Digital Journalism

ISSN: 2167-0811 (Print) 2167-082X (Online) Journal homepage: http://www.tandfonline.com/loi/rdij20

Quantitative analysis of large amounts of journalistic texts using topic modelling

Carina Jacobi, Wouter van Atteveldt & Kasper Welbers

To cite this article: Carina Jacobi, Wouter van Atteveldt & Kasper Welbers (2015): Quantitative analysis of large amounts of journalistic texts using topic modelling, Digital Journalism, DOI: 10.1080/21670811.2015.1093271

To link to this article: http://dx.doi.org/10.1080/21670811.2015.1093271

QUANTITATIVE ANALYSIS OF LARGE AMOUNTS OF JOURNALISTIC TEXTS USING TOPIC MODELLING

Carina Jacobi, Wouter van Atteveldt and **Kasper Welbers**

The huge collections of news content which have become available through digital technologies both enable and warrant scientific inquiry, challenging journalism scholars to analyse unprecedented amounts of texts. We propose Latent Dirichlet Allocation (LDA) topic modelling as a tool to face this challenge. LDA is a cutting edge technique for content analysis, designed to automatically organize large archives of documents based on latent topics, measured as patterns of word (co-)occurrence. We explain how this technique works, how different choices by the researcher affect the results and how the results can be meaningfully interpreted. To demonstrate its usefulness for journalism research, we conducted a case study of the New York Times *coverage of nuclear technology from 1945 to the present, partially replicating a study by Gamson and Modigliani. This shows that LDA is a useful tool for analysing trends and patterns in news content in large digital news archives relatively quickly.*

KEYWORDS automatic content analysis; journalism; nuclear energy; topic models

Introduction: Latent Dirichlet Allocation, Topics and Issues

The shift of news media towards online publication and archiving provides journalism scholars with new opportunities for studying journalism. At the same time, understanding the complicated dynamics of this contemporary media landscape requires an ever-larger scale of analysis, with more outlets and more content per outlet. In this article, we show how topic modelling, a relatively new method developed in the computational linguistics field, can help analyse large amounts of text without requiring manual coding, thus reducing the time and costs of such projects. For a general overview of the available methods and pitfalls for topic models, see Grimmer and Stewart (2013). In this paper, we focus on one type of topic model, Latent Dirichlet Allocation (LDA; Blei, Ng, and Jordan 2003), and demonstrate its use for journalism research. Even though topic modelling is a promising method for text analysis, with the seminal paper in computational linguistics (Blei, Ng, and Jordan 2003) published around a decade ago, it is just starting to be used in the social sciences. Political scientists, who, like journalism researchers, have both the challenge and the opportunity of newly available online archives of textual data (such as political speeches, legislative documents and social media) have started to use topic models to automatically classify these documents.

Digital Journalism, 2015
http://dx.doi.org/10.1080/21670811.2015.1093271

Notably, Quinn et al. (2010) classify speeches in the US Senate into topics using topic modelling. Lucas et al. (2015) apply topic modelling to different types of documents such as fatwas and social media posts in order to facilitate comparison between countries. Following such studies, we explore whether topic modelling can be used to classify journalistic documents into categories that are meaningful to journalism researchers.

LDA, like other topic modelling algorithms, is an unsupervised technique that automatically creates topics based on patterns of (co-)occurrence of words in the documents that are analysed. Journalistic texts are thus "coded automatically" for topics, although it is up to the researcher to interpret the results of the model and to set up the analysis in such a way that the results are useful for the study at hand. Thus, the usefulness of the technique for studying journalism crucially depends on the correspondence between topics and the constructs of theoretical interest. The goal of this article is to introduce LDA to journalism scholars and to provide a practical guide in applying the technique to their own research. Concretely, we will deal with three broad topics:

- *What is topic modelling?* The first part of this article will give a brief and mostly non-technical description of LDA.
- *How to set up an LDA topic model*: Secondly, we will describe the different parameters of the LDA topic model, and discuss issues of validity.
- *Theoretical interpretation*: The last and most important part of the article discusses how LDA topic models relate to theoretical constructs of interest to journalism researchers, especially issues and frames. Using the example of the news discourse on nuclear technology from 1945 to now, we show how LDA topics mostly correspond to the important issues in this discourse, comparing our results to the earlier study by Gamson and Modigliani (1989).

What is Topic Modelling?

Topic models are computer algorithms that identify latent patterns of word occurrence using the distribution of words in a collection of documents. The output is a set of *topics* consisting of clusters of words that co-occur in these documents according to certain patterns. In an LDA model, each document may contain multiple topics. Each of the topics has an internal consistency—the words in that topic often occur together in the documents, and/or do not appear much outside that topic. The researcher determines what this consistency refers to, and thus how the topic can be interpreted (Chang et al. 2009). It is this interpretability that determines whether topic models in general, and LDA specifically, are of use to social scientists.

In the case of journalism research, the collection of topics inferred by the model would ideally resemble a categorization of issues or frames based on substantive theory, for example the issue list used by the Comparative Agendas Project that uses categories such as Macro-economics, Foreign Affairs and Crime to categorize legislative and journalistic texts (Jones and Baumgartner 2005). However, topics are created by the LDA algorithm based on patterns of word co-occurrence in documents, which do not necessarily match theoretical concepts. It seems plausible to equate a topic with the theoretical concept "issue" or "theme", but topics could potentially also represent writing or speaking styles (e.g. words referring to emotions), events (e.g. a natural disaster)

or frames (e.g. immigrants framed as criminals), which, at least theoretically, are also formed through co-occurrence patterns of specific words. For example, in the study of Quinn et al. (2010) of topics in Congressional Record speeches, words such as violence, drug trafficking, police and prison were interpreted as being the topic *Law and Crime I: Violence/Drugs*. Another topic, *Abortion*, contained words such as baby, abortion, procedure and life (Quinn et al. 2010, 217). This study thus interprets topics as issues, whereas DiMaggio, Nag, and Blei (2013, 590–591) refer to "voices" or "frames" when interpreting different topics. However, what exactly topics represent, and if they represent different concepts given different input parameters in the model, is ultimately an empirical question.

