The Immigration Issue in the European Electoral Campaign in the UK: Text-Mining Public Debate from Newspapers and Social Media*

Paul Nulty
paul.nulty@gmail.com

Monica Poletti
m.poletti@lse.ac.uk

June 26, 2014

Abstract

In recent years, the issue of immigration has become increasingly salient in the UK political and media debate. Moreover, with the development and persistence of the economic and financial crisis within the EU, immigration has been linked to growing opposition and criticism towards the European Union. In a country in which Euroscepticism has historically been high compared to countries in continental Europe, EU immigration-related statements connected to EU free-border agreements became more widespread. For this reason, we expect immigration to be a prominent issue in the electoral campaign of the upcoming 2014 European Parliament elections in the media. By covering (potential) EU immigrants and EU immigration issues in a certain way, media tend to promote or restrain certain ideas of immigration, that might eventually affect public's views. In fact, we know from previous studies that immigration, particularly in times of economic crisis, is a challenge for society that can be framed not only in positive or negative terms, but also in economic or cultural terms.

By looking at the news coverage of several newspapers, selected according to their political orientations and including both broadsheets and tabloids, this study first considers the salience of coverage of EU immigrants and EU immigration issues in UK newspapers in the three months preceding the EU elections of May 2014. It further explores whether news coverage of different newspapers is framed in terms of economic (e.g. jobs) or cultural (e.g. identity) terms. In addition, we mine information from social media to discover how the immigration debate is framed by politically engaged members of the public on these platforms. Although inferring public opinion from social media can be problematic, it is a rewarding approach given the volume of public text available, and the ability to identify political affiliations through network connections. Understanding representation of immigration and its connections with public opinion is crucial not only in order to contribute to the scholarly debate on anti-immigration and Euroscepticism attitudes, but also to better inform public policy decisions.

*Author addresses: Department of Methodology, London School of Economics and Political Science, London WC2A 2AE. Paper prepared for presentation at the 'Text Mining and Public Policy' workshop at LSE, 27th June 2014. This paper describes work in progress and is not intended for citation. The latest versions of the code, the manuscript, and supplementary materials are available at: https://github.com/pnulty/LSEtext. This research was supported by the European Research Council grant ERC-2011-StG 283794-QUANTESS.

1 Introduction

In recent years, immigration policy-related issues have become increasingly salient in the UK public debate. This is partly connected with the development and persistence of the initially worldwide and then specifically European economic and financial crisis. Although European countries have been affected in different ways by the crisis, unemployment has been increasing virtually everywhere in Europe, including the UK. In such a context, in a country in which Euroscepticism has historically been quite high compared to other countries in continental Europe, it was relatively easy to establish a connection in the British public debate between restrictive immigration policy positions and growing opposition and criticism towards the European Union. EU immigration-related statements connected to reconsideration of EU free-border agreements to avoid other Europeans to come and work in the UK have lately become more and more widespread, particularly targeting immigrants coming from historically poorer economies such as Eastern European countries, or from countries badly hit by the crisis such as the Southern European ones. This somewhat "economic" debate added up and strengthened a pre-existing "cultural" debate about the strains and failures of UK multiculturalism and the impact of different ethnic and religious minorities values or practices (e.g. Islam) on British society.

Connection with EU issues seems to have allowed strongly anti-immigration policy statements to break through the barriers of right wing extremism and to become mainstream, together with an increasing stress in the public debate on defence of British jobs and British values. Evidence of the fact that restrictive immigration policy preferences became mainstream not only in the public debate but also among the public emerged quite palpably from the results of the 2014 European Parliament elections in the UK. The success of the previously minority right-wing populist UK Independence Party (UKIP), that based its whole electoral campaign on promoting more restrictive immigration policies (also within the EU) and possibly on leaving the EU, was unprecedented. "*26 million people in Europe are looking for work. And whose job are they after?*" was undoubtedly one of UKIP most famous poster of the electoral campaign. Although more culturally-oriented statements were also heard.

Having turned into a particularly prominent issue in the media electoral debate of the EP

elections, we choose the immigration-policy related debate as focus of our study. In any given policy area it is key to understand how the public debate is shaped and what its dynamics are. With the goal of gaining support for their policy propositions, politicians selectively emphasize features of an issue in ways that most favour their position (Riker, 1986; Schattschneider, 1960). They for instance highlight likely effects of a specific choice or its connection to important values (e.g. Jacoby, 2000, p.75). By reporting policy debates, media allows the frames of issues made by political actors to reach a wide public (Entman, 2004), and this allow citizens to make up their mind among competing arguments. Media, however, are also using frames. In covering specific issues in a certain way and in the modality of reporting statements from different politicians, different media tend to promote or restrain different ideas and policy proposals, that might also eventually affect public's views and behaviour. This is why it is important to investigate which frames are used more regularly in the political and media debate and in which media outlets they tend to be predominant.

Immigration is particularly a challenge for society in times of crisis, and the framing of immigration policy-related issues debate in the public discourse has more immediate consequences at elections time. By looking at the news coverage of several newspapers, selected according to their political orientations and including both broadsheets and tabloids, this study first considers the salience of coverage of different immigration policy-related issues in UK newspapers in the three months preceding the EU elections of May 2014. It further explores whether news coverage of different newspapers is framed in terms of economic (e.g. jobs) or cultural (e.g. identity) terms.

In addition to these substantive contributions, this study employs advanced text-mining methods for data collection and data analysis in a field in which these have been rarely used. Text from newspaper articles and social media is analysed using topic modelling to investigate the salience of sub-issues in the immigration debate.

Mass media play an important role in how the public understands and reasons about contentious social and political issues, such as immigration. They reflect and contribute to policy-issues related debate in several ways, and their role is even more important at elections time, when what issues and how the media decides to report them is particularly influential on

public opinion. Agenda-setting and framing effects might be crucial in this respect. Agenda-setting effects refers to the ability to influence how salient a topic is at a given point in time (McCombs Shaw, 1972). If an issue is covered frequently and presented as a prominent news, the public will tend to regard the issue as more important. Agenda-setting is the relative prominence given to issues/subjects (first level) or attributes of issues (second level) in the media (McCombs, 2005; Ghanem, 1997).

Framing is similar to agenda-setting in the sense that they both focus on public policy issues in the news. In addition to agenda-setting, however, framing analysis also looks at how this specific issue is talked about in the news. Any specific topic can be viewed in several ways and can be portrayed by a variety of perspectives. Stressing a particular interpretation or a particular feature of an issue makes alternative interpretations or features less prominent, and a news frame is precisely "an emphasis in salience of certain aspects of a topic" (De Vreese, 2002). Framing the issue in certain ways might not only inadvertently bias the public exposed to the frame toward a specific stance about immigration or towards certain immigration policy preferences (Bennett, 2001; Liu Kirkwood, 2006), but it could also eventually affect peoples political behaviour in the Election Day (Iyengar, Peters, Krosnic, 1984). Allen and Blinder (2013) investigate framing of immigrants in the UK media with corpus linguistic methods.

Strong frames are not necessarily "true" arguments or intellectually better ones. They can be based (and often are) on exaggerations and lies, and they can feed publics prejudices and fears. In the same way as some of the most common sources of anti-immigrant attitudes among people are related to material interests and national identity (e.g., Citrin et al., 1997; ORourke and Sinnott, 2006; Hainmueller and Hiscox, 2007; Lahav, 2004), some of the most common ways to portray anti-immigration discourse in the media seem to be by focusing either on material/economic interests or on national/cultural identity. Immigrants are therefore often portrayed either as a threat to the material world ("immigrants take British people job", "immigrants are a burden to the national social welfare system", "immigrants cost us too much money", etc.) or as a threat to their symbolic world ("immigrants take control of the UK", "traditional British values are getting lost due to immigration", "immigrants cannot even speak English properly"). The same economic and cultural dimensions are used for presenting them

as a resource. Immigrants can be portrayed as an economic advantage for the country ("immigrants do the jobs that British do not want to do", "immigrants makes the NHS more sustainable because they pay taxes") or as a cultural enrichment for the country ("immigrants enrich our culture", etc.).

Although framing effects might depend on frequency of exposure to a specific frame, perception of credibility of their sources (e.g. Druckman, 2001) and from individual characteristics of the receiver such pre-existing values (e.g. Druckman, 2001; Barker, 2005) or knowledge (e.g. Kinder and Sanders, 1999; Miller and Krosnick, 2005), in general framing effects literature suggests that if immigrants were predominantly portrayed by the media elite as a threat to the UK, than many British people will be more likely to support negative attitudes towards immigrants. Consequently they will tend to vote for anti-immigration party as well as to support policies that seek to limit the amount of immigrants entering the UK and that either exclude them from British society or that force them to adapt their values to those of the host country. Different choice and preferences would arise if immigrants were instead portrayed as a resource to the UK society .

