LDA models topics...
But what are 'topics'?

Wouter van Atteveldt and Kasper Welbers
VU University Amsterdam

Carina Jacobi
University of Vienna

Rens Vliegenthart
University of Amsterdam

Abstract

LDA topic modeling is a popular technique for unsupervised document clustering. However, the utility of LDA for analysing political communication depends on being able to interpret the topics in theoretical terms. This paper explores the relation between LDA topics and content variables traditionally used in political communication. We generate an LDA model on a full collection of front-page articles of Dutch newspapers and compare the resulting LDA topics to a manual coding of the political issues, frames, and sentiment.

In general, we find that a large number of topics are closely related to a specific issue; and that the different topics that comprise an issue can be interpreted as subissues, events, and specific journalistic framing of the issue. The relation between frames and topics is less direct, with a large amount of topics associated with each of the investigated frames while no topics were identified that really encoded just a specific frame. Finally, hardly any topic had a clear sentiment associated, with only exception for topics whose sentiment is contained in the represented issue, such as disasters. These results validate the use of LDA topics as proxies for political issues, and pave the way for a more empirical understanding of the substantive intepretation of LDA topics.

Introduction

LDA topic modeling (Blei et al., 2003) is a popular technique for unsupervised document clustering. Briefly put, LDA fits a generative model to a collection of documents that assumes that each document is composed of a number of topics, which themselves are composed of words.

In many applications, the topic in LDA models are interpreted substantively, making the implicit assumption that LDA topics are in fact 'topics' in the casual sense of the word. From anecdotal experience, this is correct for a number of topics, but other topics contain words connected with how the action or discourse is conducted rather than the substantive issue. For example, when training a topic model on statements in parliament and media there were a number of topics that were clearly substantive, but other topics dealt with ways of expressing ones opinion and also with specific terms from either venue (Van Atteveldt, 2014). Indeed, validating whether the results of an LDA can be interpreted in order to answer a specific research question is one of the more difficult aspects of doing research with topic models (Grimmer & Stewart, 2013).

This paper investigates to what extend topics do indeed match substantive political issues (e.g. economy, education, foreign affairs). Also, it investigates in what way multiple topics for the same issue are distinct, i.e. can they be seen as sub issues, events, applied frames, etc. This is achieved by running a topic model on a corpus of newpaper articles that has been manually coded for issue, valence, and frame. This allows us to connect the resulting topics to issues and see how well a topic matches one or more issues and to what extent mulitple topics together cover a complete issue.

This paper contributes to the use of LDA in two ways. First, it provides empirical grounding for the assumption that LDA topics can be interpreted as issues, and it gives some quantitative metrics on how well individual topics match certain issues when a model is trained on (unselected) newspaper data. Second, by showing that some topics are very high-precision indicators of issues, and that these topics can be identified by inspecting top keywords and articles, it opens up possible strategies to use selected topics as substantive indicators or as features in machine learning models.

By way of disclaimer, please note that this paper is rather exploratory in nature. The main goal was to get a more quantiative understanding of the substantive interpretation of LDA topics in terms such as issues and valence that are well grounded in the political communication literature. Although precision and recall figures are presented throughout the results section, these are meant to give understanding of the match between topics and other concepts. The goal of this paper is certainly not to use unsupervised LDA models as a model to predict variables for which ample coded material exist.

Method

Corpus and manual coding

The data for this paper is based on a manual coding of Dutch front page newspaper articles. All articles on the front page of *De Volkskrant* (1995–2011), *NRC Handelsblad* (1995–2011) and *de Telegraaf* (2004–2011) were collected (n=99,572 articles).

From this set, a random sample of 12,523 articles were coded manually using a coding scheme of which four variables are used in this paper:

Issue The main issue covered in each article was coded using the political issue categorisation devised in the Comparative Agendas Project (Jones & Baumgartner, 2005; Baumgartner et al., 2009). This scheme uses 28 issue categories that distinguish political issue domains such as economy, foreign affairs, and defense. It also includes categories such as ‘weather’ and ‘sport and recreation’ added specifically for media coding.[1] Each article is assigned a single topic, including a catch-all “other” category.

Political news For each article it was coded whether the article is political in nature, defined as dealing with how the authorities should, would, could, or do deal with a problem or approach an issue; or if it deals with the nomination or election of political actors.

Framing For those articles that were considered political (7,000 out of 12,523) it was further coded which journalistic frames were used in the article. For this, the coding scheme developed by Semetko & Valkenburg (2000) was used. This scheme consists of 18 yes/no questions that form scales for 5 different frames: Conflict, Human Interest, Economic Consequences, Responsibility, and Morality. This paper focuses mainly on the values of the composite scales, were an article was considered to employ a frame if at least on of the questions was answered positively.

Valence Finally, for the political articles it was coded whether they were about a valence issue, defined as any issue (such as unemployment) that can be considered uniformly good or bad for all readers. If an article is about a valence issue (1,477 out of 12,523), it was further coded whether the news was good, neutral, or bad with respect to that issue.