To illustrate what topics look like and the interpretation challenges they present, consider the example newspaper article in Figure 1. This article, which appeared after the Chernobyl disaster, criticized the Soviet Union for suppressing news about the event. In the text, each word is highlighted to indicate which topic it was drawn from. These topics were found automatically by the algorithm, but we interpreted and named the topics in a way similar to how one would interpret a factor analysis outcome.[1]

As would be expected given the content of the article, many words are drawn from a topic that we interpret as the *Nuclear Accidents* topic, indicated in dark grey with white letters. This topic includes words such as "Chernobyl", "disaster" and "radiation", but interestingly also contains the words "last Friday", probably due to the episodic nature of accident reporting. The other main topic, *Cold War*, is indicated in medium grey and contains words such as "Soviet Union" and "Gorbachev" but also "confidence" and "pledge". Finally, a number of terms such as "secrecy" and "peaceful use" of "energy" are drawn from the Nuclear Research topic, and two words are drawn from other topics.

This example highlights a number of interesting points about LDA. First, this document is split between two main topics, *Cold War* and *Nuclear Accidents*. In a coding scheme forced to have a single topic per document, it would be very difficult to choose the "dominant" topic for this article, so in our opinion this accurately reflects the nature of the article. Second, you can see that not all words are included in the analysis: most of the words in the article are not used by the algorithm (the text without highlighting), either because they are non-substantive words such as determiners or prepositions ("the" or "it"), which we excluded, or because they are too rare ("Elbe", "perish") or too common ("have" but in this context also "nuclear", since that was used

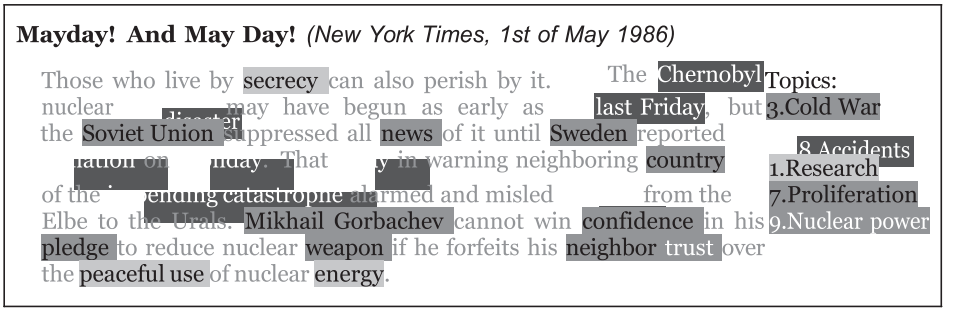

FIGURE 1
Example article with words from different topics highlighted in the text

to select the articles). Finally, it should be remembered that no *a priori* coding scheme was used by the computer, so the topics in this document were found completely automatically. The fact that the resulting topics (such as *Cold War* and *Nuclear Accidents*) make sense substantively and match the way we would define the subject or topic of articles—such as the example here—shows that to some extent our notion of "topic" matches the latent classes or co-occurrence patterns identified by the LDA algorithm.

Topics such as *Cold War* and *Proliferation* can be interpreted as issues, but also as ways of framing nuclear weapons. In such cases, a particular perspective on an issue can be signified by specific keywords, which can be seen as framing devices (Matthes and Kohring 2008). In general, if framing devices correspond to specific (latent) patterns of vocabulary use, LDA can capture these classes in specific topics, and as such LDA results can also include the frames used in a corpus of texts.

How to Set Up an LDA Topic Model

Before exploring a case study where we interpret LDA topics, we will explain how to perform an analysis using this technique. As previously mentioned, LDA processes a collection of documents and clusters the words in these documents into topics in an unsupervised way. As such, it is important to start by considering which documents should be included, and to make sure these documents consist only of the content of newspaper articles—other textual information such as author name or reader comments should be removed.

We will discuss the steps we took when conducting the analysis for our case study below. Our data consisted of 51,528 news stories (headline and lead) from the *New York Times* that mentioned nuclear power, published between 1945 and 2013. We retrieved all news stories from the *New York Times* online archive (http://developer.nytimes.com) that contained the search terms "nuclear", "atom" or "atomic" in the headline or lead.

Preprocessing: Tokenization and Lemmatization

A topic model does not analyse documents directly, but uses a so-called document–term matrix based on these documents. This matrix lists the frequency for each term (word) in each document. The first step in creating this matrix is tokenization, which means splitting the text into a list of words. Although this can be done by splitting on white space and punctuation, there are good word boundary detection tools that recognize acronyms, contractions, etc.

It is often better to begin by reducing the size of the matrix by preprocessing and feature selection, reducing computing time and improving results. An important step in preprocessing is stemming or lemmatizing. Stemming is a simple technique where the ending of words is "chopped off", leaving the stem. For example, using the frequently used Porter stemming algorithm, "weaknesses" becomes "weak", and both "failures" and "failure" become "failur". This technique does not handle irregular conjugations, for example "are" and "were" have different stems. A more powerful (and

computationally demanding) technique is lemmatizing. Lemmatizing reduces all words to their "lemma" using a lexicon in combination with regular conjugation rules. Thus, lemmatization reduces both "is" and "were" to their lemma (to) "be". For English, stemming is often sufficient, but for more richly inflected languages such as German or Dutch, lemmatization tends to give better results (Haselmayer and Jenny 2014).