2 Merging Different Approaches

The literature related to framing and framing effects is wide (e.g. Edelman, 1993; Goffman, 1974; Iyengar, 1991; Zaller 1992; Semetko and Valkenburg, 2000; Gross DAmbrosio, 2004; Vliegenthart, Schuck, Boomgarden and De Vreese, 2008; Chong Druckman, 2007; Entman, 2007). The implication of such a vast literature is that different types of frames have been identified (De Vreese, 2005: 53). Frames could for instance be generic, when they are general and applicable to different issues and in different contexts, or issue-specific, strictly related to a specific issue and to the context taken into account. Moreover, different strategies have been developed for identifying frames in the news (De Vreese, 2005). The first way is inductive, that is no predefined frames are used. The approach is explorative and it looks for the array of frames that could be found in the news without any a priori ideas (e.g. Gamson, 1992). It is usually very labour-intensive, therefore often based on small samples and quite difficult

to replicate (Hertog McLeod, 2001).The second way is deductive, namely texts are analysed using predefined frames that are searched for in the news. This approach requires to have well-defined a priori ideas of the type of frames that will be found in the news and well-defined codings of them. Although it is easier to replicate, and allows for larger sample to be dealt with, it still requires long time if carried out through manual content analysis.

Traditionally, in political science and in political communication, framing and saliency of specific issues have been investigated and detected either through interpretative textual analysis or through more systematic manual content analysis. More rarely, automated approaches using computer-aided dictionary search have been adopted. Although the latter have the advantage of being a much faster and reliable (i.e. replicable obtaining exactly the same results) approach than manual content analysis, they tended to have the disadvantage of being a more superficial and less flexible method. In computational and corpus linguistics, however, progress have been made both in agenda-setting and in framing analysis, using more complex algorithm than the past that ends up in giving more valid way to quickly identify frames within very large corpus of text.

The aim of the study is to look at the electoral discourse on immigration during the 2014 European Elections campaign in the UK using these latter type of techniques. First of all, in a deductive way, we look at whether immigration is portrayed using an economic or cultural frame in terms of the consequence for society. Frames could be either positive or negative. Thus, we also look at whether these frames are more focused on portraying immigrants as a threat or as a resource.

The type of frames presented by specific media outlet could be influenced by many factors. Given that one of the main role of media outlets is to report news, one of this factor is certainly the type of political discourse carried out by the political elites and the type of frames used by politicians during their electoral campaign. Given the predominance in the media by UKIP views on immigration, we expect that immigrants will be mostly reported as a threat to the UK society. We want to explore, however, whether it is actually the economic frame that is predominant (as UKIP leader stresses in order to avoid "racism" labelling of its party) or the cultural frame. Another relevant factor is Political Conservatism, which is associated with psy-

chological management of publics fears and uncertainty (Jost, Glaser, Kruglanski, Sulloway, 2003). Thus, in general, we expect conservative right-wing newspapers will tend to represent immigration issues in different ways than left-wing Liberal newspapers, with the former being more likely to frame immigration as a threat to the country. Another important factor is the type of audience the media outlet targets. We then expect that tabloids, which tend to be more sensationalistic than broadsheet, will favour the framing of immigrants as a threat.

Subsequently, we want to verify what the salience of these frames in the immigration discourse is in an inductive way. Although they are not identical processes, and although framing include a broader range of cognitive reasoning than agenda-setting, similarities can be found between the two processes, particularly between second-level of agenda setting and framing. Both focus indeed on the most salient aspects or sub-issues of the issue of interest (McCombs, 1997: 37; Weaver, 2007). In this study, we intend to use the second-level of agenda setting as a validation of the relevance of the deductive frames analysed before.

For doing this, we use what we can call a specific application of the automated-version of the agenda-setting approach. Instead of looking at the salience of immigration within the electoral campaign issues, which we already know, we look at the salience of different sub-issues of immigration within news about immigration. How is the immigration discourse portrayed in the media? What topics shape the public conversation on it? The relevance of sub-issues within the immigration discourse somehow relates to the way news are framed. By identifying which issues these are, and how they evolved over time, we could validate whether the two deductive-based frames used in the first part of the analysis, are actually the sub-topics that mostly shape the public conversation about immigration. This has the advantage of putting the frames into a wider picture and get a deeper sense of how the specific electoral debate has evolved over time.

3 Immigration Frames

We therefore consider two/four news frame to be of particular relevance:

- Framing immigration in terms of economic benefits
- Framing immigration in terms of economic threat

- Framing immigration in terms of cultural enrichment
- Framing immigration in terms of cultural conflict

In addition to framing of immigration in one of these frames, it is important to consider also the visibility of these frames compared to different portrayals of immigration in the news. In order to detect the prevalence of these frames, we need to decide upon an appropriate set of words that indicate when this frame is being applied.

4 Data Selection

Newspaper	Description	Articles
The Independent	Left-liberal	324
The Guardian	Broadsheet-left	392
The Times	Broadsheet centre	352
The FT	Financial news	207
The Sun	Popular tabloid	277
The Daily Mail	Popular tabloid	412
The Daily Telegraph	Broadsheet right	511
The Daily Express	Right tabloid	325

Table 1: Newspapers included in our study

We selected newspapers according to their circulation, including both broadsheets and tabloids, if available electronically via Factiva for the period we have studied. We also included newspapers of different ideological orientation. We also created a mapping between the names of the online and offline versions of the same newspaper, to allow us to merge these. We used Factiva's option to remove duplicates. Initially, we had 3950 articles, but we chose to restrict our analysis to national newspapers with large circulations. This resulted in a final set of 2800 documents.

Table 1 displays the names of the newspapers included in our study, along with the type of newspaper, their political orientation, and the number of articles talking about immigration that are included in the corpus of data from each newspaper.

5 Text-Mining Techniques

We describe the use of two quantitative text analysis techniques: Topic modelling with Latent Dirichelet Allocation (LDA), and frame analysis a model of word associations in a semantic space.

5.1 Topic Modelling

Latent Dirichelet Analysis (LDA)is a widely used method for identifying clusters of words which represent the topics present in a set of documents. LDA is a probabilistic model which treats each document as a mixture of topics, where each word in a document belongs to exactly one topic. In this model, topic structures are hidden variables which can be inferred from the observed variables — the distribution of words in the documents.

An accessible overview of the model can be found in Blei (2012). Topic modelling with LDA is a technique from computational linguistics which has been widely applied in other disciplines. To apply LDA to our collection of documents, we experimented with two freely available topic modelling software packages: *topicmodels* (Hornik and Grün, 2011), a topic modelling package in R; and *MALLET* (McCallum, 2002), a collection of statistical natural language methods implemented in Java. The results presented here were generated using *MALLET*.

5.2 Frame Analysis

While LDA has been widely applied to discover and track topics, there is less consensus on which method to use to discover how issues are framed in text. In this paper, we model our methodology on the approach taken by Sagi, Diermeier and Kaufmann (2013), who quantify how Republicans and Democrats in US Senate debates frame issues such as abortion and terrorism. In their analysis, Sagi, Diermeier and Kaufmann (2013) choose specific terms which indicate particular way of framing the issue — for example, the terms 'choice' and 'life' reliably indicate two different attitudes to the abortion debate. Similarly, the terms 'crime' and 'war' are representative of two different views of the issue of terrorism. Following their ap-

proach, we construct co-occurrence vectors for words we believe might be associated with particular ways of framing the immigration issue.

Once these indicative words have been chosen, an association between the issue of interest (in this case immigration) and each of the framing terms must be measured. We measure this association by calculating the similarity of the contexts in which the terms occur. Following Sagi, Diermeier and Kaufmann (2013), we use the Java software package *semanticvectors* (Widdows and Ferraro, 2008) to do this. First, we extract all of the contexts in which the target term or one of the framing terms occurs in the collection of articles — for each word this is a context window of length 50: 25 words preceding the term of interest and 25 words after.

For each term, we then have a set of context vectors. These are large, sparse matrices, where the columns are all of the words in the vocabulary of the corpus, and the row indicates how many times each of these words has occurred in the current context vector. The context vectors are summed and normalized, and the dimensionality of the resulting large matrix is reduced by singular value decomposition. The cosine similarity between the reduced, combined context vectors can then be used as a measure of association between the two words in the document collection. This approach falls under the umbrella of Latent Semantic Analysis (LSA), a model of word meaning whereby a word is considered to be a point in a semantic space constructed from the contexts in which it occurs (Landauer and Dumais, 1997), and similarities in word meaning can be measured in this space. The context may be an entire document (which enables a kind of word clustering not dissimilar to LDA), or a smaller window of words, as we have described in this section.

To compare framing across different variables, the document collection can be separated into sub-corpora according to the variable of interest. This was political party in the case of Sagi, Diermeier and Kaufmann (2013); for our analysis we are interested in how the issue of migration is framed by different newspapers. Therefore, we divided the corpus according to newspaper source and computed different context vectors for words in each of these sub-corpora.

6 Data Analysis

We used Factiva to download every document in the UK press from the 12 weeks preceding the European Elections that contained one of the following terms: *immigrant, immigrants, immigration*, or *immigrate* [1]. Factiva's bulk download allows 100 articles to be downloaded at once, which are retrieved as single .rtf files — each file contains 100 articles, with a header containing metadata such as source and date before each article. We used a Python script to split each article into a separate text file and label each one with its date and the newspaper in which it was published.

6.1 Text pre-processing

The topic modelling and semantic vectors methods operate on term-frequency matrices of words. For a given portion of text, the data input to the algorithm is a matrix where the columns are word types and the rows indicate the frequency of the word in the portion of text. The portion of text may be a whole document (as it is for LDA) or a context window around a term of interest (as it is for the semantic vectors approach). Representing natural language text in this way is sometimes called the 'bag-of-words' model, as the resulting matrix does not preserve the order in which the words occurred in the original text.

The *MALLET* package allows for a list of stopwords to excluded from the analysis. A standard list of stopwords includes semantically light function words such conjunctions and common prepositions. To this list, we appended some terms specific to our corpus — for example we needed to exclude the titles of the newspapers, the phrases 'copyright' and 'all rights reserved', and the search terms.

A common pre-processing step is to apply a stemmer to the word matrix so that words differing only in morphological stem are combined — for example, the counts for *earn*, *earns*, *earning*, and *earned* would be summed into a total for the stem *earn**. We did not apply this step for the LDA modelling as it makes the word clusters a little more difficult to read, and in any case words with the same stem tended to be assigned to the same topic.

[1] Due to an error while collecting the data, no articles from *The Daily Mirror* were retrieved.

6.2 Topic modelling analysis

The LDA process requires that the number of topics to be discovered is specified in advance. If the number of topics is small, the topics may be general (for example, there might be a single topic for economic words), but these topics might be indistinct and too similar to each other to be useful. We began by experimenting with different numbers of topics until the resulting word groups seemed to indicate distinct, fine-grained topics relating to the immigration debate. We found that specifying 30 topics gave the best results.

The word clusters for are indicated in Table 2 For topics that we chose for further analysis, the first column is a general label that we chose (after the fact) to refer to the topic. The second column indicates the portion of the entire corpus that each topic covered. The words in each topic are ordered by how much they contribute to the topic.

From the overall list of 30 topics, we chose a subset to examine in more detail. Some of the topics seem to be indistinct mixtures of function words or words with particular relevance to a single story, and we did not examine these further. Instead, we compared some of the topics relevant to cultural and economic framing of the immigration debate, both across the corpus as a whole and within particular newspapers and particular dates of publication.

Some LDA implementations require that each topic is of equal prevalence across the corpus as a whole. However, we use the hyperparameter optimization option in *MALLET* which allows for different topics to have different prominences across the corpus. Therefore, we can measure the prominence of interesting topics for the complete set of articles, shown in Figure 1. From the six topics that we chose to focus on, it seems as though economic issues were more prominent in the debate during this time than cultural issues.

The most prominent topic is an economic topic and relates to benefits claim by immigrants. The third most prominent topic relates instead to immigration and business. However, the the controversial campaign of UKIP, often labelled as "racist" by leftist newspapers, was quite prominent in the news. Also, the topic of different ethnic and religious minorities. Two prominent topics are those connected to "schools" (e.g. immigrants children cannot speak English, or teachers need more training and support to be able to deal with immigrant children at school) and challenges faced by women in different religions. Again, the salience of the topic

No.	Label	Prominence	Top five words
27	Media/Debate	0.39536	public political media social left
9		0.38159	time dont back good day
26	Numbers Debate	0.25435	uk migration year eu britain
29	EP Election	0.18485	ukip party labour voters vote election
3	Tories	0.17853	cameron minister tory david prime
5	Benefits	0.15959	tax benefits work pay money
22	UKIP Campaign	0.12362	ukip farage party racist nigel
24	Health	0.12003	care health report public government
12	Media	0.09385	news newspapers thesun english ea
13	Justice	0.09246	court year police home human
15	Family	0.09175	family children mother home parents
1	UK and EU	0.09065	eu european europe britain cameron
18	-	0.08713	britain country british world english
2	Asylum	0.08557	home office uk asylum case
6	Local issues	0.0775	london local council east south
20	Business	0.07714	business uk market prices house
25	Minorities	0.0748	ethnic cent white minority groups
19	-	0.0692	mail damonl newspapers daim uk
4	Clegg-Farage Debate	0.06529	clegg farage debate nick nigel leader
28	Scotland	0.05902	scotland scottish border uk independence
23	Schools	0.05901	school english schools students children
17	Brokenshire	0.0582	british workers minister brokenshire labour
14	-	0.05685	love sex men women book man
7	Arts	0.05039	film music theatre play story band
21	-	0.04717	police man men french found
0	-	0.04626	war book world books benn
11	-	0.04242	bbc radio tv show series street programme
10	Women+Religion	0.03672	women church muslim christian country
16	-	0.02815	gardiner mp labour asked facebook night
8	-	0.02568	wine food world huffington fish made

Table 2: All 30 topics identified in the collection of articles. The third column shows the Dirichelet prior, which approximately indicates how prominent that topic is in the corpus as a whole. The final column shows the five words that contribute most to the topic, in order of importance.

does not necessarily give a good indication of the sentiment or attitude of the article towards immigrants.

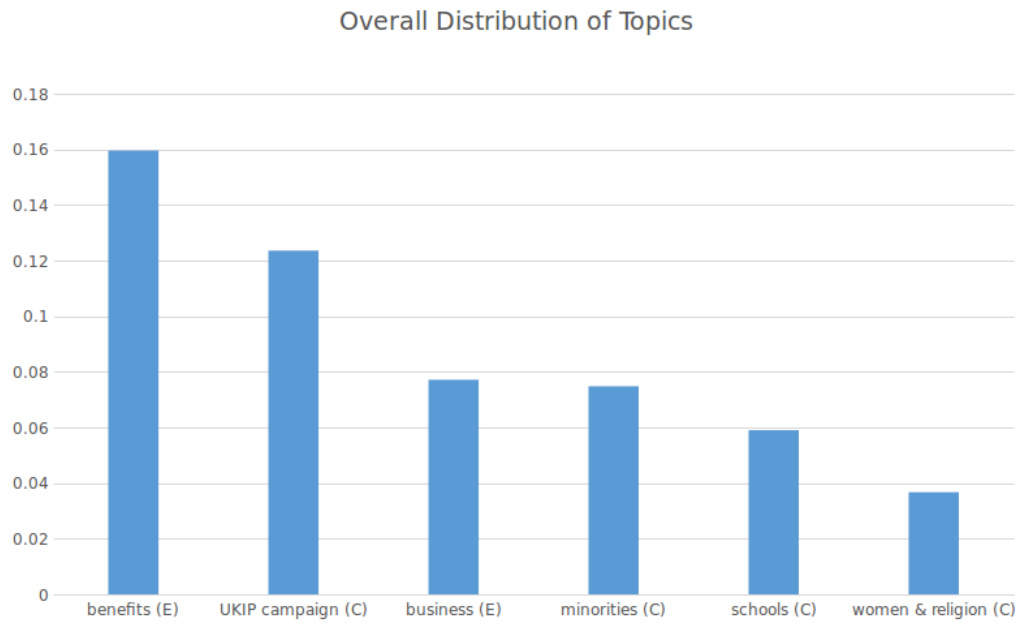

Figure 1: The Dirichelet prior for each topic across the whole corpus. This number is roughly proportional to the overall portion of the collection assigned to each topic (McCallum, 2002), but, unlike the topic-document scores, is not an exact fraction.

We then measured the average prominence of each of the topics in each of the newspapers. In the figures below, the y-axis indicates the mean proportion of an article from the newspaper indicated on the x-axis that is composed of words assigned to the topic of interest. The x-axis shows seven major daily UK newspapers, approximately listed in order of their broad ideological stance.

6.3 Economic Topics

Figure 2 shows the distribution of the 'Benefits and Welfare' topic across seven newspapers. Although it is possible for the same word to occur among the most prominent words for multiple topics, within any particular document, each word is assigned to only one topic. This means that for a given document, we can say what proportion of the document is assigned to each topic. These charts show the mean topic proportion for each newspaper. The sum of the topic proportions across all 30 topics is 1. That is, on average, 8.6% an article from *The Daily*

Express is composed of words assigned to the 'Benefits and Welfare' topic, compared to 3.6% for an article from *The Guardian*.

Figure 3 shows the topic proportions for the 'Business' topic. The papers traditionally considered in the 'broadsheet' category have higher proportions of text about this topic.

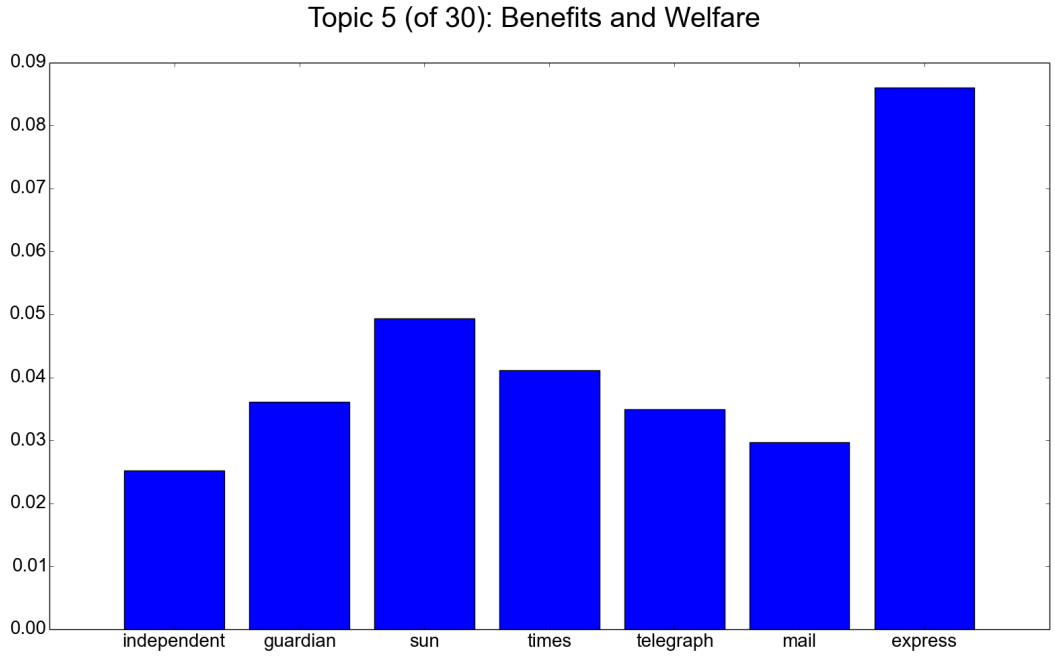

Figure 2: The benefits and welfare topic by newspaper source. The top twenty words in this topic were:*tax, benefits, work, pay, money, million, benefit, government, jobs, welfare, income, cost, taxes, state, public, budget, system, paid, pounds, billion.*

The Daily Express stands out as giving particular attention to the issue of benefits and taxation in connection with immigration. The centrist and broadsheet newspapers seem to cover the 'Business' topic more than the others.

We initially included the Financial Times in the article collection, but the results indicated that it was largely composed of the Business topic, as in figure 4.

In a way this is a useful kind of face-validation — we expect that the FT will cover business issues more than other sources. In addition, the distribution of the other papers in this model is very similar to their distribution in figure 3, despite the fact that it is a different run of the LDA model, where the topics will be assigned different numbers and composed slightly differently due to the extra articles and different values for the random initialization of the topic prior.

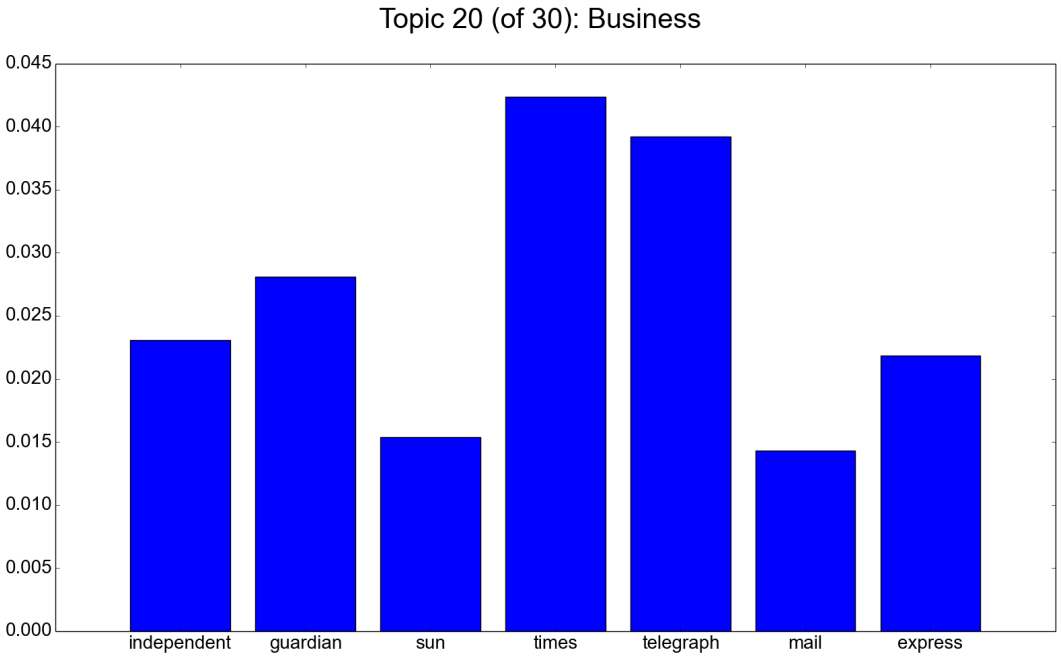

Figure 3: The 'business' topic by newspaper source, excluding the Financial Times The top twenty words in this topic were:*business uk market prices house year company economy cent growth million london investment businesses housing bank companies capital years high*

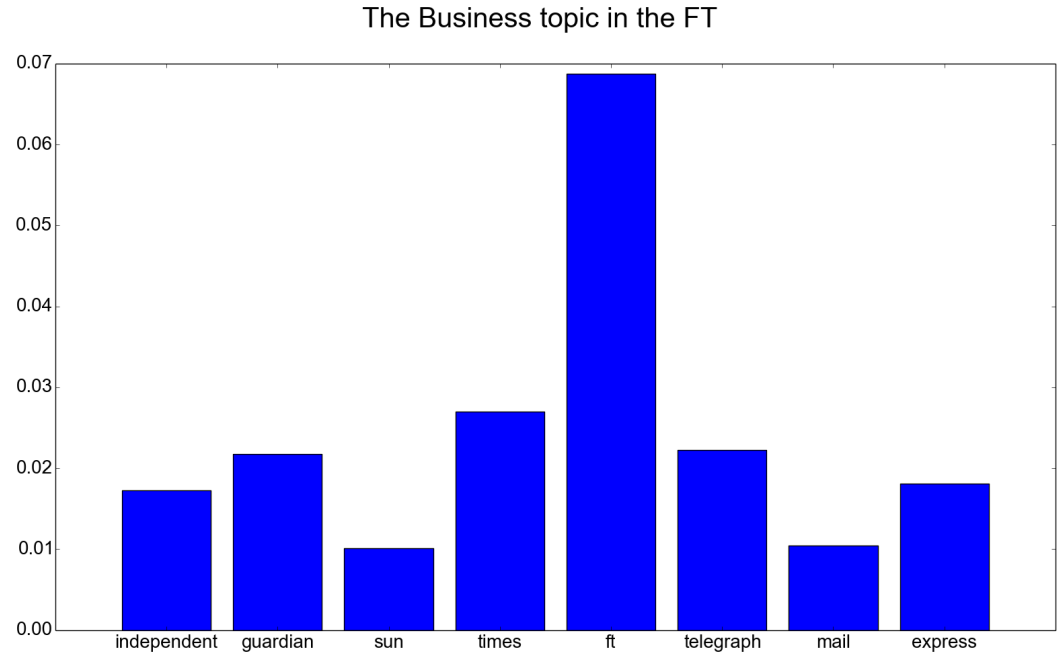

Figure 4: The *business* topic by newspaper source including the Financial Times. The top twenty words in this topic were:*business uk market prices house year company economy cent growth million london investment businesses housing bank companies capital years high*

6.4 Cultural Topics

Figures 5 and 6 show the topic proportions for the 'Minorities' and 'Women and Religion' topics respectively.

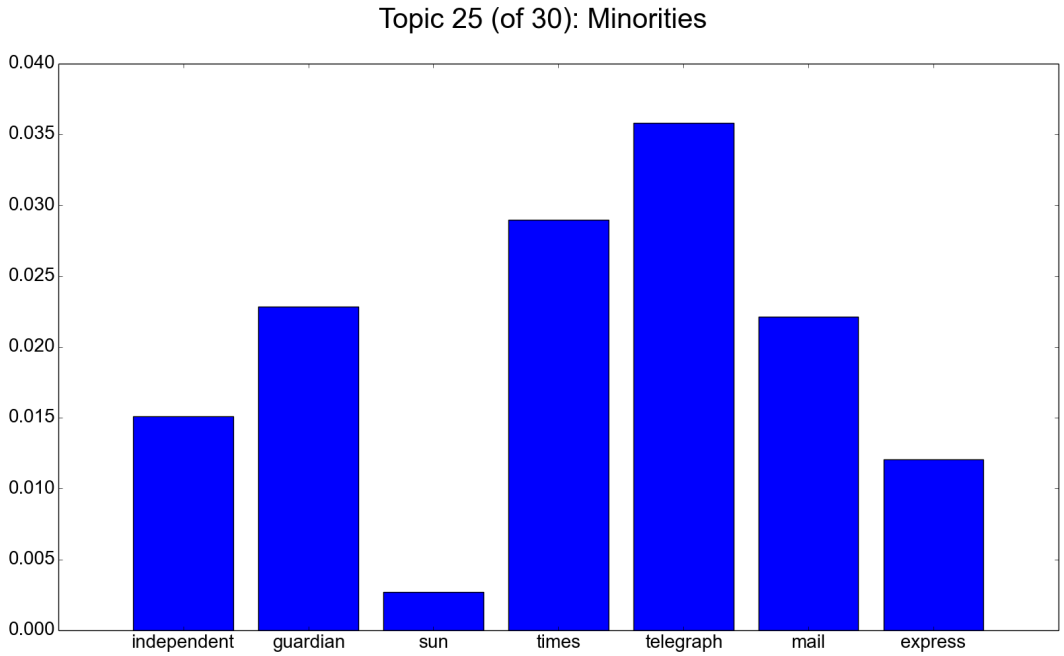

Figure 5: The *minorities* topic by newspaper source. The top twenty words in this topic were:*ethnic cent white minority groups black population study britain british minorities english university report research communities class found born online*

6.5 Topics over time

We also observed how the prominence of topics varied over time. We divided the articles according to the week in which they were published, which allows us to see the average proportion of a document that is composed of a given topic for in a give week. The x axis corresponds to the week of the year in which the article was published (we collected articles from the 22nd February to 22nd May 2014). The results for this time-series analysis show strong spikes at particular points, indicating that large amounts of the text written about a particular topic was driven by a single story or closely related group of stories.

The story-driven nature of the topic prominence makes analysis here problematic — rather than showing that a particular agenda is set and influences later coverage, we simply

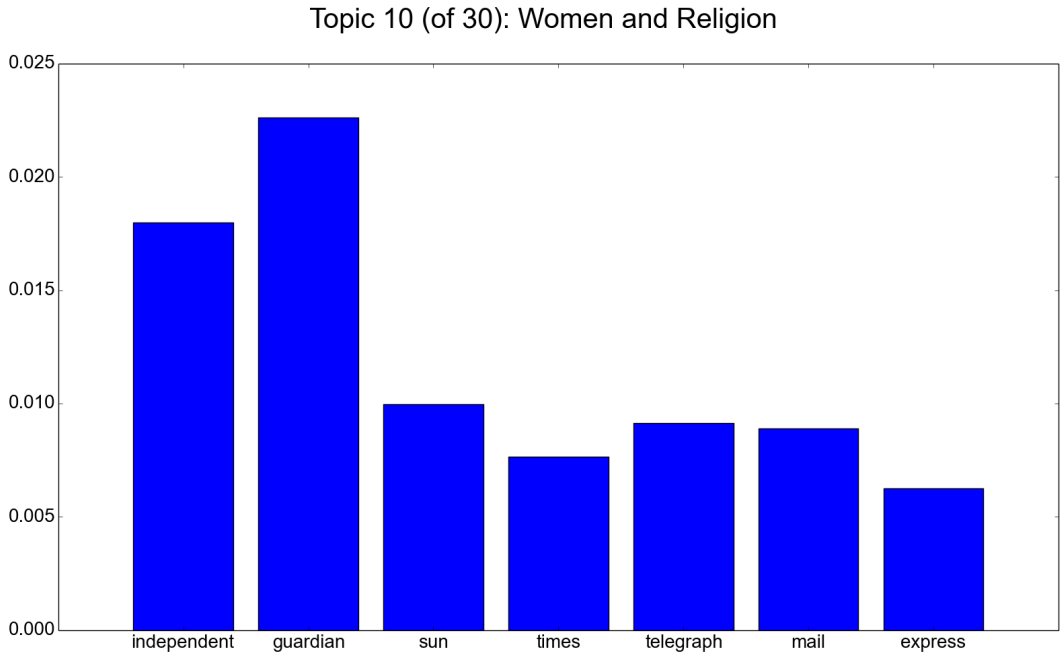

Figure 6: The *Women and religion* topic by newspaper source. The top twenty words in this topic were: *women church muslim christian country gay violence female faith muslims girls religious archbishop uk culture catholic britain england jewish countries*

see topics spike as stories break and then fall away again. However, it does serve as another confirmation of topic validity, as the timing of the rise in certain topics coincides with identifiable news events — for example, topic 17 clearly refers to two stories about James Brokenshire, the Minister of State for Security and Immigration, who generated a surge of coverage on 6th of May following his first major speech, in which he commented on the employment of cheap immigrant workers by middle-class families. This topic is shown in 7. The speech was badly received when several newspapers reported that Conservative politicians including David Cameron employed eastern-european nannies.

We can also observe a spike corresponding to the debates between Nick Clegg and Nigel Farage. This is shown in Figure 8. The rise in the last week seems to be due to this topic being incorrectly assigned some words associated with a story involving Prince Charles' comments on Ukraine, which occurred on 20th May, two days before the elections.

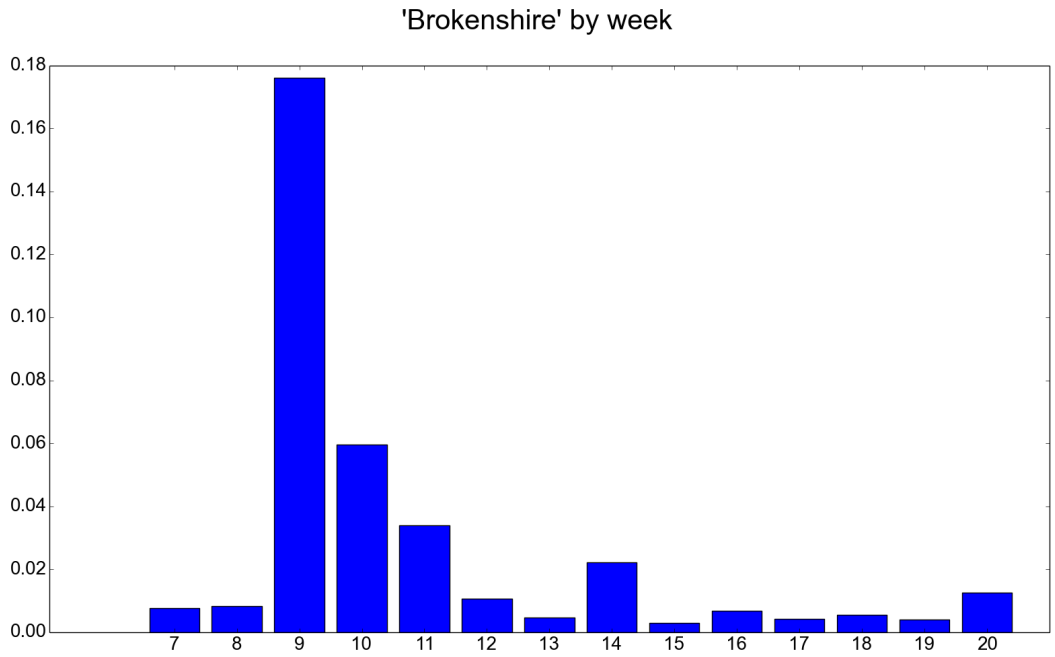

Figure 7: The topic referring to coverage of immigration minister James Brokenshire's speech on 5th March. Week number 9 in our dataset begins on 3rd March. The top twenty words in this topic are: *british workers minister brokenshire labour nanny home foreign secretary cable speech metropolitan cheap elite cameron james business vince wealthy jobs*

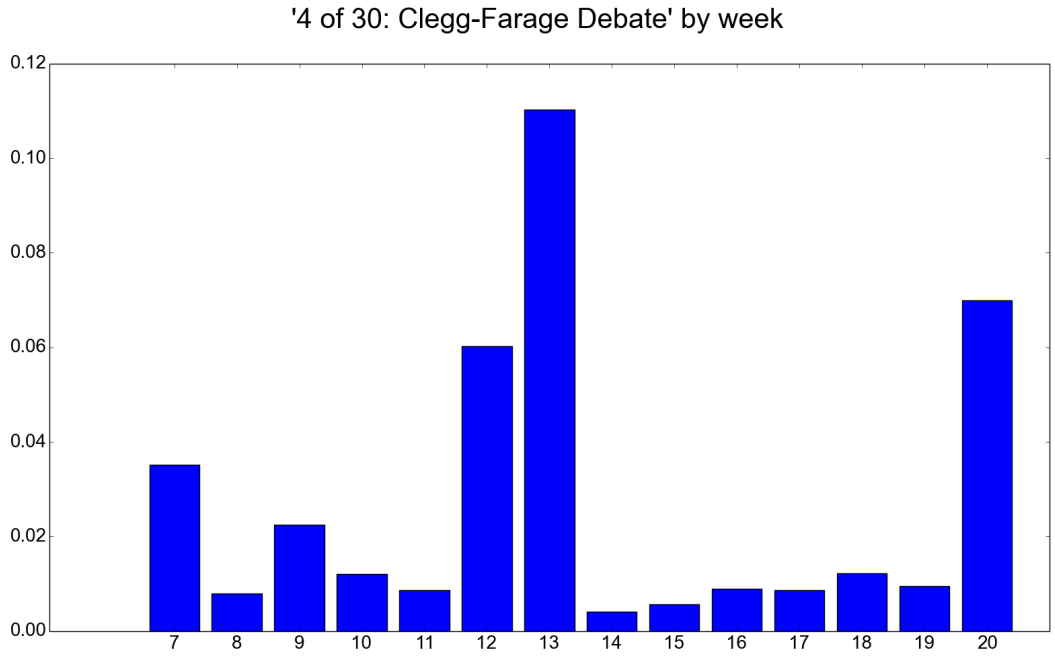

Figure 8: The topic corresponding to the debates between Nick Clegg and Nigel Farage. Week 12 and 13 corresponds to 24th March to 6th April. There were debates on 26th March and 2nd April. The words most associated with this topic were: *clegg farage debate nick nigel leader putin eu prince lib ukraine europe britain debates russian charles british european prime dem*

6.6 Frame Association Analysis

It is difficult to select terms that precisely delineate the framing of immigration issues in the same way that 'choice' and 'life' define the abortion debate. Nonetheless, we compiled a list of words to capture aspects of the economic and cultural debate around immigration in the UK. We then measured the association score for each of these words with the word stem *immigr**. For this analysis, we again removed stopwords from the term-frequency matrix, and the semantic vectors package also performs word stemming. We compared the association scores across three broadsheet, three tabloid, and a group of three broadly right/centre-right newspapers. The charts below show the results.

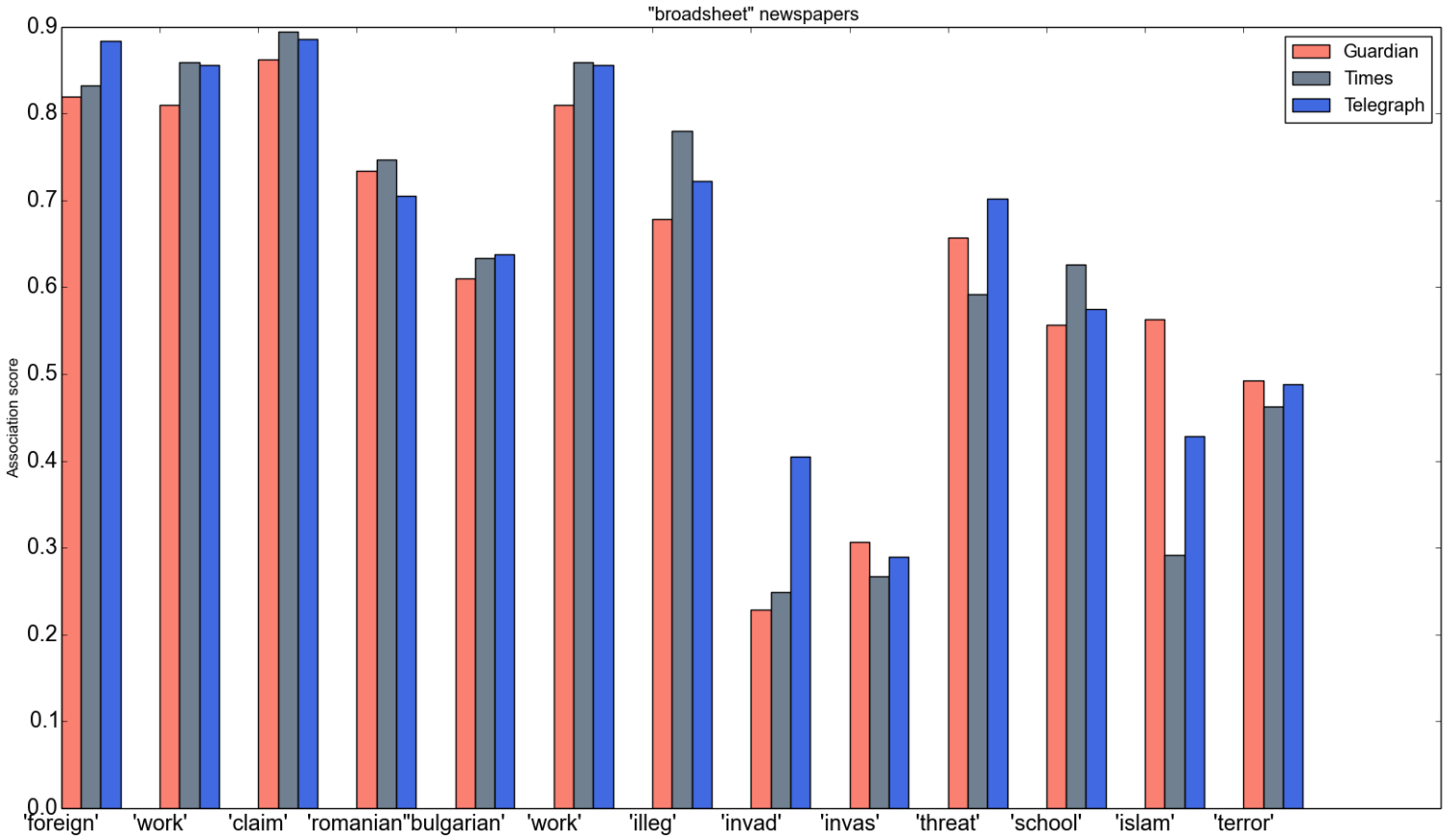

Figure 9: Association scores indicating framing of the immigration debate in broadsheet newspapers.

These results are difficult to interpret. It is not appropriate to compare the overall asso-

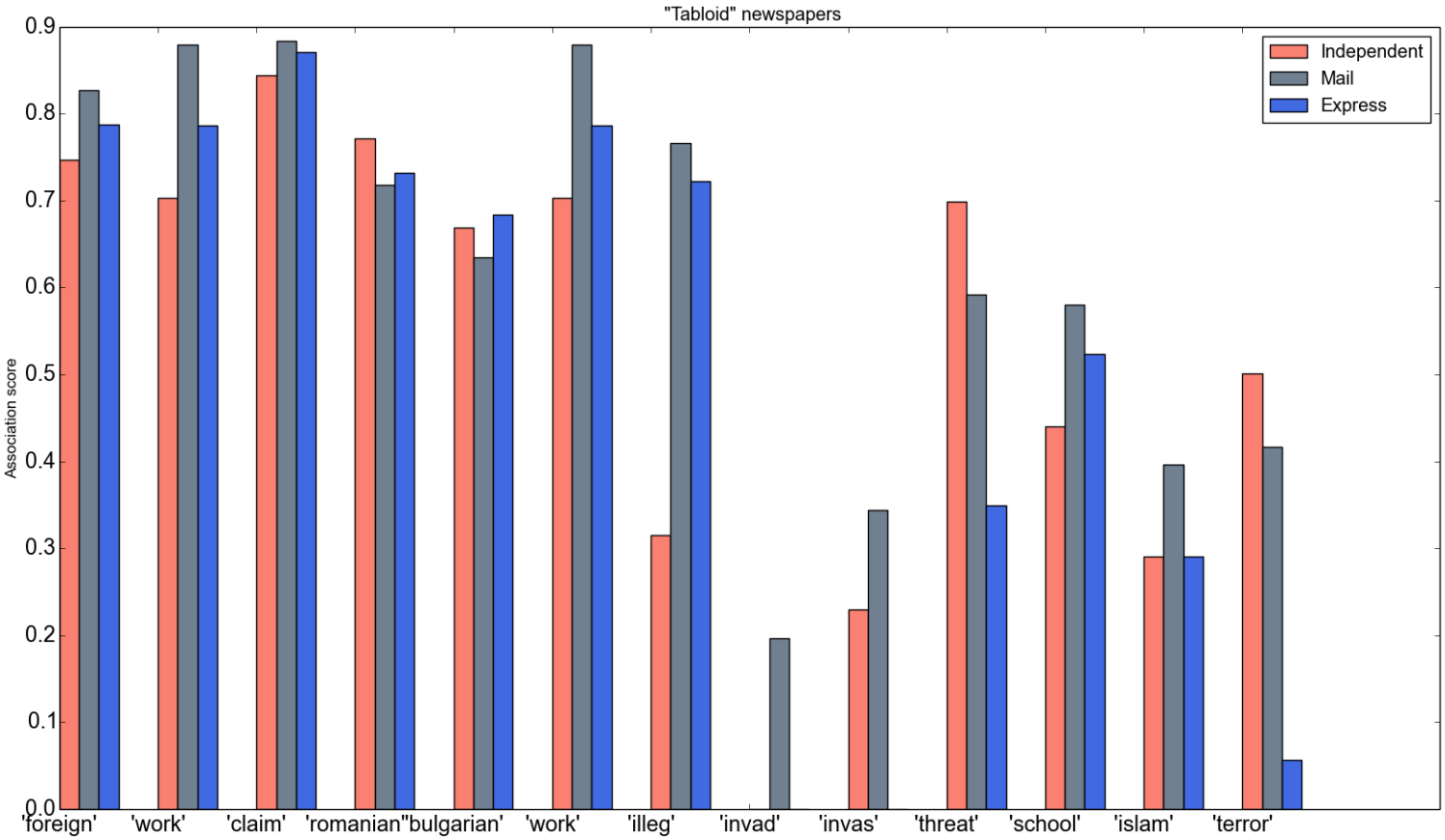

Figure 10: Association scores indicating framing of the immigration debate in tabloid newspapers.

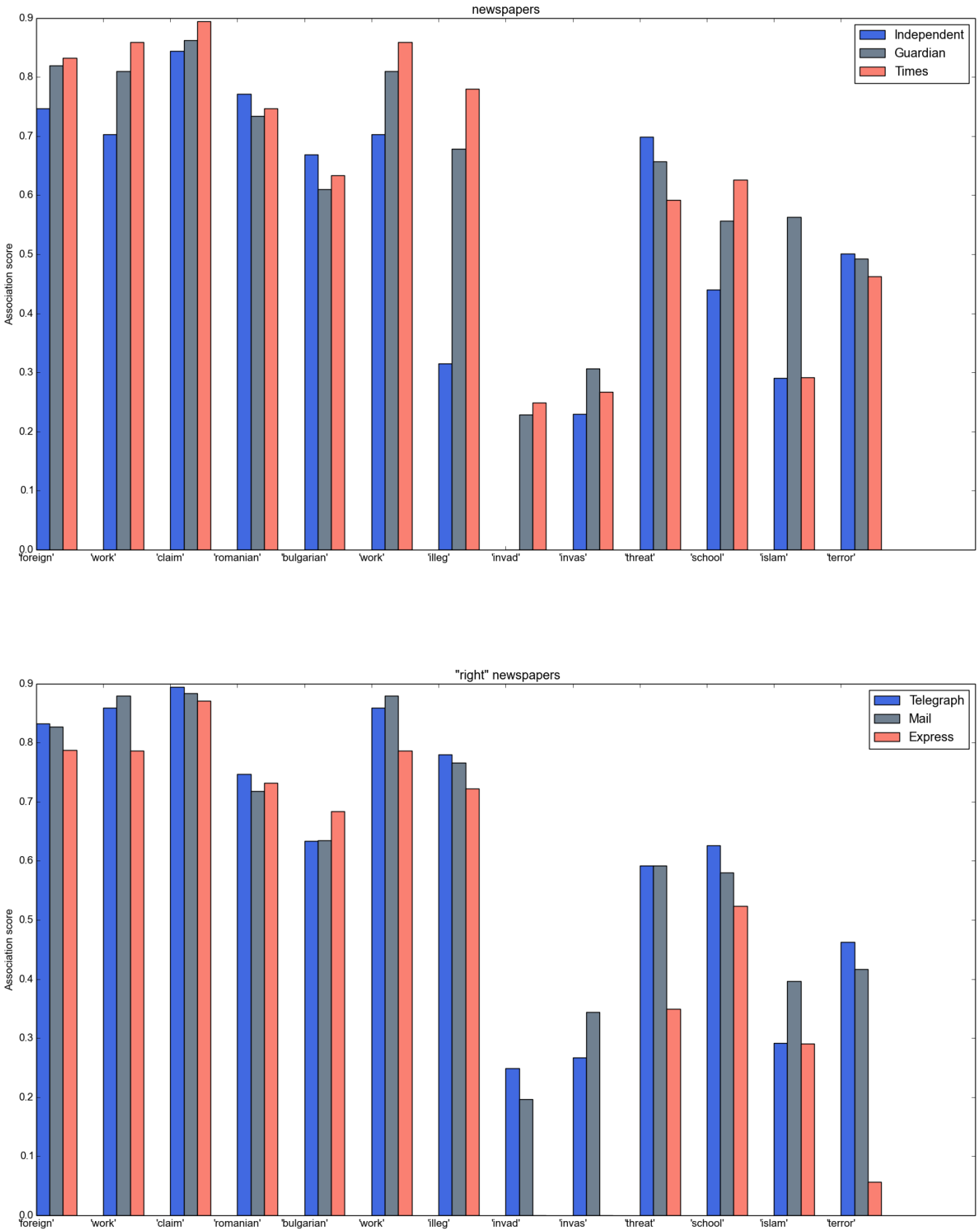

Figure 11: Association scores indicating framing of the immigration debate.

ciation score for words across all documents, as the general frequency of the word will effect its association — work will have a higher frequency than Bulgarian simply because it is a more frequent term overall. Comparing lexical relations among terms with asymmetric overall frequencies is an ongoing problem in natural language processing.

We might expect that right-wing newspapers which are generally perceived to be opposed to immigration would be more likely to frame immigration as *invasion*, yet the results indicate that *The Daily Express* does not use the stem *invad** or *invas** within 50 words of *immigr**. We confirmed this result with a simple text search of the 352 Daily Express articles, and found that the terms invade or invasion did not occur in the documents.

Furthermore, when the term did occur in other newspapers, we found that it was often used in a sarcastic manner to caricature the language of opponents to immigration, e.g.:

> predictions of a mass invasion of the uk by foreigners were as accurate as saying five million scots will invade orpington if edinburgh splits from the uk — *Daily Mail, 5th April 2014*

Closer inspection of articles in The Guardian containing the stem terror* again highlight the difficulty of selecting specific terms to identify particular ways of framing the debate.

> she spoke of the terror she feels at the prospect of going back to her home country — *Guardian, 24th March 2014*

One of the most noticeable differences in language association is the framing of the immigrants as 'illegal', which is prominent in all newspapers except for The Independent. We verified this with a manual search of Independent articles, and found only a handful of (8/324) articles which described immigrants as 'illegal', compared to dozens in each of the other sources.

7 Social Media: Topics in tweets

We collected posts to the microblogging platform Twitter which mentioned the terms *immigrant, immigrants, immigration*, or *immigrate*. Problems implementing a connection to twitter's streaming service meant we were only able to collect these tweets in the 8 eight days

leading to election day, rather than 12 weeks as we did with the newspaper articles. Nonetheless, we stored 430,095 tweets containing these terms over these days. We discarded re-tweets, tweets containing links, and tweets of fewer than twenty characters, and were left with 68,959 posts.

Applying LDA to tweet data could be problematic, as each document (tweet) is necessarily of very short length. We experimented with treating tweets from the same user as though they were in the same document, but this didn't substantially change the topics discovered. From a total of 68,959 posts, there were 26,947 unique users.

We are interested in tweets about the UK immigration debate — isolating social media posts relating to a particular region is an ongoing problem in information retrieval. Less than 1% of tweets contain GPS geotagging information. Where geographic location is associated with a tweet, this is most usually through the user-specified 'location' field, where the user can enter any string, but this information is often vague or unreliable. In addition, location alone is no guarantee that the content of a particular tweet refers to a local issue.

To solve this problem, we applied topic modelling iteratively to discover keywords which allowed us to filter the tweets. First, we found topics within the overall set of tweets. We chose a larger number of topics (40) as it is essential that the word groups are as unmixed as possible.

Topics found in the overall dataset are shown in table 7

Then, we identified topics within these that referred to the UK immigration debate. Finally, we filtered the corpus of tweets using words from this topic in an attempt to isolate tweets on this topic, and re-applied topic modelling to this reduced dataset to find sub-topics within the immigration debate. Using a small list of terms taken from the UK and EU related topics in the overall corpus, we calculated a simple ratio of these terms to the total number of terms in each the tweet, to measure the extent to which a tweet was was related to the UK and the EU elections.

After applying this filter, we were left with 9458 posts. From these, we discovered 20 sub-topics indicated in **??**. We have removed the filter words from these topics.

From the word clusters in 7, it appears that we have succeeded in restricting the dataset to tweets about UK and EU issues. However, despite experimenting with different numbers of

Prominence	Top five words
0.04252	ukip eu vote party policies
0.02549	racist ukip policy eu control
0.02201	jobs british taking benefits job
0.01662	good amp day work customs ive working
0.01487	reform gop act america love immigrationreform
0.01465	country illegal amp support immigrate wouldnt
0.01344	laws law illegal reform amp amnesty borders
0.01082	wage jobs time wages amp low illegal
0.01064	labour ukip eu amp mass policy tories
0.01032	housing services consultants canada nhs
0.01001	illegal aliens year states million
0.00894	eu uk thing free economy movement c
0.00845	card poor illegal green system
0.00844	problem song illegal make big
0.00827	bill german wife history vote gop
0.00797	english speak language learn fucking
0.00792	back issues illegal def calling
0.00783	vote england irish country god amp
0.00741	white house illegal citizenship amp
0.00726	stop great country talking theyre
0.00715	care health americans pay education illegal
0.00708	agencies undocumented military debate antiimmigration
0.00701	tho illegal caribbean haitians amp minister
0.00674	open door policy amp ur crime
0.00673	jonathanhoenig illegal latvia countries major tb
0.00669	years colorado springs american illegal usa l
0.00666	coming food news whyimvotingukip sick bad
0.00638	border desert bitch migo illegal back
0.00638	children learn culture british english
0.00623	amp illegal human trafficking social
0.00612	world half boehner reform colonise
0.00609	racism antiimmigrant fear totally
5.00E-03	work family families jensan business
0.00543	today uk figures wasnt reform sad
0.00481	plans writers population hate save held
0.0043	illegal alien police officer murdered
0.00404	amp easier noneu regulations artists
0.00359	deportation wa seattle hillary solicitor
0.00312	welcomeus timeisnow fwdus reform

Table 3: All 40 topics identified in the total corpus of tweets

Prominence	Top five words
0.05339	election voting party vote policies country
0.04442	uncontrolled vote country uk racist britain
0.03188	jobs housing amp taking nhs vote
0.02712	racist talk bbcqt bnp make labour
0.02706	policy system points racist stop
0.02329	vote control problem talking blame
0.02298	amp nigel german wife figures mass
0.02124	vote policies england english romanian
0.02019	labour vote britain job bbc uncontrolled
0.01888	fear answer economy antiimmigration plan
0.01803	voters issues truth debate tories politicians
0.01725	parties elections conservatives question welfare
0.01713	policy racist amp support tory school
0.01605	anti free movement uk thing agree uncontrolled
0.01537	uk mass open door vote left policy
0.01483	uk controlled time canada borders open
0.01436	tax ur hate voftheppl work brits promise
0.01387	english speak racist white learn javid
0.01213	uk learn asian culture rise tory secretary
0.00675	noneu easier regulations artists bird petrel

Table 4: 20 Topics from tweets filtered by UK and EU keywords, after removing the filtering words

topics, the resulting word clusters are not as clear or easily labelled as those which emerged from the newspaper articles. We suspect that this may be due to the very short document length. It is clear that the most prominent topics deal with the election itself, and fears over availability of jobs and housing in the context of immigration, but this method must be refined, or more data collected, before a deeper analysis is possible.

8 Discussion and Conclusion

In this paper we have presented data extracted from UK media and social media texts in run up to the 2014 EU elections. We show that quantitative techniques can successfully isolate topics from these articles and track the variation in topic prominence over time or according to news source. We note that economic topics have more prominence in the corpus overall than cultural topics. We observe several notable differences in topic prominence across variables.

The 'benefits and welfare topic' (the most prominent topic that we investigated in-depth),

received noticeably greater attention in The Daily Express than in other newspapers.

The 'business' topic received greater coverage in newspaper traditionally described as 'broadsheet' as compared to 'tabloid'.

The two papers generally considered to be most 'left-liberal' of those we analysed, The Guardian and The Independent, gave noticeably more attention to the topic that we have labelled as 'Women and Religion'.

A great deal of the variation in topic prominence seems to be due to spikes when a particular story breaks, rather than a consistent agenda-setting by the media. For example, the topic of the immigration minister's first speech is prominent in the week of the event, but does not seem to return as an issue in the debate closer to the election date.

Issue framing proves to be a more difficult case than topic tracking. Word association alone, even within a narrow context window, does not provide a sufficiently fine-grained view of the sentiment and attitude adopted by those writing on the issue. This paper is a work-in-progress, and further analysis is needed to create a fuller picture of the debate. The bag-of-words model is not always sufficient to capture the tone of public debate, and both more sophisticated quantitative techniques (natural language processing) and validation through qualitative methods will be necessary.

References

Allen, W and S Blinder. 2013. "Migration in the News: Portrayals of Immigrants, Migrants, Asylum Seekers and Refugees in National British Newspapers, 2010 to 2012." *Migration Observatory report, COMPAS, University of Oxford* .

Blei, David M. 2012. "Probabilistic topic models." *Communications of the ACM* 55(4):77–84.

Hornik, Kurt and Bettina Grün. 2011. "topicmodels: An R package for fitting topic models." *Journal of Statistical Software* 40(13):1–30.

Landauer, Thomas K and Susan T Dumais. 1997. "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." *Psychological review* 104(2):211.

McCallum, Andrew Kachites. 2002. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu.

Sagi, Eyal, Daniel Diermeier and Stefan Kaufmann. 2013. "Identifying Issue Frames in Text." *PLoS one* 8(7):e69185.

Widdows, Dominic and Kathleen Ferraro. 2008. Semantic Vectors: a Scalable Open Source Package and Online Technology Management Application. In *LREC*.

- Barker, D.C. (2005). Values, frames, and persuasion in presidential nomination campaigns. Political Behaviour, 27, 375-94.

  Blei, D., Ng. A., and Jordan, M. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993-1022.

- Chong, D. and Druckman, J. N. (2007). Framing Theory. Annual Review of Political Science, 10, 103-26.

- Citrin, Jack, Donald Green, Christopher Muste, and Cara Wong (1997). "Public Opinion Toward Immigration Reform: The Role of Economic Motivations, Journal of Politics, 59 (3): 858-81.

- De Vreese, C.H. (2005). News framing: Theory and typology, Document Design, 13(1): 51–62.

- Druckman, J. N. (2001). On the Limits of Framing: Who can frame? Journal of Politics, 63(4), 1041-1066.

- Edelman, M. (1993). Contestable categories on public opinion. Political Communication, 10, 231-242.

- Entman, R. (2004). Projections of Power. Framing News, Public Opinion, and US Foreign Policy. Chicago:

- University of Chicago Press.

- Entman, R. M. (2007). Framing bias: Media in the distribution of power. Journal of Communication, 57, 163-173

- Gamson, W. A. (1992). Talking Politics. New York: Cambridge University Press.

- Ghanem, S. (1997). Filling in the tapestry: The second level of agenda setting. In: M. McCombs, D. L. Shaw, and D. Weaver (Eds.), Communication and democracy (pp. 314). Mahwah, NJ: Erlbaum.

- Goffman, E. (1974). Frame Analysis: An essay on the organization of experience. New York: Harper & Row.

- Gross, K. and DAmbrosio, L. (2004). Framing emotional response. Political Psychology, 25, 1-29.

- Hainmueller, Jens and Michael J. Hiscox (2007). "Educated Preferences: Explaining Individual Attitudes toward Immigration in Europe, International Organization, 61(2): 399-442.

- Hertog, J. K., & McLeod, D. M. (2001). A multiperspectival approach to framing analysis: A field guide. In S. D. Reese, O.H. Gandy, & A. E. Grant (Eds.), Framing public life (pp. 139162). Mahwah, NJ: Lawrence Erlbaum.

- Iyengar, S. (1991). Is anyone responsible? How television frames political issues. Chicago: University Chicago Press.

- Jacoby, W. G. (2000). Issue framing and public opinion on government spending. American Journal of Political Science, 44, 750-767.

- Jost, J. T., Glaser, J., Kruglanski, A. W., and Sulloway, F. G. (2003). Political conservatism as motivated social cognition. Psychological Bullettin, 129 (3), 339-375.

- Kinder, D. R. & Sanders, L. M. (1990). Mimicking political debate with survey questions: the case of white opinion on affirmative action for blacks. Social Cognition, 8, 73-103.

- Lahav, Gallya (2004). "Public Opinion toward Immigration in the European Union: Does it Matter?, Comparative Political Studies, 37(10): 1151-1183.

- Laver, M., Benoit, K. and Garry, J. (2003). Extracting Policy position from political texts using words as data. American Political Science Review, 97 (2), 311-31.

- McCombs, M. (1997). New frontiers in agenda setting: Agendas of attributes and frames. Mass Communication Review, 24 (1 & 2), 3252.

- McCombs, M. (2005). A look at agenda-setting: Past, present and future. Journalism Studies, 6, 543557.

- McCombs, M. and D.L. Shaw (1972) The Agenda-Setting Function of the Press, Public Opinion Quarterly, 36(2): 176–87.

- Miller, J. M. & Krosnick, J.A. (2000). News media impact on the ingredients of presidential evaluations: politically knowledgeable citizens are guided by a trusted source. American Journal of Political Science, 66: 581-605.

- O'Rourke, Kevin H. and Richard Sinnott (2006). "The determinants of Individual Attitudes towards Immigration", European Journal of Political Economy, 22(4), 838-61.

- Riker, W. (1986). The Art of Political Manipulation. New Heaven: Yale University Press.

- Schattschneider, E. E. (1960). The semisovereign People. New York: Holt, Rhinehart, Winston.

- Semetko, H. and Valkenburg, P. M. (2000). Framing European Politics: A Content Analysis of Press and Television News. Journal of Communication, 93-109.

- Vliegenthart, R., Schuck, A, Boomgarden, H. G. and De Vreese, C. (2008). News coverage and support for European Integration, 1990-2006. International Journal of Public Opinion Research, 20 (4), 415-439.

- Weaver, D. H. (2007). Thoughts on Agenda Setting, Framing and Priming. Journal of Communication, 57, 142-147.

- Zaller, J. (1992). The nature and origin of mass opinion. New York: Cambridge University Press.