Latent Dirichlet Allocation

All articles were preprocessed using the Alpino phrase structure grammar parser (Van Noord, 2006). All lemmata with a substantive part of speech (noun, adjective, or verb) were selected that occured in at least 25 documents and in less than 20% of all documents (to remove uninformative words). This resulted in a vocabulary of 7,181 terms, which was used to fit an LDA topic model with 500 topics using the R package `lda`.

We decided to use a large number of topics in order to be able to capture sub-issues of the larger issues and to prevent ‘crowding out’ the smaller issue categories: the largest issue category (20: democracy and elections) occurred 1,400 times, while the smallest category (21: land management) occured only 53 times in 12,523 documents.

To simplify the analysis, the assignment of topics to documents was dichotomized using a cutoff value of 15% of all words in the document. To determine this threshold value, the list of topics was inspected manually to identify a topic that closely matched an issue category based on the top words in the issue. The selected topic (topic 4: *agriculture, cow, pig, animal*) was found to be a high-precision indicator of the issue Agriculture (precision weighted by occurrence: .82, Pearson correlation between topic occurrence (weight) and issue 0.67, p<.001). Moreover, the topic did not score highly for any other issue (next-best correlation: 0.02) and since agriculture is a small and distinctive topic, and not many other topics were correlated with the issue (next-best correlations were 0.3 and 0.2). Figure 1 shows the precision/recall curve for this topic with various cutoff values. As can be seen, recall climbs steeply until a cutoff of 15%-20% is reached, after which recall plateaus while

[1] See `http://www.policyagendas.org/page/topic-codebook` or the Appendix for a list of codes and descriptions. Note that we use the ‘additional topics’ listed on the website, but restricted coding to the major issue cateogies.

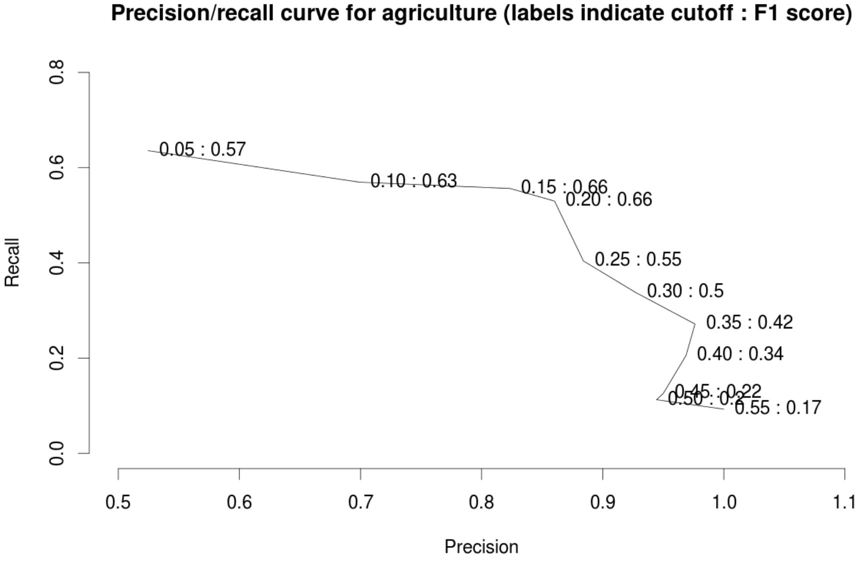


Figure 1. Precision/recall for topic 4 on issue Agriculture

precision drops. From this, a cutoff value of 15% was determined to give a good indication of a topic being a substantial part of a document. This value was confirmed informally in the precision/recall curves of other strong topic:issue combinations. Finally, all topics that occurred in less than 10 documents (n=174 out of 500) in the dichotomised data frame were removed. These topics, such as 222 (*own, take, happen, give, together*) contained words that were used in a lot of topics, but did not figure prominently in any.

Since now both manually coded variables and the LDA topics are dichotomous variables, the standard metrics of precision and recall can be calculated on any topic:variable combination, which form the basis of the exploratory results presented below.

Results

Topics and Issues

The first question to be answered is whether topics correspond to political issues. Since there are many more topics than issues, we focus on the predictive precision of topics for issues. Figure 2 shows a histogram of the precision of each topic for its best matching issue, summarised in table 1. As can be seen, almost all topics match an issue to some degree. Quite a number of topics (133) has more than 50% precision on its best topics, and 31 topics even had more than 75% precision. Thus, if we take a cutoff of 50% precision as meaning that a topic 'matches' an issue, the first finding is that around one third of LDA topics matches a political issue coded at the document level.