Feature Selection

The next step is feature selection. A moderately large corpus can typically contain more than 100,000 unique words, with most of these words occurring in only a few documents. Applying an LDA model on all words in a corpus is both computationally expensive and not very useful, as most words have distribution patterns that do not contribute to meaningful topics. For example, some words are too frequent to be informative—words like "the" and "have" generally occur in every document regardless of topic. A useful technique to filter out words that are too rare or too common is to use *tf-idf* (term frequency–inverse document frequency), which assigns a low score to words that are either very rare or very frequent. Another option is to have a minimal frequency cut-off to filter out the rare words and use a list of common stop words (and/or cap the inverse document frequency) to filter out overly common words.

Lemmatization software is often combined with POS-tagging. POS (part of speech) tags indicate the type of word, e.g. verb, noun or preposition. For topic modelling, it is often best to only use specific parts of speech, especially nouns, proper nouns and, depending on the task and corpus, adjectives and verbs. This automatically filters out the most common stop words, which tend to be determiners or prepositions (with the exception of common verbs like "to be" and "to have").

For the analysis presented here, we used the lemmatizer and POS-tagger from Stanford's corenlp suite (De Marneffe, MacCartney, and Manning 2006), and selected all nouns, verbs, adjectives and proper nouns. We filtered out all terms with a frequency of less than 20 and which occurred in more than 25 per cent of documents, and we removed all terms that contained a number or non-alphanumeric characters, yielding a total vocabulary of 8493 terms.

Choosing Parameters

After creating a sufficiently small and relevant document–term matrix on which to run the model, there are some more choices the researcher needs to make before running the model. Although one of the main advantages of a topic model is that no *a priori* coding schemes need to be supplied, there are certain parameters that need to be set. In particular, the number of topics (K) needs to be specified, which indicates into how many topics the LDA model should classify the words in the documents. There is no default or simple rule of thumb for this parameter. The trade-off is comparable to factor analysis: the goal is to describe the data with fewer dimensions (topics) than are actually present, but with enough dimensions so that as little relevant information as possible is lost. We first discuss and use the perplexity measure, which is a commonly used computational indication for the correct amount of topics (Blei, Ng,

and Jordan 2003). However, we stress that this should be only used to make an initial selection of models with an acceptable amount of information loss, and that interpretability of topics is a more important criterion for social science purposes.

A second parameter is the *alpha* hyperparameter, which affects how many topics a document can contain.[2] A common default value for the alpha is 50 divided by the number of topics. Substantively, a lower alpha leads to a higher concentration of topic distributions within documents, meaning that documents score high on a few topics rather than low on many. Accordingly, if the goal is to assign one or a few topics per document then it makes sense to use a low alpha (Haselmayer and Marcelo 2014).

Perplexity

From a computational perspective, a good indication of the right number of topics is that number with which the model best predicts the data. This is comparable to goodness-of-fit measures for statistical models.[3] For topic models such as LDA, a commonly used indicator is perplexity, where a lower perplexity indicates a better prediction (Blei, Ng, and Jordan 2003). To calculate the perplexity, we first train an LDA model on a portion of the data. Then, the model is evaluated using the held-out data. This routine is repeated for models with different numbers of topics, so that it becomes clear which amount leads to the lowest perplexity.

Figure 2 shows the perplexity of different models for our data. We trained the LDA models using 30,000 of the 48,604 documents, and then calculated the perplexity of each model over the remaining 18,604 documents. We varied the number of topics

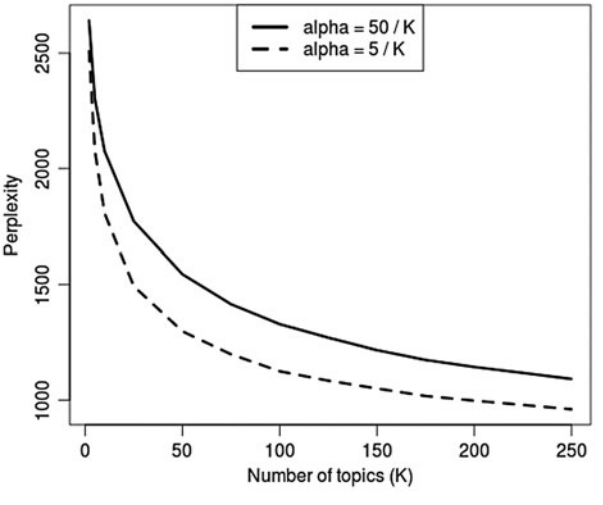

FIGURE 2

Perplexity of LDA models with different numbers of topics and alpha

Notes: The line graph shows how perplexity decreases (and model fit thus increases) as the number of topics increases. The number of topics that corresponds to a great change in the direction of the line graph is a good number to use for fitting a first model. For example, $K = 25$ for our model with alpha = $50/K$.

used: 2, 5, 10, 25, and then on to 250 in steps of 25. The results show that perplexity decreases as the number of topics increases, implying that a model with more topics is better at predicting the held-out data. At around 25 and 50 topics, the decrease in perplexity for additional topics becomes notably less. This is one way to interpret the right number of topics, similar to the interpretation of the elbow in the scree plot of a factor analysis. The other way is to look at the number of topics with the lowest perplexity. We can see that this point will be somewhere beyond 250.

However, having the "right" number of topics in a mathematical sense does not say anything about the interpretability of the topics produced. In fact, mathematical goodness-of-fit measures and human interpretation may lead to contradicting conclusions on the best number of topics, given that especially with a high number of topics, the computer algorithm may find nuances that are not semantically meaningful to humans (Chang et al. 2009). Also, as we use topics to answer substantive questions about the documents we study, it is important that the topics that result from the analysis contribute towards answering these questions, instead of providing the best prediction of the data. For the analysis presented in this article, we thus looked both at perplexity and interpretability when deciding on the number of topics to use. Judging from the perplexity, a good choice is probably between 25 and 50 topics. However, to facilitate comparing our results with Gamson and Modigliani (1989), we will first use a simpler model with $K = 10$ topics, and then show the differences between this model and a model with $K = 25$ topics.

