the *internal political processes* of the EU is thus underresearched. This paper aims to address this gap in the literature by presenting

a new dataset that captures MEP Twitter activities in the EP, and the manner in which the online social network they find themselves interacting within evolves over time.

2.3 Analysis of Twitter Data

Many researchers have become interested in exploring network structures within the Twitter platform, given the potential for the platform to facilitate both online conversation and the rapid spread of information. Java et al. (2007) provided an initial analysis of the early growth of the platform, and also performed a small-scale evaluation that indicated the presence of distinct Twitter user communities, where the members shared common interests as reflected by the terms appearing in their tweets. Kwak et al. (2010) performed an evaluation based on a sample of 41.7 million users and 106 million tweets from a network mining perspective. The authors studied aspects such as: identifying influential users, information diffusion, and trending topics.

To examine the content of user interactions on Twitter, Shamma et al. (2009) performed an analysis on tweeting activity during the 2008 US presidential debates. The authors demonstrated that frequent terms reflected the topics being discussed, but the use of informal vocabulary complicated topic identification. As an alternative to analysing all terms present in tweets, some researchers have focused specifically on hashtags in tweets. It has been observed that hashtags can lead to the formation of ad-hoc groupings around certain themes and topics (Shi et al., 2014). From a content analysis perspective, hashtags often represent informal "labels" for tweets (Ma et al., 2014), and can potentially mitigate the dif-

ficulties of handling multi-lingual tweet corpora[4]. As a result, hashtags have been used as key features in a number of tasks. For event detection, the emergence of hashtags exhibiting "bursty" behaviour can potentially be indicative of breaking news events (Cui et al., 2012), while for topic discovery the co-occurrence of hashtags can provide useful topic indicators (Wang et al., 2014). As an example in the political domain, Kalmeijer (2014) applied a spectral clustering approach to tags in tweets posted by members of the Dutch parliament, in order to identify topics of interest and to investigate the differences between content posted by politicians from distinct parties. As with other text clusterings tasks, synonymy remains an issue in content analysis via hashtags. Often users will use different tags to label similar content, rather than converging on a single hashtag. To identify groups of semantically-related tags, Muntean et al. (2012) applied k-means to both the hashtags and terms appearing in tweets.

What is clear from the literature review detailed above is that the role of Twitter as a political communication tool has become an important topic of study in the fields of political science and data analytics. While we have a growing understanding of how Twitter is used during election campaigns, less attention has been paid to the use of Twitter as a communication tool outside election time. Methodological developments in the field of content-analysis have the potential to provide new insights into how political figures use Twitter to communicate their day-to-day activities in political systems. This study aims to demonstrate this in the context of the EP.

[4]This is especially salient in the multi-lingual context of the EP.

2.4 Topic Modelling

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. A topic model typically consists of k topics, each represented by a ranked list of strongly-associated terms (often referred to as a "topic descriptor"). Each document in the corpus can also be associated with one or more topics. 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).

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; Greene and Cross, 2015). NMF is an unsupervised approach for reducing the dimensionality of non-negative matrices. Given a document-term matrix $\mathbf{A}$, the goal is to approximate this matrix as the product of two non-negative approximate factors $\mathbf{W}$ and $\mathbf{H}$, each with k dimensions, which can be interpreted as k topics. Like LDA, the number of topics k to generate is chosen beforehand. The values in $\mathbf{H}$ provide term weights which can be used to generate topic descriptions, while the values in $\mathbf{W}$ provide topic memberships for documents. One of the advantages of NMF methods over existing LDA methods

is that there are fewer parameter choices involved in the modelling process.

2.5 Ensemble Clustering

In the machine learning literature, it has been shown that combining the strengths of a diverse set of clusterings can often yield more accurate and stable solutions (Strehl and Ghosh, 2002a). Such ensemble clustering approaches typically involves two phases: a *generation* phase where a collection of "base" clusterings are produced, and an *integration* phase where an aggregation function is applied to the ensemble members to produce a consensus solution. Generation often involves repeatedly applying a "base" algorithm with a stochastic element to different samples selected at random from a larger dataset. The most frequently employed integration strategy has been to use the information provided by an ensemble to determine the level of association between pairs of objects in a dataset Strehl and Ghosh (2002a); Fred (2001). The fundamental assumption underlying this strategy is that pairs belonging to the same natural class will frequently be co-assigned during repeated executions of the base clustering algorithm. Other strategies have involved matching together similar clusters from different runs of the base algorithm.

While most of this work has focused on producing disjoint clusterings (*i.e.* each item in the dataset can only belong to a single cluster), researchers have considering combining probabilistic clusterings (Punera and Ghosh, 2007) and factorisations produced via NMF (Greene et al., 2008). In the latter case, the approach was applied to identify hierarchical structures in biological network data.

3 Theory

To date the majority of studies of MEPs as communicative actors have focused on either their communication strategies internal to the Parliament (speeches and Parliamentary questions), or their use of communication tools like Twitter *during* election campaigns (Obholzer and Daniel, 2016). Here we are interested in how MEPs utilise Twitter as a communication tool once they have been elected to office. Specifically, we are interested in the type of issues that garner attention, and what drives this attention. The agenda-setting literature and the punctuated equilibrium model of agenda dynamics is a useful place to start.

3.1 Theorising Attention Dynamics

The punctuated equilibrium model suggest that policy agendas are generally characterised by long periods of stability and gradual evolution, which are then interrupted by dramatic realignments from time to time (Baumgartner and Jones, 1991; Jones and Baumgartner, 2005; Baumgartner and Jones, 2002). The cognitive limitations of political actors and the constraints on policy change imposed by political institutions contribute to agenda stability. In contrast, the reactions of political actors to new information or events in combination with cascade effects in interest mobilization can lead to sudden realignments of issue attention. Essentially, the extended periods of stability in issue attention are a result of negative feedback processes, while sudden punctuations in attention are a result of positive feedback mechanisms that act in short bursts and force the policy agenda into a new equilibrium (Jennings and John, 2009). In this study we aim to account for both mechanism types to provide an account of how MEP attention to

topics evolves over time. One window through which we can examine the implications of the punctuated equilibrium model is by considering the varying effects of different types of events on MEP attention to different issues.

3.2 Events as a driver of issue attention

An event can be defined as something that happens at some specific time and place, and the unavoidable consequences it implies (Yang et al., 1999). As Woolley (2000) points out, it is almost impossible to know with any confidence the true universe of events. Instead, we are reliant on the documented reports and reactions of actors generated by such events.[5] Twitter is a particularly useful source for such data, as there is little limitation on users ability to react to events, unlike in traditional media outlets where editorial control is present. In the context of Twitter, an event can thus be formally defined as a real-world occurrence e with (1) an associated time period T_e and (2) a time-ordered stream of Twitter messages, of substantial volume, discussing the occurrence and published during time T_e (Becker et al., 2012).

Events impact upon agenda topics as they can trigger changes in MEP attention to a given topic over time. We can differentiate between two distinct periods in which MEP attention to a topic can be affected by a topic-relevant event, relating to the time before and the time after said event. Before an topic-related event, MEP attention to said topic can be assumed to be evolving according to an established trend or equilibrium, most likely established by a set of stable MEP- and institutional-level variables. Upon the occasion of a topic-relevant event, this

[5]The existence of these reports is of course predicated upon information about an event reaching a potential report writer

trend or equilibrium will be impacted upon by characteristics of the event itself. Whether or not an event is expected is a key consideration in determining the likely effect the event might have on MEP attention.