The links between topics and issues can also be visualised as a bipartite graph, where the nodes are topics or issues, and edges are formed between a topic and the issues for which it is a $>$ 25% precision predictor. Note that since each document is coded as being about

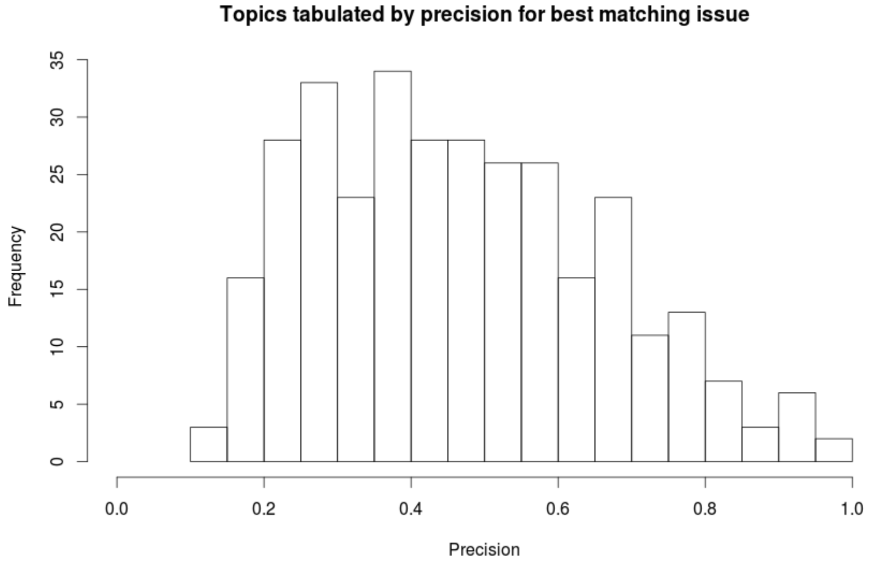


Figure 2. Histogram of topics by precision

Precision	<25%	25-50%	50-75%	>75%
Frequency	47	146	102	31

Table 1: Frequency of topics by precision for best matching issue

a unique political issue, a topic can be a $>$ 50% predictor for only one issue, and a $>$ 25% precision predictor for at most three.

The resulting graph is shown in 3. The red numbered nodes are the issues, and the unlabeled nodes are the topics. The size of each node is proportional to the number of documents that the topic or issue occurs in. The topics are coloured according to how well they predict the political nature of a topic, with blue nodes indicating that most ($>$ 80%) of the documents the topic occured in were political, while yellow nodes had little ($<$ 20%) political content.

What is immediately clear for this picture is that, unsurprisingly, the number of topics connected to an issue is strongly correlated with the size of the issue ($\rho = 0.9, p < .001$). Also, because each article was coded with a single issue, most of the topics are only connected to one issue above the precision threshold of 25%: the topics are located in clouds around each issue, and only a small number of topics bridge between the issues.

Predicting issues with multiple topics

The next question is whether a combination of issues forms a good approximation of a specific issue. The temporal correlation made above shows that at least for some issues the temporal correlation of the sum of defense-related topics are a very good approximation of the defense issue. For a more statistical measure, we can see whether the occurence of any of the topics related to an issue is a good predictor of the issue. Please note that