Alpha

Regarding the alpha hyper-parameter, since the data used for our analysis cover a long history of textual data concerning an issue that involves various events and viewpoints, it makes sense to define several clearly distinguishable topics. In addition to better scores for the lower alpha, this is a substantive argument for us to use the lower alpha of five divided by K instead of the default.

Tool Support for LDA

The easiest way to get started with LDA is through the open-source statistical package R.[4] Although specialized software for topic models is available, such as MALLET (McCallum 2002) or the Stanford Topic Modeling Toolbox (Ramage et al. 2009), an advantage of using R is that it is a statistical package that many social scientists already use for other analyses. However, LDA works the same no matter the software used to run the model. The tm package in R can automatically generate the document–term matrix from texts and includes options for stemming and feature selection (Meyer, Hornik, and Feinerer 2008). The topicmodels package can directly fit an LDA model from the document–term matrix object created by tm. For more sophisticated preprocessing, we use the Stanford corenlp suite (De Marneffe, MacCartney, and Manning 2006), which contains a collection of modules for grammatical analysis. For our case study, we uploaded our documents in the Web-based content analysis toolkit AmCAT.[5] For the preprocessing and the analysis itself, we used the statistical computing environment R,

that can connect to AmCAT directly to retrieve the documents there.[6] AmCAT uses the xtas extensible text annotation suite to automatically perform the preprocessing in a scalable manner (De Rooij, Vishneuski, and De Rijke 2012).[7]

Nuclear Topics: 1945 to Now

In order to demonstrate the use of LDA to explore the topics in a given set of documents, and show the change in these topics over time, we have performed LDA on *New York Times* articles dealing with nuclear technology. The famous study by Gamson and Modigliani (1989) shows how the framing of the issue "nuclear power" in the news changed over time since 1945. We were interested to see whether the topics found by performing an LDA over newspaper articles from this period show similar changes over time. Additionally, we extended the research period to 2013, to include more recent coverage of nuclear technology.

The question is thus whether the change in culture surrounding nuclear issues found by Gamson and Modigliani is expressed in a change in word use over time that is captured by LDA. Gamson and Modigliani (1989) identified seven "packages" or frames in newspaper and television coverage of nuclear energy between 1945 and 1989: *Progress*, *Energy Independence*, *Devil's Bargain*, *Runaway*, *Public Accountability*, *Not Cost Effective* and *Soft Paths* (Gamson and Modigliani 1989, 24–25). If we compare the outcome of an LDA analysis with the results of Gamson and Modigliani's study, to what extent do we find similar results? Note that for a number of reasons we do not expect perfect correspondence between the LDA topics and Gamson and Modigliani's packages, even if we disregard the difference between manual and automatic analysis. First, these packages were identified by examining not only the text of news stories, but also images and cartoons, with a focus on editorial content, whereas our analysis is performed on the lead paragraphs of news stories only. Second, whereas Gamson and Modigliani only analyse nuclear power, our investigation deals with the whole coverage of nuclear technology, including nuclear weapons. Finally, our investigation covers the post-Cold War period as well, while Gamson and Modigliani of course only analyse the discourse until the 1980s. Nonetheless, it is interesting to compare our findings to theirs since it shows to what extent the results of an automatic topic modelling approach compare to those of a well-known and very thorough manual analysis.

As discussed above, for reasons of comparability to Gamson and Modigliani's seven packages, we will first focus on an analysis using 10 topics. Table 1 shows the topics that resulted from this analysis, including our interpretation and the 10 words that represent a topic most strongly. To facilitate interpretation we also made a topic browser,[8] which is a tool to interactively explore the results of a topic model (Gardner et al. 2010). Our topic browser features two extra pieces of information in addition to top words. One is the top documents assigned to a topic in which the topic words are highlighted. The other is a semantic network, or semantic map, which visualizes the co-occurrence of top words, thus showing topic coherence and facilitating interpretation. For our discussion, we categorized the topics into: (1) topics that show a pattern in their use over time; (2) topics that are more or less continuously present; and (3) three topics that turned out to be irrelevant for our case study.

TABLE 1
LDA results on US nuclear discourse, 10 topics

Topic	Interpretation	Most representative words
Topics with temporal patterns		
1	*Research*	atomic, Energy, WASHINGTON, scientist, energy, bomb, Commission, United, research, weapon
3	*Cold War*	United, States, Union, Soviet, soviet, weapon, arm, missile, President, treaty
7	*Proliferation*	Iran, United, North, Korea, program, weapon, States, official, country, China
8	*Accidents/ Danger*	plant, power, reactor, Island, Nuclear, accident, Commission, official, waste, safety
Continuous topics		
5	*Weapons*	test, submarine, Japan, first, Navy, year, explosion, missile, ship, bomb
9	*Nuclear Power*	power, plant, company, year, energy, percent, utility, cost, Company, reactor
10	*US Politics*	war, President, weapon, Mr., year, military, policy, world, Reagan, House
Irrelevant topics		
2	*Summaries*	New, new, year, government, official, York, people, business, President, state
4	*Book Reviews*	week, life, book, man, woman, John, year, New, family, University
6	*Films & Music*	Street, West, Theater, Mr., Sunday, East, show, New, tomorrow, p.m.