3.2.1 Expected v Unexpected Events

In general the punctuated equilibrium model has been applied to contexts where agenda evolution is constrained by cognitive limitations, institutions, and the restricting nature of policy-making processes. Plenary agendas/debates, legislative outputs, and the conclusions of EU Council meetings have all been considered and have been found to be subject to these types of constraints. Political actors engagement with issues through social media can be expected to be less affected by such constraints, as there are few formal limits to what can be said through this medium. In an online environment like Twitter, one can expect new salient information and events to spread quickly, and for information cascades to amplify the effects of such information on the issues that are being paid attention to. To explore these mechanisms, we differentiate between two types of events based upon actor's ability to anticipate them.

The first class of event that is expected to impact upon MEP attention relates to events that are salient and set to occur on a date known well in advance. In a political context, elections and referenda are usually the prototypical example of such an event, where an election/referendum date is known in advance, and the actors have an interest in the election/referendum outcome. If we consider these types of events in the context of the punctuated equilibrium model, we would expect little attention to election/referendum topics before they are announced,[6] and a gradual

[6]Or expected to be announced.

increase in attention to the topic as the election/referendum date approaches and the associated campaigns intensify. The expected nature of the event allows actors to form expectations and gradually adjust these expectations to new information as it becomes available. The effect of this type of event on aggregate MEP attention is likely to be gradual and to dissipate once the event has occurred, as a new equilibrium (government/constitutional/institutional choice) has been established and concerns directly relating to the election/referendum result fade away.

The second class of event that will impact upon MEP attention is that which is salient and unexpected. Unexpected events cannot be anticipated by actors and as a result should have a much more immediate impact on issue attention, provided information about the unexpected event gets to the relevant actors. We should observe a stable trend in topic attention up to the point where an unexpected topic-relevant event occurs and a sudden interruption in this trend at the point when the event occurs. The effect of a topic-relevant event in the period after its occurrence can be short-term or long-term. Short-term effects peter out quickly and result in little overall change in attention dynamics beyond the immediate period following the event. They will be characterised by a leptokurtic distribution of MEP attention surrounding the immediate time of the event.

Long-term effects on the other hand result in a re-alignment of MEP attention priorities. Such realignments can be positive-sum, increasing the overall usage of Twitter and level of attention MEPs pay to the set of topics (a *usage effect*), or they can be zero-sum and characterised by a *substitution effect* where attention transfers from one topic to another but the overall level of attention to all topics does not change. If such realignments of attention occur, we should expect to observe a stable trend in attention before the event, an interruption in this trend at

the time of the event, and a new trend being established after the event.

We predict both usage and substitution effects on issue attention due to events to be present in MEP Twitter data. To date capturing such dynamics at a very fine-grained level has been difficult. For the purposes of this paper, we simply explore the degree to which topic attention evolves over time and interpret this evolution in the light of the theoretical arguments just presented. In future work we plan to explicitly explore the conditions under which usage and substitution effects might occur.

4 Methods

4.1 Constructing the Dependent Variable: Overview

In this section we propose a new method for topic modelling, which involves applying ensemble learning in the form of two layers of NMF, in order to produce a robust and accurate final set of topics. This method builds on previous work on dynamic topic modelling involving the combination of topics from different time periods (Greene and Cross, 2015). Topic models are particularly useful for uncovering latent patterns in text use across large corpora of text, and thus serve to unveil how MEP attention to different topics evolve over time.

Fig. 1 shows an overview of the method, which can naturally be divided into two steps, following previous ensemble approaches:

1. *Ensemble generation*: Create a set of *base topic models* by executing multiple runs of NMF applied on different subsets of documents drawn from the overall corpus.

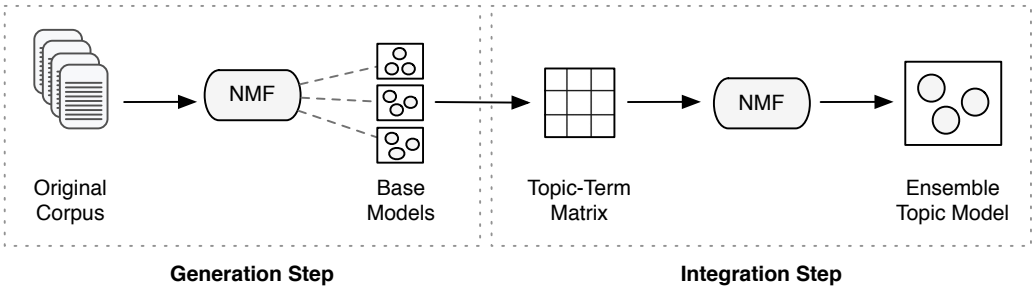

Figure 1: Illustration of the two steps involved in Ensemble Topic modelling: generation and integration.

2. *Ensemble integration*: Transform the base topic models to a suitable intermediate representation, and apply a further run of NMF to produce a single *ensemble topic model*, which represents the final output of the method.

We now discuss each of these steps in more detail.

4.2 Ensemble Generation

It has frequently been shown that supervised ensemble learning algorithms are most successful when constructed from a set of accurate classifiers whose errors lie in different parts of the data space (Opitz and Shavlik, 1996). Similarly, unsupervised ensemble procedures typically seek to encourage diversity with a view to improving the quality of the information available in the integration phase (Topchy et al., 2005). Therefore, in the first step of our approach, we create a diverse set of r base topic models – *i.e.* the topic term descriptors and document assignments will differ from one base model to another. Diversity is encouraged in two ways. Firstly, for each base topic model we randomly select a subset of 80% of documents from our original corpus. Secondly, we apply NMF to the sample of documents, where the starting factors are randomly initialised. In each case we

use a fixed pre-specified value for the number of topics k. After each run, the $\mathbf{W}$ factor from the base topic model (*i.e.* the topic-term weight matrix) is stored for later use. Note that in our experiments we use the fast alternating least squares implementation of NMF introduced by Lin (2007).

4.3 Ensemble Integration

In the second step, we create a new representation of our corpus in the form of a topic-term matrix $\mathbf{M}$. The matrix is created by stacking the transpose of each $\mathbf{W}$ factor generated in the first step. This results in a matrix where each row corresponds to a topic from one of the base topic models, and each column is a term from the original corpus. Each entry M_{ij} holds the weight of association for term *i* in relation to a single topic from a base model.

Once we have created $\mathbf{M}$, we apply the second layer of NMF to this matrix to produce the final ensemble topic model. To improve the quality of the resulting topics, we generate initial factors using the popular Non-negative Double Singular Value Decomposition (NNDSVD) initialisation approach of Boutsidis and Gallopoulos (2008). As an input parameter to NMF, we specify a final number of k' topics. While this value can be set to be the same as the number of topics k in the base models, in practice we observe that an appropriate value of k' may be larger than this the due to the ensemble approach being able to capture topics that only appear intermittently among a diverse set of base topic models.

The results of this process can then be considered at different levels of granularity. Here we are concerned with individual MEP attention to a given topic and how this evolves over time, so we construct a measure where the unit of anal-

ysis is an individual MEPs contribution to a given topic on a given week. This allows us to consider the evolution of an MEP with regards to different topics at the individual and aggregate level in the analysis that follows.

4.4 Interpretation

The output of our ensemble topic model not only identifies the top terms for each topic but also the top weeks in which that topic was relevant, the top MEPs who contributed, and the top member states involved. These are calculated as a percentage weight of association, generated from the $\mathbf{W}$ factor from the ensemble topic model. In this factor each MEP document has a weight for every ensemble topic. MEP documents can be divided up across a number of associated dimensions: the specific week in time, the MEP in question, the member state of the MEP, and their parliamentary group. We can sum the weights for a topic across all of these dimensions. By dividing by the total across all topics, we can calculate the percentage contribution of each time point, MEP, country or group to a given ensemble topic.

5 Data

5.1 Data Collection

After the European Parliament election 2014, we compiled a curated list of MEPs with active Twitter accounts, based on information available on the official website of the Parliament[7], and also by manually inspecting a range of existing user lists

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

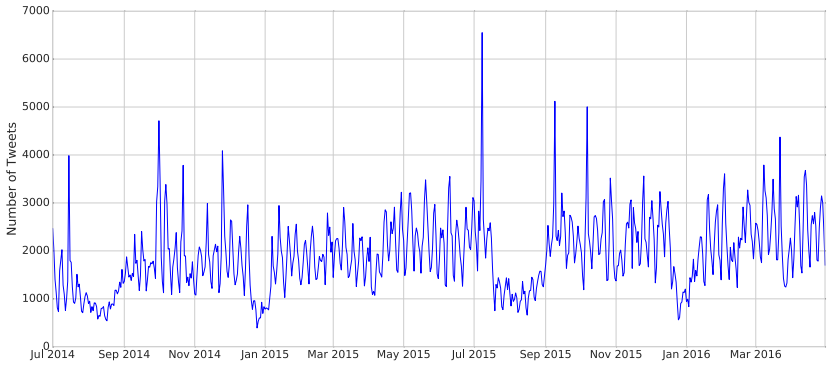

Figure 2: Number of tweets per day by all MEPs during July 2014 – April 2016.

available on Twitter. Our list was subsequently updated in October 2015, yielding 584 active public accounts corresponding to sitting MEPs. As of June 2016, 570 of these accounts are still active and publicly-accessible. Tweets were collected for the 584 accounts using the Twitter REST APIs[8], from the commencement of the 8th European Parliament on 1 July 2014, up until 30 April 2016. This yielded a corpus of 1,289,214 tweets, of which 48.87% are retweets (*i.e.* reshares of posts by other users), and 12.95% are replies (*i.e.* part of a conversation with another user). Approximately 12.95% of MEP tweets are geotagged with the user's location information, which is relatively high when we consider that only approximately 1% of tweets by the general public are geotagged (Jurgens et al., 2015).

Fig. 2 shows a plot of the number of tweets per day during this period, where the noticeable troughs correspond to the Parliament's summer break during the month of August. The largest spike in tweeting activity occurred on 8 July 2015, when 6,551 tweets were posted by MEPs. This corresponds to the day that Greek Prime Minister Alexis Tsipras addressed the European Parliament on plans aimed

[8] `https://dev.twitter.com/rest/public`

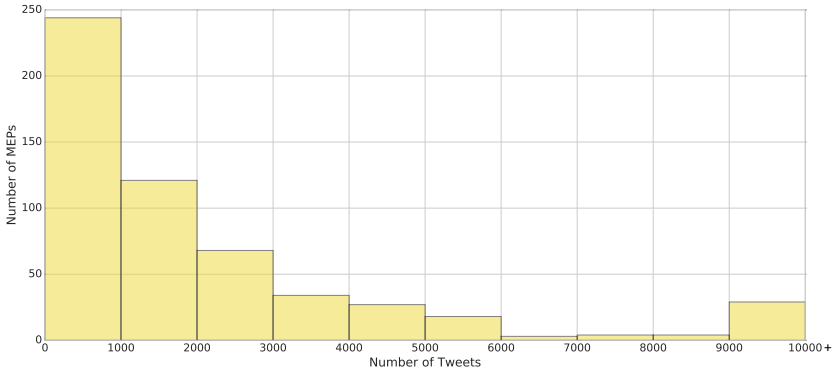

Figure 3: Distribution of number of tweets posted by all MEPs during the period July 2014 – April 2016.

at resolving the Greece's debt crisis. Fig. 3 illustrates the distribution of the number of tweets posted per MEP during the period covered by our study. The mean number of tweets per user is 2093.7, while the median is 1226. A small cohort of 24 MEPs posted over 10,000 tweets during the 34 month period, while 11 MEPs tweeted ten or less times.

5.2 Anglophone Data

Having collected the data, we inspected the language metadata provided by the Twitter API for each tweet. As we might expect, we see that English is the most common language used by MEPs, accounting for 506,742 tweets (39.31%). The next most prevalent languages were French (10.79%), Spanish (10.52%), Italian (9.65%), and German (6.44%). Applying text mining methods to multi-lingual corpora is extremely challenging, as it necessitates pre-processing the documents from each language using a separate, appropriate set of tools, while also ensuring that all languages are treated equally. For the current study, we chose to focus on English-language tweets. When examining the distribution of languages on a

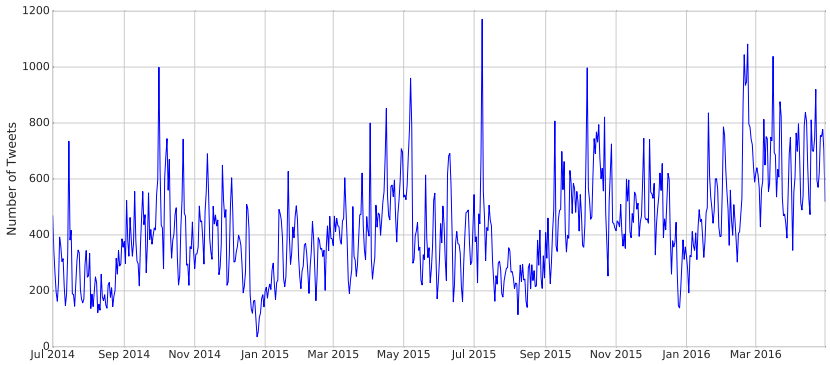

Figure 4: Number of tweets per day in English by MEPs from Anglophone member states during July 2014 – April 2016.

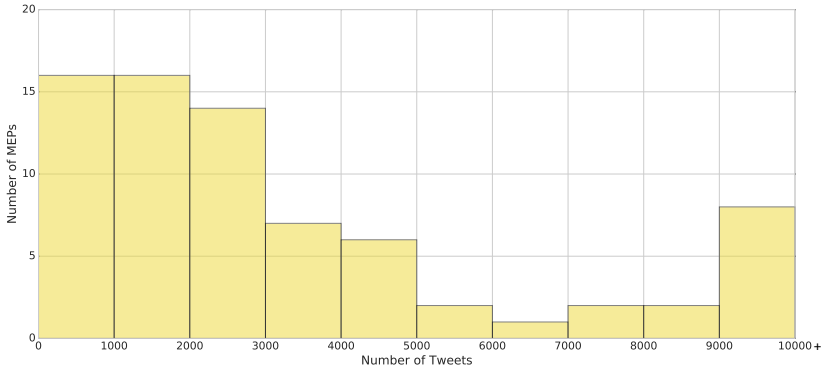

Figure 5: Distribution of number of tweets in English posted by MEPs from Anglophone member states during the period July 2014 – April 2016.

per-country basis, we noted that out of the 28 member states, MEPs from Ireland, the United Kingdom, and Malta tweet predominantly in English.

To provide a coherent case study, we consider the content produced by MEPs from these three states alone. This corresponds to 311,337 tweets posted by 82 MEPs from our complete curated list. We observe that MEPs from these countries occasionally post tweets that are either partially or completely written in their native tongues (*i.e.* Irish, Welsh, and Maltese), but not correctly annotated as such by Twitter. Therefore, we applied a further language filtering process to remove these tweets. This left a total of 285,364 relevant tweets of which 157,881 (55.33%) are

retweets and 35,216 (12.34%) are replies.

As we see from Fig. 4, the tweeting activity over time for this subset of MEPs broadly corresponds to that of the full set (Fig. 2), with similar peaks and troughs. However, Fig. 5 suggests that this subset includes a disproportionate number of frequent tweeters. In particular, four UK-based MEPs (Margot Parker, Julie Ward, David Coburn, Julia Reid) and one Irish MEP (Nessa Childers) posted over 10,000 tweets during the time period considered in this study.

5.3 Data Pre-processing

The full set of English-language Anglophone tweets was pre-processed as follows. Firstly, all links and user mentions were stripped from the tweet text. At this point, the tweets for each MEP for a given week were concatenated into a single weekly "MEP document". The justification for this is that individual tweets are short and often contain very little textual content that is useful from the perspective of topic modelling. However, by combining multiple tweets from the same user into a single, longer document, we can perform topic modelling more effectively. After creating these documents, we apply standard text pre-processing steps:

1. Find all unigram tokens (*i.e.* individual words) in each MEP document, through standard case conversion and string tokenisation. These tokens include both ordinary words and hashtags.
2. Remove single character tokens, emoticons, and tokens corresponding to generic stop words (*e.g.* "are", "the") and Twitter-specific stop words (*e.g.* "rt", "mt").
3. Remove documents containing < 3 tokens.
4. Construct a document-term matrix based on the remaining tokens and docu-

ments. Apply TF-IDF term weighting and document length normalisation. The resulting dataset consisted of a total of 6,445 MEP documents from 96 weeks, represented by 91,163 distinct terms.

5.4 Independent Variables

In order to assess whether a given topic is affected by expected or unexpected events, we manually view topic distributions over time and identify significant changes in MEP attention. In most cases, the event in question is very easy to identify, with events like the Brussels/Paris terrorist attacks and elections/referenda being public knowledge. In situations where a particular spike in MEP attention cannot immediately be identified, we use the topic keywords entered into an internet search engine and consider Tweet documents to identify likely event candidates. In later work we plan to develop a more robust event-detection technique to automate this process.

In order to identify the level of governance to which a given topic is relevant, we hand code topics based on author knowledge of the multi-level governance system of the EU. Once again, in many cases, the relevant level of governance is obvious (Commission appointments are relevant to the EU level; national elections are relevant to the national level; regional elections are relevant to the regional level). In some cases, the relevant level of governance is less clear and are based on judgment calls, having considered the terms describing the topic and a set of associated tweets.

For our MEP-level controls, we take party group and national party memberships from the European Parliament website. We use the individual-level esti-

mates of MEP left-right position and pro-/anti-EU integration position constructed by Barberá et al. (2015) from MEP Twitter network positions.

6 Results

6.1 Overview

We applied the ensemble topic modelling approach presented previously to the Anglophone dataset, consisting of 6,445 MEP documents across 96 weeks. To generate the ensemble, we generate 100 base topic models as described in Section 4.2, each containing $k = 50$ topics. We then integrate these models as described in Section 4.3 to produce an ensemble topic model with $k' = 60$ topics. The results of our topic model are displayed in Table 1.

We manually assign topic labels based on the top-10 most-associated set of terms for each topic. As can be seen in Table 1, there is a large amount of variation in the topics detected, and we can see examples of topics relating to expected events (Brexit referendum - Topics 23, 29, 32, 41, 42, 48) and unexpected events (Brussels and Paris attacks - Topics 17, 46). We can also see examples of topics relating to all levels of the multi-level polity that is the EU. The international level is represented by topics relating to the COP21 agreement and Israel/Palestine (Topics 10, 53). The EU level is represented by a multitude of topics including the Commission and Commission appointments (Topics 8, 37, 59). The national level is the most prominent level addressed by MEPs with much attention dedicated to national elections in the UK and Ireland (Topics 1, 12-16, 18, 26, 30, 36, 39, 40, 47, 52), and to the Brexit referendum (Topics 23, 29, 32, 41, 42, 48). The sub-

national level is also represented with topics relating to the Scottish independence referendum (Topics 4, 38), the London Mayoral election (Topic 18), and the UK regions (Topic 54). In order to explore the results of our topic detection technique in more detail and examine whether our expectation about the impact of events and MEP characteristics on attention to different topics find any support, we begin by exploring a number of case study topics.

6.2 Case Studies

The first of these relates to the day-to-day activity of the European Parliament at the EU-level. Figure 6 outlines the top 30 terms (either individuals words or hashtags) associated with this topic, which help us identify what the topic is about. We can deduce from these terms that they seem to be associated with MEP Tweets about debates, meetings, events and talks - i.e. the day-to-day business of Parliamentarians. Figure 7 plots the attention each MEP pays to this topic over time. Attention is captured as the percentage of weight put on that topic each week. We also plot the mean level of attention for all MEPs to unveil trends over time. What can be seen is that attention to this topic is very stable at about 3-4% of the MEP Twitter output each week. We can also observe small dips in attention to this topic around Christmas time and during the summer recess of Parliament in July/August, which are exactly the times when MEP involvement in these activities are likely to be reduced.

In contrast to the rather constant level of attention paid to this EU-level day-to-day politics topic, unexpected exogenous events like the terrorist attacks in Paris and Brussels are very different in character. Figure 8 and 9 relate to MEP

Topic number	Topic label	Top-10 words associated with topic
0	Day to day politics	meeting forward looking event tonight meet morning speaking visit discussion
1	UK election - UKIP	ukip farage nigel party immigration voteukip election nhs rochester leader
2	Malta	malta metsola maltese roberta libya josephmuscat valletta migrationeu mediterranean gozo
3	Ireland GAA	ireland cork irish kerry northern gaa tourism congrats best delighted
4	Scottish referendum 1	scotland scottish scots salmond edinburgh glasgow independence vote referendum alex
5	TTIP	ttip isds trade public debate services against deal nhs mep
6	Grexit	greece greek tsipras debt syriza grexit euro bailout eurogroup eurozone
7	UK Labour	labour nhs meps tories ed miliband working party workers zero
8	EP hearings	ephearings2014 commissioner designate hearing hogan canete hill vella ephearing2014 moscovici
9	Education	arts manchester nw julie young support children cumbria education rights
10	Israel/Palestine	gaza israel palestine israeli gazaunderattack peace palestinian iraq children ukraine
11	British economy	yorkshire leeds pnr security longtermplan plan economy harrogate britain humber
12	UK election - Greens	green hinkley greens nuclear renewables bristol greenerin party votegreen2015 molly
13	Welsh assembly elections	wales welsh cardiff plaid16 funds rhondda mymep neath efa euro
14	UK Local elections - UKIP	thurrock ukip grays aveley tim aker local crossing council essex
15	Irish election - FG1	irish times dublin gael fine fg election enda kenny piece
16	Irish election - FG2	ge16 leadersdebate fg gael fine election enda right2change kenny recovery
17	Brussels attack	brussels thoughts airport brusselsattacks victims attacks explosions safe metro easter
18	London Mayoral election	london syed4london mayor mayoral candidate conservative syed housing londoners kamall
19	EP vote	vote meps voted against parliament conflictminerals voting votes report strasbourg
20	Refugee crisis	refugees refugeecrisis refugee refugeeswelcome crisis soteu syrian asylum relocation eplenary
21	Agriculture	dairy farmers milk agriculture sector farming agri food farm cap
22	UK steel industry	steel saveoursteel industry action dumping workers uk meps chinese jobs
23	Brexit - Dudley fight	etheridge dudley bill mep ukip video sedgley articles council birmingham
24	Gender equality	women iwd2016 gender equality violence internationalwomensday rights girls pioneers men
25	Thank yous	week thanksall mentions followers reach mention came best audience growing
26	Ireland election - Sinn Fein	fein sinn carthy sf irish dublin matt mep conference monaghan
27	Britian in the world	uk trade economy isis foreign security policy world british eu
28	SNP	snp sturgeon coburn nicola scottish holyrood david oil gay forth
29	Brexit - In 1	intogether wildlife stay britain strongerin keep air pollution lib campaign
30	UK election - Labour 1	ge2015 votelabour labourdoorstep team campaigning vote tory campaign mp manifesto
31	Christmas	christmas merry happy xmas wishing santa flood festive floods message
32	Brexit - In 2	labourinforbritain euref strongerin rights campaign britain johnson remain labourin alan
33	Migration	migration immigration mediterranean migrationeu migrants migrant live net policy states
34	Panama papers	mizzi pn marlene muscat panama busuttil panamapapers cachia comodini konrad
35	MEP - Miriam Dalli	dalli miriam mep autism malta youth libya emissions climatechange unemployment
36	UK election - SNP	votesnp ge15 snp scotland leadersdebate westminster adoption alyn ge2015 mediterranean
37	EU Commission	juncker commission president parliament claude jean strasbourg meps group euro
38	Scottish referendum 2	indyref voteyes labourno lab14 bettertogether salmond ukraine glasgow independence nothanks
39	UK elections - Leaders	cameron david renegotiation britain tory deal immigration tories miliband leadersdebate
40	UK elections - Labour 2	lab15 fringe lab14 speech brighton stand rights speaking meps social
41	Brexit - Out 1	brexit leave voteleave remain deal obama leaveeu vote euref britain
42	Brexit - SNP	bothvotessnp sp16 scotland night euref manifesto announced aye angus seafood
43	Business	funding business businesses deadline innovation apply dorset available growth smes
44	Tax evasion	tax taxjustice avoidance luxleaks evasion dodging pay public panamapapers havens
45	MEP - Luke 'Ming' Flanagan	ming luke flanagan roscommon vinb ireland irish cannabis dail td
46	Paris attacks	paris isis parisattacks police france attacks attack oldham terror terrorism
47	UK Election - Polls	lab con poll yougov ukip ld ldem grn lead tories
48	Brexit - Out 2	leave referendum campaign borders britain control leader corbyn british saynoeutour
49	UK regions	west midlands birmingham mids north oldham wba south royton coventry
50	Global warming	energy climate change gas warming global energyunion memo coal emissions
51	Ibrahim Halawa	ibrahim freeibrahim halawa egypt boylan lynn trial irish egyptian release
52	UK election - Outcome	congratulations conservative delighted candidate election voteconservative conservatives excellent fantastic mp
53	COP21 agreement	cop21 climate paris climatechange change agreement cop21paris emissions parisagreement carbon
54	UK region - North east	north east newcastle trade jude ccs jobs durham latest teesside
55	Water charges	water right2water charges irishwater dublin protest boylan roaming lynn fg
56	Migration crisis	calais migrants migrant crisis britain illegal police border bbc germany
57	Digital single market	digital dsm digitaloptimism market single copyright ft data tech digitalsinglemarket
58	MEP - Amjad Bashir	amjad bashir mep trade bradford ukipconf14 rotherham muslim lincs conservative
59	UK Commission appointment	hill lord clacton financial commissioner trade econ capital heywood markets

Table 1: Top 10 topic descriptors for the 60 topics detected by applying ensemble topic modelling on the Anglophone MEP dataset.

Twitter activity about the terrorist attacks that took place in Brussels on March 22nd 2016. The substantive content of the topic can once again be discerned from the top terms associated with it (Fig. 8). These include the type of attack carried out ("explosions"), and the locations of the attacks ("airport","metro").

Figure 6: Top 30 terms associated with the day-to-day politics topic. The size of the term is proportional to the term's rank in the topic descriptor.

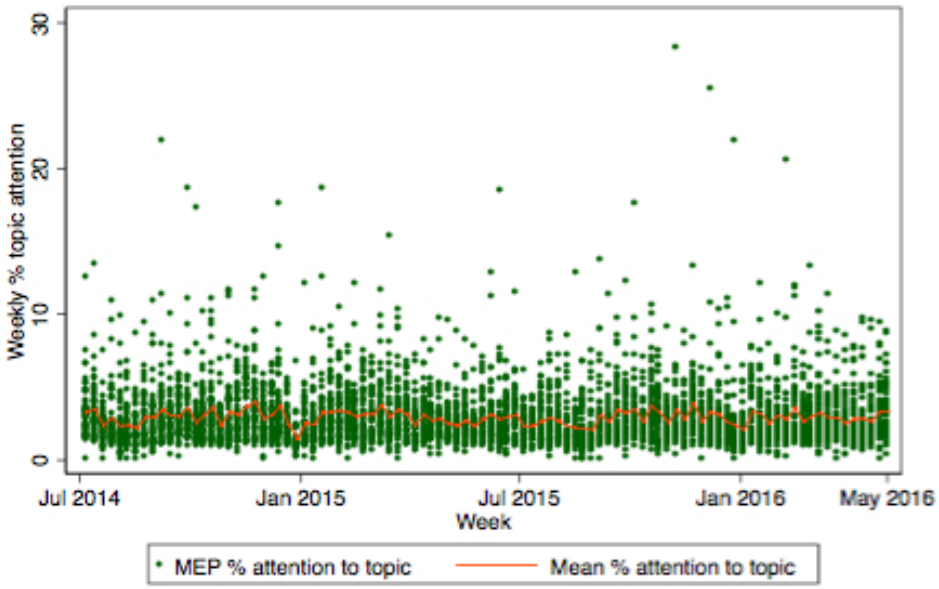

Figure 7: Timeline of topic relating to day-to-day EU business.

We can also discern MEPs using Twitter to express their thoughts for those affected ("thoughts","families"). Looking at Figure 9 in more detail, it clearly has a very different time signature. For almost the entire period close to no MEP

Figure 8: Top 30 terms associated with the Brussels attacks.

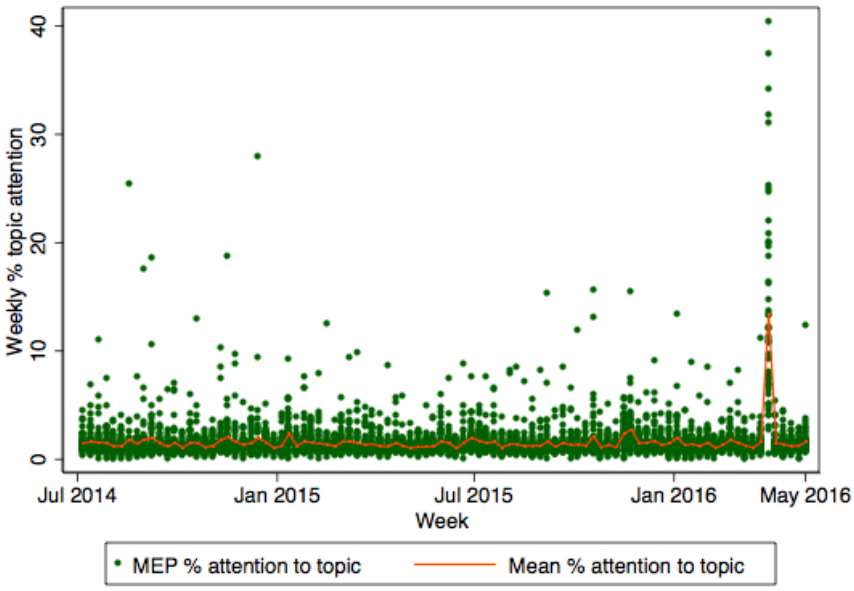

Figure 9: Timeline of topic relating to the Brussels attacks.

attention is paid to the topic because the attacks had not occurred.[9] The spike in MEP attention coincides with the date of the attack and is a result of MEPs

[9]Note that attention to the topic is not zero at any point as terms like 'Brussels', 'airport', and 'parliament' are likely to come up in Tweets relating to other topics over the time period considered.

Figure 10: Top 30 terms associated with the Paris attacks.

directly commenting on it.

In the same time period, Paris was also affected by a series of terrorist attacks. Our topic model unveils a topic relating to MEP Twitter activity commenting on these attacks. In Figure 10, we clearly see this with the top terms relating to this topic including “parisattacks”, “terrorism”, “jesuischarlie”, and “charliehebdo”. When we consider how attention to this topic evolved over time (Figure 11), the first obvious spike in MEP attention coincides with the attacks on the Charlie Hebdo satirical magazine on January 7th 2015. We see a second spike in MEP attention on November 13th 2015 when a coordinated set of attacks happened across Paris. Interestingly, we also see a spike in attention to this topic at the time of the aforementioned Brussels attacks, which was later revealed to be carried out by the same terrorist cell.

Both the Brussels and Paris attacks can be considered unexpected exogenous shocks, which are associated with sudden spikes in MEP attention that quickly dissipate. Scheduled salient events like referenda are expected to be associated

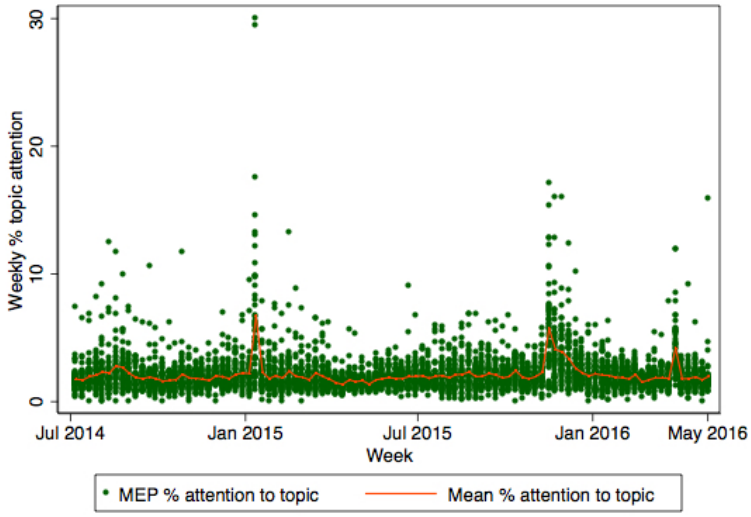

Figure 11: Timeline of topic relating to the Paris attacks.

with very different patterns in MEP attention due to the fact that they are announced in advance. Using the set of terms associated with each topic, we were able to discern 6 topics relating to the UK referendum on EU membership that takes place on June 23rd 2016 (Topics 23, 29, 32, 41, 42, 48). A clear increase in attention to Brexit topics is discernible in Figures 12 and 13, with much more of increase in the 'vote leave' side in Figure 13. While attention to these topics was very low for the first 10 months of out analysis, we begin to see some changes in these patterns around the time the Conservative government in the UK won the general election in May 2015. Holding a referendum on EU membership was one of the campaign promises of the conservative party in that election. After the election win we see a gradual increase in attention to Brexit from both the 'in' and 'out' sides of the debate right up to the end of our time series. Overall, this is in line with our expectation that expected salient political events will garner more and more attention as the date of the event approaches. We aim to update our

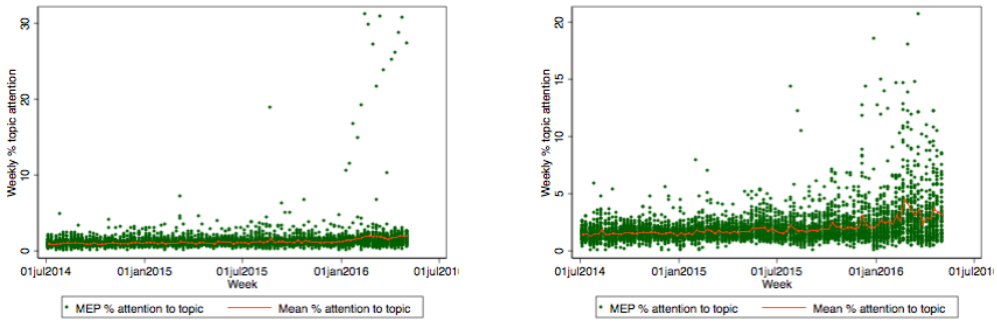

Figure 12: Remain side of Brexit debate (Topics 29 & 32)

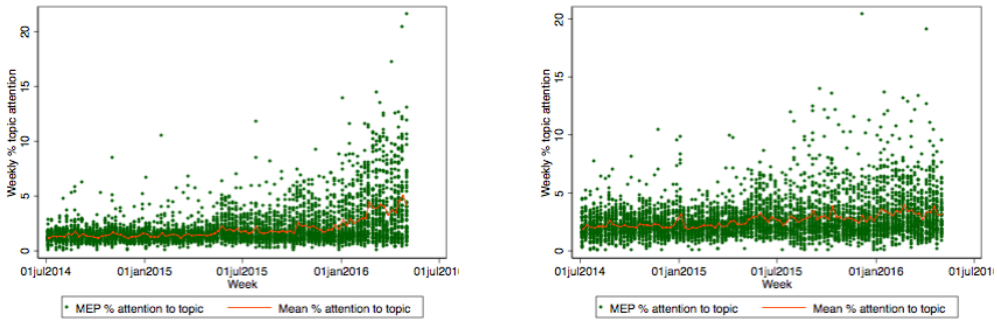

Figure 13: Leave side of Brexit debate (Topics 41 & 48)

results once the outcome of the referendum is known.

7 Analysis

Up until this point we have provided descriptive results of the outputs of our topic model, but the real advantage of this approach to providing a measure of MEP attention to different topics is that it allows us to test explanations of what might cause the variation observed. In order to examine the determinants of MEP attention we consider how MEP party membership and ideology structure the attention paid to different topics.

Figure 14 plots the coefficients of two distinct models of the determinants of

total MEP attention to our detected topics over the entire periods.[10] The dependent variable in these models is a sum of the weight for each MEP for each topic for the entire period considered. The continuous nature of this variable with a right-skewed distribution implies a generalized linear model with a log link is appropriate. Topic-level fixed effects are excluded from the Figures.

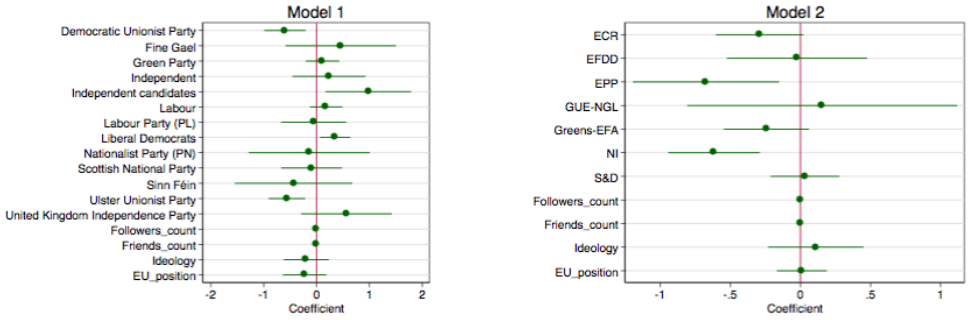

Figure 14: Coefficient plots for Model 1 and 2.

Model 1 explores the determinants of MEP topic attention accounting for MEP party national affiliation, MEP engagement with Twitter, and MEP ideology at the individual level. The Conservative party in the UK is the baseline category of national party affiliation against which other parties are compared. In general, MEPs tend not to contribute to topics more or less than the conservative MEPs, with the exception of independents and Liberal Democrats who tend to contribute more, and Democratic Unionists and Ulster Unionists in Northern Ireland who tend to contribute less. Individual-level ideology and position on EU integration do not have a significant effect on topic contributions, while each additional friend on Twitter leads to a substantively small but significant increase in topic contributions. This is probably due to the fact that those with more friends on the social network are more engaged users who tweet more often.

[10]We plan to undertake a full time series analysis of agenda dynamics in due course.

Model 2 accounts for the same set of MEP characteristics but this time accounts for MEP party group affiliation. This time the baseline party group is the ALDE. We observe very little difference between MEP contributions across the Party groups except for MEPs belonging to the EPP group and the Non-Inscrits, who tend to contribute less to topics overall than MEPs from the ALDE group. Further analysis is required in order to explain why the EPP and Non-Inscrits tend to contribute to topics less than the ALDE.

Once again, the analysis presented here is preliminary. In the next version of the paper we plan to add a significant number of MEP-level control variables to this analysis (committee assignments, gender, age, experience in Parliament etc.).

8 Conclusions

In this paper we have presented our preliminary examination of the content of MEP tweets on Twitter during the 8th European Parliament. We have introduced a new form of topic model based on the concept of ensemble learning, which takes the form of two layers of Non-Negative Matrix Factorisation (NMF). This method can provide a robust and informative account of MEP attention to different issues over time as reflected by their activity on social media. As a case study, we applied this method to a set of weekly English language tweet documents from MEPs from Anglophone countries in the EU (UK, Ireland, Malta), built from a total of over 285k raw tweets. The resulting topics demonstrate how the issues addressed by MEPs through the Twitter platform evolve over time, responding to internal and external stimuli as predicted by punctuated equilibrium theories of agenda dynamics.

We plan to expand the project in a number of directions. The most obvious shortcoming of what we present here is the focus on English language tweets in isolation. It will be necessary to expand our topic models to other European languages in order to provide a more complete account of the dynamics of MEP attention to different issues on Twitter. This will require handling tweets from each language using appropriate pre-processing techniques. In addition, we plan to examine approaches to automatically identify an optimal number of topics for the corpus, rather than relying on manual selection.

To conclude, this paper has demonstrated the usefulness of an ensemble topic modelling approach to unveiling the issues that MEPs tend to tweet about.

Acknowledgments. This research was partly supported by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

References

Aiello, L. M., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., Göker, A., Kompatsiaris, I., and Jaimes, A. (2013). Sensing trending topics in twitter. *IEEE Transactions on Multimedia*, 15(6):1268–1282.

Alexandrova, P., Carammia, M., and Timmermans, A. (2012). Policy punctuations and issue diversity on the european council agenda. *Policy Studies Journal*, 40(1):69–88.

Alexandrova, P., Timmermans, A., Carammia, M., and Princen, S. (2013). Mea-

suring the european council agenda: Introducing a new approach and dataset. *European Union Politics*, page 1465116513509124.

Anstead, N. and O'Loughlin, B. (2015). Social media analysis and public opinion: the 2010 UK General Election. *Journal of Computer-Mediated Communication*, 20(2):204–220.

Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, 23(1):76–91.

Barberá, P., Popa, S., and Schmitt, H. (2015). Prospects of ideological realignment(s) in the 2014 ep elections? analyzing the common multidimensional political space for voters, parties, and legislators in europe. In *MPSA Conference 2015*.

Baumgartner, F. R., Breunig, C., Green-Pedersen, C., Jones, B. D., Mortensen, P. B., Nuytemans, M., and Walgrave, S. (2009). Punctuated equilibrium in comparative perspective. *American Journal of Political Science*, 53(3):603–620.

Baumgartner, F. R. and Jones, B. D. (1991). Agenda dynamics and policy subsystems. *The journal of Politics*, 53(04):1044–1074.

Baumgartner, F. R. and Jones, B. D. (1993). *Agendas and Instability in American Politics*. University of Chicago Press.

Baumgartner, F. R. and Jones, B. D. (2002). *Policy dynamics*. University of Chicago Press.

Becker, H., Iter, D., Naaman, M., and Gravano, L. (2012). Identifying content for planned events across social media sites. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 533–542. ACM.

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

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

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A., and Menczer, F. (2011). Predicting the political alignment of Twitter users. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom)*, pages 192–199.

Cui, A., Zhang, M., Liu, Y., Ma, S., and Zhang, K. (2012). Discover breaking events with popular hashtags in twitter. In *Proc. 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*, pages 1794–1798. ACM.

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*, 41(6):391–407.

Dowding, K., Hindmoor, A., and Martin, A. (2015). The comparative policy agendas project: theory, measurement and findings. *Journal of Public Policy*, pages 1–23.

Ecker, A. (2015). Estimating policy positions using social network data cross-validating position estimates of political parties and individual legislators in the polish parliament. *Social Science Computer Review*.

Fred, A. (2001). Finding consistent clusters in data partitions. In *Proc. 2nd International Workshop on Multiple Classifier Systems (MCS'01)*, volume 2096, pages 309–318. Springer.

Ghaemi, R., Sulaiman, M., Ibrahim, H., and Mustapha, N. (2009). A Survey: Clustering Ensembles Techniques. In *Proceedings of World Academy of Science, Engineering AND Technology*, volume 38, pages 2070–3740.

Gibson, R. K. (2015). Party change, social media and the rise of 'citizen-initiated'campaigning. *Party politics*, 21(2):183–197.

Greene, D., Cagney, G., Krogan, N., and Cunningham, P. (2008). Ensemble Non-negative Matrix Factorization Methods for Clustering Protein-Protein Interactions. *Bioinformatics*, 24(15):1722–1728.

Greene, D. and Cross, J. P. (2015). Exploring the political agenda of the european parliament using a dynamic topic modelling approach. In *5th Annual General Conference of the European Political Science Association (EPSA'15)*.

Java, A., Song, X., Finin, T., and Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. In *Proc. 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65, New York, NY, USA. ACM.

Jennings, W. and John, P. (2009). The dynamics of political attention: public

opinion and the queen's speech in the united kingdom. *American Journal of Political Science*, 53(4):838–854.

John, P. and Bevan, S. (2012). What are policy punctuations? large changes in the legislative agenda of the uk government, 1911–2008. *Policy Studies Journal*, 40(1):89–108.

Jones, B. D. (1994). *Reconceiving decision-making in democratic politics: Attention, choice, and public policy*. University of Chicago Press.

Jones, B. D. and Baumgartner, F. R. (2005). *The politics of attention: How government prioritizes problems*. University of Chicago Press.

Jones, B. D. and Baumgartner, F. R. (2012). From there to here: Punctuated equilibrium to the general punctuation thesis to a theory of government information processing. *Policy Studies Journal*, 40(1):1–20.

Jungherr, A. (2014a). The Role of the Internet in Political Campaigns in Germany. *German Politics*, (ahead-of-print):1–8.

Jungherr, A. (2014b). Twitter in Politics: A Comprehensive Literature Review. *Available at SSRN*.

Jurgens, D., Finethy, T., McCorriston, J., Xu, Y. T., and Ruths, D. (2015). Geolocation prediction in twitter using social networks: a critical analysis and review of current practice. In *Proc. 9th International AAAI Conference on Weblogs and Social Media (ICWSM'15)*.

Kalmeijer, J. (2014). Hashtag clustering to summarize the topics discussed by dutch members of parliament. Unpublished.

King, A., Orlando, F., and Sparks, D. B. (2011). Ideological extremity and primary success: A social network approach. In *MPSA conference*.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a social network or a news media? In *Proc. 19th international conference on World Wide Web (WWW'10)*, pages 591–600. ACM.

Larsson, A. O. (2015). The eu parliament on twitter—assessing the permanent online practices of parliamentarians. *Journal of Information Technology & Politics*, (ahead-of-print):1–18.

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

Lin, C. (2007). Projected gradient methods for non-negative matrix factorization. *Neural Computation*, 19(10):2756–2779.

Lorenzo-Rodríguez, J. and Madariaga, A. G. (2015). Going public with a private profile? analyzing the online strategies of 2014 european parliament election candidates. In *Annual Meeting of the Midwest Political Science Association, Chicago*.

Ma, Z., Sun, A., Yuan, Q., and Cong, G. (2014). Tagging your tweets: A probabilistic modeling of hashtag annotation in twitter. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 999–1008. ACM.

Muntean, C. I., Morar, G. A., and Moldovan, D. (2012). Exploring the meaning

behind twitter hashtags through clustering. In *Business Information Systems Workshops*, pages 231–242. Springer.

Nulty, P., Theocharis, Y., Popa, S. A., Parnet, O., and Benoit, K. (2015). Social media and political communication in the 2014 elections to the european parliament.

Obholzer, L. and Daniel, W. T. (2016). An online electoral connection? how electoral systems condition representatives social media use. *European Union Politics*.

Opitz, D. W. and Shavlik, J. W. (1996). Generating accurate and diverse members of a neural-network ensemble. *Neural Information Processing Systems*, 8:535–541.

Punera, K. and Ghosh, J. (2007). Soft Cluster Ensembles. In *Advances in Fuzzy Clustering and Its Applications*. Wiley.

Shamma, D., Kennedy, L., and Churchill, E. (2009). Tweet the debates: Understanding community annotation of uncollected sources. In *Proceedings of the 1st SIGMM workshop on Social media*, pages 3–10. ACM.

Shi, B., Ifrim, G., and Hurley, N. (2014). Be In The Know: Connecting News Articles to Relevant Twitter Conversations. *arXiv preprint arXiv:1405.3117*.

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

Strandberg, K. (2013). A social media revolution or just a case of history repeating

itself? the use of social media in the 2011 finnish parliamentary elections. *New Media & Society*, page 1461444812470612.

Strehl, A. and Ghosh, J. (2002a). Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617.

Strehl, A. and Ghosh, J. (2002b). Cluster ensembles - a knowledge reuse framework for combining partitionings. In *Proc. Conference on Artificial Intelligence (AAAI'02)*, pages 93–98. AAAI/MIT Press.

Theocharis, Y., Barbera, P., Fazekas, Z., and Popa, S. A. (2015). A bad workman blames his tweets? the consequences of citizens' uncivil twitter use when interacting with party candidates. *The Consequences of Citizens' Uncivil Twitter Use When Interacting with Party Candidates (September 5, 2015)*.

Topchy, A., Jain, A., and Punch, W. (2005). Clustering ensembles: Models of consensus and weak partitions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1866–1881.

Vergeer, M., Hermans, L., and Sams, S. (2013). Online social networks and microblogging in political campaigning the exploration of a new campaign tool and a new campaign style. *Party Politics*, 19(3):477–501.

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*, pages 375–384. ACM.

Wang, Y., Liu, J., Qu, J., Huang, Y., Chen, J., and Feng, X. (2014). Hashtag graph

based topic model for tweet mining. In *Proc. IEEE International Conference on Data Mining (ICDM'2014)*, pages 1025–1030.

Woolley, J. T. (2000). Using media-based data in studies of politics. *American Journal of Political Science*, pages 156–173.

Yang, Y., Carbonell, J. G., Brown, R. D., Pierce, T., Archibald, B. T., and Liu, X. (1999). Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems*, (4):32–43.