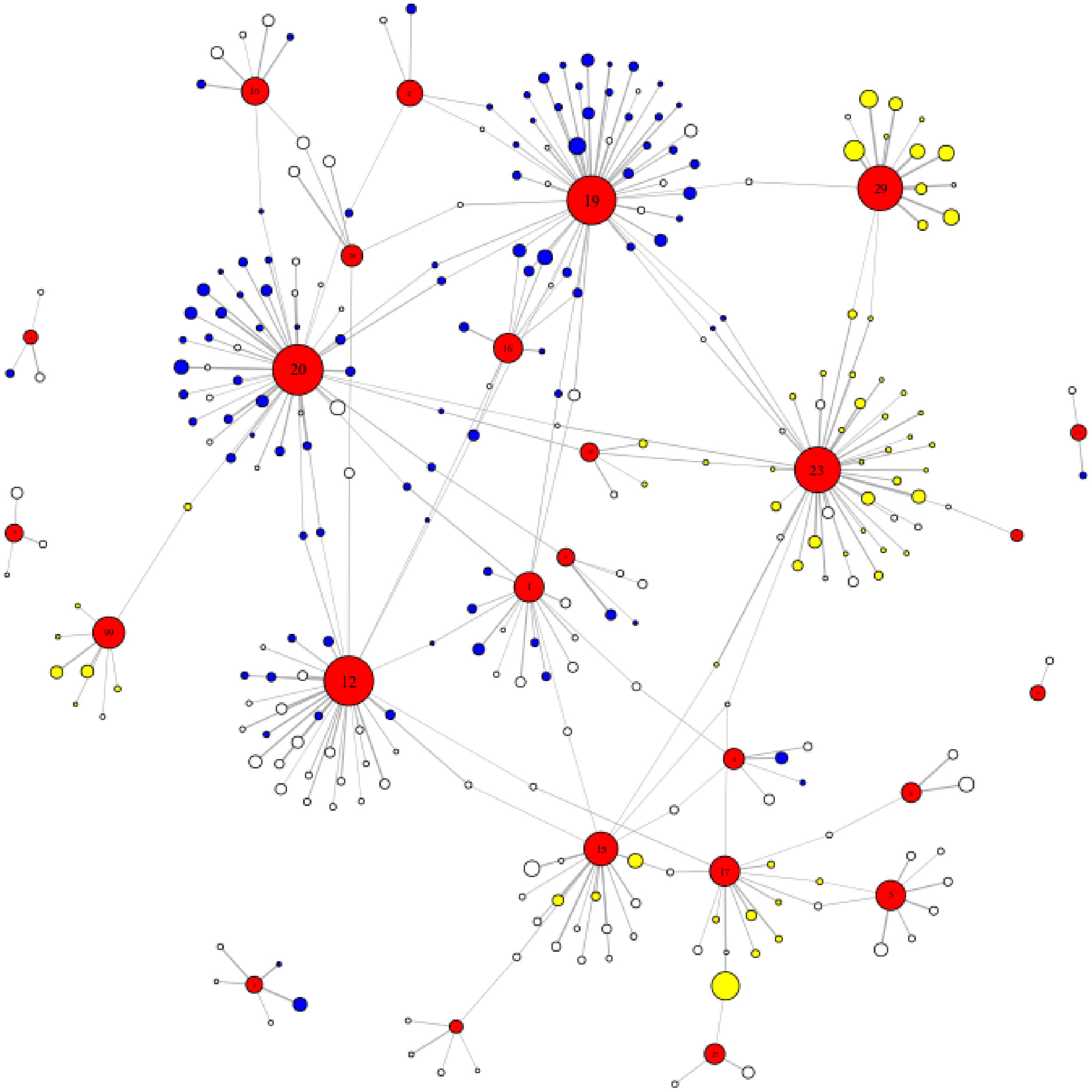


Figure 3. Network diagrams of topics and issues

we are not primarily interested in using topics from a coded dataset to predict issues in a machine learning sense. If there is sufficient coded data, it is unlikely that unsupervised topic models aligned with issues using coded data will do a better job of predicting than a supervised algorithm. However, if the different high-precision topics can be combined into a predictor with decent recall, then we know that the topics provide good coverage of the issue. Moreover, if the issues can be determined based on inspecting the top-words and top-articles for each topic (which is a common way of validating topic models), then we can also use topic models as a very cheap approximation of manual coding without requiring any training material.

Figure 4 shows the overall precision/recall curves for four different issues. For all issues except for Sport (which is easiest to predict), there is a strong tradeoff between precision and recall, with precision dropping quickly as the cutoff for including issues is decreased. In general, the bend in the curve inidicating the optimal cutoff for maximising the F1-score seems to occur around a cutoff of 25%, with F1 score ranging from around .8 for Sports to .5 for Politics and Foreign Affairs.

This is summarised in table Table 2. This table gives the number of topics with precision higher than 25% as well as the precision, recall, and f-score achieved and the total number of documents manually coded with that topic. As can be seen, the overall macro-average precision, recall, and F1-score are all around 0.5, with slightly better precision than recall. Some issues were predicted quite a lot better, such as Sport (F1=0.79), Agriculture (0.69) and Education (0.65). For some issues, only few or even no topics had more precision than the cutoff value of 25%, and scored very low. However, also some of the larger issues such as defense and economy scored substantially lower than the other big issues. This is quite likely caused by these issues being strongly related to other issues, especially defense and foreign affairs.

Political nature

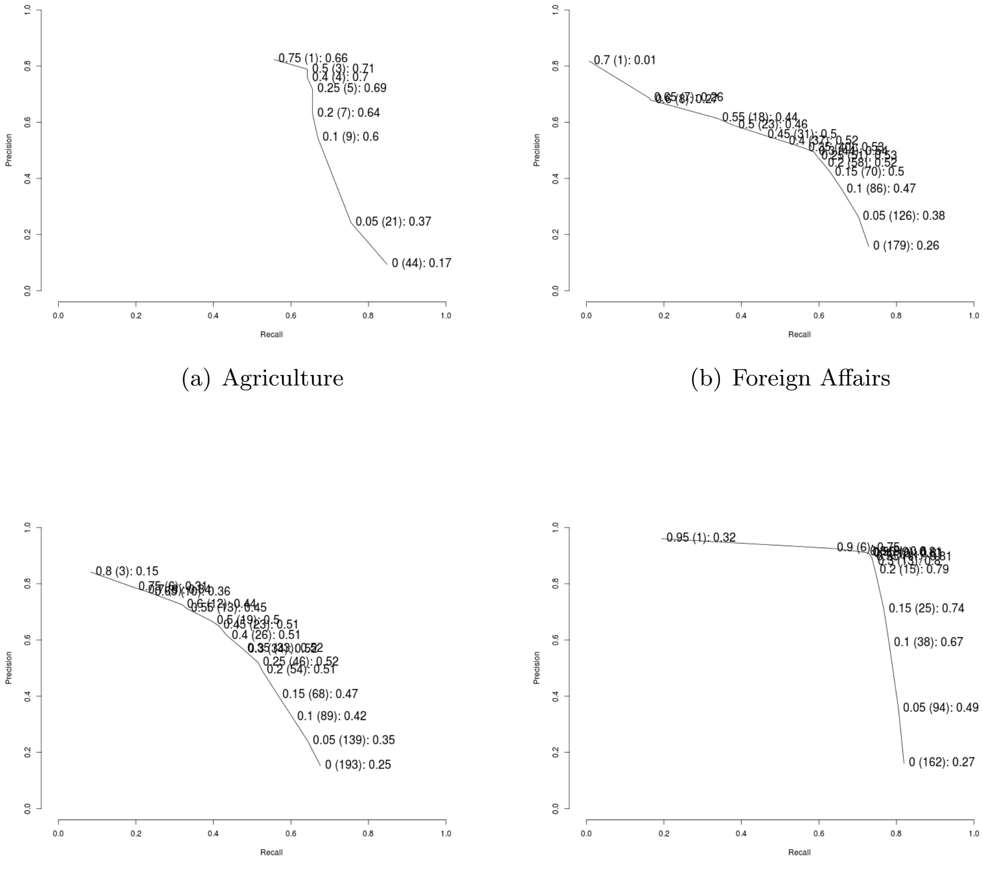
As described earlier, all articles were manually coded with a single dichotomous variable indicating whether it was political in nature. A large number of topics were very strong predictors of the (non-)political nature of topics, with 42 topics being >90% political and another 44 over 80%. As was already clear from the network diagram presented above, the distinction between political and non-political is strongly related to the issue of an article. In fact, out of the 42 strongly political topics, 26 were >50% indicators of a limited range of topics. Unsurprisingly, 12 topics were connected with Democracy and Elections and another 9 were connected with Foreign Affairs. Within these issues, the political nature also did not vary much: of all articles coded as either Foreign Affairs or Democracy and Election where 93% political compared to 55% average. Thus, the fact that these topics are political in nature is determined by their connection to specific issues more than that they show a specific political aspect of an issue.

At the other end of the scale we see a similar picture. 32 topics occured in articles that were over 90% non-political, with another 28 topics over 80%. As for the political topics, of the 32 non-political topics 22 were >50% precision indicators for a limited range of topics dominated by arts (n=11) and sports (n=6).

There are exceptions to this pattern. Topic 40 containing mostly political party names (*PvdA, CDA, VVD, cabinet, parliament, D66*) did not strongly occur with any specific issue

Table 2: Predictive accuracy per issue

Issue		# Articles	# Topics	Precision	Recall	F1-score
20	Democracy	1400	46	0.52	0.52	0.52
12	Crime	1349	34	0.53	0.44	0.48
19	Foreign Affairs	1325	51	0.48	0.59	0.53
23	Arts	1152	44	0.54	0.48	0.51
29	Sport	1118	15	0.85	0.75	0.79
15	Business	628	21	0.52	0.58	0.55
99	Other	560	8	0.53	0.23	0.32
17	Science	496	15	0.45	0.65	0.53
1	Economy	489	17	0.41	0.47	0.44
3	Health	488	8	0.62	0.38	0.47
16	Defense	470	11	0.41	0.42	0.41
10	Transportation	414	7	0.55	0.42	0.47
2	Civil rights	366	5	0.37	0.13	0.19
28	Accidents	256	5	0.50	0.54	0.52
5	Labour	239	6	0.41	0.43	0.42
27	Weather	229	3	0.33	0.74	0.45
6	Education	212	3	0.72	0.60	0.65
30	Obituaries	187	5	0.49	0.27	0.34
9	Immigration	171	5	0.49	0.48	0.49
7	Environment	169	3	0.40	0.26	0.32
4	Agriculture	151	5	0.72	0.66	0.69
13	Social affairs	131	2	0.48	0.18	0.27
14	Community planning	123	1	0.37	0.09	0.14
31	Religion	108	3	0.46	0.39	0.42
8	Energy	94	5	0.37	0.35	0.36
18	Foreign trade	80	1	0.36	0.05	0.09
24	Decentral gov.	65	0	0.00	0.00	0.00
21	Public planning	53	0	0.00	0.00	0.00
	Total	12523	329	0.53	0.48	0.49



(a) Agriculture

(b) Foreign Affairs

(c) Politics

(d) Sports

Labels indicate *cutoff* (*number of topics*): *F1-score*

Figure 4. Precision/recall for different topics

(highest precision was 18% for Democracy and Elections). So, this topic seems to be used as a 'mix-in' topic for multiple issues to indicate a political context of these issues.

Valence

Out of all articles, only 1,477 articles were coded as being about valence issues. 9 topics were a >50% predictor of valence, and these were dominated by bad news about foreign affairs (n=7), for example topic 143 about humanitarian aid (*help, keep, UN, aid worker, humanitarian, promise, Red Cross*) and 436 about the Middle Eastern conflict (*Israel, Hamas, Hezbollah, Syria, Palestinian, Lebanon*). The other >50% valence topics were a topic about crime and a topic about obituaries (*person, human, family, picture, princess, queen*). Interestingly, the issue with the highest proportion of valence issues was civil rights (31% valence from manual coding), but the only topics connected to that issue with a moderate connection to valence were also connected with foreign affairs, e.g. topic 22 about the Arab Spring in Lybia (44% valence, 39% civil rights, 19% foreign affairs; *Lybia, leader, far, Tripoli, colonel*). All topics mentioned so far were not just connected to valence, but specifically to bad news. This is not suprising as bad news was coded almost four times as frequently than good news (987 vs. 286 articles). One of the three topics related to good news was about the global economy (topic 299, 26% good news; *IMF, country, minister, billion, financial, international, economy*), while the economy topic with the highest precision on bad news only had 17% precision on bad news and dealt mainly with domestic economic news (topic 487; *percent, ecnomic, growth, quarter, CBS*).

Topics as subissues

As was shown above, many topics are high-precision indicators of a single political issue, and the larger issues had multiple indicator topics, ranging from very few topics for small issues such as 21 (public planning, 0 topics) or 18 (foreign trade, 1 topic) to 51 topics for issue 19 (foreign affairs). This raises the question of how the different topics that are connected to the same issue are distinct from each other. In this section, we will explore how these topics can differ in terms of political nature, substantive focus, temporal focus, frames, and links with other topics.

Political nature From the graph presented above it was clear that for some issues there is a distinction between political and non-political topics. If we look at the bigger issues, we see that an issue such as 29 (sports and recreation) has mainly non-political topics, while 19 (foreign affairs) and 20 (democracy and elections) are almost exclusively political in nature. Differences within an issue are found, for example Issue 15 (business) is a mix of non-political and neutral topics and issue 12 (crime) is interesting because it has both political and neutral topics. For example, for crime the topics range from very political (topic 33, 93% political: *minister, interior, justic, Donner, Hirsch Ballin, CDA*) to topics with hardly a political angle, such as topic 360 (95% crime, 39% politics: *police, criminal, Amsterdam, murder*), although it should be noted that even the articles with the least political topic still had a political angle in 30% of cases. In fact, none of the issues had both highly political and highly non-political topics except for the bridging topics between foreign affairs and sports and entertainment. Thus, the issue category to a large extend determines the political nature of articles, as confirmed by a chi-squared test of political nature and issue ($\chi^2 = 5414, df = 54, p < .001$). However, within each issue some of the topics are more political, while other topics are more descriptive, such as reports on individual crimes.

Substantive focus Another distinction between different topics is if an issue is really a collection of mostly different subissues. For example, issue 28 (fires and accidents) has two strongly connected topics: 477 (*fire, person, disaster, fire brigade, victim*) and 240 (*car, driver, hit, drives, accident*), which clearly cover the two separate subissues in the issue name. Issue 29 (sport and entertainment) has clear topics for competition soccer, international soccer, cycling, and olympics. More politically relevant, an issue such as crime has distinct topics for discussing crimes and victims, police, and the justice system.

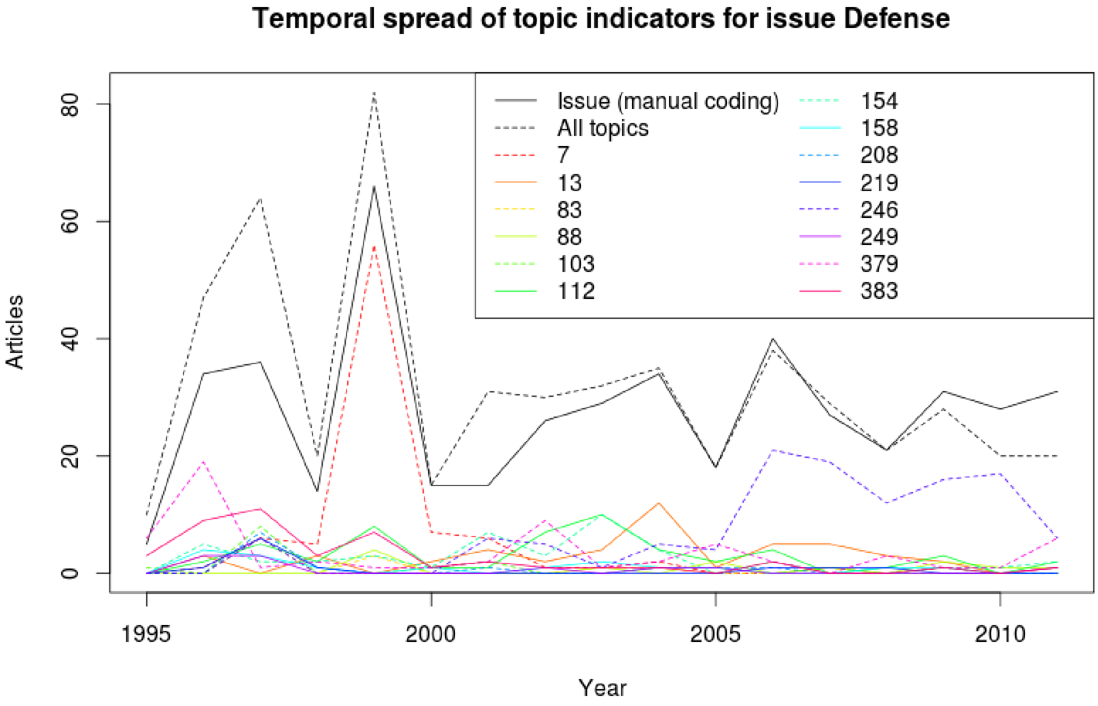


Figure 5. Occurence of topics related to issue 4: Defense

Temporal focus It is conceivable that different topics refer to the same issue in different temporal contexts. If a topic refers to a specific event, such as the war in Iraq, it can be expected that for example the foreign affairs issue is broken down into separate topics that are prominent in different time frames. Figure 5 shows the distribution of defense articles and of all topics that were a >25% precision indicator of defense. It is quite clear that the sum of topics is a better correlate with the issue ($\rho = .90$) than any individual topic (highest $\rho = 68$ for topic 7). Also, a number of topics are clearly time-bound. Topic 7 peaks in 1999 and represents the intervention in Kosovo (*Kosovo, Serbian, NATO, Belgrado, Yugoslavia*) in which Dutch fighter jets participated and which generated a lot of news because of the lack of UN mandate. Topic 379 deals with Bosnia and the Dutch mission in Srebrenica and peaks in 1995 and 1996 (the fall of Srebrenica) and later again in 2002 when the final report of the parliamentary investigation lead to the resignation of prime minister Kok and in 2005 with the attention around the 10th anniversary of the genocide and the tribunal case against Milosovic. The final clear temporal topic is 246 (*Afghanistan, military, Dutch, Uruzgan*), which deals with the Dutch mission in Uruzgan.

The "missing peak" in this picture is Iraq. which is represented in topic 478 as noted above. This topic, however, is is connected with foreign affairs (precision 60%) rather than defense (17%) because there was no Dutch participation in the actual fighting. There is a small defense issue related to the Dutch participation in the stabilisation force after the war which peaks in 2004 (topic 13: *defense, military, Dutch, minister, send, Iraq*).

Frames For the issue of defense, there was not a great deal of variation in the use of framing. Almost all topics are political, and almost all topics are frames using the conflict and responsibility frames. The only exception to this is topic 16 (*investigation, testimony, possible, minister*) which dealt with the investigation of the behaviour of Dutch soldiers in Srebrenice and was mainly frames from a human interest or emotional perspective. This is not the case for other issues. For example, for crime (issue 12), there are clear topics that deal with responsibility/solution such as topic 360 about murder mentioned above (95% responsibility) and topic 449 about the criminal justice system (71% responsibility; *lawyer, case, justice, prosecution, court, judge*) and topic 252 about the police (93% responsibility; *police, crime, corps, minister, national*). Other topics have a more human interest angle such as topic 11 that deals with (murder) victims (84% human interest; *man, murder, police, girl, body, woman*) and 281 about sexual abuse (75% human interest, *child, sexual, victim, abuse, violence*) Two topics were mainly conflict oriented: topic 373 about criminal cases (89% conflict; *trial, judge, accusation, persecutor*) and a relatively small topic about political reorganisation of the justice system (n=8, 100% conflict, *committee, advice, report, verdict, important*).

Bridging topics The network graph presented in Figure 3 shows a number of issues that have >25% precision on multiple topics. In fact, out of 326 topics there were 60 topics connected to two issues, and 4 connected to three.

Most of these connections make substantive sense. There is a large number of topics connecting foreign affairs (19) and defense (16), that are related to foreign defense missions, such as topic 246 (*Afghanistan, military, Dutch, Uruzgan*). Economy (issue 1) is connected mainly with foreign affairs and politics, for example in topic 277 about election coverage of economic plans (50% economy, 30% democracy and elections; *extra, money, cabinet, billion spending*). Sports and recreation (issue 29) is connected with arts and culture (23) through topics 421 about (music) festivals (*visitor, photograph, events, organisation, party, audience*) and 269 about TV shows (*series, image, start, story, episodes*).

Looking at individual 'bridging' issues can explain the less obvious links. For example, the link between 12 (crime) and 15 (business) is a topic dealing with fraud, especially the financial scandal with Dutch supermarket chain Ahold (*Ahold, financial, turns out, former, important, Van der Hoeven*). The large yellow circle at the bottom is the actually the topic that occurs prominently in most articles (n=429): the weather, which connects issues 27 (weather and natural disasters) and 17 (science and technology). Finally, topic 379 forms a link between three issues: 12 (crime), 16 (defense) and 19 (foreign affairs). This topic is a very coherent topic about the failure of the Dutch troops to defend Srebrenica and the tribunal cases against the perpetrators of the war crimes committed in the aftermath (topic 379:*Bosnian, Srebranica, Serbia, Muslim, Mladic, Sarejevo, Karadzic*)

Note that because articles could only be coded with a single issue category, precision for topics that are 'in between' issues will always be low, and this can also partly explain the relatively low F1-scores reported above. However, this does not mean that the topics do not correspond to real political issues, but rather that it does not match the artefacts introduced by forcing the coder to pick a single topic, while many newspaper articles in fact deal with multiple topics. For example, a lot of newspaper articles about military intervention will talk about defense as well as foreign affairs, and a lot of election news coverage also has a substantive focus. A topic that exemplifies how news stories do not always fit single issue categories is topic 48, which deals with the publication of the Islam-critical film Submission by politician Hirsi Ali and film maker Van Gogh and the subsequent murder of Van Gogh (*Wilders, film, Hirsi Ali, murder, Van Gogh, call, react, Islam, threaten*). This relatively small topic (31 articles) connects civil rights (26%), democracy (26%), crime (13%) and art (13%), and presumably most of these articles actually contain a mix of these issues.

The exploration presented above shows that the different topics highlight specific aspects of issues, with topics differentiating either between specific substantive topics, different temporal events, or different framing. However, it is also clear from those examples that the topics are primarily substantively different topics, and that the framing is connected more to the substance. For example, the human interest framing of crime occurs together with a substantive focus on victims, while the conflict framing co-occurs with a substantive focus on criminal cases. Thus, rather than stating that the topics are frames *per se*, they should be seen as substantive subtopics that may correlate with a preponderance of a certain frame.

Conclusion

This paper explores the relationship between LDA topics and content variables that are traditionally used in political communication: issues, valence, and framing. By running a topic model on a fairly large corpus of manually annotated newspaper articles (n=12,538), we were able to show to what extent the occurrence of LDA topics in articles matches the manual coding of these variables.

Although the results presented above are only a first step in exploring this relation, it yielded a number of interesting findings. First, around a third of the 500 topics in the model were strongly connected (precision > 50%) to a single issue, and the majority had a precision of more than 25% for at least one issue. Second, especially the larger issues were covered by more than one topic, and together these topics provide decent indicators for the issue (average F1 around .5). Third, within an issue the different topics play different roles also depending on the issue. For example, for defense the topics were related to specific temporal events, while for e.g. crime the topics differed in their substantive focus (victims, police, criminal cases). Overall, we can conclude that a large number of LDA topics correspond to clear substantive issues and subissues that match categories that have shown to be politically relevant in the literature.

These findings have some positive implications for using LDA models in political communication research. For one, it shows that at least in the case of general newspaper coverage many LDA topics can indeed be interpreted as substantive issues, and it is often immediately clear from the top words which topics are related to which issue. Thus, one can directly use LDA topics in lieu of manual coding if the accuracy presented here is sufficient for the task. For this, it should be remembered that it is generally not needed to get very high per-document accuracy. For example, the issue Defense had a relatively bad overall F1 score of .41, but the per-year correlation between the LDA topics connected to defense and the manually coded issue was .92. So, a researcher interested in long-term shifts in issue attention can safely use these topics as a measurement of attention for defense.

A second implication is that it can be very interesting the use LDA topics to bootstrap machine learning of issues. Machine learning and active learning has been used with success in the comparative agendas project (Hillard et al., 2008), but still requires a large amount of training material, especially to cover the less frequent topics. Since LDA topics are high-precision indicators of issues, training an LDA model on a large set of unannotated texts should yield features that will improve machine learning performance for smaller training sets.

This paper has many limitations and functions more as an exploration of an interesting question than as a definitive answer to it. For one, all results are based on a single LDA model. Although a quick validity check suggests that the results do not change drastically on re-running the model or when varying the number of topics, this should be explored more thoroughly and especially the influence of the number of topics should be investigated. Also, the findings are presented very qualitatively, and it would be much stronger to develop quantitative metrics for some of the interpretations and run them on a wider range of models.

A final limitation is that the manually coded variables used for this paper were mainly coded at the article level, with only a single value allowed. Since LDA expressly uses multiple topics in an article, this is not a very good match. Moreover, it can be assumed that multiple topics is in fact a better model for news stories, which are seldom limited to a single political issue domain; this has been noted as a problem with agenda coding. It would be very interesting to extend or replicate this study to different domains and with documents that have been coded at a more detailed level of analysis. One possibility is to use election manifestos, which have been coded at the sub-sentence level in the Comparative Agendas Project using the same issue categories described in this paper. It would also be interesting to run LDA models on specific subsets of documents, for example all documents on a certain issues, and see whether the LDA topics match substantive sub topics or are more linked to temporal or even coincidental clusters.

References

Baumgartner, F. R., Breunig, C., Green-Pedersen, C., Jones, B. D., Mortensen, P. B., Nuytemans, M., & Walgrave, S. (2009). Punctuated equilibrium in comparative perspective. *American Journal of Political Science*, *53*(3), 602-âĂŞ619.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993–1022.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*, 267–297.

Hillard, D., Purpura, S., & Wilkerson, J. (2008). Computer assisted topic classification for mixed methods social science research. *Journal of Information Technology and Politics*, *4*(4), 31–64.

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

Semetko, H. A., & Valkenburg, P. M. (2000). Framing European politics: A content analysis of press and television news. *Journal of Communication*, *50 (2)*, 93–109.

Van Noord, G. (2006). At last parsing is now operational. In P. Mertens, C. Fairon, A. Dister, , & P. Watrin (Eds.), *Verbum ex machina, actes de la 13e conference sur le traitement automatique des langues naturelles* (pp. 20–42). Louvain-la-Neuve, Belgium: Presses Universitatires de Louvain.