Topics That Show Some Pattern in Their Use Over Time

Firstly, we found a number of topics that have a strong temporal dimension, that is, they are strongly present in the news in some years or decades, but not in others. In that respect, these topics are most similar to the shifting packages found by Gamson and Modigliani. Figure 3 shows the change in occurrence over time for the four topics discussed below, which are discussed in chronological order.

News stories in which topic 1, which we labelled *Research*, deal with research on nuclear technology, including both energy research and nuclear weapons research. This topic is most strongly present in the early part of the data-set, and its usage sharply decreases over time, especially from the 1980s onwards. In terms of temporal focus, and in the focus on possibilities created by nuclear research, this topic is comparable to Gamson and Modigliani's Progress package, although the latter did not include nuclear weapons research as their research was focused on nuclear energy.

In topic 3, *Cold War*, the words "United", "States", "Soviet", "Union" and "weapon" immediately suggest that this is a topic about the US–Soviet conflict. Lower-ranking words include variations on the "weapon" theme as well as more diplomatic terms such as "agreement" and "proposal". The topic occurs most frequently between the mid-1950s and the mid-1980s, with a peak in the early 1960s that can be easily identified as the Cuban missile crisis.

Finally, the last topic in this category is topic 8, *Nuclear Power– Accidents/Danger*. Although the top words in this cluster, "plant", "power" and "reactor", are not very informative, peripheral words like "accident", "safety" and "radioactive" show how to

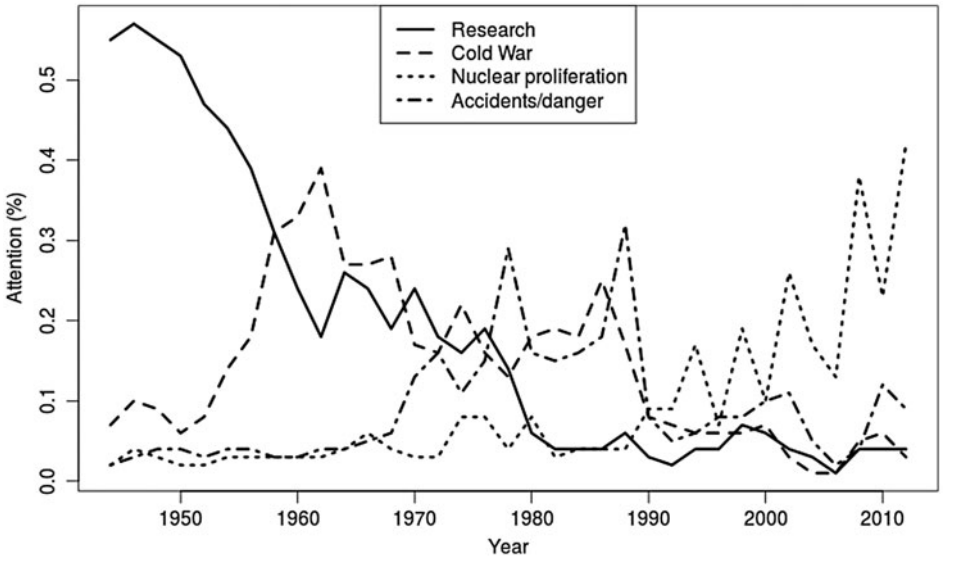

FIGURE 3
Occurrence of topics that have a strong temporal component

interpret it. Indeed, Chernobyl and Three Mile Island are both included in this topic, which shows peaks in news attention in both 1979 and 1986. Articles on smaller nuclear accidents or discussions on nuclear hazard in general also contain this topic. Interestingly, although this topic shows a small peak around the 2011 Fukushima reactor, most of the coverage of that event is classified in topic 5 (nuclear weapons), discussed below. This topic is most closely related to Gamson and Modigliani's *Runaway* and *Devil's Bargain* packages, as it focuses on the negative qualities of nuclear energy, and both the topic and the *Runaway* frame occur in roughly the same time period.

Finally, the articles in topic 7, *Nuclear Proliferation*, deal with nuclear weapons in countries such as North Korea, Iran and Pakistan, and US actions or policies against the possession of these weapons. Peaks of attention for these topics occur in the 1990s and especially the 2000s, with events in North Korea and later Iran as triggers. Although the topic occurs after the period covered by Gamson and Modigliani, semantically it resembles their *Runaway* package, in the sense that the technology is no longer under control and now poses a danger to its very inventors.

More or Less Continuously Present Topics

These topics do have some fluctuation in their occurrence over time, but this fluctuation shows no clear trend.

In topic 5, *Nuclear Weapons*, the words associated with the topic seem to be mostly related to the development and testing of nuclear bombs, especially nuclear submarines. This topic is most strongly present in the 1950s through the early 1970s, with a peak in 2011 after the Fukushima incident. This latter peak is possibly a confusion caused by the prevalence of ocean-related words in both the tsunami coverage and the discourse on nuclear submarines.

Topic 9, *Nuclear Power*, is a second cluster with words related to economics. Similar to topic 8, "power" and "plant" form the top words of a cluster. However, this time the peripheral words show a different focus: "company", "corporation", "year" and "percent" suggest that these articles are about nuclear energy as a business, and report on the earnings of companies that operate in this sector. Over time, we see a peak in the 1970s (the oil crisis) and in the 1980s.

The final relevant topic, topic 10, *US Politics*, concerns the policies of the United States on its own nuclear weapons and defence. The main peak for this topic is in the early 1980s, with the Rearming America programme.

Irrelevant Topics

These topics have nothing to do with nuclear power in terms of their content, but appear in our results anyway because we included all articles mentioning the word "nuclear", including book reviews and news summaries. In topic 2, news summaries and items from the section "Inside The Times" are clustered together as one topic. These tend to focus on New York and on politics (local, domestic or foreign). However, they do not deal with nuclear issues directly, so we discard this topic. Topic 4 consists of short book reviews, of which one or more use words related to the nuclear topic. Again, these do not deal with nuclear issues in the news directly. Lastly, topic 6 represents news stories on films or concerts, some of which have to do with nuclear power, others show up in articles that mention the nuclear issue elsewhere.

As this overview shows, not all topics contain useful results and the topics are not ordered in a way that makes it easy to distinguish the useful from the non-useful (as usefulness is of course something determined by the researcher, not the computer). However, the irrelevant topics were clearly distinguishable, which makes it easy to discard them altogether or ignore them in further analysis. Although it might seem annoying at first that such topics are also generated by the analysis, this is actually quite useful. Since most data-sets contain a degree of noise (such as book reviews or sports results), LDA can be used as an inexpensive way to locate those articles that are relevant for answering a particular research question from a larger sample.

Our purpose here, however, was to compare the topics we found using LDA with the outcome of the framing research by Gamson and Modigliani (1989). Compared to the frames or interpretive packages found in their analysis, the topics of our LDA analysis seem to be more concrete and specific. Similar to Gamson and Modigliani's study, we found that topics change over time, but not all of them increase or decrease in a linear way. Also, topics seem to either cluster a number of related events together (nuclear proliferation talks, nuclear accidents) or represent issues that are continuously present over a longer period of time (the economics of nuclear energy, nuclear weapons tests). It is not possible, however, to deduct a particular viewpoint or frame from a topic directly—for example, we found no clear "anti-nuclear" cluster, whereas Gamson and Modigliani found multiple frames that are critical of nuclear energy. It is quite likely that the coverage of nuclear accidents and danger is predominantly covered from an anti-nuclear perspective, but even in this case there is a clear difference between the "issue" or event being covered (nuclear accidents) and the frame with which it is

covered. That said, for some of the topics, such as the *Research* and *Accidents* topics, we do see that the temporal pattern is similar to that identified for the *Progress* and *Runaway* packages, respectively.

To see whether increasing the number of clusters helps find word patterns that are more fine-grained and more frame-like than those representing issues, we will explore what happens if we increase the number of topics from 10 to 25.

Granularity: How Many Topics?

Changing the number of topics (K) changes what is called the granularity, or level of detail, of the model. The higher the granularity, the more detailed the analysis. Using a larger number of topics implies higher granularity: each topic then represents more specific content characteristics.

If a model with 25 topics is compared to a model with 10 topics, some of the topics in the model with 25 topics tend to blend together in the 10-topic model. The topics in the model with 50 topics can thus be considered to represent smaller grains of the 10 topics. As an example, consider the events in Chernobyl and Three Mile Island. Both events are clearly separated by time, space, the actors involved and various circumstances. Yet the two are related by the nature of the event, and thus in vocabulary used: a malfunction of a nuclear reactor, with awful consequences to health and the environment. In a model with many topics, these events can be distinguished in different topics, whereas in a model with fewer topics, these events are blended together, representing a broader theme. In our data, we saw this specific example in a comparison of models with 10 and 25 topics.

Table 2 gives an overview of the relevant topics from the 25-topic model. Each "detailed" ($K = 25$) topic is listed below the "broader" ($K = 10$) topic it resembles most. This similarity is computed by determining in which documents each $K = 10$ and $K = 25$ topic occurs, and then calculating the cosine similarity of these occurrence vectors. So, two topics with perfect similarity (1.0) would occur in exactly the same (relative) frequency in all documents. For instance, topic 24 which we labelled *3 Mile Island*, is most similar to topic 8 above it (with cosine similarity 0.70).

In the $K = 10$ model, coverage of Chernobyl is clustered together with coverage of the Three Mile Island incident, to form a cluster about nuclear accidents (topic 8), and coverage of Fukushima was split between this topic and the nuclear weapons topic (topic 5). However, in the $K = 25$ model, different connections between these events are found, and Chernobyl and Fukushima end up in the same cluster, while Three Mile Island gets its own cluster. Finally, topic 8 ($K = 10$) is highly similar to topic 15 ($K = 25$), which deals with power plant construction and especially the Shoreham nuclear power station. This power station was constructed on Long Island in the 1980s, but never actually used, as local residents objected in the wake of the Chernobyl disaster. As shown in Figure 4, these three constituent topics in the $K = 25$ model together trace the $K = 10$ *Accidents/Danger* topic (the per-year correlation between the $K = 10$ topic and the sum of the $K = 25$ topics is 0.95, $p < 0.001$), but each are focused around a specific time-point.

TABLE 2
Most similar $K = 25$ topics for each relevant $K = 10$ topic

Topic	Similarity
1: Research	
9: Research	0.68
8: Universities	0.39
17: Scientific Development	0.27
21: Nuclear Weapon Materials	
3: Cold War	
7: Rearming America	0.61
18: Cold War	0.61
20: NATO	0.50
8: Accidents/Danger	
24: 3 Mile Island	0.70
15: Power Plant Construction (Shoreham)	0.62
2: Chernobyl & Fukushima	0.32
7: Nuclear Proliferation	
6: Iran	0.63
4: North Korea	0.58
14: Iraq	0.42
3: India & Pakistan	0.25
21: Fissile Materials	0.25
5: Nuclear Weapons	
11: Nuclear Submarines	0.53
2: Chernobyl & Fukushima	0.39
19: Nuclear Weapons	0.31
23: Nuclear Power	0.52
10: US Politics	
1: Nuclear War Threat	0.68
5: US Politics	0.46
19: Nuclear Weapons	0.36
12: Protests	0.28

Topics from the 10-topic model ($K = 10$) are set in italics, with the most similar topics from the 25 topic model ($K = 25$) listed beneath each $K = 10$ topic. Some $K = 25$ topics occur twice when they were similar to more than one $K = 10$ topic. Similarity is based on whether topics occur in the same documents (calculated as cosine similarity). Example: Topic 3 from the $K = 10$ model, interpreted as the Cold War topic, is similar to three topics from the $K = 25$ model: topic 7 on Reagan and Rearming America (sparking the 1980s' arms race); topic 18 on the Cold War; and topic 20 on NATO.

Discussion: Best Practices

This article showcased a relatively recent tool, Latent Dirichlet Analysis, or LDA. LDA is an unsupervised topic modelling technique that automatically creates "topics", that is, clusters of words, from a collection of documents. These topics may represent issues that recur over time, related events or other regularities in articles.

We think LDA can be a valuable tool in any large-scale content analysis project. For preliminary analysis, LDA can very quickly give a rough overview of what kind of topics are discussed in which media or time periods. However, the best way of proving the statistical, internal and external validity of LDA and of topic models in general is still under discussion (Chang et al. 2009; DiMaggio, Nag, and Blei 2013; Ramirez et al. 2012). We would advise journalism scholars to start by making a perplexity plot for different numbers of topics, and then look at the models where the perplexity decrease drops

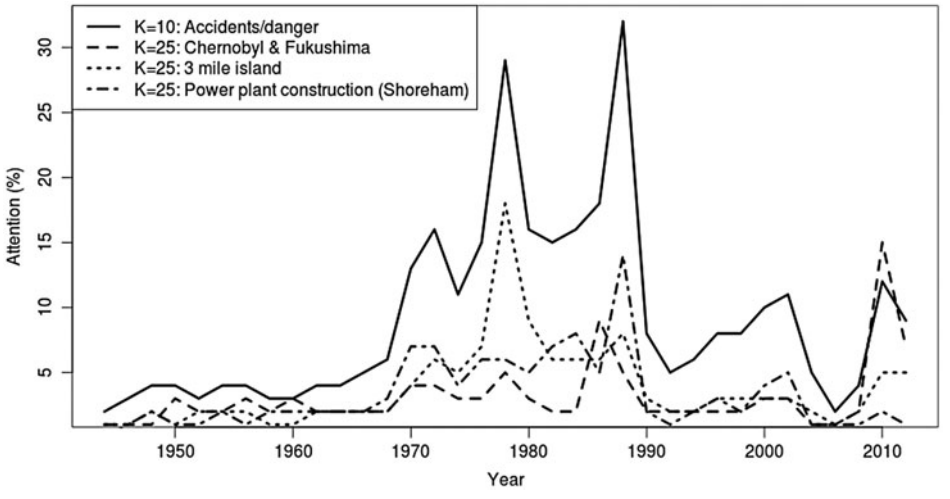

FIGURE 4
Occurrence over time of detailed (K = 25) topics that constitute the *Accidents/ Danger* topic from the K = 10 model

off. However, the researcher should also manually inspect these topic models, and for each topic decide on the correct interpretation by looking at the words in the topic, in which media and time frames it occurs, and also at the documents that are most indicative of a specific topic. For this purpose, we advise using a topic browser.[9] Manual interpretation reveals which topics are closely related to the theoretical quantities the researcher is interested in. Also, the researcher can decide to combine multiple topics that are semantically related and/or to remove irrelevant topics such as the book reviews identified in our case study. This should be followed by a formal internal validity evaluation of these topics by checking a sample of automatically coded articles or by comparing to a sample of manually coded articles (for a more elaborate discussion of validity checks of LDA topic models, see DiMaggio, Nag, and Blei (2013). Even though the validity may be insufficient for completely automatic analysis, it can be used as a form of semi-automatic coding by having the manual coder check whether the coding is correct, which is much quicker (and cheaper) than fully coding each document. Furthermore, even if the topics cannot be immediately used to answer the substantive research question, they can enhance the subsequent manual or automatic content analysis in numerous ways. First, by inductively showing which topics occur, it can help the researcher to create or improve the codebook by suggesting new codes and examples and by showing which codes might need to be combined or split up. Second, if the researcher is interested in doing automatic keyword analysis, the word lists per topic can offer inspiration for keywords and synonyms that might otherwise be left out. Third, LDA can be used to quickly find good example documents for infrequent topics or documents that use multiple topics, such as the example document in Figure 1. This can be used for manual coder training and evaluating the codebook, but also for creating the training data for subsequent machine learning, where rare topics usually give the worst performance because of the lack of training documents. Finally, as shown by the three "irrelevant" topics that were derived from our analysis, LDA can be used to filter out categories of texts that are not relevant from an overall sample.

The comparison in this paper also shows two limitations of LDA for analysing journalistic texts. First, not all topics represent substantive word clusters, but also other consistent patterns of co-occurrence such as genre and writing style. This is most dramatically shown by "irrelevant" topics such as book reviews, but can be considered beneficial by allowing the researcher to quickly discard such clusters of documents. However, it is also possible that shifts in word use, such as from "atomic" to "nuclear", caused documents to belong to different topics in the 1950s as compared to the 2000s even though they are substantively similar. The same would hold for writing style differences between different media and especially different formats (e.g. print media versus television or online). Dynamic Topic Models (Blei and Lafferty 2006) and Structural Topic Models (Roberts et al. 2014) are two extensions of LDA that explicitly deal with shifts over time and between groups of documents, respectively.

Second, the topics identified in this study did not approach what we could call a "frame" in the sense that Gamson and Modigliani (1989) or Entman (1993) would use this concept—as coherent interpretative packages. Similarly, the topics did not represent explicit valence or sentiment—there were no clear pro- or anti-nuclear topics. Although attempts to combine LDA and sentiment analysis in one model have been made (Lin and Yulan 2009), since news coverage is difficult for automatic sentiment analysis in general (Balahur et al. 2013), such methods will likely not yield sufficiently valid results for journalism studies in the near future. Following DiMaggio, Nag, and Blei (2013, 603), we suggest further research can combine LDA as described here with frame and sentiment analysis using other methods, e.g. using machine learning (e.g. Burscher et al. 2014).

By showcasing LDA and by showing some best practices for running and interpreting model results, this article contributes to the adoption of topic modelling in the practice of journalism research, which is a useful technique for every digital journalism scholar to have in their toolbox to deal with the very large data-sets that are becoming available. Although the burden of making sense of the results is still on the researcher, LDA offers a quick and cheap way to get a good overview of a collection of documents so that substantive questions can be answered immediately, especially about broad patterns in topic use over time or between media. Additionally, it is very helpful for performing preliminary analysis before venturing on a more traditional (and expensive) automatic or manual content analysis project.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the authors.

NOTES

1. Factor analysis is a dimensionality reduction technique: given a set of observed variables, a smaller set of factors is calculated that preserve as much information as possible in a lower-dimensional space. This is often used in the field of psychology as a measurement of latent, unobserved causes for certain observations. For instance, if a single factor largely explains the results for a set of questions

relating to anxiety, the factor can be interpreted as a measurement of anxiety. Similarly, a topic in topic modelling can be interpreted and named based on what the main words have in common.

2. Technically, the alpha hyper-parameter controls the concentration of the Dirichlet distribution regarding the distribution of topics over documents. In Bayesian statistics, a hyper-parameter is a parameter that controls distributions such as the Dirichlet distribution. The term hyper-parameter is used to distinguish them from the parameters of the topic model that is the result of the analysis. For a good explanation of the role of hyper-parameters, we suggest the introduction to the Dirichlet distribution by Frigyik, Kapila, and Gupta (2010).
3. A goodness-of-fit measure describes how similar the predicted or expected values of a model are to the actual observed values. An example is the R^2 measure in linear regression, which indicates what proportion of variance of the dependent variable is explained by the independent variables.
4. See http://www.r-project.org.
5. See http://amcat.nl.
6. See http://github.com/amcat/amcat-r for the relevant R code. The R scripts that were used for our analysis can be downloaded from http://github.com/AUTHOR/corpus-tools.
7. See http://github.com/AUTHOR/xtas for the xtas modules for corenlp and other lemmatizers.
8. The topic browser can be found at http://rpubs.com/Anonymous/78,706.
9. Our R script for creating a topic browser is available at http://github.com/vanatteveldt/topicbrowser.

REFERENCES

Balahur, Alexandra, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik Van Der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. 2013. "Sentiment Analysis in the News."*Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010)*: 2216-2220. Valletta, 19–21 May 2010.

Blei, David. M., and John D. Lafferty. 2006. "Dynamic Topic Models." In *Proceedings of the 23rd International Conference on Machine Learning*, 113–120. Pittsburgh, 2006. doi: 10.1145/1143844.1143859.

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

Burscher, Björn, Daan Odijk, Rens Vliegenthart, Maarten de Rijke, and Claes H. de Vreese. 2014. "Teaching the Computer to Code Frames in News: Comparing Two Supervised Machine Learning Approaches to Frame Analysis." *Communication Methods and Measures* 8 (3): 190–206. doi: 10.1080/19312458.2014.937527.

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David M. Blei. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." In *Advances in Neural Information Processing Systems*, 288–296.

De Marneffe, Marie-Catherine, Bill MacCartney, and Christopher D. Manning. 2006. "Generating Typed Dependency Parses from Phrase Structure Parses." In *Proceedings of Lrec* 6: 449–454.

De Rooij, Ork, Andrei Vishneuski, and Maarten De Rijke. 2012. "Xtas: Text Analysis in a Timely Manner." Paper presented at *Dir 2012: 12th Dutch-Belgian Information Retrieval Workshop*. Ghent, 23–24 February 2012.

DiMaggio, Paul, Manish Nag, and David Blei. 2013. "Exploiting Affinities between Topic Modelling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding." *Poetics* 41:570–606.

Entman, Robert M. 1993. "Framing: Toward Clarification of a Fractured Paradigm." *Journal of Communication* 43 (4): 51–58.

Frigyik, Bela A., Amol Kapila, and Maya R. Gupta. 2010. "Introduction to the Dirichlet Distribution and Related Processes". Technical Report UWEETR-2010-0006. Seattle: University of Washington Department of Electrical Engineering.

Gamson, William A., and Andre Modigliani. 1989. "Media Discourse and Public Opinion on Nuclear Power: A Constructionist Approach." *American Journal of Sociology* 95: 1–37.

Gardner, Matthew J., Joshua Lutes, Jeff Lund, Josh Hansen, Dan Walker, Eric Ringger, and Kevin Seppi. 2010. "The Topic Browser: An Interactive Tool for Browsing Topic Models." In *Proceedings of the Neural Information Processing Systems (NIPS) Workshop on Challenges of Data Visualization*: 5528–5235. Whistler, December 11, 2010.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–297. doi:10.1093/pan/mps028.

Haselmayer, Martin, and Marcelo Jenny. 2014. "Measuring the Tonality of Negative Campaigning: Combining a Dictionary Approach with Crowd-Coding." Paper presented at *Political Context Matters: Content Analysis in the Social Sciences*. University of Mannheim.

Jones, Bryan D., and Frank R. Baumgartner. 2005. *The Politics of Attention: How Government Prioritizes Problems*. Chicago, IL: University of Chicago Press.

Lin, Chenghua, and Yulan He. 2009. "Joint Sentiment/Topic Model for Sentiment Analysis." In *Proceedings of the 18th ACM conference on information and knowledge management* (5), 375–384. