COHESIFY
The Impact of EU Cohesion Policy on European Identification

EU Cohesion policy in the media: A computational text analysis of online news, user comments and social media

Juan Miguel Carrascosa, Carlos Mendez and Vasiliki Triga

Research Paper 12

Work package 4 – Task 4.1 & 4.3

Cyprus University of Technology; University of Strathclyde, European Policies Research Centre

The COHESIFY project (February 2016-April 2018) has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 693127

This research paper investigates Cohesion Policy in the mass media by applying computational text analysis to a novel media dataset. Specifically, structural topic modelling and sentiment analysis is applied to online news, user comments and social media at multiple territorial levels. The dataset includes 4,000 news stories, 33,000 user comments, 3,700 posts and 19,500 tweets from Facebook and Twitter respectively, as well as comments and reactions. We discover a two-level hierarchy of descending sentiment on Cohesion policy news stories, whereby international media use more negative sentiment than EU web-native media at one level, and the national media in turn use more negative sentiment than regional level sources at the domestic level. The sentiment of user comments on news articles varies across our country cases, being mainly neutral or positive in Spain and overwhelmingly negative in the United Kingdom even in pro-European news sources. Finally, social media content on Facebook and Twitter is largely neutral, and dominated by official policy channels and stakeholders. We conclude that a territorially-targeted media strategy is needed to improve public appreciation of Cohesion policy, along with more emotive and topical social media activity in order engage and connect with citizens.

TABLE OF CONTENTS

Introduction

The mass media play a critical role in the European Union's (EU) public sphere by reflecting EU debates and informing citizens about the EU. Research shows that the media not only represent news but also redefine and reshape news, which can impact on citizens' attitudes to the EU (de Vreese & Boomgaarten, 2006; de Vreese & Semetko, 2004; Vliegenthart et al 2008), their European identity (Bruter 2003; 2009) and voting behaviour (Banducci and Semetko 2003; Giebler et al. 2018).

Despite the increasing acknowledgement of the significance of the mass media for the European public sphere and EU integration, there is scant research on media coverage of specific EU policies. Exceptions include of the Euro/EMU (de Vreese et al. 2001) and Common Foreign and Security Policy (de Vreese and Kandyla 2009). Furthermore, much of the comparative literature on mainstream news has focused on transnational or national news without paying attention to regional news sources (cf. Perez 2013; Hepp et al 2016). This is surprising given that EU institutions have made it a priority to communicate in partnership with national, regional and local opinion leaders and stakeholders since the 2002 White Paper on European Governance and, more recently, through a range of initiatives to reconnect with citizens locally in the aftermath of the crisis.

EU Cohesion Policy is a highly relevant case for exploring media coverage across time and space given its high visibility, subnational reach and political salience. It is one of the most visible EU policies with direct impacts on people's daily lives through infrastructure investments, business grants and training for people across all regions of the EU. With its pioneering multilevel governance model and partnership principle, Cohesion Policy is credited with encouraging the participation and empowerment of subnational governments in regional policy and the EU decision-making more generally, as well as encouraging local and civic mobilisation at all levels. Finally, Cohesion Policy accounts for a major share of the EU budget (€350 billion out of €1 trillion) and is a classic 'redistributive' policy involving transfers of funding from richer states to poorer states and a high level of politicisation in funding decision-making. This is most salient in the media during the periodic budget negotiations on the EU's multi-annual financial framework, typically pitting net payers against net beneficiaries seeking to maximise their net budget returns, with final decisions often taken in 'marathon' EU summit negotiations by heads of state.

Set against this background, this research paper aims to uncover territorial and temporal patterns in the EU Cohesion Policy coverage and tone in the online news and social media. To do so, structural topic modelling and sentiment analysis is applied to an original Cohesion Policy media dataset at multiple territorial levels. The dataset covers 4,000 Cohesion Policy-related news stories and 33,000 user comments. On social media, the dataset includes more than 3,700 posts and 19,500 tweets from Facebook and Twitter respectively as well as comments and reactions.

The research paper is organised as follows. The next section describes the media monitor system used to collect the data. The methodology sections sets out both the data collection methodology and computational text analysis techniques used for analysis the media distinguishing topic modelling and sentiment analysis. The results are then presented for each media type: news media, user comments and social media. The conclusion presents the key findings and policy implications.

Methodology and data

In this section, we describe our media monitor system and data collection methodology for gathering (i) thousands of articles across several web media sources (ii) the User-Generated-Content (i.e. users' comments) associated to these set of articles and, (iii) the information from Online Social Networks (OSNs) where we focus on Facebook and Twitter given their dominance in the social media ecosystem. The steps include the selection of keywords and sources and the formulation of an *EU Relevant Metric* to identify relevant articles and filter out *noisy* articles. After presenting the media monitor datasets that we will use for our analysis, we set out the computational text analysis techniques employed, comprising structural topic modelling and sentiment analysis.

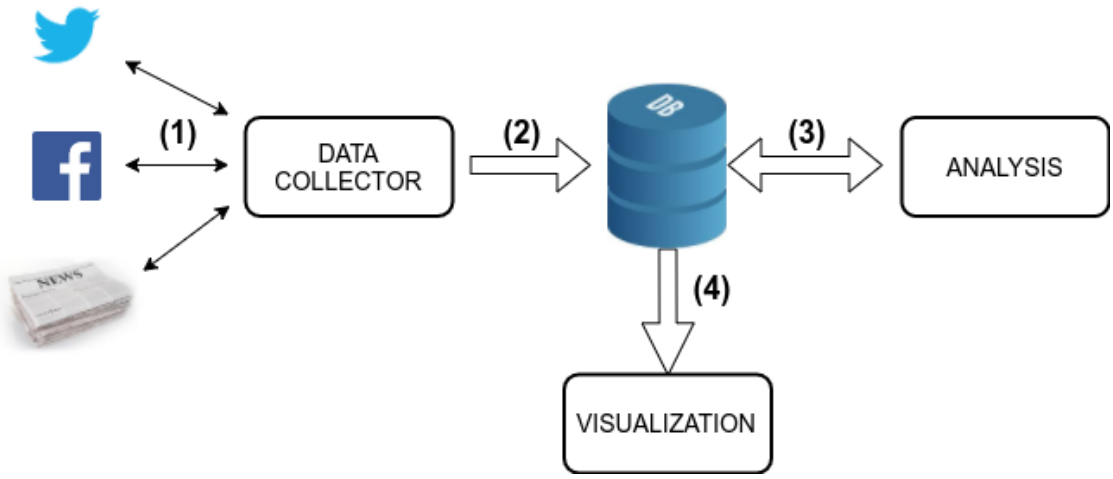
The Media Monitor System

The high level architecture of the media monitor system is depicted in Figure 1. The system consists of four modules: the data collector, a centralized database and the analysis and visualization modules. These modules are detailed below.

Data collector

This component is responsible for collecting web media and social media data and stores all the information in a database. It takes as inputs a set of keywords and a set of sources (i.e. webpages from mass media) and, with the help of a task manager, executes different crawlers depending on the target. Specifically, the data collector is composed of four different monitors/crawlers:[1] Facebook Monitor, Twitter Monitor and a Web Media Crawler - which is also used for collecting user-generated-content. Figure 2 visualizes the composition of this module.

Figure 1: Architecture overview of the media monitor system



[1] Technically it is possible to use either term: monitor or crawler. However, in this research paper we refer to "monitor" when we use an external API to collect data while we use the term "crawler" when we have implemented our own spider for data collection.

Database

We use a MongoDB database to store all the information collected by the previous module. Since the data structure needs of each crawler are completely different, each crawler from the data collector stores the information in a different table. Importantly, the centralized database also stores the processed data from the analysis module, which is used for evaluation and visualization purposes.

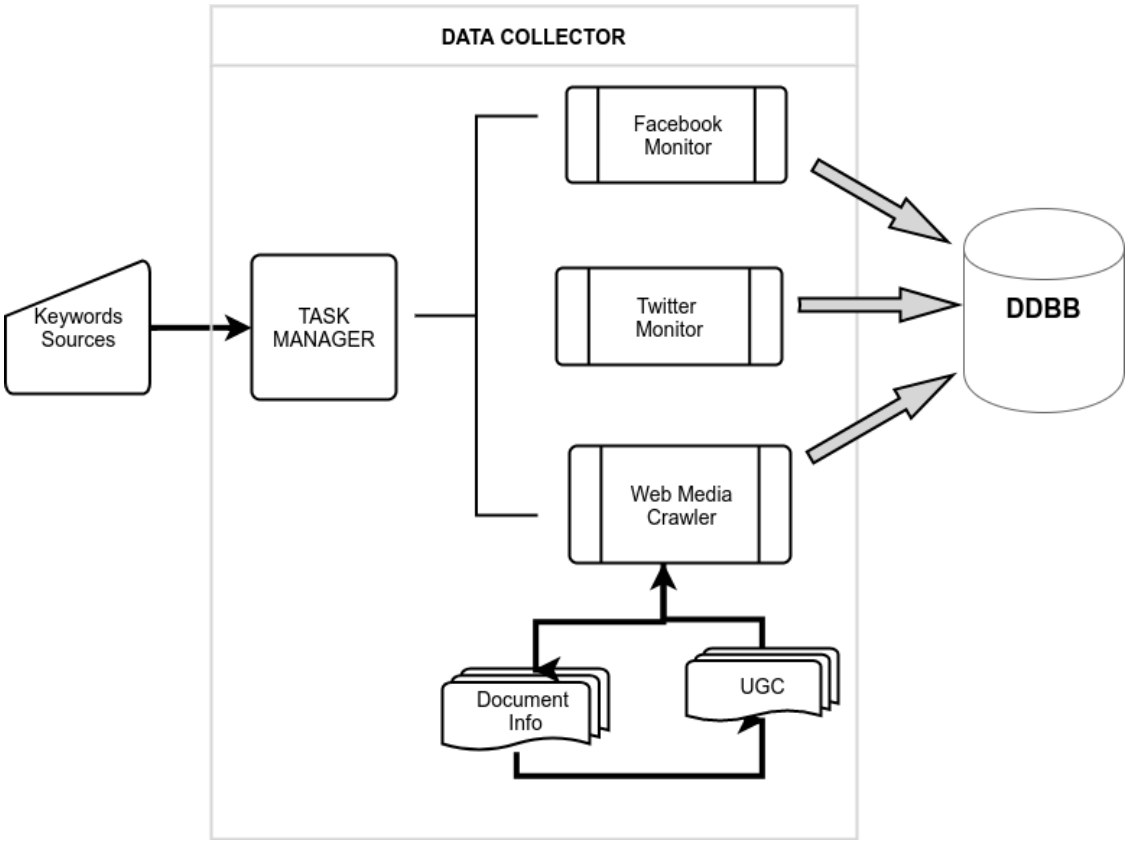
Analysis module

This module is in charge of all the computational text analysis over the information stored in the database. The specific details are explained in more detail below where we cover aspects such as the data pre-processing phase, topic modelling detection and the sentiment analysis approach to evaluate the polarity (i.e. positive, negative or neutral) of our corpus.

Visualisation module

An essential part of the media monitor system is to provide a visualization tool that allows the analyst to explore and interact with the data. This is especially important for the topic modelling where human inspection of topics and their interpretation is crucial. In addition, the system must provide interactive visualization of the results obtained from the Analysis module. Specifically, this requirement involves interacting with the database and the implementation of a completely reactive environment for the analyst. To that end, we designed a web application developed in R (i.e. Shiny) to facilitate the analysts' interaction with the data and processed results. To give but two very simple examples, a reactive environment allows the analyst to interactively alter variable thresholds (for instance, the polarity when calculating sentiment) or the relevance assigned to a topic (when inspecting the results of topic modelling). Facilitating this type of interaction with the data and ability to check how robust findings are when altering key parameters is absolutely crucial to the evaluation and interpretation task.

Figure 2: Data Collector



Data collection

Selection of keywords and sources

From the perspective of mass media, the EU's Cohesion Policy is not a common topic such as, say, Sport or the Economy. For instance, a web media article covering the construction of a new museum in a city receiving EU funds might use labels such as: locality, public works and the purpose of the museum. They are unlikely to use labels related to how the museum was funded. For this reason, searches using the labels from the mass media would be of little value and more likely to hinder the search for relevant articles. Because of this, we adopted a keyword search approach. This involved defining a harmonised set of keywords that could be used in multiple languages to aid us in the search for relevant articles across the vastness of the Internet.

With the help of a group of experts in Cohesion Policy we identified a group of keywords to cover as many cases as possible without overloading the system or including *false friends* across languages. Table 1 summarizes the list of words used. It should be noted that for each word (or group of words) we have used the different endings of the words too (e.g. fund, funds or funding).

Table 1: List of keywords

English	Spanish
Cohesion Policy	Política de Cohesión
Cohesion Fund	Fondo de Cohesión
European Social Fund	Fondo Social Europeo
European Regional Development Fund	Fondo Europe de Desarollo Regional
Structural Funds	Fondos Europeos
European Structural and Investment Funds	Fondos Estructurales y de Inversión Europeos
EU regional policy	Política Regional UE
European Union regional policy	Política Regional de la Unión Europea
European Funds	Fondos Europeos
Interreg	Interreg

In a similar way to the selection of keywords, a list of mass media were identified by country experts as sources. We strove to make sure that the initial media list varied in terms of ideological profile -i.e., the inclusion of both right and left leaning sources, as well as media sources that were rooted in different territorial levels, i.e., mostly national audiences or regional audiences. Thus, we include both left and right leaning media sources at the national level (e.g. The Telegraph and The Guardian for the UK, and El País and El Mundo for Spain) as well regional sources (e.g. The Scotsman or La Voz de Galicia) for each case. In addition, we have a transnational category of sources whose audience is more international in scope. Here we distinguish between a more international focus and a more EU-focused set of sources. Table 2 shows the complete list of news media grouped by level (i.e. Transnational,[2] United Kingdom and Spain).

[2] We refer as Transnational those web media sources related to Europe or with a more international audience.

Table 2: List of sources

	International	United Kingdom	Spain
Inter/National	The Economist	The Times	El Pais
	Reuters	The Telegraph	ABC
	Politico	The Independent	El Mundo
	Huffington Post	The Guardian	La Vanguardia
	Financial Times	The Daily Mail	
Regional/European	EurActiv	WalesOnline	El Periodico
	EU Observer	South Wales Argus	El Correo Bilbao
	Bloomberg	The Irish News	Diario Vasco
		The Scotsman	Diario Navarra
		Herald Scotland	Diario Sur
			La Voz de Galicia
			La Nueva España
			La Verdad
			Hoy (Extremadura)

Collecting Data

- Web Media:

To collect the articles from several sources and keywords we have implemented an automated tool to make queries to Google Search News from the beginning of 2007 up to end of 2017. The output of these queries is a list of URLs from different sources. In the next step, the data collector automatically visits every URL and processes the html storing the relevant meta-data from the given webpage. Note that some meta-data, such as the main content of an article, requires its own heuristic for identifying the body content and removing the non-related one. The specific dates of data collection along the number of articles included in each of the use cases are shown in Table 5.

The search for news through the use of keywords can generate a number of problems including duplication problems and very low volume of articles for some versions of the digital newspapers, which contain little or no information relevant to our case. For these reasons, a filter is applied to eliminate duplicate documents and those documents whose web source contains a reduced number of articles.

- User-Generated-Content:

Most web media sources these days provide additional functionalities that allow readers to interact with the content of the webpage. Some of this interaction can take the form of sharing an article in an online social networks, assigning likes to the content, or to comment on the content and, therefore, start a conversation with other users. We are

interested in this latter type of interaction. We refer to this as User-Generated-Content (UGC) in web media, i.e. the discussion generated by readers' comments for a given article.

Obtaining UGC poses a number of technical challenges in view of the diverse way in which this is handled by different media sources. Furthermore, it is important to note that the majority of articles contain no comments. Thus, to make the data collection feasible and to make more valid inferences, we limit ourselves to media sources that have a reasonable volume of UGC. For instance, it would make little sense to collect two comments from a media source that has only three articles in our corpus. We would be able to infer very little from such data.

- Social Media:

Here we describe the steps for the two OSNs we focus on:

Facebook: allows for data collection via an API. While the Facebook API offers a functionality for collecting all the post associated to a Facebook Page, its API cannot be used to search for posts containing a specific keyword or set of keywords. Therefore, we had to manually search for and select a list of Facebook Pages for the two languages, English and Spanish, which appeared relevant to Cohesion Policy (see Table 3). We could then make the corresponding queries to Facebook in order to get the posts from these pages. To do that, we created a Facebook application and, with the help of a python code, we automated the process of collecting all the posts from the given list. Once we have the posts, we need to remove, or rather, to identify only the posts which are about Cohesion Policy. In addition, due to the characteristics of Facebook, posts can contain a link to a video or a photo, which cannot be analysed in our computational text analysis. Therefore it is necessary to process all the posts and filter out those non-relevant for our analysis. To achieve this goal, we consider that a post is relevant if contains at least one of the keywords. Finally, for those remaining posts, we obtain additional meta-information (i.e. comments, published date, author) and reactions (i.e. number of likes or shares) related to the posts and for every comment in the post. Table 4 shows the different fields we collect from the original post and the comments in a post.

In sum, the steps taken were as follows:

1. Collect Posts (using Facebook API + python for automatization).
2. Filter non-relevant posts.
3. Collect information and reactions associated to the posts.

Table 3: List of Facebook pages

English		Spanish
SocialEurope	TheEconomist	FondosEuropeosEnAndalucia
EUobserver	EuropeanYouthEU	FondosEuropeosEnCastillayLeon
EuropeanCommsion	ScientistForEU	FondosEuropeosParaLaCultura
Euractiv	EUTogether	FondosEstructuralesCastillaLaMancha
EUinmyRegion	InterregCE	FondoEuropeoDeBecas
InterregEurope		

Table 4: Facebook information from posts

Comments	Message	Author	Sum of reactions
Reactions	Number of comments	Number of shares	Number of likes
	Number of loves	Number of wows	Number of hahas
	Number of sads	Number of angrys	Number of special

Twitter: provides several APIs to access the information in the OSN. Our media monitor system makes use of the REST API to collect all the tweets for the list of keywords defined in Subsection and the hashtags associated to them. We query the Twitter API from a Twitter application that we created. Unfortunately, Twitter imposes a maximum date limit to collect tweets from the past. This limitation reduces our potential dataset for longitudinal analyses since we only collect tweets over the past 6 months. Several fields are returned by Twitter in each query. For simplicity, we only mention those relevant for our analysis:

1. Text: message for the tweet
2. retweet_count: number of retweets
3. created_at: published date

Measuring the relevance of an article

Given the huge volume of articles in the web, a key challenge for our methodology is to define a meaningful and easy to understand and measure metric for helping us to identify relevant articles for our study and filter out *noisy* ones. The relevance of an article is basically defined by the number of instances of relevant keywords in a given text. The idea behind this metric is to give articles a weight that allow us to classify them as highly related to Cohesion Policy or with little significance. At the same time, it helps us to eliminate

possible articles that have sneaked into our crawler because of a fail in the Google Search.[3] In these cases, the value of our metric will be equal to zero.

It is important to underscore why this metric is important for analysing the corpus of data collected. The keyword search can generate lots of non-relevant articles. Indeed, we found that this also occurred when using other third-party software, such as LexisNexis. The high number of non-relevant articles generated by third-party software is precisely the reason why we developed our own crawlers for data gathering. By calibrating our EU relevance metric we can achieve a higher degree of precision -that is we can define the threshold so as to increase precision (the probability that a given article is relevant) even at the expense of filtering out some potentially relevant articles. In other words, it is more important for us that the corpus is relevant to the topic of Cohesion Policy even if we exclude some relevant articles by having a high threshold.

Let us first introduce some definitions used in our metrics:

(i) We define the intersection of keywords and content for an article as:

$$Intersection_{Keywords}$$

(ii) For the content of an article and after removing the stop words (i.e. common words, prepositions) we define the length of this content-filtered as:

$$Length_{contentFiltered}$$

(iii) Finally, we define the total number of instances that the word Europe shows up in body content of the articles as:

$$EU_{count}$$

Using these definitions we formally expressed our EU Metric Value as follows:

$$EU_{Metric} = \frac{Intersection_{Keywords}}{Length_{contentFiltered}} * EU_{count}$$

Dataset Summary

In this subsection we provide an overview of the data collected by our media monitor system to help the reader get an idea of the universe of the topic. Table 5 summarizes the number of articles with a EU Metric value greater than zero, the UGC as comments from the users from the media sources with highest number of articles and also, the info from social media as tweets from Twitter and posts from Facebook. In addition to the volume in each dataset, we also include the reactions as the sum of comments, likes or shares - to name some of the meta-information included. Finally, to provide an idea of the community talking and reacting to this topic, the table shows the number of unique users in each dataset.

[3] Google Search returns results based on the whole content of the webpage. This means that if the keyword is present in the webpage but not in the body content (e.g. at the bottom as part of the signature of the source or in a sidebar linking to a different article) it would be collected by our system.

Table 5: Dataset overview

	Volume	Stats	Unique Users
Web Media - Articles	4,092	33,183	N/A
Web Media - UGC	33,183	N/A	7,945
Social Media - Facebook	3,601	60,132	2,321
Social Media - Twitter	19,653	37,886	13,298

Limitations

Social media and web media have some limitations that are commonly attributed to API limits or time windows. In the previous sections some of these limitations have been briefly mentioned for the different cases. In this subsection we list some of the problems and limitations of the data collection:

- Google Search News: the main problem is the bias towards recent articles. Google tends to report newest article first and for queries from a specific interval (e.g. only articles from 2006) the output is quite low. This issue is not necessarily a limitation itself since many newspapers (e.g., smaller regional sources) did not have a digital version many years ago. On the other hand, Google policy is to return only a subset of relevant articles. We are aware of this possible bias in our analysis and plan to implement individual crawlers for several sources to avoid this time limitation and increase the volume of articles in the dataset.

- User-Generated-Content: each webpage is developed using different technologies and structure. It implies that a single crawler cannot be used to collect the UGC from all the articles. For example, it is quite often to have the discussion of the users injected in a separate iframe inside the webpage or even in a different webpage. Also, the way the comments are shown differs in terms of structure or options. How webpages manage conversations from hundreds of users and comments is a key challenge when our system needs to collect all the information automatically. In addition, some webpages require users to be logged in to get access to the comments. All these limitations increase the technical complexity for our system and therefore for reasons of efficiency and feasibility we decided to gather UGC data for only for those webpages with the highest number of articles in our database.

- Social Media: despite the availability of APIs to extract user data from OSNs, the use of this service is commonly limited in time. This is due to the nature of how OSNs store and provide access to their data. For example, in Twitter free-queries imply only a subset of the available tweets (i.e. there is a paid service) for historic data. In the case of Facebook, although time is not a limitation since it is possible to collect all the posts from a Facebook page, the limitation is in terms of flexibility because the API does not allow for keyword searches.

Computational Text Analysis

In this section, we show the main functionalities of our analysis module. The topic modelling analysis whose objective is to extract the most significant topics from a set of documents and (ii) the sentiment analysis to analyze the opinion in the articles from web media, the comments from users and also the posts and tweets from Facebook and Twitter.

Topic Modelling

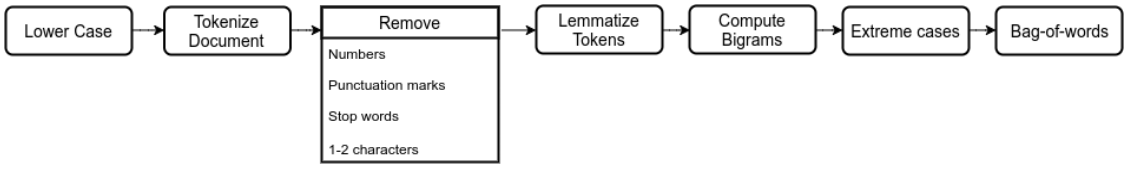
In the fields of machine learning and natural language processing (NLP), the analysis and modelling of documents for the extraction of the most relevant topics has increased dramatically in the last years. Latent Dirichlet Allocation (LDA) (D. M. Blei, Ng, and Jordan 2003) is a generative probabilistic model that is widely used for topic modelling. The basic idea behind LDA is that documents are represented as mixtures of topics, where each topic is characterized by words with certain probabilities.

The key concepts and terms of topic modelling are:

- LDA: technique that automatically discovers topics from documents in a corpus.
- Corpus: contains several documents.
- Documents: contains multiple topics in different proportions.
- Topics: clusters of similar words.

Next we describe in detail the operations undertaken during each phase to run our LDA model.

Figure 3: Pre-processing steps



- Pre-processing:

The pre-processing phase is a key step in any computational text analysis. In our case, this phase is divided in several steps shown in Figure 3: Pre-processing steps. To help the reader gain a better understanding of these steps, we use an example text that is part of an original document in our database. Let us define as our document the following raw data:

> The European Commission has suspended more than £45 million of payments intended for Scotland after it identified irregularities in spending. The money from the European Social Fund, intended for projects across the country, has now been frozen by Brussels until the Commission is satisfied that the Holyrood administration has resolved the problems.

In the first and second step we transform this raw data into lower case to homogenize every words and then tokenize the document:

['the', 'european', 'commission', 'has', 'suspended', 'more', 'than', '£', '45', 'million', 'of', 'payments', 'intended', 'for', 'scotland', 'after', 'it', 'identified', 'irregularities', 'in', 'spending', 'the', 'money', 'from', 'the', 'european', 'social', 'fund', 'intended', 'for', 'projects', 'across', 'the', 'country', 'has', 'now', 'been', 'frozen', 'by', 'brussels', 'until', 'the', 'commission', 'is', 'satisfied', 'that', 'the', 'holyrood', 'administration', 'has', 'resolved', 'the', 'problems']

Third, it is time to remove tokens that do not tend to be useful in the model. For example, remove numeric tokens, punctuation marks and tokens that are only one or two characters. Also, in this step we leverage well-known stop words dictionary for English and Spanish to reduce even more our tokenize document. This stop word list contains, among others, common words such as prepositions: and, the, but, etc. In this step we have reduced our tokenize document from 54 to 26 tokens but keeping all the meaning of the document.

['european', 'commission', 'suspended', 'million', 'payments', 'intended', 'scotland', 'identified', 'irregularities', 'spending', 'money', 'european', 'social', 'fund', 'intended', 'projects', 'across', 'country', 'frozen', 'brussels', 'commission', 'satisfied', 'holyrood', 'administration', 'resolved', 'problems']

In the fourth step we use the WordNet lemmatizer from NLTK. A lemmatizer is preferred over a stemmer in this case because it produces more readable words that are easy to read and understand and this is very desirable in topic modelling.

Then, we compute the bigrams for the document. Bigrams are sets of two adjacent words. For example, if the words "European" and "Union" appear together frequently, it means they are probably a bigram: "European_Union". Note that this probability cannot be measured using a single document, so this step is applied over the whole corpus (i.e. document collection). The output of these steps is shown in the next box:

['european_commission', 'suspended', 'million', 'payment', 'intended', 'scotland', 'identified', 'irregularity', 'spending', 'money', 'european_social', 'fund', 'intended', 'project', 'across', 'country', 'frozen', 'brussels', 'commission', 'satisfied', 'holyrood', 'administration', 'resolved', 'problem']

From the previous frame we can observe the effect of the lemmatizer in several cases: "irregularities -> irregularity" or "problems -> problem". Further, two bigrams show up in our example: "european_commission" and "european_social".[4]

Later, we apply an extra step called "Extreme Cases" where we remove rare words and common words. This is done based on the frequency of the word from the documents. This filter is applied because words that appear in most documents (i.e. more than 60% of documents) do not contribute to the topic model with any additional information, since the word is present everywhere. The same occurs with words in few documents (i.e. 10 documents) where the model cannot learn about these extreme cases.

Finally, we create a dictionary representation of the documents and convert these pre-process tokenized documents into a document-term matrix (i.e. bag-of-words, vectorised representation of the documents) to work with.

[4] Our media monitor system works only with bigrams but we plan to update it to include trigrams too. For instance, one probably trigram in our example would be "european_social_fund" instead of a bigram "european_social" and a unigram "fund"

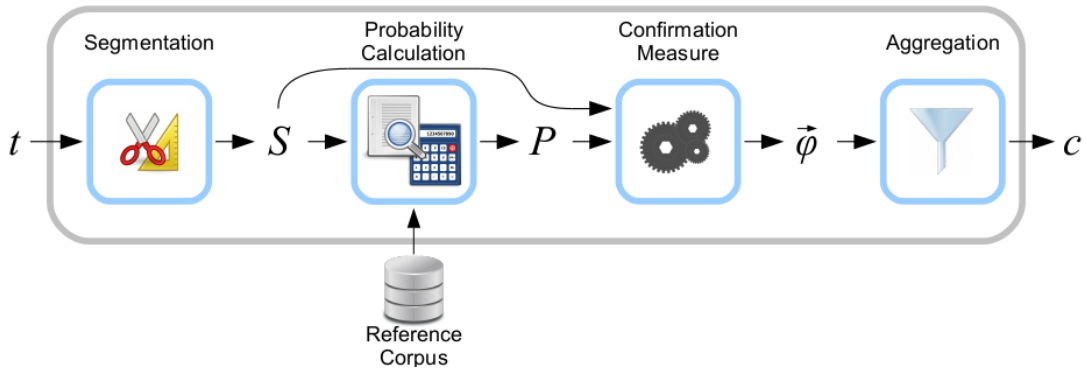
- Training:

Once we have pre-processed all the documents in our corpus and created our dictionary and document-term matrix, is time to start training our LDA model. One of the main parameters in LDA model is the number of topics we are looking for. But, how many significant topics are there in our collection of documents? This value depends directly on the nature of our corpus and our goal in topic modelling. In some cases, we are interested in high-level topics (e.g. economy, sports, etc.) but in other cases we could be searching for more specific topics. Ultimately, the decision of selecting K, where K refers to the number of topics in our model, depends on our research objectives. Our system includes an automatic tool to compare and help us to choose the optimal number of topics, topic coherence, defined below.

- Topic coherence:

Topic coherence refers to the optimization of number of topics in our LDA model (i.e. coherence metric-plot). To aid optimal selection of Topics, for human interpretability, we leverage the topic coherence pipeline from Röder, Both, and Hinneburg (2015) (see Figure 4) which is implemented in Python's gensim package. This pipeline needs to select a coherence measure among several options. To choose the best coherence measure we run the model using the most common ones: u_mass and c_v. We then manually inspect the results in order to select the best output in terms of human interpretability. More details about the impact of these coherence measures can be found in Röder, Both, and Hinneburg (2015).

Figure 4: Overview of the topic coherence pipeline



- Structural topic modelling of social phenomena One increasingly popular variant of LDA, developed by social scientists, is Structural Topic Models (STM) (Roberts et al. 2014). It draws on all the features of LDA, but also allows the analysts to incorporate contextual variables during the model fitting process. This is an important feature for social scientists, since analysts investigating social phenomena are generally interested group effects, such as ideology (e.g. left-wing versus right-wing), size (e.g. small country versus large country), level (e.g. national versus subnational), resources (rich versus poor) -to name but a few contextual variables that have a grouping element. STM allows us to preserve contextual meta-information for subsequent analysis of estimated effects of such grouping variables.

In terms of our empirical analysis, we only apply the STM-variant of LDA to our news media corpus. This makes sense since we have meta-information about the sources, e.g., the country or level in which the media source is rooted, the sources popularity, its ideological leanings -if they have any, etc.. Such contextual information can therefore be incorporated, where relevant, during the model fitting process. This is critical when trying to make inferences about estimated effects from changing from one group to another, for instance, from the regional to the national level. Indeed, given that our subject of analysis relates to Cohesion Policy, this grouping variable is deemed of crucial importance.

However, for our analysis of social media and user-generated content, this makes less sense since we have little meta-information about the sources, i.e., the individuals posting messages or tweeting. To that end, we apply the standard LDA and take advantage of its better optimisation for the task at hand.

In terms of the analytical steps in STM are the same as LDA. Essentially, there is a need for a manual human inspection of the outputs. Grimmer and Stewart refer to this as the need to "validate, validate, validate" the outputs (Grimmer and Stewart 2013). The validation process involves various iterations of human inspection of the topics and documents. When the corpus is large, as in our case, the visualization module is a crucial component for facilitating human inspection.

Sentiment Analysis

Sentiment analysis (also known as opinion mining) leverages techniques from natural language processing, text analysis or similar computational linguistic to determine the polarity (positive, negative or neutral) for a text. Typically, most sentences in a document do not express any opinion and are considered neutral (i.e. objective). On the other hand, sentences that are subjective promulgate an opinion. This opinion (or sentiment) can be positive or negative depending on the polarity of the words used. Sentiment analysis can be conducted at various different levels: (i) Document-level (articles from web media); (ii) Paragraph-level; (iii) Sentence-level (Facebook posts, tweets or user's comments) or even (iv) at the Word-level.

- Background:

According to Musto, Semeraro, and Polignano (2014) the state-of-the-art for sentiment analysis can be broadly classified into two categories: (i) supervised approaches based on a classification model using a set of labeled data and (ii) unsupervised (or lexicon-based) approaches that infer the sentiment of a text combining the polarity of the words (or the phrases) which compose it.

For this research paper we do not have the scope compare both approaches. To begin with a supervised approach would have required creating a training dataset based on human coding of texts. Thus, we will focus our discussion on the main advantages and disadvantages from a point of view of viability and time-efficiency of our approach.

One of the main differences between both approaches is that unsupervised methods do not require any ground truth for training the classification model. The ground truth would imply coding hundreds of articles in a variety of European languages. Although our system is

presently working for Spanish and English we plan to extend it to cover more countries/languages. So, getting a ground truth for most languages would be costly in terms of human resources and time. Note also that we have a plurality of sources, from lengthy web articles, to comments, Facebook posts and Tweets, which would make the coding/annotating of texts even more multi-layered and time consuming. Since unsupervised methods rely on external lexical resources (i.e. dictionary) containing words and its polarization (positive, negative, neutral) or numerical sentiment score, they can be more easily deployed to our different types of data.

- Our approach:

Based on the strengths and weaknesses discussed in the previous section we decided to include in our media monitor system a lexicon-based approach for conducting the sentiment analysis due in large part to its simplicity. In addition, to increase the effectiveness of our method we include a rule-based approach to manage negation words, idioms, intensification or emoticons. The sentiment score goes from -1 most extreme negative; +1 most extreme positive.

- Limitations:

As we have discussed above, the main limitations of our approach is the need of a good dictionary for each language to compute polarity. In short, sentiment analysis is well-developed for the English language, much less so for other languages. This limitation can be reduced with the help of a translator from the input language to English. However, this could be hampered by varying translator performance across languages. More generally, we plan to improve the media monitor system by including exploring additional approaches to the sentiment analysis.

Results

In this section, we apply the methods and modelling strategies explained above to explore the results across different cases and using different types of media sources. To facilitate interpretability we have split the results into three main sources: (i) News Media and (ii) User-Generated-Content. For these two sources of data we further disaggregate by looking at three cases: the Transnational level, the United Kingdom and Spain. For our third media source, (iii) Social Media, the grouping can only be done at the level of language and we correspondingly analyze Facebook and Twitter in English and Spanish.

News Media

In this section we present the main findings of applying Structural Topic Modelling (stm) to the English and Spanish news media. Note that the STM is a form of LDA modelling that allows us to preserve meta-data, i.e. grouping variables of interest such as media source, or territorial level, when running our models and estimating effects.

Since one of our concern with variations across media sources we apply a first filter that removes media sources with too few articles (<20 documents per media source) for the

subsequent modelling process. All in all, the analysis covers a corpus of approximately 4,000 articles in English and Spanish.

In terms of disaggregating per language, for English we run separate analyses:

(i) Transnational media sources (where we distinguish between EU web natives, such as Euobserver, and international press, such as The Economist).
(ii) UK national media sources (such as The Guardian and the Daily Mail) and UK regional media sources (such as The Scotsman).

For the Spanish language, we do not have a transnational level since we only include media sources from Spain. Thus, our breakdown is as follows:

(iii) Spanish national media sources (such as El Mundo and El País) and Spanish regional media sources (such as La Voz de Galicia).

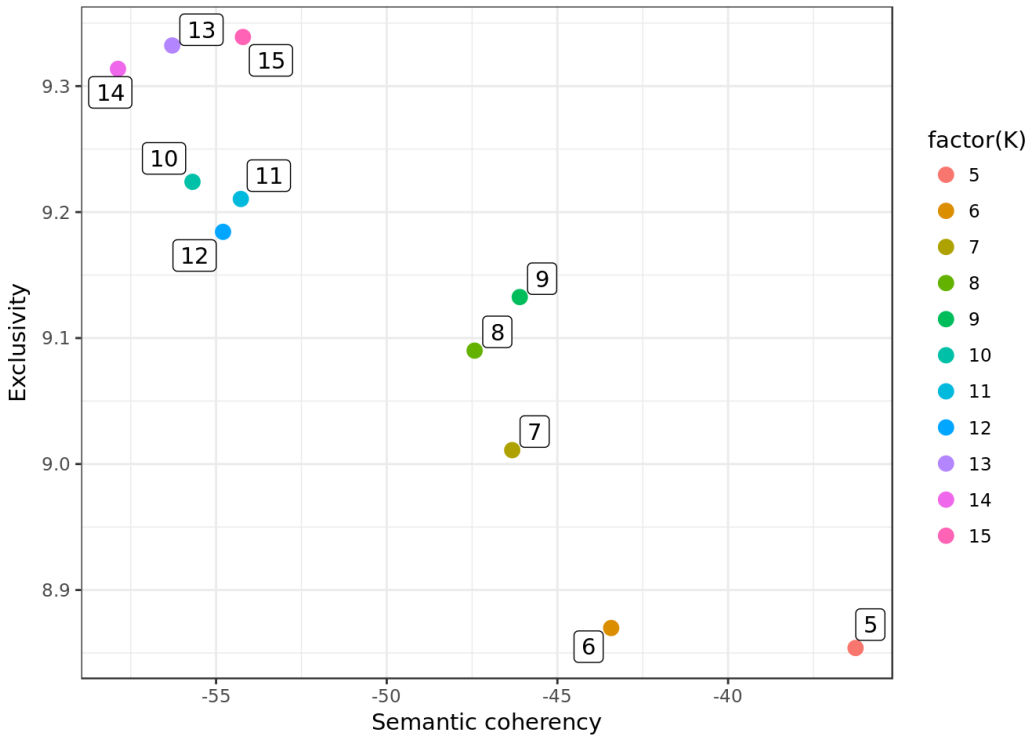
For each of the three cases we follow a similar approach. We begin with the (i) search for the optimal K topics (i.e., selecting the most appropriate number of topics); (ii) we then inspect the key words that are associated with each topic using different metrics and criteria. This step allows us to assign short labels to each topic to facilitate interpretability. Note, that we have also manually inspected the articles associated with the topics as a check on label assignment (iii) we show the estimated proportions for each topic; (iv) we estimate the effect of how the topics are treated by varying the level of territoriality (e.g. regional versus national) and lastly (v) we show the evolution of topics proportions over time.

We now look at each of the three categories in turn.

Transnational

We begin by running a search for the most appropriate number of topics for running the STM. We select models that maximise semantic coherency and exclusivity.

Figure 5 Search for K topics

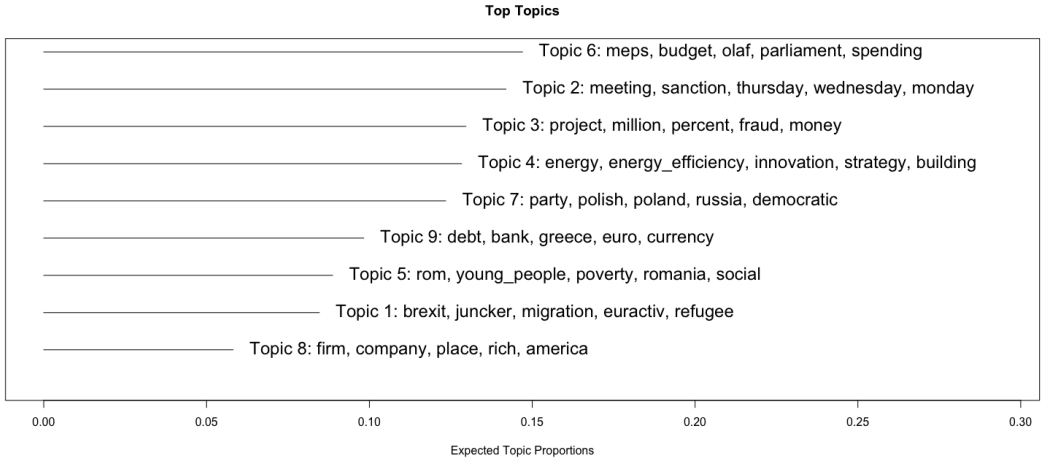


From Figure 5, the K=9 model seems the most optimal since it is closest to the upper right quadrant. We therefore run a K=9 stm model. We also create a dummy variable called "level". This variable takes on the value "EU" if the sources are the web native EU sources (i.e., Politico, EUobserver and EUractiv). These are typically web only sites, so-called web natives. The "Int" value is assigned to other sources that are printed and not exclusively focused on the EU (i.e, a print media source that has a global focus rather than a specifically EU focus). In practice, after filtering, the "Int" category comprises mainly The Economist media source.

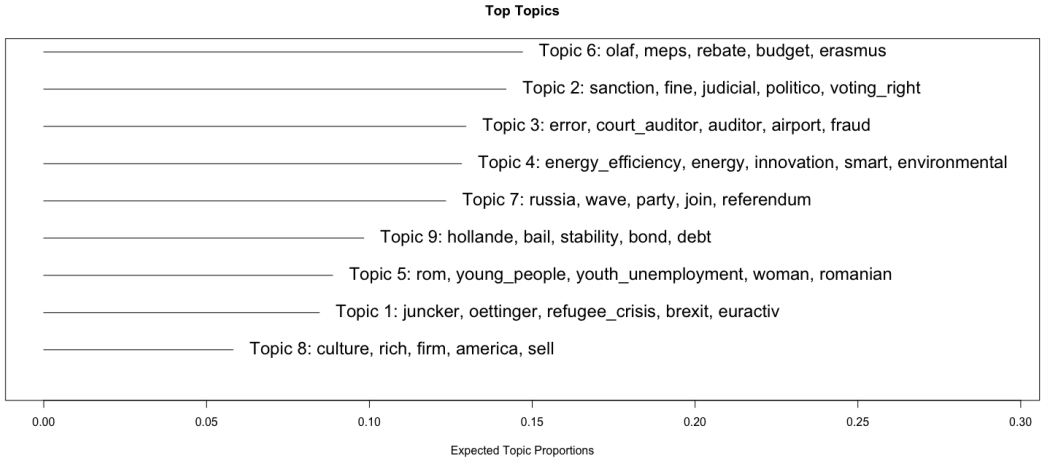
Summary of the 9 topic model

We now explore some of the key words associated with the 9-topic model. The FREX metric is a useful metric for evaluating topic quality through a combination of semantic coherence and exclusivity of words to topics. In the plots below we vary the weight on this key metric (from a broader threshold to a more exclusive threshold) in order to probe the five most important keywords associated with the topics. This provides us with a first glimpse of the topics identified.

- FREX summary (weight =0.6)



- FREX summary (weight =0.1)



We can already see that the Frex metric reveals some meaningful keywords associated with the topics. To probe further, we look at the 9 topics and augment the number of key words. We use 4 metrics and assign some labels that seem to capture the topics that have been identified. Note we have also conducted a manual inspection of the articles associated with the topics before assigning short labels.

In terms of the metrics, the most useful one is the FREX and to a lesser extent Score. The LIFT metric can produce very unique words -giving higher weight to words that appear less frequently in other topics. Whereas the Prob metric simply lists the most probable words in the topic -this criterion is not always so useful but it is worth reporting.

-(Refugees / Migrants) Topic 1 Top Words: Highest Prob: country, european, fund, cohesion_policy, also, project, brexit, member_state, investment, future, migration, greece, refugee, juncker, solidarity, euractiv, policy, budget, funding, europe FREX: brexit, juncker, euractiv, migration, refugee, solidarity, cohesion_policy, oettinger, grant, cooperation, refugee_crisis, migrant, asylum_seeker, yesterday, turkey, objective, cohesion, investment_plan, trust, czech Lift: michael, refugee_crisis, oettinger, commissioner_corina, juncker, relocation, investment_plan, brexit, euractiv, asylum,

commission_vice, terrorism, jyrki_katainen, yesterday, migration, asylum_seeker, turkey, refugee, hungary_poland, told_euractiv Score: michael, cohesion_policy, refugee, brexit, migrant, euractiv, greece, oettinger, asylum_seeker, refugee_crisis, juncker, asylum, investment_plan, migration, relocation, grant, financial_instrument, told_euractiv, digital, emmanuel_macron

- (Conditionality) Topic 2 Top Words: Highest Prob: commission, member_state, rule, commissioner, also, meeting, country, percent, brussels, european_commission, national, official, proposal, state, minister, idea, wednesday, monday, sanction, issue FREX: sanction, meeting, thursday, wednesday, monday, fine, court, rule, idea, rompuy, tuesday, forward, deficit, judicial, legal, commissioner, treaty, pact, council, diplomat Lift: excessive, voting_right, rotating_presidency, judicial, politico, suspended, suspension, fine, breach, sanction, jose_manuel, council_president, compliance, rompuy, light, pact, barroso, gathering, lisbon_treaty, spring Score: excessive, sanction, rompuy, finance_minister, voting_right, percent, suspension, fine, meeting, court, commission, diplomat, tuesday, member_state, thursday, judicial, treaty, euobserver_brussels, cabinet, barroso

- (Irregularities) Topic 3 Top Words: Highest Prob: fund, project, money, percent, member_state, commission, million, report, billion, also, funding, country, year, regional, european, european_commission, fraud, national, programme, part FREX: fraud, error, million, project, airport, percent, money, auditor, report, court_auditor, regional_policy, funded, fund, total, allocated, aimed, erdf, regional, funding, amount Lift: raising, error, court_auditor, auditor, erdf, airport, earmarked, national_authority, fraud, funded, waste, johannes_hahn, audit, lithuania, allocated, aimed, regional_authority, million, regional_policy, properly Score: raising, error, airport, court_auditor, percent, fraud, auditor, million, erdf, waste, regional_policy, project, member_state, fund, commission, bulgaria, money, transport, latvia, baltic

- (Carbon/Environment/Urbanism) Topic 4 Top Words: Highest Prob: region, europe, european, energy, investment, policy, city, need, building, strategy, cohesion_policy, citizen, regional, energy_efficiency, innovation, also, level, economy, project, social FREX: energy, energy_efficiency, innovation, strategy, building, sustainable, city, region, urban, sector, environment, environmental, coal, goal, smart, green, water, climate_change, citizen, climate Lift: provides, energy_efficiency, smart, energy, climate_change, emission, coal, innovation, environmental, urban, pillar, sustainable, governance, fuel, dialogue, water, committee_region, strategy, climate, environment Score: provides, energy_efficiency, energy, cohesion_policy, urban, smart, building, coal, innovation, city, climate_change, region, sustainable, pillar, committee_region, strategy, dialogue, industry, emission, transport

- (Employment/Youth/Social) Topic 5 Top Words: Highest Prob: social, people, romania, country, work, european, poverty, rom, scheme, young_people, year, report, also, employment, number, help, education, unemployment, youth_unemployment, according FREX: rom, young_people, poverty, romania, youth_unemployment, romanian, woman, social, education, child, unemployment, employment, scheme, housing, people, work, training, european_social, health, unemployed Lift: newspaper, rom, young_people, youth_unemployment, woman, youth, romanian, unemployed, poverty, child, youth_employment, skill, million_people, labour_market, education, housing, training, inclusion, romania, danish Score: newspaper, rom, young_people, youth_unemployment,

poverty, woman, child, unemployed, romanian, skill, migrant, romania, labour_market, young, youth, european_social, housing, bulgaria, bulgarian, employee

- **(Budget)** Topic 6 Top Words: Highest Prob: budget, meps, member_state, billion, parliament, spending, also, money, commission, european_parliament, payment, year, deal, negotiation, national, olaf, fund, increase, government, talk FREX: meps, budget, olaf, parliament, negotiation, spending, payment, next_year, deal, european_parliament, paper, bill, seven_year, talk, farm, farmer, rebate, office, increase, student Lift: olaf, erasmus, janusz_lewandowski, rebate, anti_fraud, annual_budget, meps, common_agricultural, multi_annual, strasbourg, financial_framework, budget, next_seven, cycle, negotiating, cover, tabled, investigation, seven_year, plenary Score: erasmus, olaf, budget, meps, rebate, payment, farm, janusz_lewandowski, anti_fraud, spending, parliament, farmer, investigation, multi_annual, member_state, strasbourg, tabled, cohesion_policy, annual_budget, percent

- **(Eastern Europe)** Topic 7 Top Words: Highest Prob: country, poland, party, government, europe, european, polish, member, germany, political, also, election, prime_minister, want, hungary, brussels, like, year, leader, power FREX: party, polish, russia, poland, democratic, join, referendum, membership, election, warsaw, hungarian, democracy, prime_minister, communist, wave, pole, ukraine, campaign, russian, border Lift: wave, poll, eurosceptic, russia, ally, nationalist, russian, pole, party, join, referendum, military, schengen, membership, soviet, joining, democratic, candidate, warsaw, polish Score: wave, poland, russia, polish, party, warsaw, schengen, poll, russian, hungarian, prime_minister, refugee, pole, democratic, poland_hungary, hungary, ally, independence, communist, referendum

- **(Business/Industry)** Topic 8 Top Words: Highest Prob: country, company, place, state, region, government, economy, poland, city, say, firm, local, year, people, world, market, even, business, investment, many FREX: firm, company, rich, place, america, culture, technology, asset, world, china, american, land, university, industry, product, local, food, value, century, sell Lift: culture, sell, rich, firm, america, land, century, asset, owned, american, giant, globalisation, product, property, grow, food, plant, base, stake, china Score: culture, firm, america, technology, china, rich, asset, owned, sell, american, century, industry, industrial, poland, building, city, farmer, property, land, local

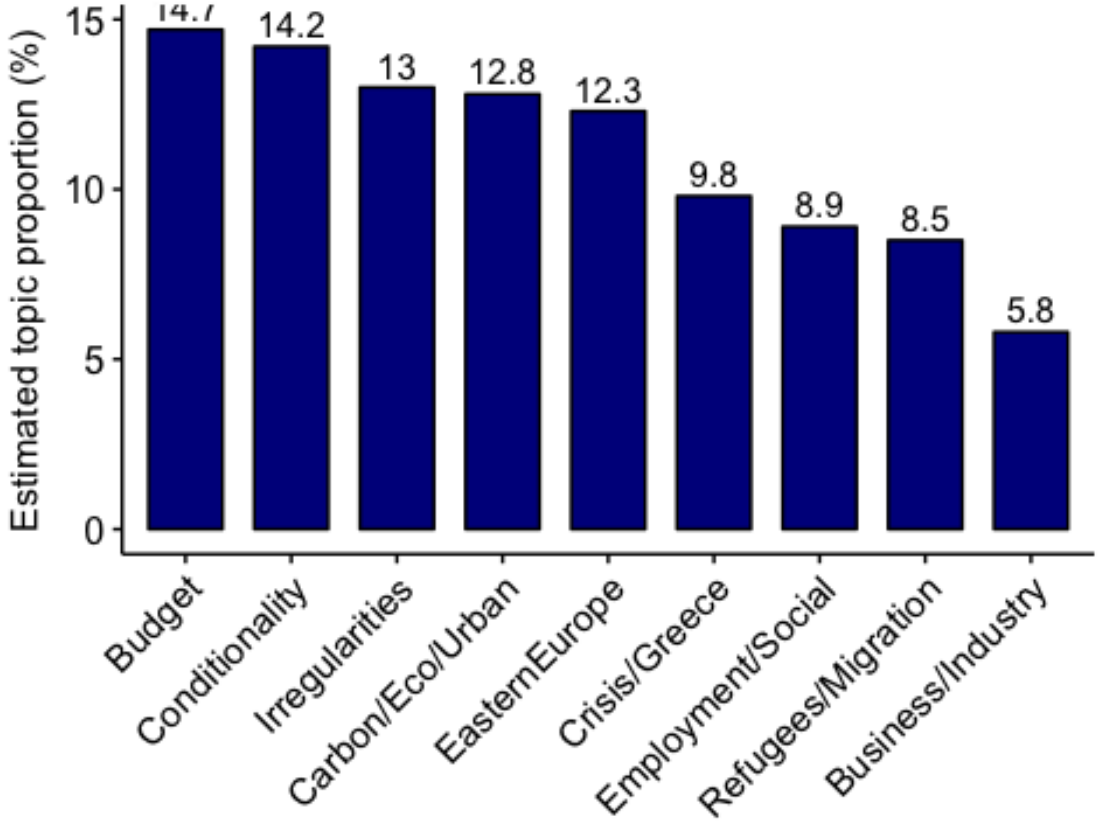
-**(Crisis/Greece)** Topic 9 Top Words: Highest Prob: greece, country, european, billion, euro, bank, debt, economy, europe, government, eurozone, growth, year, greek, fiscal, crisis, portugal, leader, germany, currency FREX: debt, bank, currency, hollande, greece, portugal, euro_zone, fiscal, euro, greek, bail, austerity, eurozone, stability, merkel, bond, growth, bailout, loan, lending Lift: bail, hollande, lending, stability, bond, euro_zone, bailout, debt, international_monetary, bank, currency, rescue, debt_crisis, merkel, portugal, greek, austerity, austerity_measure, central_bank, fiscal Score: stability, greece, euro_zone, bail, hollande, debt, bond, greek, currency, lending, eurozone, bank, bailout, portugal, fiscal, merkel, euro_area, finance_minister, central_bank, single_currency

Estimating topic proportions

To summarise, the 9 topics seem to cover a coherent array of issues. In Figure 6 we plot the estimated topic proportions. There is a degree of overlap in the top three issues, which deal with budgetary politics, conditionality and irregularities. Of the main thematic priorities, it

is the low carbon economy broadly understood that is the most likely topic to feature in our transnational media sample.

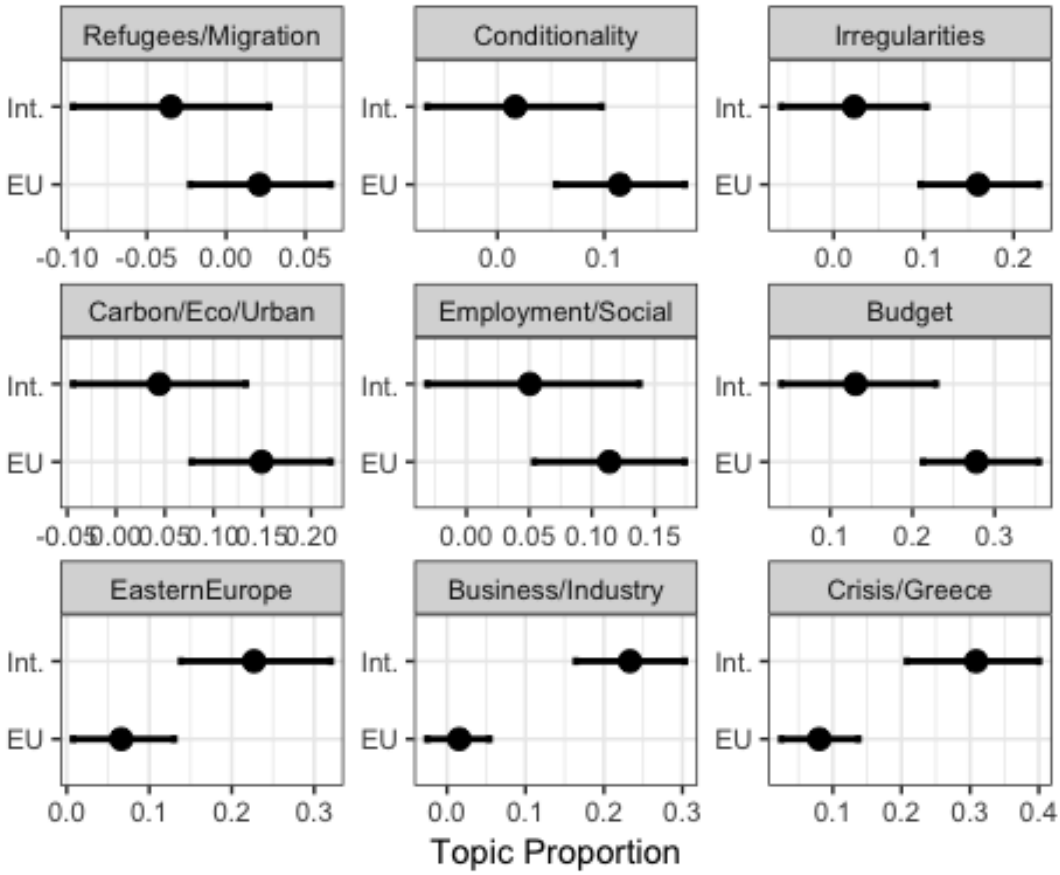
Figure 6: Summary of topic proportions with assigned labels.



Effects of type of media source on topic discussion

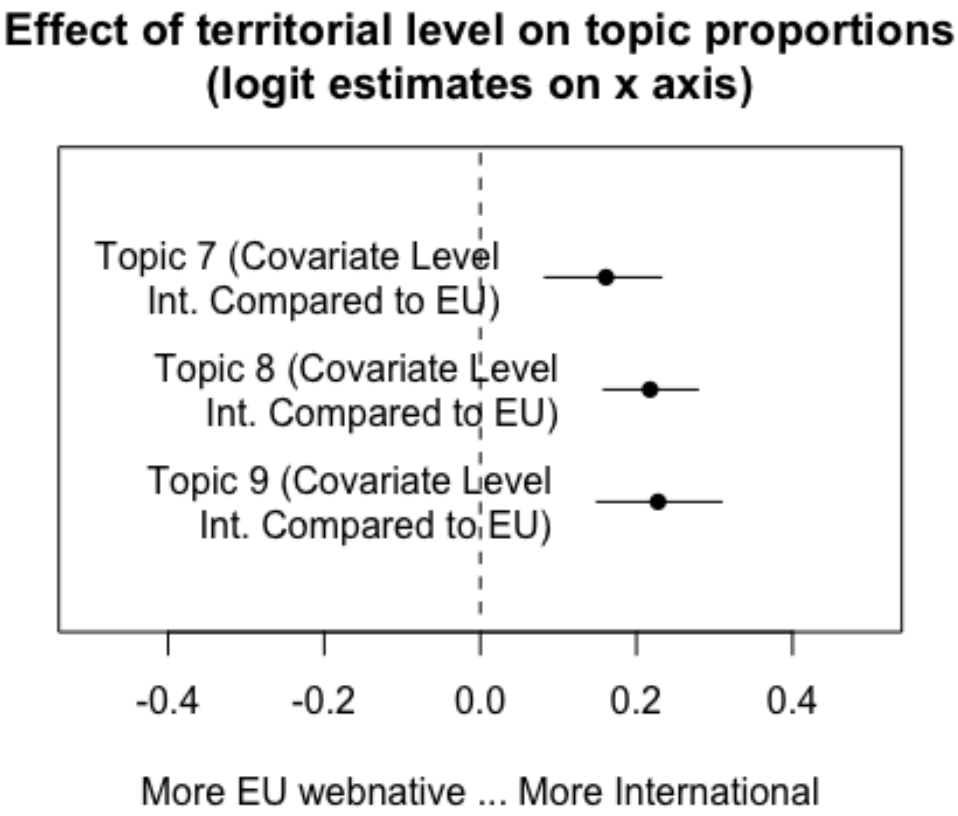
One of the benefits of fitting an STM is that it allows us to include meta-data parameters in our models. Below we make use of two additional variables, level (a binary variable which distinguishes between "EU" (web-native media sources) and "Int" (international source).

Figure 7: Level of media source and estimated topic proportions (with 95% confidence intervals)



In Figure 7 and 8 we can see the effect of level on estimated topic proportions when comparing the two distinct media. Basically, The Economist ("Int.") is more likely to discuss topics related to Eastern Europe, Business/Industry, Crisis/Greece and Irregularities than the "EU web natives". We can see that there are categories with the estimates and the 95% confidence interval error bars overlap -such as Refugees/Migration- and others where there is a clear differentiation between levels of media -such as the Crisis/Greece- category.

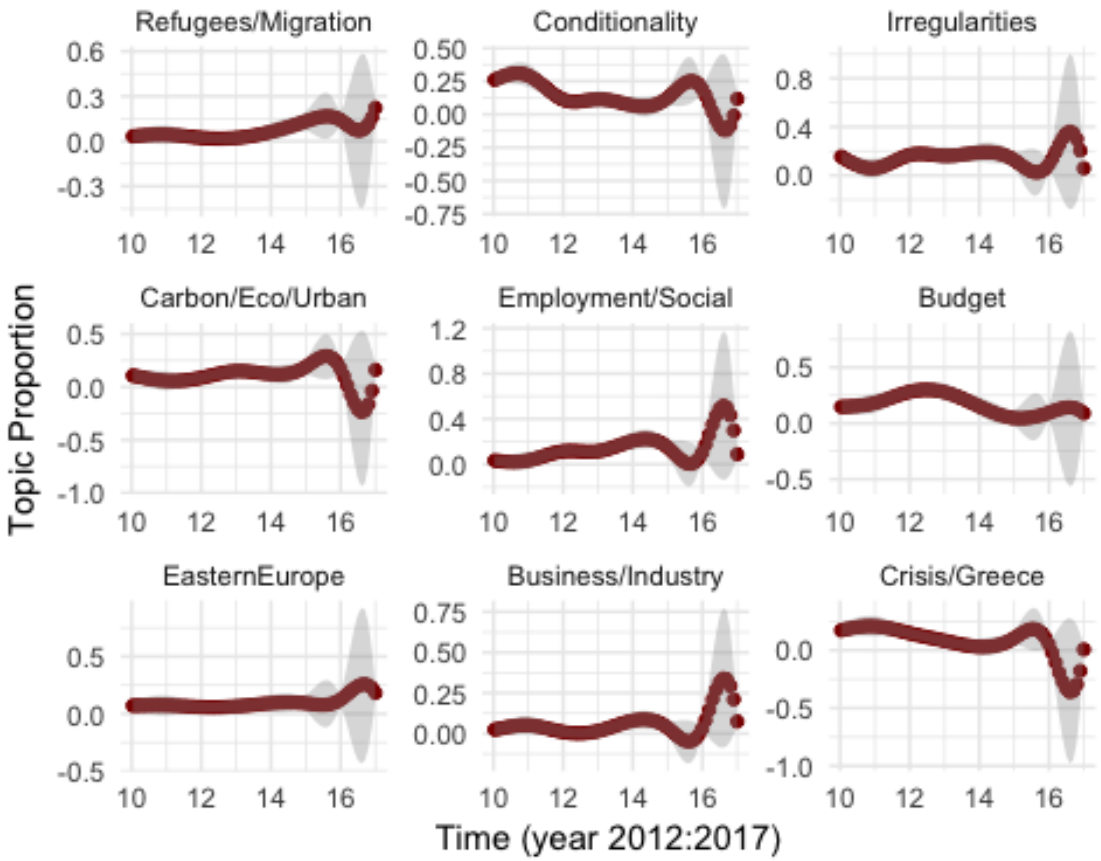
Figure 8: Estimated effect of shifting from 'EU' to 'Int' level



Modelling time

We can also make use of the time variable per topic. In Figure 9 we can see that the Crisis/Greece topic has dropped while topic such as Refugees/Migration and Employment/Social affairs are on the increase.

Figure 9: Estimated topic proportions over time

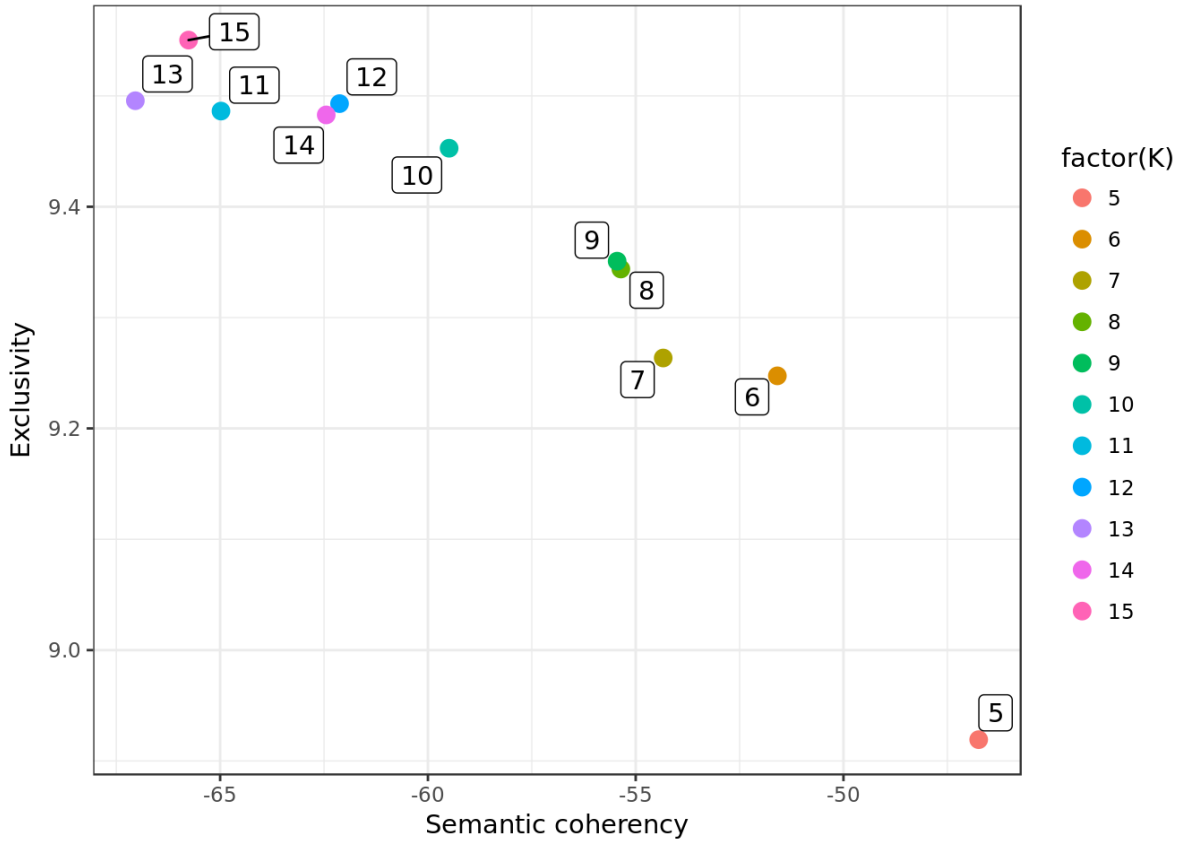


Spain

Search for K topics

As with previous cases, we begin by running a search for most appropriate K number of topics for fitting the stm.

Figure 10: Selecting K number of topics for Spain media.

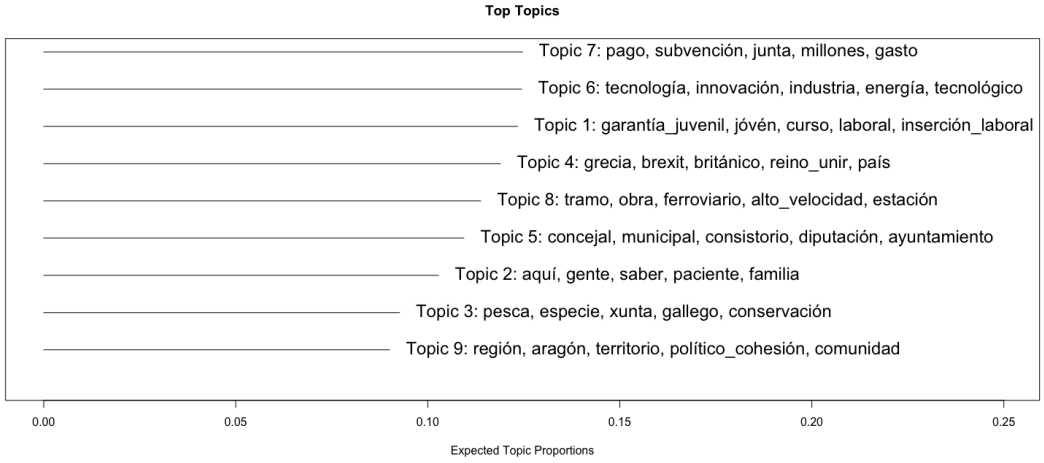


As with the two previous examples, a K=9 model seems the most optimal (see Figure 10). We fit an stm model with K=9 topics and create a dummy variable called "level". This takes on the value "National" if the sources are national (e.g. El Pais, El Mundo etc) or "Regional" if the source is a regional newsprint (e.g. La Voz de Galicia, etc.).

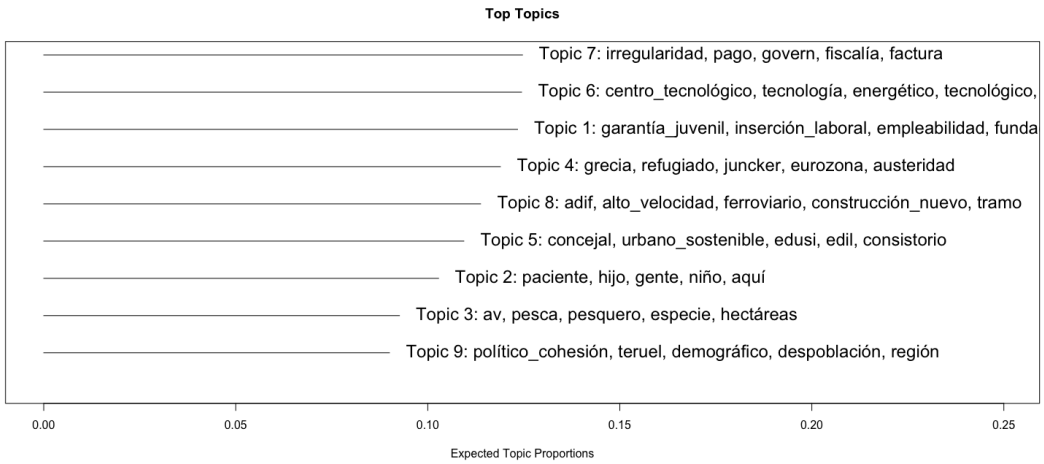
Summary of the 9 topic model

We apply the same analysis as above by first making use of the Frex metric to explore the five most important words associated with the 9-topic model, while varying the Frex weight.

- FREX summary (weight =0.6)



- FREX summary (weight =0.1)



We can see some of the topics likely to emerge by glancing at the frex keywords. We can also tell that topic 2 seems to contain some less coherent words (from a Cohesion Policy perspective). Below we augment the number of key words and metrics, as well as conducting an inspection of the actual article content, to assign some category to the topics.

- (Employment / Youth) Topic 1 Top Words: Highest Prob: empleo, formación, programa, jóvén, persona, empresa, curso, laboral, alumno, social, trabajo, año, fondo_social, poder, europeo, iniciativa, euros, mujer, centro, profesional FREX: garantía_juvenil, jóvén, curso, inserción_laboral, laboral, formación_profesional, social_europeo, persona_discapacidad, desempleados, juvenil, alumno, formación, empleabilidad, mercado_laboral, fondo_social, discapacidad, empleo, menor_año, inserción, escuela_taller Lift: autoempleo, búsqueda_trabajo, contrato_formación, experiencia_laboral, grado_medio, situación_desempleo, certificar_profesionalidad, cualificación_profesional, empleo_joven, estudiar_trabajar, fundación_telefónico, garantía_juvenil, inserción_laboral, jóvén_desempleados, lanzaderas, lanzaderas_empleo, ninis, operativo_empleo, perfil_profesional, realizar_práctica Score: garantía_juvenil, jóvén, empleo, alcalá, inserción_laboral, alumno, mujer, curso, lanzaderas, fundación_once, escuela_taller,

laboral, discapacidad, formación, cruz_rojo, mercado_laboral, participante, fundación_telefónico, empleabilidad, persona_discapacidad

- (Health/R+D) Topic 2 Top Words: Highest Prob: hacer, poder, decir, año, ahora, querer, persona, trabajar, llegar, solo, problema, aunque, mismo, llevar, trabajo, explicar, bien, saber, cómo, caso FREX: gente, aquí, paciente, saber, niño, hijo, familia, cosa, cómo, vivir, gitano, pensar, vida, hablar, cambiar, profesor, parecer, mundo, siempre, creer Lift: marido, parálisis, cáncer, paciente, madre, enfermo, relatar, hijo, compañero, sonar, hija, academia, profesora, amigo, gana, policía, síntoma, casar, gente, niño Score: parálisis, paciente, gitano, mujer, vivir, aprender, hospital, cáncer, gente, niño, hablar, decir, padre, aquí, universidad, hijo, problema, enfermedad, enfermo, nadie

- (Environment/Rural) Topic 3 Top Words: Highest Prob: agua, zona, poder, hacer, galicia, especie, pueblo, medio_ambiente, fondo_europeo, pesca, llegar, parte, xunta, comarca, conservación, gallego, además, encontrar, aunque, primero FREX: pesca, especie, xunta, gallego, av, conservación, agua, pesquero, montaña, galicia, hectáreas, museo, monte, barco, pueblo, flota, forestal, rural, natural, siglo Lift: pico, caminar, av, fauna, cofradía, parque_nacional, sector_pesquero, galega, ganadero, embarcación, buq, pesquero, molino, kilo, pesca, finca, miño, restaurar, parque_natural, hectáreas Score: caminar, pesca, av, agua, concello, museo, especie, xunta, conservación, pesquero, pueblo, residuo, zona, captura, planta, miño, árbol, natura, flota, galicia

- (EUaffairs) Topic 4 Top Words: Highest Prob: europeo, país, españa, poder, político, europa, gobierno, bruselas, deber, nuevo, hacer, crisis, económico, decir, millones, acuerdo, comisión, comisión_europeo, partido, aunque FREX: grecia, británico, brexit, reino_unir, país, ministro, alemania, polonia, déficit, euro, partido, negociación, bruselas, refugiado, crisis, sanción, juncker, rajoy, deuda, alemán Lift: angela_merkel, arancel, banco_central, canciller, ecofin, eurogrupo, juncker, luis_guindos, ministro_economía, ministro_finanza, monetario, objetivo_déficit, pierre_moscovici, político_exterior, presidente_francés, primer_ministro, socialdemócrato, austeridad, eurozona, grecia Score: bruselas, escaño, brexit, eurozona, grecia, juncker, país, merkel, guindos, eurogrupo, refugiado, monetario, multa, británico, reino_unir, zona_euro, parlamento_europeo, comisión_europeo, berlín, primer_ministro

- (Community/Urbanism) Topic 5 Top Words: Highest Prob: proyecto, ayuntamiento, ciudad, municipal, municipio, alcalde, fondo_europeo, diputación, local, barrio, espacio, iniciativa, concejal, actuación, presentar, consistorio, propuesta, plaza, urbano, zona FREX: concejal, consistorio, municipal, diputación, ayuntamiento, alcalde, edil, ciudad, alcaldesa, barrio, edusi, equipo_gobierno, cultural, muralla, estrategia_desarrollo, pamplona, municipio, urbano_sostenible, urbano, urbanismo Lift: carballo, integrar_edusi, danza, grupo_municipal, integrar_dusi, tercero_convocatoria, turístico_cultural, urbano_sostenible, casco_histórico, edusi, festival, espectáculo, concejal, desarrollo_urbano, interreg_españa, edil, smart_city, institución_provincial, teniente_alcalde, equipo_gobierno Score: ayuntamiento, carballo, municipal, alcalde, concejal, consistorio, ciudad, muralla, edusi, dusi, municipio, barrio, urbano_sostenible, espectáculo, edil, equipo_gobierno, urbano, diputación, danza, casco_antiguo

- (R+D/Innovation) Topic 6 Top Words: Highest Prob: proyecto, empresa, nuevo, sector, innovación, desarrollo, investigación, desarrollar, sistema, programa, permitir, universidad,

objetivo, través, producto, poder, además, industria, tecnología, centro FREX: tecnología, innovación, tecnológico, energético, energía, industria, pymes, industrial, producto, investigación, empresarial, científico, emprendedor, inteligente, competitividad, investigador, universidad, centro_tecnológico, internacionalización, aplicación Lift: desarrollo_industrial, pyme, fundación_biodiversidad, mejora_competitividad, promoción_exterior, sector_industrial, startups, energía_turismo, aplicación_móvil, centro_tecnológico, instituto_fomento, biotecnología, renovable, software, monitorización, innovación_tecnológico, instituto_tecnológico, info, pesca_alimentación, desarrollo_tecnológico Score: desarrollo_industrial, innovación, tecnología, pymes, empresa, investigación, producto, centro_tecnológico, universidad, energético, científico, emprendedor, internacionalización, fabricación, industria, competitividad, investigador, fundación_biodiversidad, economía_circular, energía

- (Irregularities/Spending) Topic 7 Top Words: Highest Prob: millones, millones_euros, euros, ayuda, fondo, subvención, presupuesto, público, gobierno, junta, caso, según, fondo_europeo, poder, dinero, gasto, recibir, empresa, plan, andalucía FREX: pago, junta, subvención, millones, gasto, cuenta, sindicato, factura, partida, denunciar, ministerio, cantidad, dinero, irregularidad, hacienda, psoe, expediente, justificar, fraude, portavoz Lift: audiencia_nacional, melilla, juzgar_instrucción, investigados, investigadar, tribunal_cuenta, devolución, sumario, ccoo, imputados, cofinanciar, diligencia, fiscalía_anticorrupción, fiscalía, irregularidad, algeciras, reclamación, olaf, govern, auditoría Score: melilla, millones, millones_euros, subvención, pago, irregularidad, euros, psoe, cuenta, junta, fraude, dinero, denunciar, juez, factura, andalucía, olaf, sindicato, contrato, inversión

- (Transport/Infrastructure) Topic 8 Top Words: Highest Prob: obra, nuevo, proyecto, euros, construcción, zona, instalación, edificio, centro, actuación, tramo, infraestructura, servicio, millones_euros, inversión, además, trabajo, prever, permitir, acceso FREX: tramo, obra, ferroviario, alto_velocidad, aparcamiento, estación, licitación, carretera, túnel, construcción, tráfico, adif, conexión, metros_cuadrados, avenida, plazo_ejecución, autovía, trazado, instalación, puerto Lift: construcción_nuevo, ejecución_obra, glorieta, metros_longitud, obra_construcción, ramal, redacción_proyecto, vial, anchura, aseos, colector, conectará, diario_oficial, estacionamiento, estación_depuradora, inicio_obra, rampa, adif, superficie_metros, licitado Score: obra, diario_oficial, tramo, aparcamiento, metros, trazado, adif, ferroviario, instalación, peatonal, alto_velocidad, túnel, kilómetros, metros_cuadrados, carretera, aseos, plazo_ejecución, avenida, ministerio_fomento, vial

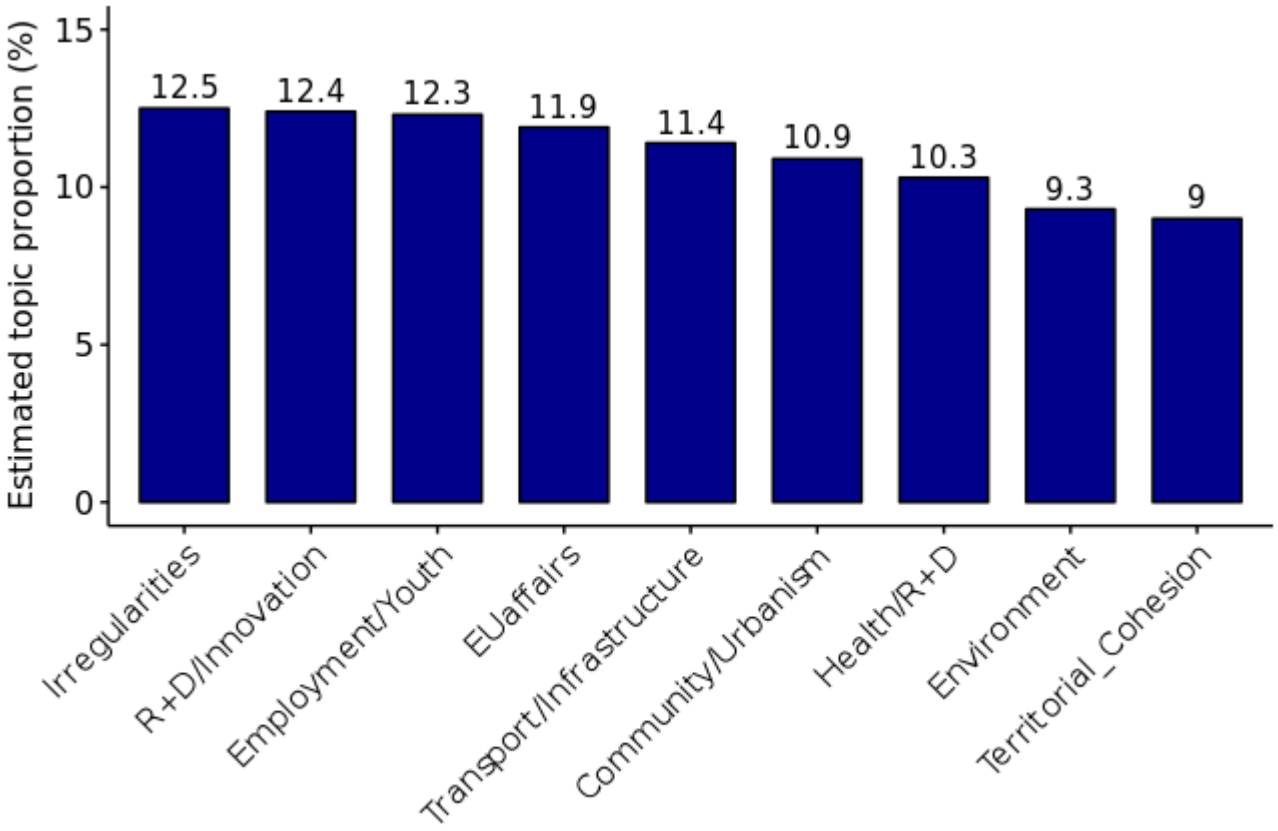
- (Territorial cohesion) Topic 9 Top Words: Highest Prob: región, europeo, comunidad, político, españa, presidente, europa, territorio, fondo, inversión, unión_europeo, medida, financiación, aragón, deber, estrategia, objetivo, económico, regional, población FREX: región, político_cohesión, aragón, territorio, teruel, despoblación, euskadi, demográfico, canarias, comunidad, territorial, prioridad, cooperación, país_vasco, extremadura, vasco, castilla_mancha, regional, comunidad_autónomo, estrategia Lift: tesis, región_español, político_cohesión, aquitania, menos_desarrolladar, teruel, demográfico, banda_ancho, despoblación, canario, comité_región, euskadi, cohesión_social, recalcado, grupo_trabajo, cupo, canarias, dispersión, soria, país_vasco Score: tesis, político_cohesión, despoblación, aragón, región, demográfico, teruel, convergencia, comisión_europeo, estrategia,

inversión, bruselas, comité_región, banda_ancho, autonomías, canarias, fondo_estructural, político_regional, político, europa

Estimating topic proportions

We now look at the estimated proportions per topic. What is interesting to note in terms of topic proportions for the case of Spain is the relatively high proportion in the 'irregularities' category. In practice, this topic includes a broad set of issues related to both irregularities and spending.

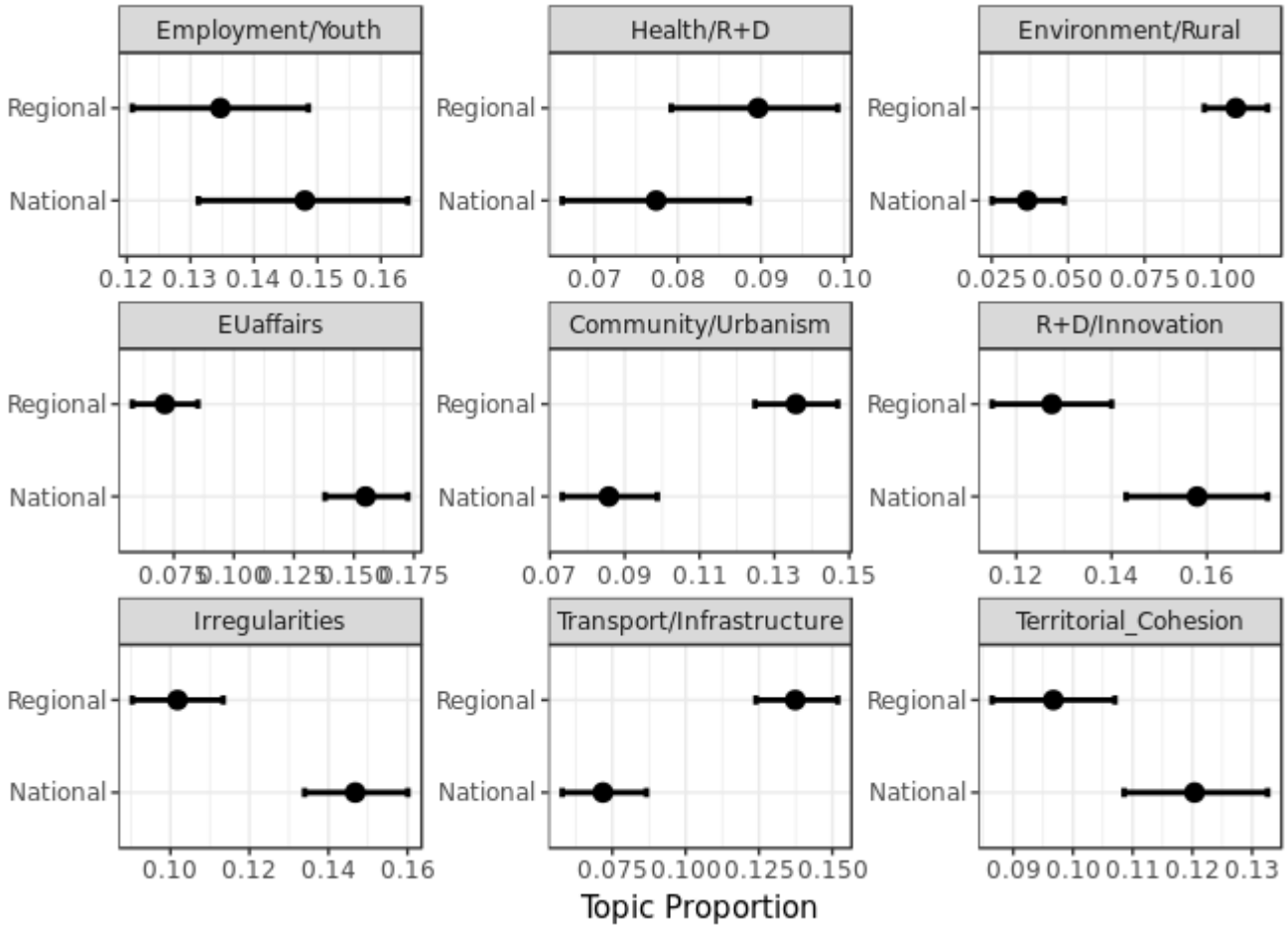
Figure 11: Summary of estimated topic proportions with assigned labels



Effects of type of media source on topic discussion

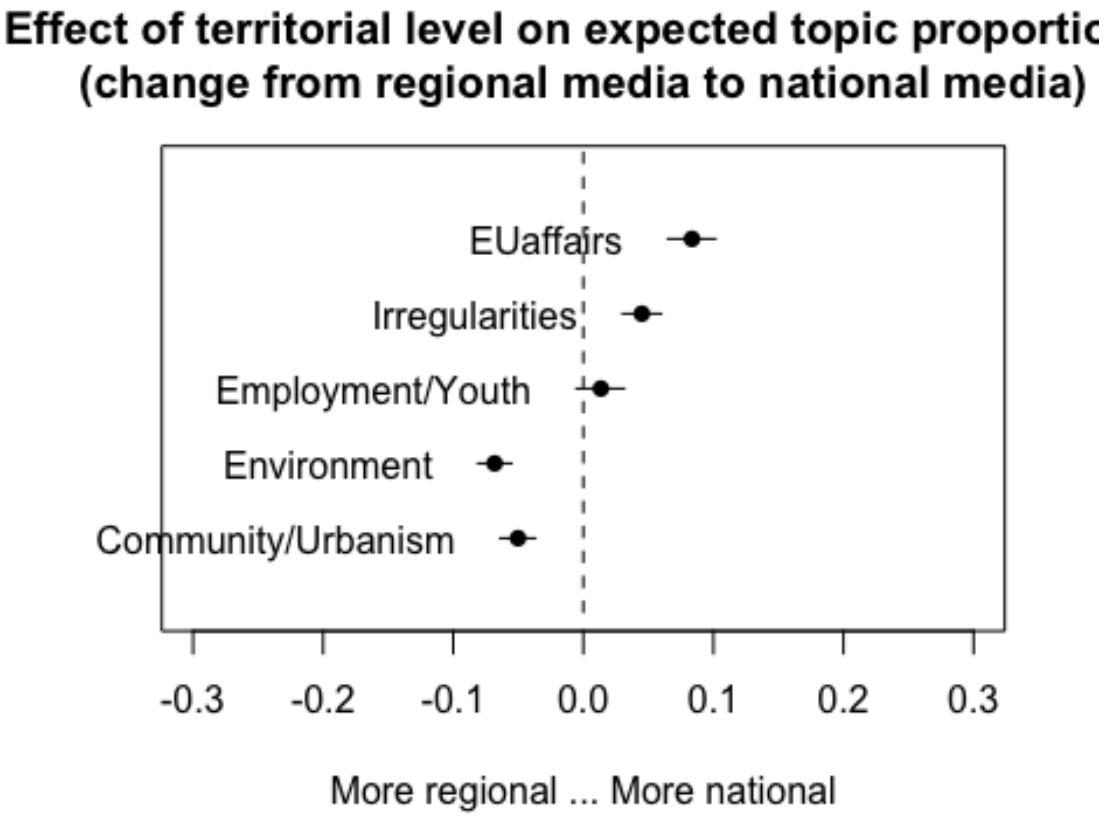
We now look at the effects of territorial levels on media discussion. As can be seen in Figure 12 there are significant differences in the expected topic proportions among the two classes of media.

Figure 12: Level of media source and estimated topic proportions (with 95% confidence intervals)



While the topic of Employment/Youth appears to exhibit no significant differences across territorial levels, the remaining topics do differ significantly. The plot below picks a few topics for further contrasting. The basic message is that there is no statistically significant difference when discussing Employment and Youth, but there are positive coefficients in relation to EU affairs and the Irregularities topics, meaning that the national level is more likely to talk about these topics than the regional level. The inverse is the case for the Community/Urbanism and the Environment/Rural topic - the negative coefficient indicates that this is more likely to be a topic promoted by the regional sources.

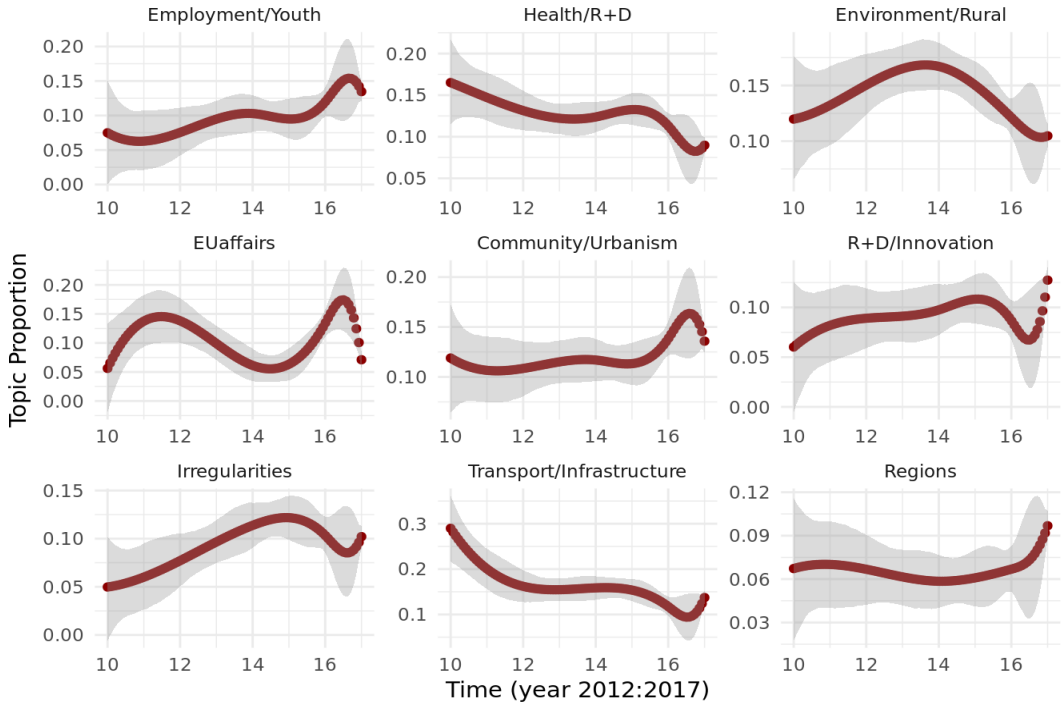
Figure 13: Effect of territorial level on expected topic proportions (change from regional media to national media)



Modelling time

Looking at the evolution of topics over time in Figure 14, the most noteworthy point is the increase in the talk about Spending and Irregularities as well as the Employment/Youth topic. As mentioned above, the irregularities category is more encompassing than the short label suggests, and includes broader issues related to spending.

Figure 14: Spain time modelling

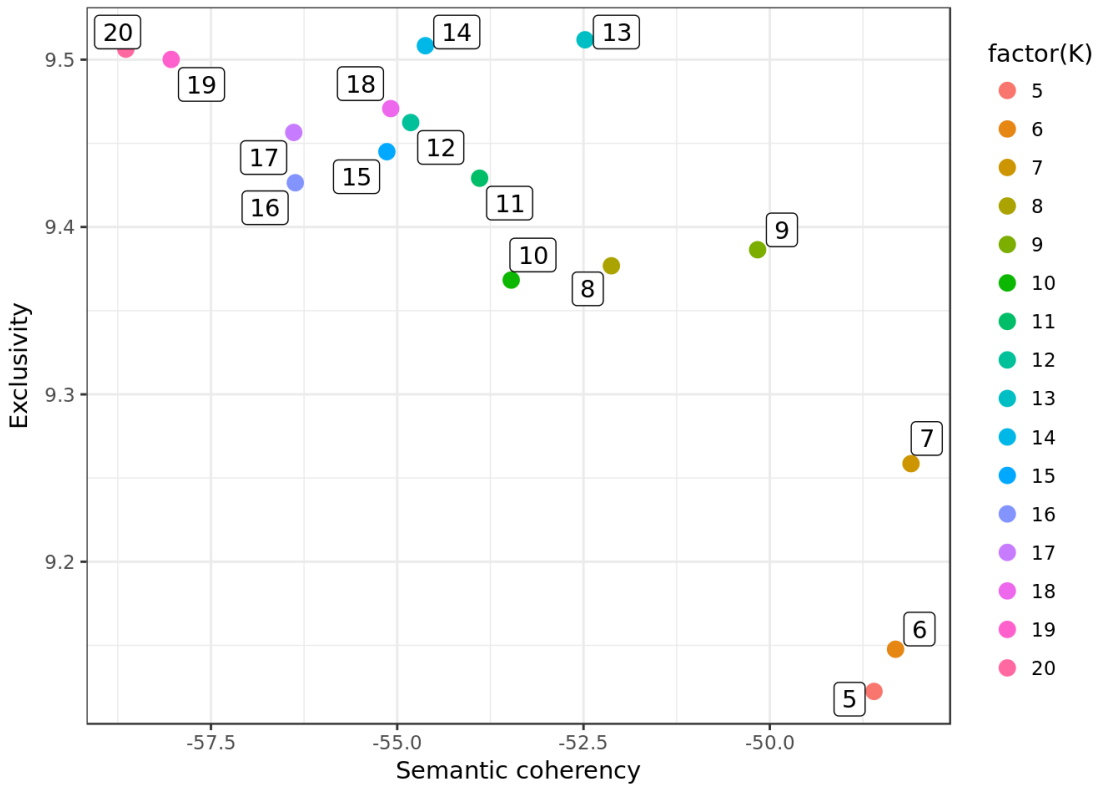


United Kingdom

Search for K topics

As with previous case, we begin by running a search for most appropriate K number of topics for fitting the stm.

Figure 15: Selecting K number of topics for United Kingdom media.

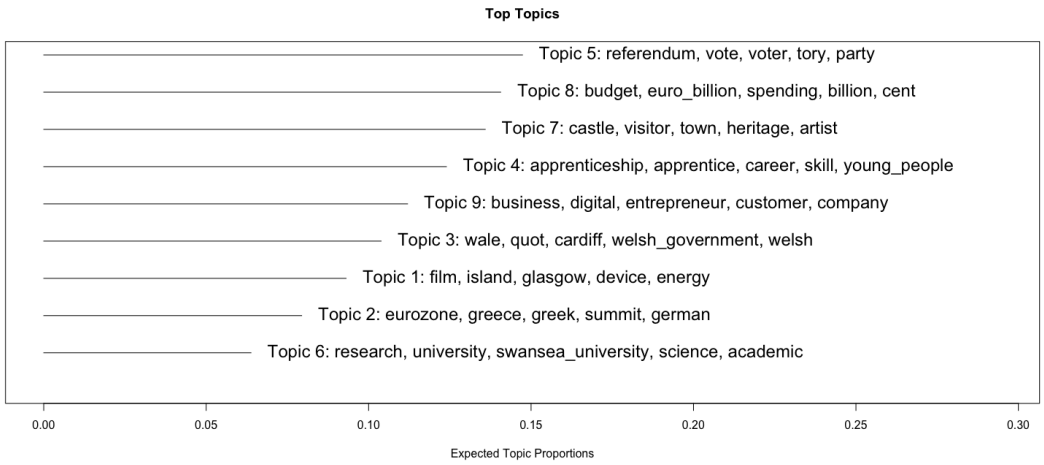


As shown in Figure 15, K=9 model also seems the most optimal. We fit a K=9 stm model and create a dummy variable called "level". This takes on the value "National" if the sources are national (e.g. Guardian, Daily Mail etc) or "Regional" if the source is a regional newsprint (e.g. Scotsman, Walesonline, etc.).

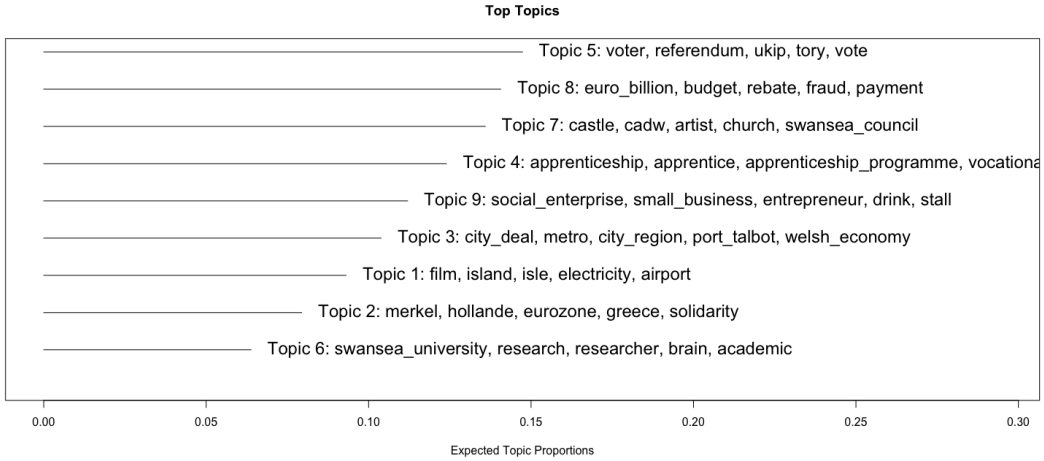
Summary of the 9 topic model

Applying the same analysis as above, the Frex metric is used first to explore the five most important words associated with the 9-topic model, while varying the Frex weight.

- FREX summary (weight =0.6)



- FREX summary (weight =0.1)



The FREX words give us a reasonably good hint as to what the issues per topic relate to. These are examined in more detail by expanding the number of keywords shown and looking at some additional metrics. As above, we assign some labels based on our inspection of the articles.

- (Energy) Topic 1 Top Words: Highest Prob: million, project, scotland, power, energy, film, council, company, investment, site, funding, development, island, industry, city, glasgow, first, device, community, wave FREX: film, island, glasgow, device, electricity, isle, wave, energy, cornwall, land, million, screen, belfast, power, green, scotland, nuclear, marine, studio, renewable_energy Lift: airport, renewables, wind_farm, grid, isle, turbine, film, heating, heat, island, highland_island, wind_turbine, electricity, device, screen, renewable_energy, cable, tidal, glasgow, acre Score: airport, site, device, island, electricity, scotland, film, boat, power, cornwall, acre, scottish, glasgow, scottish_government, belfast, turbine, grid, million, renewables, city_council

- (Euaffairs) Topic 2 Top Words: Highest Prob: european, country, europe, greece, greek, eurozone, leader, bank, , growth, crisis, also, summit, meeting, euro, june, president, market, german, economic FREX: eurozone, greece, greek, summit, german, merkel, hollande, crisis, poland, president, bank, germany, fiscal, debt, june, finance_minister, meeting, angela_merkel, , euro Lift: quota, council_president, french_president, german_chancellor, hollande, international_monetary, band, merkel, single_currency, solidarity, angela_merkel, eurozone, polish, xe7ois, greece, financial_market, greek, central_bank, athens, summit Score: greece, band, merkel, greek, hollande, eurozone, angela_merkel, solidarity, german_chancellor, central_bank, euro, quota, french_president, polish, juncker, refugee, german, brussels, council_president, fiscal

- (Wales) Topic 3 Top Words: Highest Prob: wale, pound, funding, welsh_government, quot, welsh, fund, project, cardiff, investment, business, area, work, region, scheme, government, valley, need, transport, read FREX: wale, cardiff, quot, city_deal, metro, welsh_government, valley, welsh, pound, transport, south_wale, city_region, station, rail, welsh_economy, port_talbot, morgan, local_authority, assembly, plaid_cymr Lift: city_deal, metro, stimulate, edwina, alun, skate, carwyn_jones, port_talbot, city_region, welsh_economy, plaid_cymr, public_transport, hart, lifetime, davy_, rail, morgan, bristol, cardiff, valley Score: wale, stimulate, quot, welsh_government, metro, cardiff, city_deal,

welsh, pound, plaid_cymr, south_wale, port_talbot, valley, welsh_economy, city_region, across_wale, station, swansea, rail, county_borough

- (Employment/Training) Topic 4 Top Words: Highest Prob: work, skill, training, wale, people, apprenticeship, year, career, young_people, employment, learning, support, apprentice, course, working, opportunity, employer, college, help, also FREX: apprenticeship, apprentice, career, skill, young_people, employer, learning, training, qualification, charity, woman, college, disability, apprenticeship_programme, european_social, education, job_growth, adult, employment, course Lift: training_provider, vocational, childcare, coleg, julie_james, year_old, apprenticeship_programme, careerswales, apprenticeship, apprentice, tutor, learner, gender, mental_health, girl, qualification, female, career, placement, visitwww Score: childcare, apprenticeship, apprentice, apprenticeship_programme, learner, vocational, wale, quot, skill, disability, julie_james, qualification, career, careerswales, job_growth, learning, employer, training, young_people, coleg

- (Brexit) Topic 5 Top Words: Highest Prob: britain, people, brexit, say, referendum, government, british, party, country, vote, labour, europe, european_union, london, want, leave, european, many, time, remain FREX: referendum, vote, voter, tory, party, leave, labour, brexit, ukip, vote_leave, immigration, election, remain, membership, politics, campaign, political, independence, british, poll Lift: minority, vote_remain, jeremy_corbyn, voted_leave, voter, nationalist, constituency, leave_campaign, chaos, downing_street, ukip, boris_johnson, referendum, foreign_secretary, getty_image, labour_party, supporter, tory, vote, right_wing Score: minority, brexit, referendum, tory, voter, britain, cameron, vote, vote_leave, ukip, prime_minister, election, manifesto, immigration, westminster, david_cameron, political, treaty, poll, scottish

- (R+D/Innovation) Topic 6 Top Words: Highest Prob: university, research, project, student, funding, innovation, science, industry, pound, swansea_university, also, campus, centre, academic, professor, impact, global, facility, manufacturing, opportunity FREX: research, university, swansea_university, science, academic, professor, campus, student, innovation, collaboration, manufacturing, researcher, scientist, brain, global, facility, impact, engineering, vice_chancellor, knowledge Lift: brain, researcher, laboratory, swansea_university, scientist, scientific, research, vice_chancellor, academic, collaborative, university, science, professor, collaboration, teaching, undergraduate, higher_education, campus, convergence, excellence Score: brain, research, swansea_university, university, campus, student, innovation, scientist, researcher, academic, vice_chancellor, professor, collaborative, swansea, engineering, science, horizon, manufacturing, nbsp, teaching

- (Heritage/Local) Topic 7 Top Words: Highest Prob: project, year, castle, people, also, local, site, town, work, visitor, community, centre, area, pound, council, building, park, life, part, heritage FREX: castle, visitor, heritage, artist, town, garden, church, historic, street, stone, attraction, cadw, history, village, exhibition, resident, park, sculpture, trail, century Lift: ancient, artwork, swansea_council, cadw, grade_listed, heritage_tourism, medieval, heritage_lottery, castle, sculpture, artist, church, gallery, bike, stone, exhibition, display, heritage, visitor, restoration Score: castle, medieval, cadw, site, swansea_council, heritage_tourism, artist, sculpture, heritage, swansea, garden, artwork, church, visitor, exhibition, park, attraction, town, merthyr, heritage_lottery

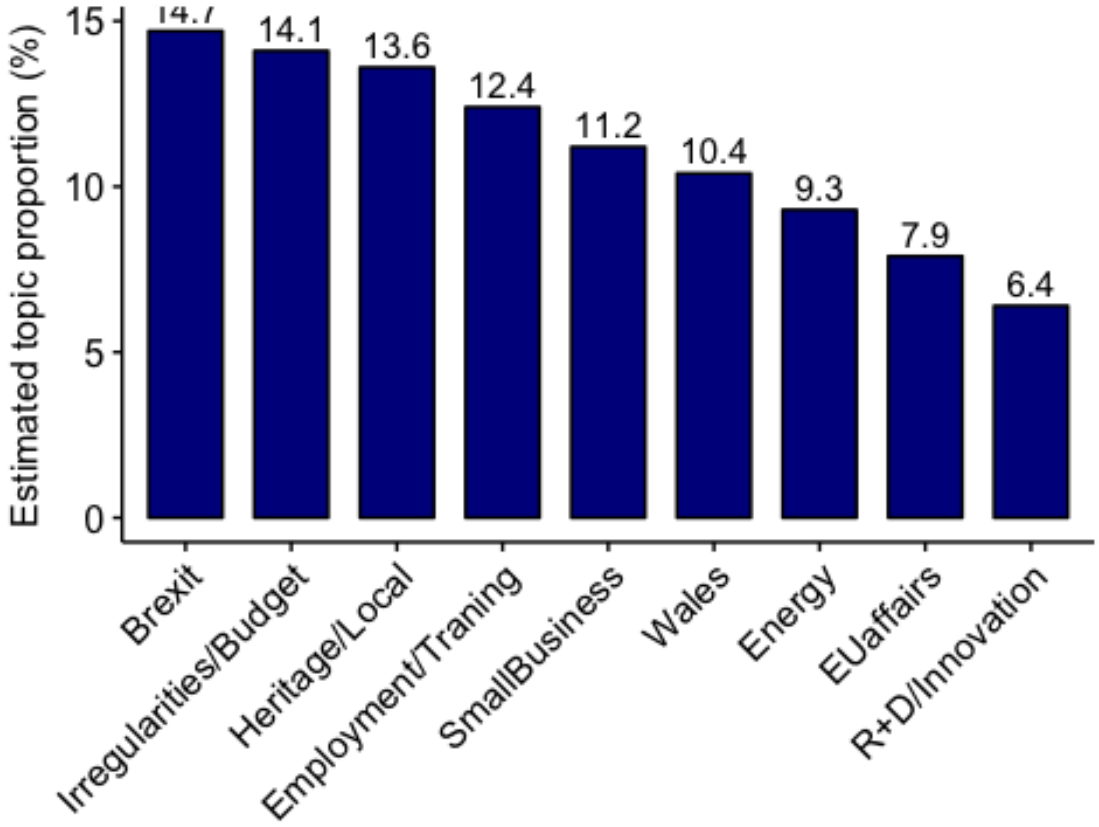
- (Irregularities/Budget) Topic 8 Top Words: Highest Prob: government, budget, britain, funding, fund, money, european, year, billion, cent, spending, also, report, brexit, programme, country, euro_billion, million, brussels, cost FREX: budget, euro_billion, spending, billion, cent, payment, northern_ireland, fraud, irish, rebate, member_state, cash, brussels, cut, account, commission, report, contribution, legislation, justice Lift: noted, rebate, european_court, auditor, euro_billion, whitehall, freeze, fraud, common_agricultural, contributor, audit, budget, expenditure, calculation, directive, payment, spending, republic, justice, public_spending Score: euro_billion, noted, brexit, brussels, rebate, northern_ireland, fraud, member_state, britain, single_market, euro, european_court, scottish, cent, budget, billion, auditor, cameron, devolved, regulation

- (SmallBusiness) Topic 9 Top Words: Highest Prob: business, company, support, market, service, help, local, growth, also, say, year, digital, funding, product, enterprise, time, social, programme, make, job FREX: business, digital, entrepreneur, customer, social_enterprise, small_business, waste, enterprise, company, product, grow, broadband, market, food, online, drink, advice, beer, loan, firm Lift: stall, operative, social_enterprise, drink, expanding, small_business, entrepreneur, small_medium, premise, superfast_broadband, beer, entrepreneurial, manufacturer, commerce, sized, broadband, manufacture, digital, customer, supplier Score: stall, business, social_enterprise, entrepreneur, pontypool, small_business, broadband, operative, entrepreneurial, superfast_broadband, loan, product, premise, company, customer, drink, software, site, digital, innovation

Estimating topic proportions

A reasonably coherent grouping of words associated with topics is found and checked for consistency by inspecting the content of some of the articles assigned to the topics. The estimated topic proportions in Figure 16 reveal some interesting findings. For example, the topic of Brexit is the most popular - a fact that is not altogether too surprising given the saliency of this issue in the UK press.

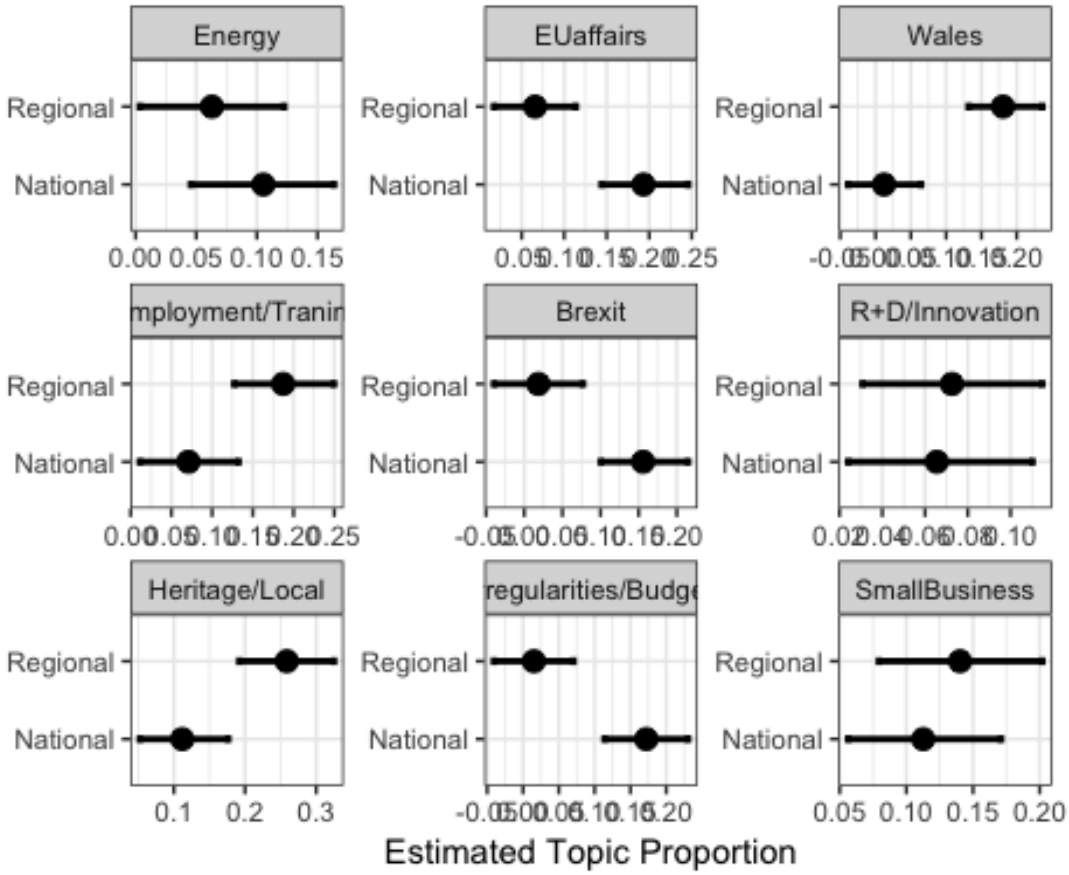
Figure 16: Summary of estimated topic proportions with assigned labels



Effects of type of media source on topic discussion

We now investigate how the level variable affects estimated topic proportions. This allows us to explore the extent to which media from different territorial levels focus on particular topics. As can be seen in the Figure 7 and 18, the level of territoriality in which a media sources is rooted generates significant differences in the expected topic proportions.

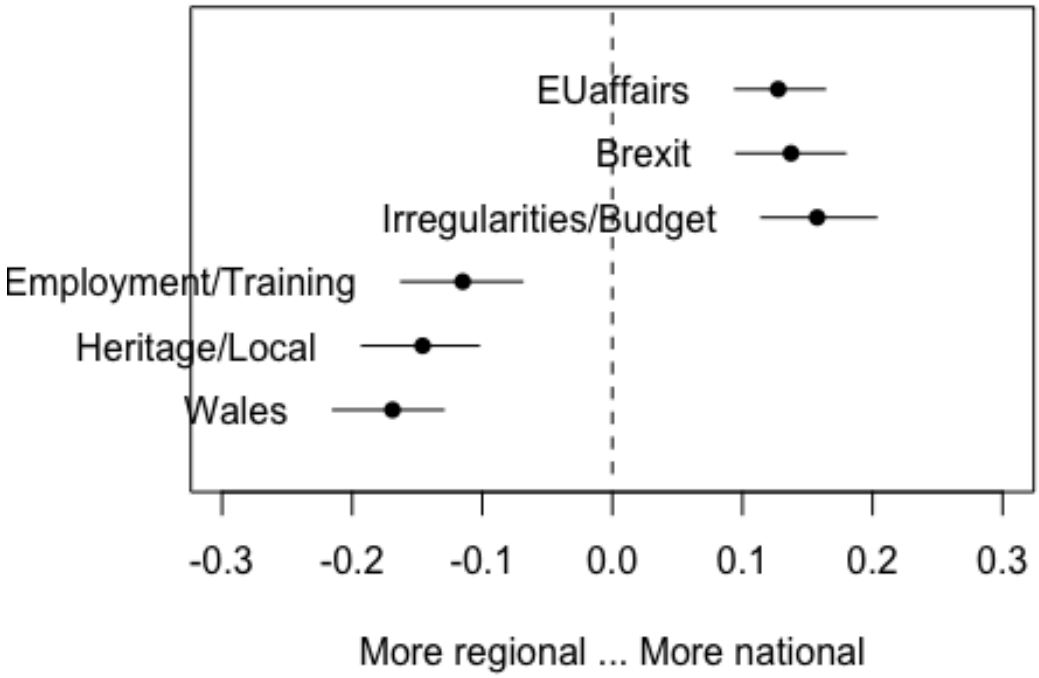
Figure 17: Level of media source and estimated topic proportions (with 95% confidence intervals)



In the plots we can see that some topics proportions clearly overlap (Energy and R+D/Innovation). There is no statistically significant difference in estimated topic proportions for these topics, whose error bars overlap. Others exhibit some clear differences (compare Employment /Training with Brexit).

Figure 18 includes a few of the most interesting topics for further comparison of topic focus across territorial levels. Positive coefficients indicate a statistically significant differences in estimated topic proportions (meaning more likely to be discussed by national media), while negative coefficients indicate more regional discussion of these topics.

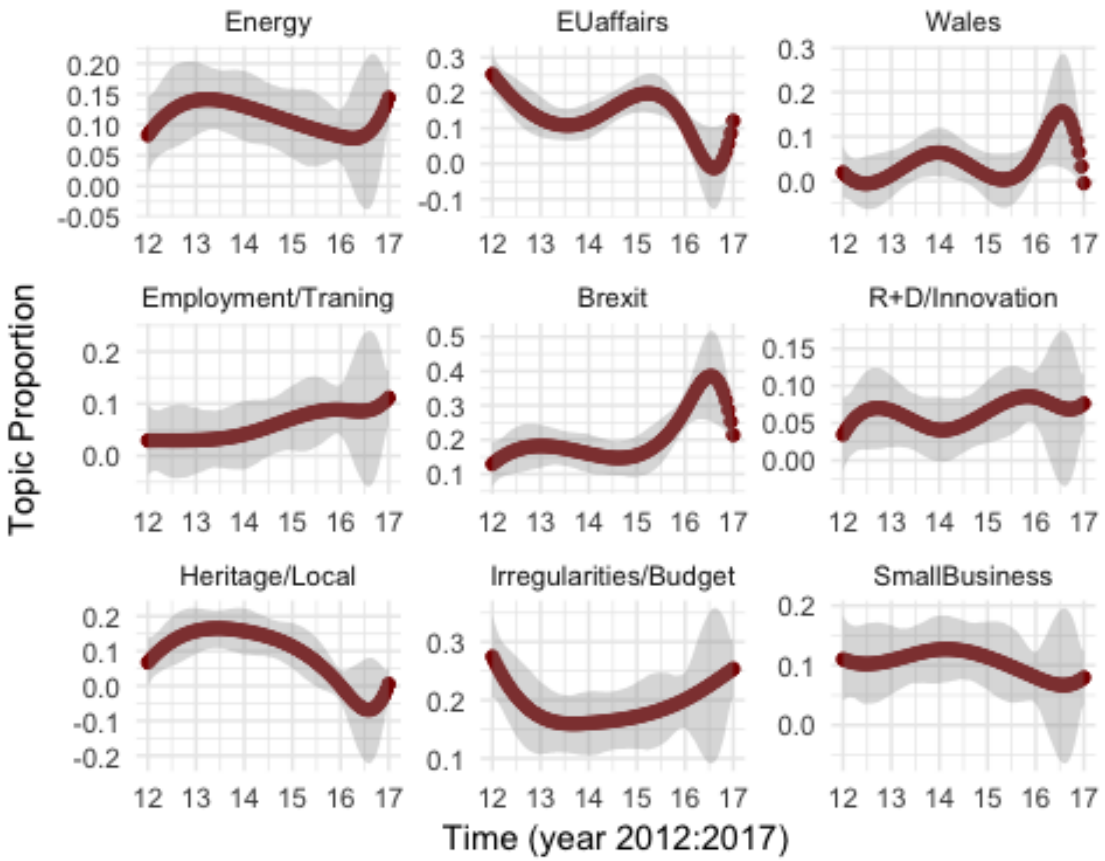
Figure 18: Effect of territorial level on expected topic proportions (change from regional media to national media)



Modelling time

The time modelling in Figure 19 reveals some clear and consistent patterns. Most notably, the Brexit topic is clearly in ascendency while other topics are more stable over time, such as the of R+D/Innovation or the topic of SmallBusiness and entrepreneurship.

Figure 19: UK time modelling



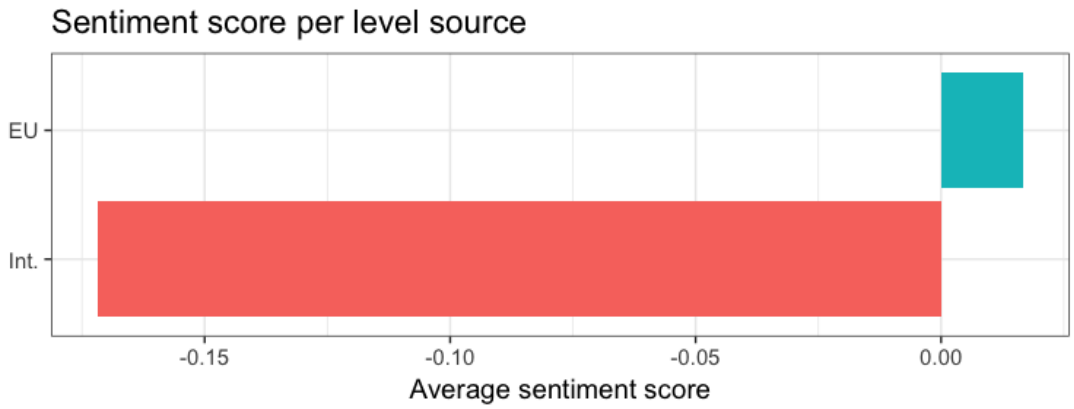
Sentiment analysis

In this last subsection we apply sentiment analysis to investigate the overall sentiment associated with the different media. Note that all media sources with less than 20 articles were filtered out.

Comparison of sentiment (transnational media)

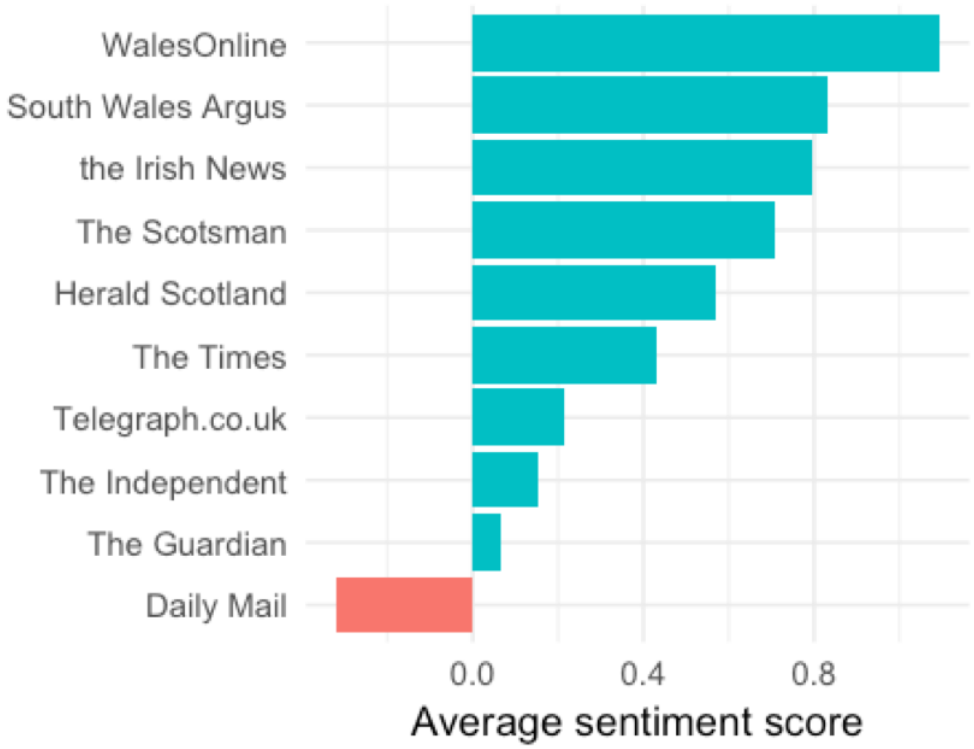
When comparing the sentiment in the content of the articles, we find that the international press (i.e. The Economist) tends to have more negative sentiment than the EU web natives. This is shown in Figure 20, there is large difference between the EU natives and international press with the latter using more negative words.

Figure 20: Sentiment analysis (transnational level)



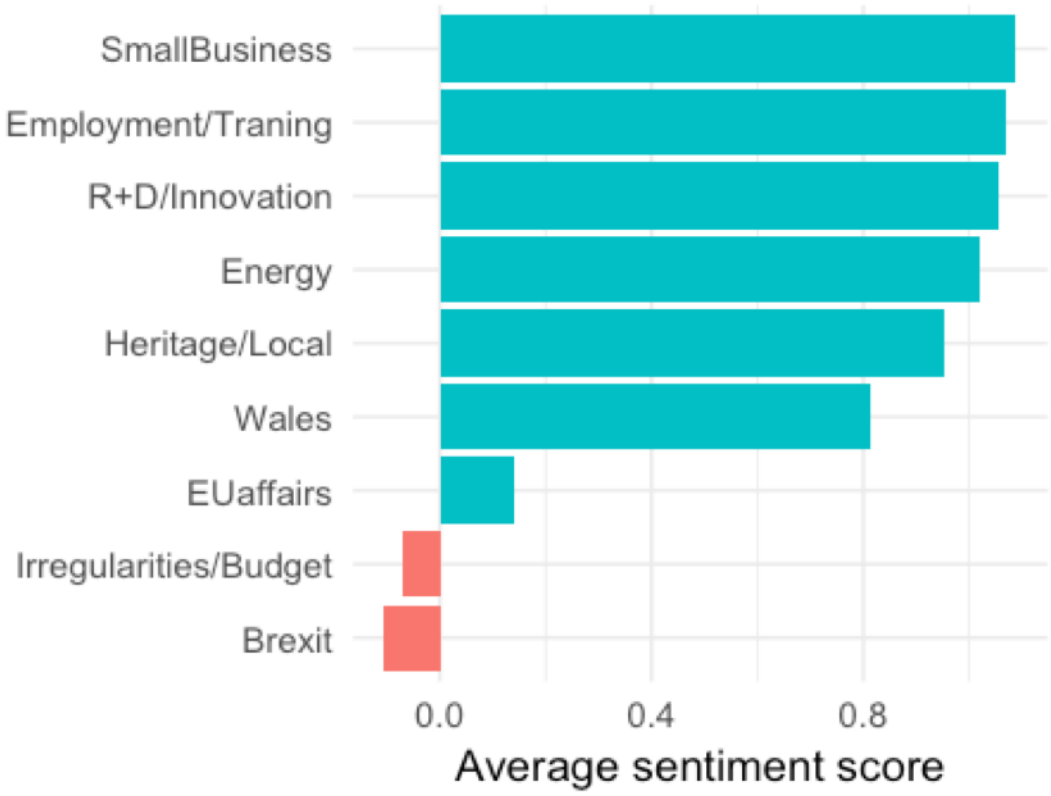
Turning to the United Kingdom case, Figure 21 shows that the regional newspapers are the most positive (the first 5 rows are all regional sources). The national sources are less positive, with the Daily Mail being the newspaper with most negative sentiment in its content dealing with Cohesion Policy. This last finding is hardly surprising given this particular newspaper's notoriously Eurosceptic position. More surprising is the relatively low positive score for the Guardian given that it is a left-leaning and pro-EU newspaper which would be expected to be more positive about a redistributive policy supporting investment in poor regions and communities.

Figure 21: Average sentiment score per media (United Kingdom)



We also take a look at the sentiment per topic for the UK case in Figure 22. The distribution of sentiment per topic appears to be coherent, with Brexit and the Irregularities/Budget topic both being overall negative compared to the other topics which convey much more positive sentiment.

Figure 22: Sentiment per topic (UK)



User-Generated-Content

Table 6 shows a summary of the total number of comments collected from the articles in the web media. The last line shows the number of comments labeled as relevant to the EU Cohesion Policy. We obtain this set of comments after applying a preprocessing in which we check for the occurrences of significant keywords and remove non relevant comments to Cohesion Policy. We observe the low percentage of relevant comments among the comments analyzed. Our hypothesis for this low percentage is that most articles tackle several topics besides those related to EU Cohesion Policy. Thus, for instance, an article about the construction of a high school partly funded with Cohesion funds could generate many comments or discussions about the area where it is built, the architecture of the building, but not necessarily about the funding source itself. To apply sentiment analysis to the whole corpus of comments would thus give a very distorted interpretation. As we can see Table 6 there is a very lively community discussing topics in which there is a mention of Cohesion Policy (especially for the two national cases, Spain and the UK) but very little of that discussion specifically pertains to Cohesion Policy. We now look at the individual cases more closely.

Table 6: Summary of user generated content

	Transnational	Spain	United Kingdom
Number of comments	631	20,218	12,334
Unique users	317	4,286	6820
EU Relevant comments	25	79	353

Transnational

Despite the low number of Cohesion Policy relevant comments in the transnational news sources (see Figure 23), we consider it still useful to apply sentiment analysis. The results are shown in Figure 24. The dotted red line corresponds to a neutral sentiment value among comments for a given article. What is immediately noteworthy is how little of the commentary is positive (or at least above the red line). Most commentary is actually negative from a sentiment analysis perspective. A good example of the most negative comments were from the article entitled "Brussels to UK: Give us clarity on divorce bill or Brexit talks will stall" from the Politico news media source. While the most positive comments related to an article about standing up to the present Hungarian Prime Minister.

In Figure 23 we also show a list of article headlines ordered by the number of comments and colored by topics from our STM model. Here we can observe how articles associated to the topic "Eastern Europe" are the ones which attract more attention from the community -this is a broad category dealing with issues related to Hungary and Poland which can (and do) generate lively debates among readers. Indeed, this topic overlaps to a certain extent with another closely related topic, Conditionality, which also elicits lots of reactions.

Figure 23: Example of UGC top comments from Transnational sources

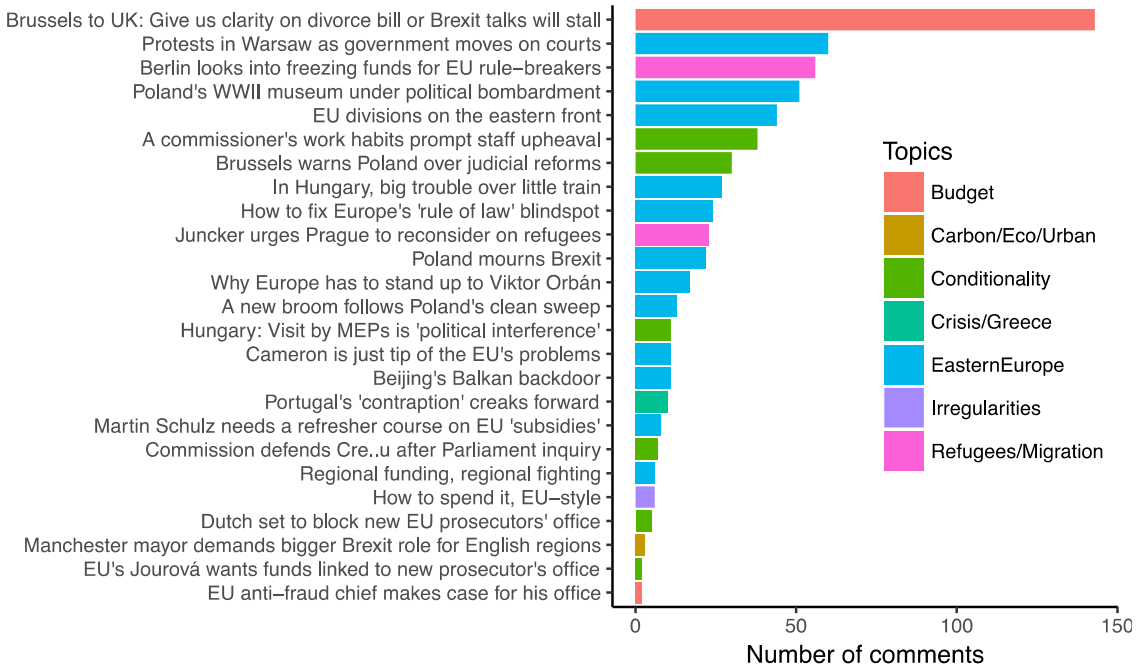
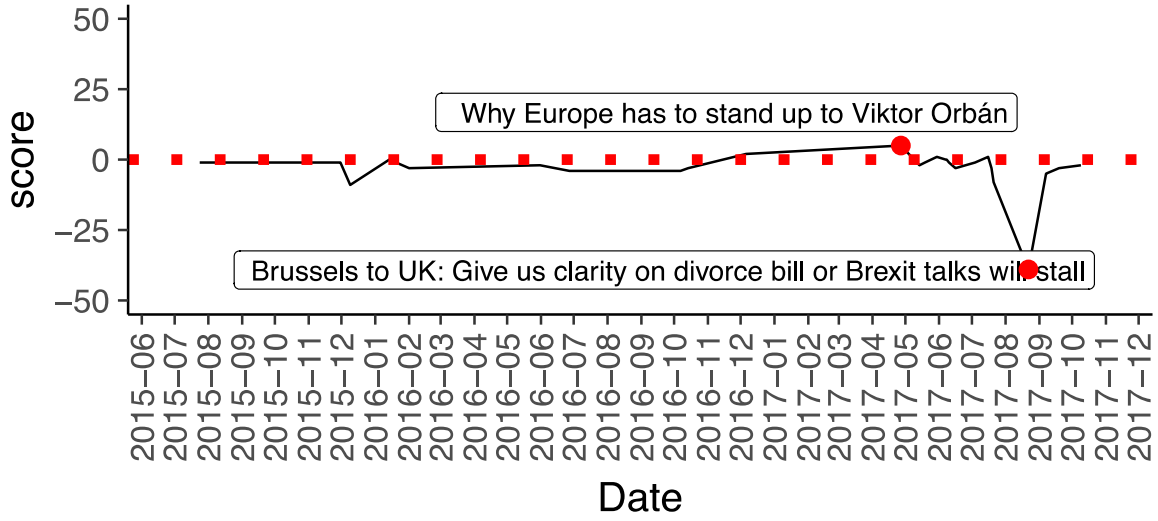


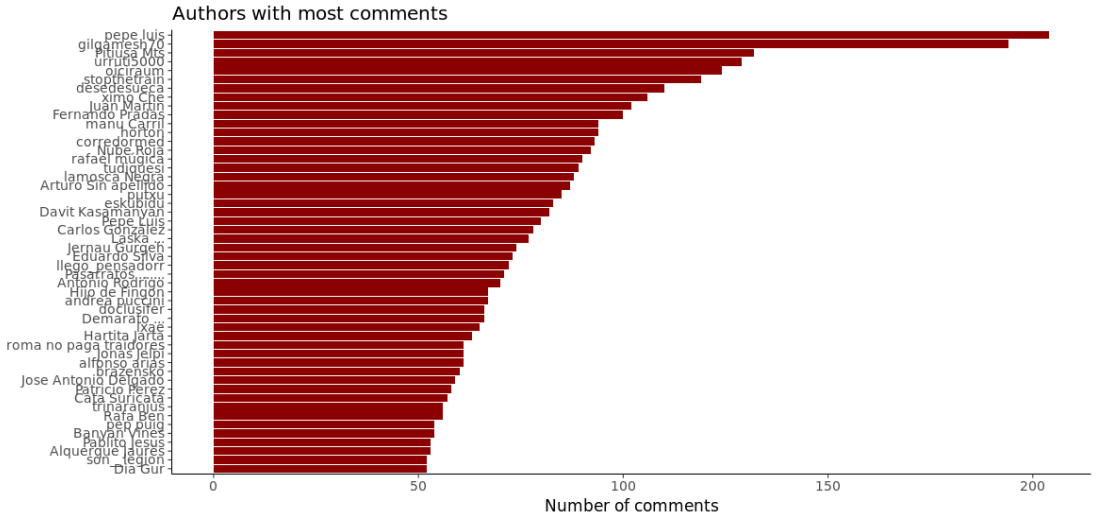
Figure 24: Example of UGC sentiment from Transnational sources.



Spain

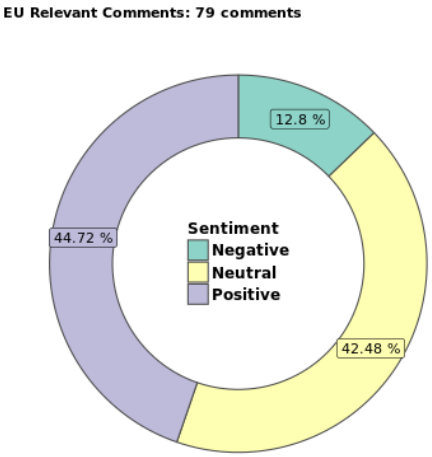
For the case of Spain, we collected comments from different news media sources from Spain. Figure 25 displays the list of top authors in terms of highest number of comments in sources from Spain. Note that there exists a large community of users with more than 50 comments in the news media.

Figure 25: Distribution of users with the highest number of comments from Spain news media.



Our focus however is on that commentary that is relevant to Cohesion Policy. We begin by noting how little of the overall discussion is directly related to Cohesion Policy. We identified under one hundred comments out of approximately 20,000. In Figure 26 we show the distribution of EU relevant comments based on the results from the sentiment analysis. We can clearly see that most of the comments are either neutral or positive. Indeed, most sentiment -nearly half- is actually positive. Only a few comments, barely over 10 percent, appeared to have a bias towards negative opinion.

Figure 26: Distribution of sentiment analysis over EU relevant comments from UGC in spanish



In Figure 27 shows the articles with the highest number of comments and colored according to the classification obtained in our STM model. We observe how articles from the higher level topic "EUaffairs" are more likely to generate commentary among the community of readers. This is not too surprising given that this topic tends to encompass salient issues.

On the other hand, articles related to topics labeled as "Employment/Rural", "Territorial_Cohesion" or "R+D" do not generate lively interest from the community.

Figure 27: Articles with the highest number of comments from Spain news media. Colors reflect the topic label from our STM model.

United Kingdom

We now focus on the case of the UK and the accompanying UGC data that was collected. Although the absolute number of comments was lower than in the Spanish case, the ratio of relevant comments was higher for the UK. Nearly three percent of the total commentary was identified as being relevant to Cohesion Policy. As in the Spanish case, in Figure 28 we see a similar distribution among some of the top commentators, albeit with less absolute numbers. This suggests a lively discussion among the community of commentators. An exploratory sentiment analysis conducted for two news media sources containing the bulk of the commentary (see Figure 29) revealed a very different picture to Spain. In the UK, there was an overwhelming negative bias associated with articles relevant to Cohesion Policy. In fact, very few comments were associated with neutral -let alone positive sentiment. Furthermore, this applies to both sources analysed, the Daily Mail and The Guardian. It is not altogether too surprising that the Eurosceptic, Daily Mail, source would have disproportionately negative bias. However, the same appears to apply to the more EU sympathetic Guardian. Both exhibited overwhelmingly negative bias. Figure 29 flags a

couple of comments to give an idea of the content of the articles generating the commentary.

Figure 28: Distribution of users with highest number of comments from UK news media.

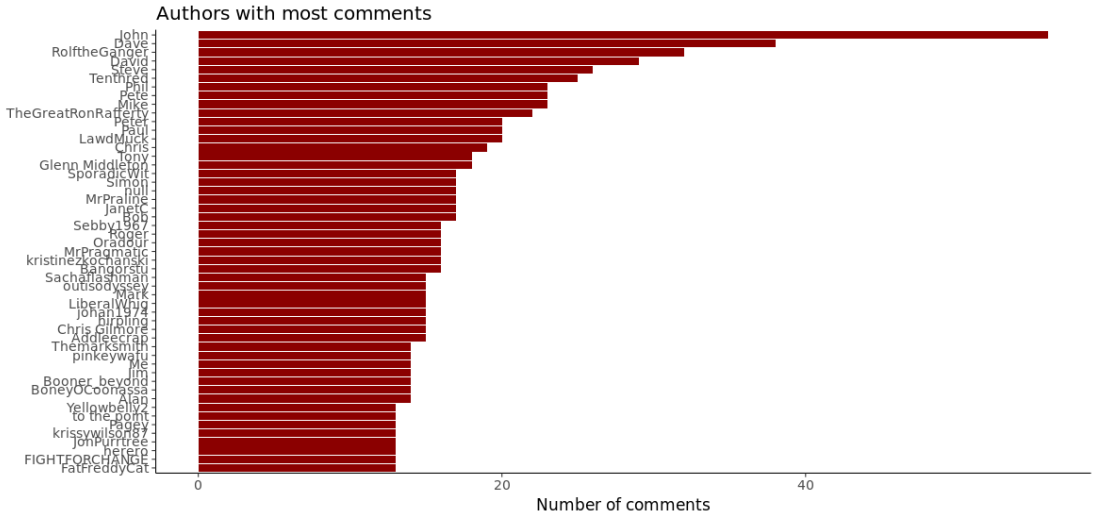
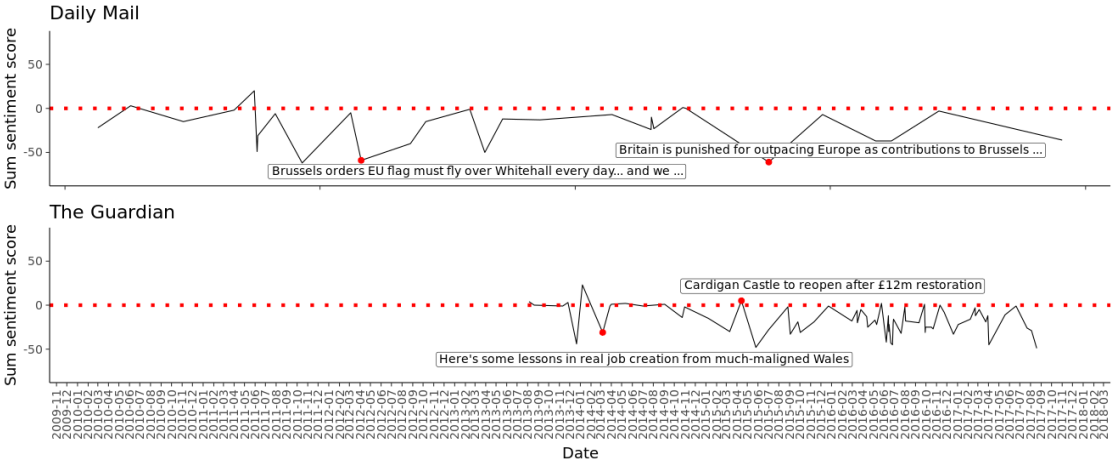


Figure 29: Example of UGC sentiment from United Kingdom news media sources.



In the next stage, we reorder the articles by the number of comments and plot the results in Figure 30 . The comments and articles they relate to are grouped by different colours to reflect the topic label assigned from the STM model. It is quite clear how to topics in particular, generate the major share of commentary, the topic broadly defined as "Brexit" and the "Irregularities/Budget" topic.

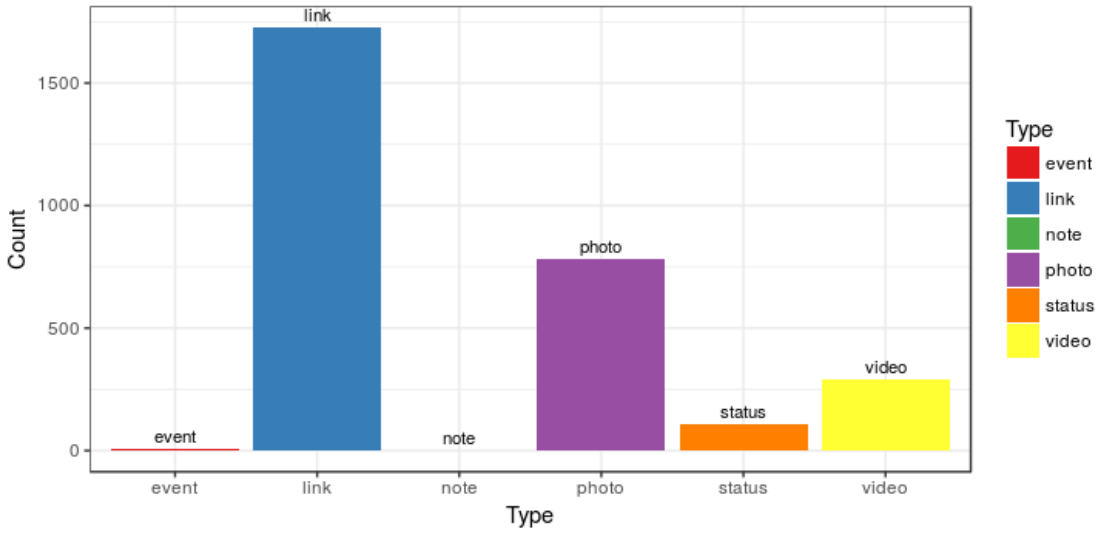
Figure 30: Articles with the highest number of comments from UK news media. Colours reflect the topic label from our STM model.

Social Media

In this subsection we move from comments collected in articles from the different news media sources to the content and attributes of posts and tweets from Facebook and Twitter respectively. Note that our analysis of social media is conducted by language instead of geographical area. This is because OSNs do not provide specific geo-location. For instance, tweets collected from a given English keyword could belong to a user in the UK, Ireland or another EU country. For our analysis of both languages we discuss the results from the two OSNs in parallel.

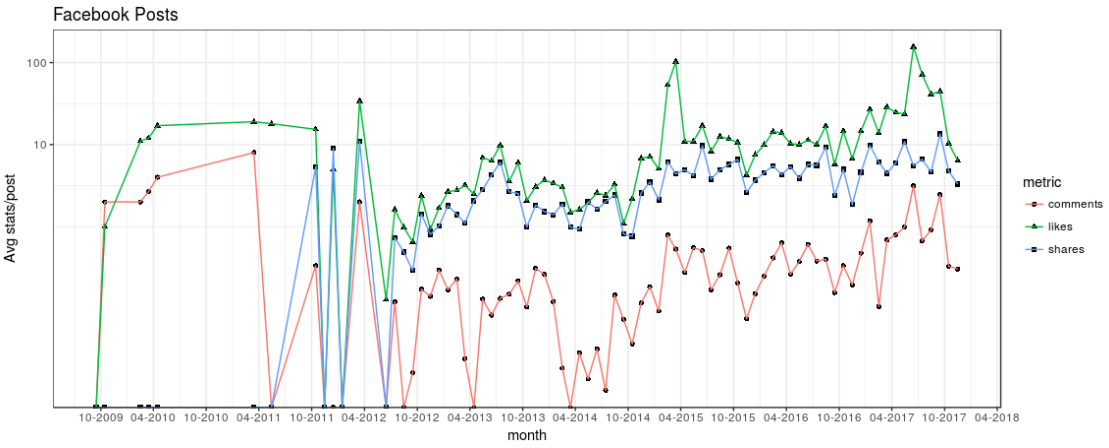
To illustrate the diversity of content in social media, in Figure 31 we show the different types of posts collected from Facebook related to EU Cohesion Policy. We can observe that most of posts belong to the category "link", which means the post contains a link to an external sources (e.g. news media source). It is worth noting that posts of the type "photo" were not filtered in our pre-processing phase because, in addition to the photo, the post contains a descriptive paragraph related to EU Cohesion Policy. The same approach was applied to the "video" category and other similar visual content.

Figure 31: Overview of posts' types from Facebook data.



We can probe additional meta-information related to Facebook posts by examining the evolution of the average stats per post (e.g. comments, likes, shares) of our Facebook dataset (see Figue Figure 32). Taking into consideration that during the earlier years (roughly between 2009-2011) Facebook pages related to EU Cohesion Policy barely existed, we can observe since 2012 a general increase in the trend towards interacting with topics related to Cohesion Policy, as exhibited by the rising number of comments, likes and shares. At the same time, it is also clear that users are more likely to share or click-like the content than to comment on it.

Figure 32: Timeline showing evolution of a number of Facebook summary statistics.



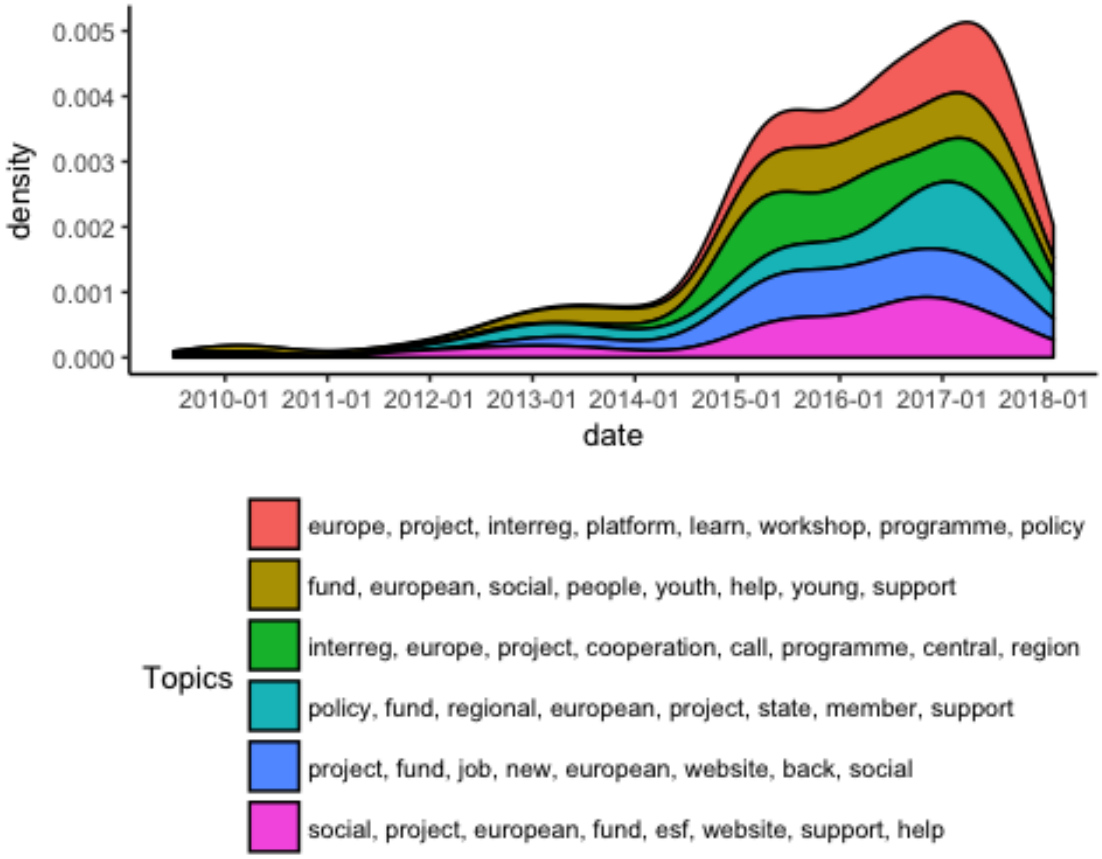
In the following subsections, we discuss the main findings from an analysis of Facebook and Twitter sources in the two languages, English and Spanish. Here we follow the same procedure: we first run an exploratory LDA model on the posts (or tweets) and then compute the sentiment.

English

Facebook

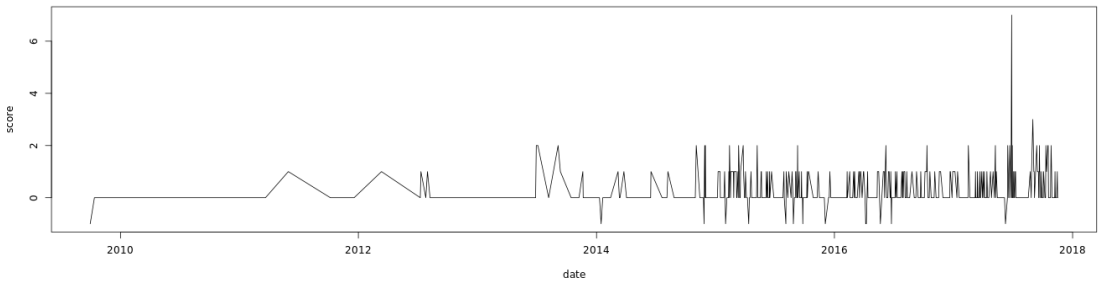
Figure 33 shows the evolution of the topics. As commented previously, we observe a low volume of content in the first years. However, since around 2014 the density has increased dramatically across all topics. Regarding the detected topics, the first topic (red) is defined by words such as program or workshop related to European projects. Secondly, aid for young people is the main topic covered in topic 2 (yellow). Finally, in the last topics we highlight words like website or social. A more exhaustive analysis would need to be carried out to specify the coherency of the remaining topics.

Figure 33: Density distribution LDA modelling over Facebook posts content (English case).



Moving onto the evolution of sentiment over time, as shown in Figure 34, we find that most of the posts can be considered neutral. This is rather intuitive and a closer inspection reveals that most of the posts are of an "informative" nature and therefore contain mainly objective statements. Interestingly, for those posts that do exhibit an element of subjectivity, these clearly have a positive sentiment bias. Indeed, there are relatively few peaks in the negative area.

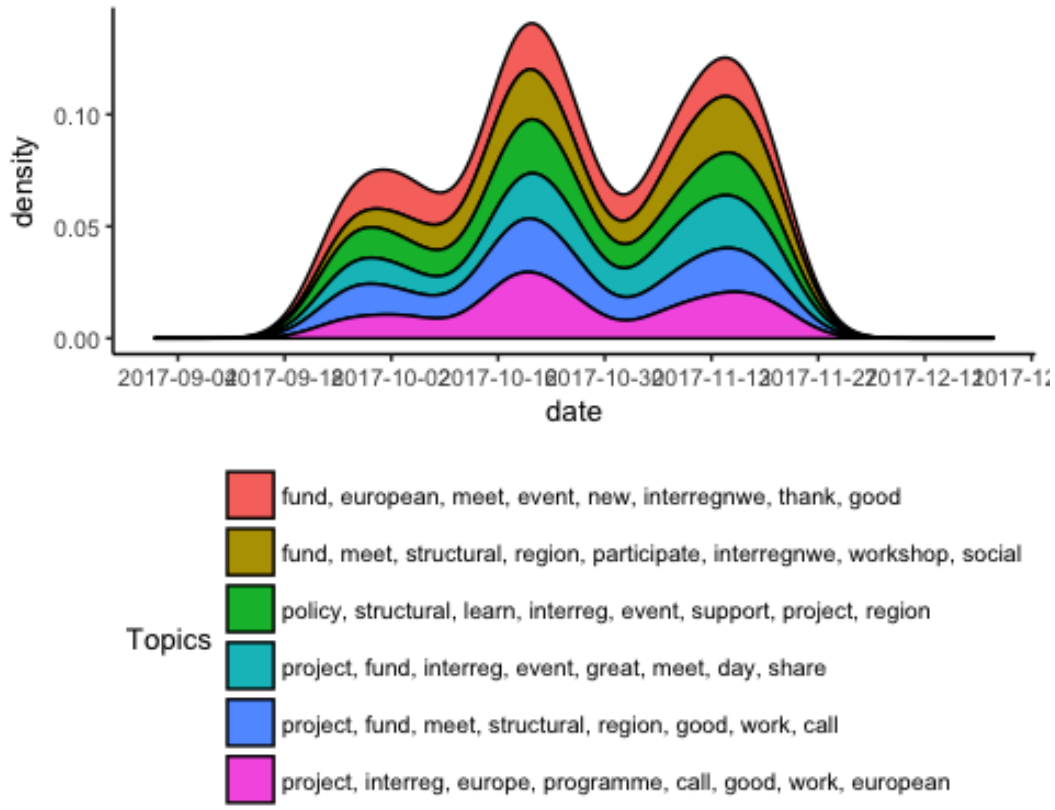
Figure 34: Facebook sentiment analysis (English case)



Twitter

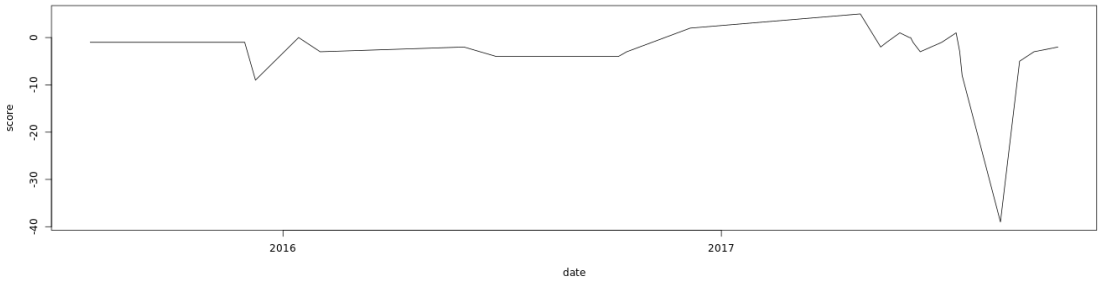
The results of the LDA modelling of Twitter content in English is shown in Figure 35. It is important to take into account that, given the constraints in acquiring historic Twitter data, our window is a rather small one of approximately six months of Twitter activity. Nonetheless, we can already observe two peaks in the density distribution mainly provoked by the last topic in (purple) for the first peak and topics 2 (yellow) and 4 (light blue) for the second one. Among the keywords shown, we can see the topic 2 contains specific keywords such as "participate" or "workshop" while topic 4 makes reference to the keyword "meet" all of which suggest that the increment seemed related to the announcement of a event(s) or workshop(s) related to EU Cohesion Policy.

Figure 35: Density distribution LDA modelling over Twitter content (English case).



Regarding the sentiment analysis, we plot the results in Figure 36. It shows that most of the sentiment of the tweets analyzed are neutral -as appeared to be the case for Facebook. Although overwhelmingly neutral in sentiment, there are nonetheless a few punctuated negative peaks where the feeling is clearly negative. For example, the abrupt fall at the end of the figure (end of 2017) coincided with a large number of negative tweets. Further exploratory analysis would need to be conducted to identify the reason for this increased negativity.

Figure 36: Twitter sentiment analysis (English case)

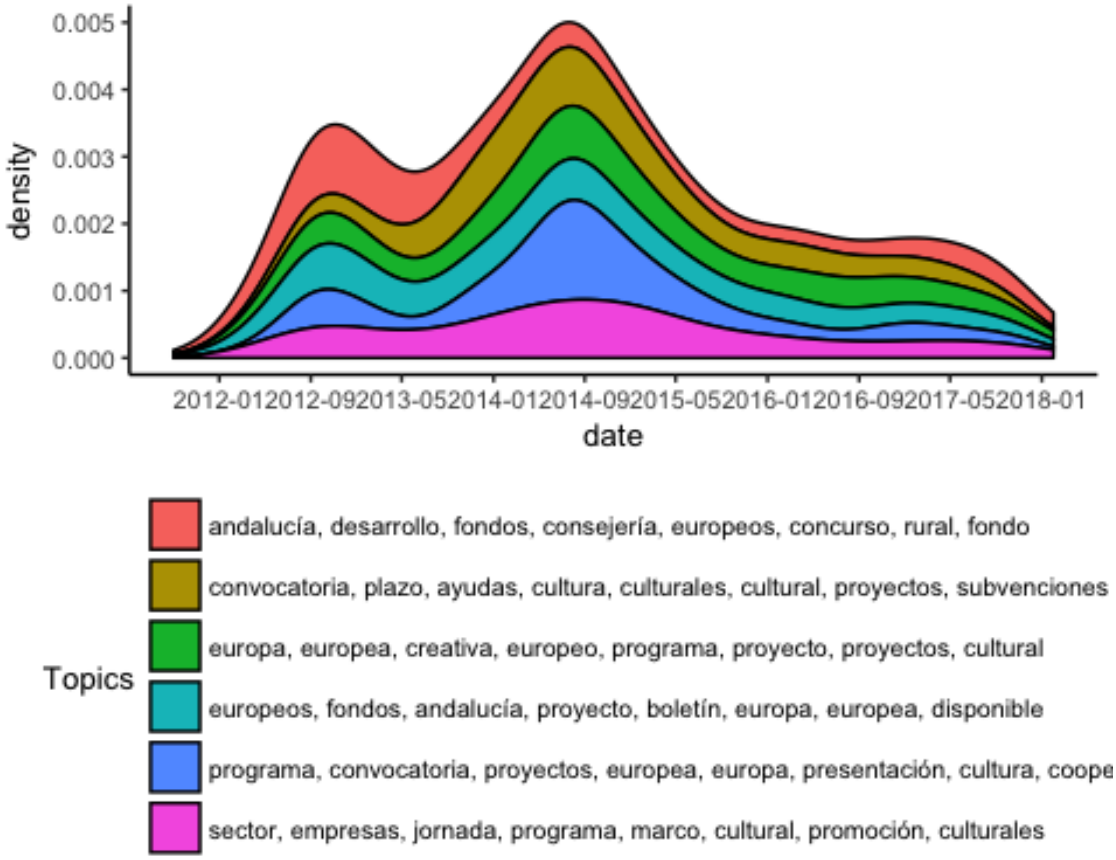


Spanish

Facebook

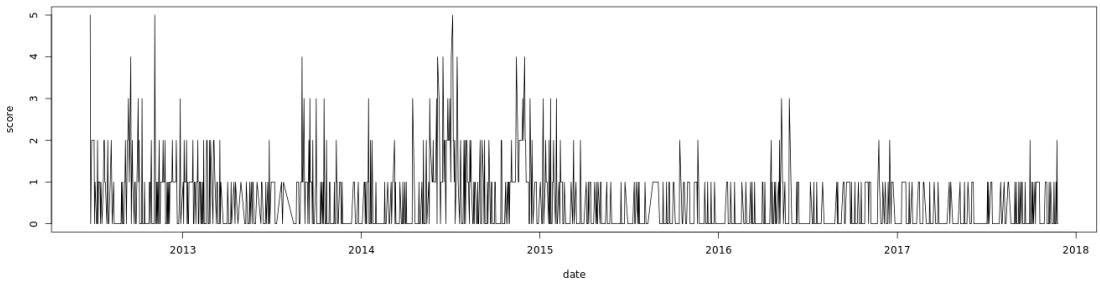
As with the English case, we begin by running a LDA model (see Figure 37) on the Facebook sources. The main difference with the English case is the higher density of posts during 2014 where topic 5 (dark blue) had a significant increment over several months. This topic contains words such as "programa" (program) or "convocatoria" (call) and seemed to be mostly linked to a call for EU co-funded projects. Also, we can see how topic 2 (yellow) contains the words "convocatoria" (call), "subvenciones" (subsidies), "ayudas" (grants) or "plazo" (time limit) which are clearly related to European grants.

Figure 37: Density distribution LDA modelling over Facebook content (Spanish case)



Similar to the English Facebook case, a sentiment analysis for the Spanish Facebook posts reveals a mostly positive bias (see Figure 38). Our hypothesis for this mostly positive association is that the sources of the post activity are mostly official or semi-official Facebook groups with a connection to the policy community dealing with Cohesion Policy. This is clearly the case for the Spanish Facebook groups analysed.

Figure 38: Facebook sentiment analysis (Spanish case)



Twitter

Turning to the Twitter data, and bearing in mind the short window of activity analysed - the last six months - we first plot the results of the LDA modelling for the Spanish language case as can be seen in Figure 39. A manual inspection of the keywords in the different groups did not reveal any clear clustering of themes. We believe that this is probably due to the limited time window. Although the volume of Twitter data might be relatively large in volume we have insufficient historical data to track the evolution of topics over time at this stage. Nonetheless, it seems that topics 1 and 4 (red, light-blue) are about research. A more detailed analysis should be conducted after augmenting the dataset to cover at least a minimal two or three-year time span.

Turning to the sentiment analysis applied to the Twitter data in Spanish. A slightly more positive picture than the UK case, which was punctuated with some negative sentiment even though the majority was mostly neutral. In the Spanish language case, most of the sentiment was positive as can be seen in Figure 40. While this finding holds for now further analysis using different sentiment approaches (e.g. dictionaries) should also be conducted to evaluate how robust these findings are using different approaches. It may be that the model tested is insufficiently optimized for short texts in Spanish and generates bias towards positive sentiment.

Figure 39: Density distribution LDA modelling over Twitter content (Spanish case).

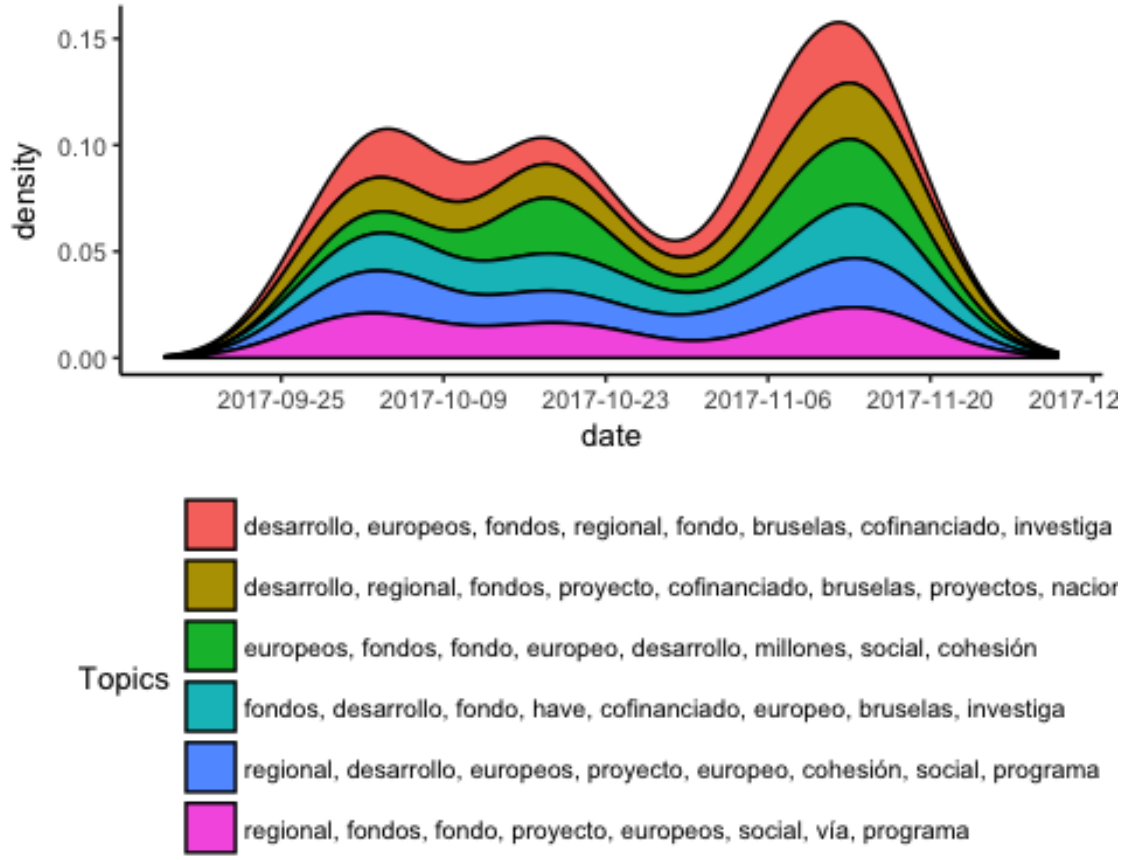
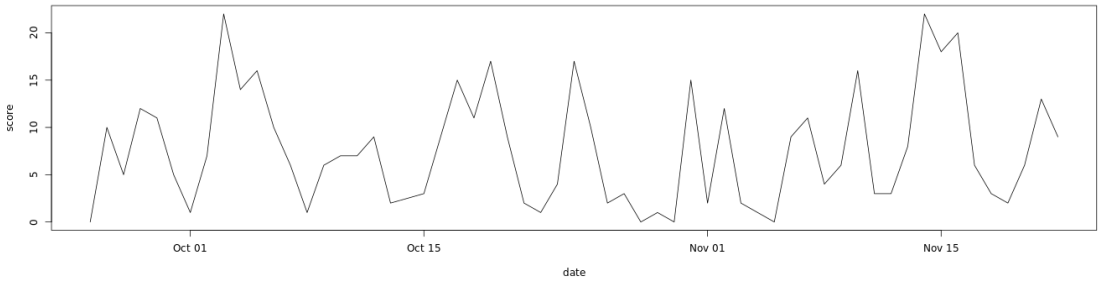


Figure 40: Twitter sentiment analysis (Spanish case)



Conclusions

This research report has analysed Cohesion Policy in the mass media by applying Computational Text Analysis to EU-wide media in English and Spanish. In this concluding section, we discuss the key findings across the three media sources: news media, user generated content and social media.

News media

The news media analysis filtered our news corpus to identify articles that had a high relevance to EU Cohesion Policy (approximately 4,000 news articles) and applied a structural topic model to uncover the most prevalent topics/issues discussed. We found a relatively high degree of topic convergence in both Spanish and English in terms of the main topics discussed. The degree to which topics were emphasised across different media obviously varied when further disaggregating the data. Thus, for English we distinguished between UK media and Transnational media (i.e., EU sources in English and International press such as The Economist) while for Spanish language we only included media from Spain.

The topic convergence at the broader level largely dovetail Cohesion Policy thematic objectives/priorities. We found that major topics discussed included low-carbon economy, R+D and Innovation, Employment and Training, Entrepreneurship, Transport and Infrastructure, and Local Development and Cultural Heritage. We also found that Cohesion Policy was frequently mentioned in connection with broader EU political themes. In all cases, broader themes were uncovered which tended to bundle EU affairs (primarily articles about the Eurozone crisis or the Migration crisis); EU budgetary politics related to issues such as Conditionality, and a separate topic related to Spending Irregularities.

Our analysis controlled for the level of territoriality in which media sources were rooted. We distinguished between regional and national sources when comparing the UK and Spain and between EU web native sources and International press when looking at the transnational media. Statistically significant differences in the estimated proportions of topics discussed emerged when controlling for territorial level. By and large, there were different foci across the levels. The national level media (as well as the International media) tended to focus on higher level issues such as EU affairs, budgetary politics and Irregularities while the regional media's focus was more congruent with the EU's programmatic priorities and discussed topics such as Energy, Environment, Local Development and Cultural heritage.

A further exploratory sentiment analysis (on the English content) also revealed that international media tended to use more negative sentiment compared to the EU web-native media. For the country cases, a similar pattern emerged. However, this time it was the national media that used more negative sentiment whereas the regional level sources, on balance, used more positive words in documents related to the EU's Cohesion Policy.

User-Generated Content (UGC)

We refer to UGC as the comments associated to articles from News Media. We collected more than 33,000 comments from 11,000 distinct users. We observed a very active

community of users in the news media analysed with hundreds of users with more than 25 comments.

It is important to note that most comments of UGC were not directly associated to the EU Cohesion Policy. Of the total, we identified 457 comments (1,4 percent) which we labelled as significantly related to our topic of interest. The volume was more limited in absolute numbers for the Transnational media, although it contained a higher ratio of relevant comments (nearly 4 percent of the total). In short, the highest volume of the UGC was from Spanish and UK media sources.

A further exploratory sentiment analysis revealed that these comments are mainly neutral or positive in the case of Spain. However, UGC from United Kingdom showed a much higher bias towards negative opinion -this was not only the case for the Eurosceptic source analysed (Daily Mail) but also the case for the nominally more pro-European source (The Guardian).

Social Media

Social Media analysis was performed at the language level over the two main Online Social Networks (OSN), Facebook and Twitter. We analysed more than 3700 posts and 19500 tweets from Facebook and Twitter, respectively. In addition, we collected meta information (i.e. such as comments and reactions) generated by these social media sources. This increased the numbers to 63,000 for Facebook and more than 52,000 for Twitter. In terms of unique users posting in both OSN we had 2321 users for Facebook and 13298 in Twitter, which represents an average of 1,55 posts/user in Facebook and 1,47 posts/user for Twitter.

Applying sentiment analysis revealed that by far the majority of posts and tweets do not express any opinion (i.e., they are mostly objective statements such as ones that provide information about, say, an event or a call). Ultimately, the largely neutral/positive sentiment associated with the social media sources analysed is not too surprising since these social media groups were mainly composed of official channels or policy communities with some connection to Cohesion Policy.

These findings have important policy implications. First, we conclude that a territorially targeted media strategy is needed to improve public understanding and appreciation of Cohesion policy, given the significant variations in topic coverage and sentiment across media at different territorial levels. Second, the Cohesion policy community needs to engage in more emotive and topical social media activity and two-way exchanges in order connect with citizens, instead of circulating factual information that does not encourage public engagement. Finally, we have demonstrated the benefits of applying computational text analysis techniques to analyse big data in Cohesion policy. EU and national institutions with responsibilities for evaluating the impact of Cohesion policy communication strategies on the mass media could benefit from applying similar techniques.

References

Banducci, S. A., & Semetko, H. A. (2003). Media and mobilization in the 1999 European parliamentary election. In M. Bond (Ed.), Europe, parliament and the Media (pp. 189–204). London: Federal Trust for Education and Research.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (Jan): 993–1022.

De Vreese Claes H, H. A. S., Jochen Peter (2001) 'Framing Politics at the Launch of the Euro: A Cross-National Comparative Study of Frames in the News', Political Communication 18(2): 107–122.

De Vreese C.H. and Boomgaarden Hajo G. (2006) 'Media Effects on Public Opinion about the Enlargement of the European Union', JCMS: Journal of Common Market Studies 44(2): 419–436.

De Vreese Claes H. and Kandyla Anna (2009) 'News Framing and Public Support for a Common Foreign and Security Policy', JCMS: Journal of Common Market Studies 47(3): 453–481.

Giebler, H., Kritzinger, S., Xezonakis, G. and Banducci, S. (2017) 'Priming Europe: Media effects on loyalty, voice and exit in European Parliament elections', Acta Politica 52(1): 110–132.

Grimmer, Justin, and Brandon M Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–97.

Hepp, A. et al. (ed.) (2016) The communicative construction of Europe: cultures of political discourse, public sphere and the Euro crisis, Basingstoke: Palgrave Macmillan.

Hoffman, Matthew, Francis R. Bach, and David M. Blei. 2010. "Online Learning for Latent Dirichlet Allocation." Edited by J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta. Curran Associates, Inc., 856–64. http://papers.nips.cc/paper/3902-online-learning-for-latent-dirichlet-allocation.pdf.

Musto, Cataldo, Giovanni Semeraro, and Marco Polignano. 2014. "A Comparison of Lexicon-Based Approaches for Sentiment Analysis of Microblog Posts." *Information Filtering and Retrieval* 59.

Pérez, F. S. (2013) Political Communication in Europe, London: Palgrave Macmillan UK, available at http://link.springer.com/10.1057/9781137305138 (accessed August 2016).

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58 (4): 1064–82.

Röder, Michael, Andreas Both, and Alexander Hinneburg. 2015. "Exploring the Space of Topic Coherence Measures." In *Proceedings of the Eighth Acm International Conference on Web Search and Data Mining*, 399–408. ACM.

Topix. 2017. "Automated Topic Discovery: A Tutorial." https://topix.io/tutorial/tutorial.html.

Vliegenthart, R., Schuck, A. R. T., Boomgaarden, H. G., Vreese, D. and H, C. (2008) 'News Coverage and Support for European Integration, 1990–2006', International Journal of Public Opinion Research 20(4): 415–439.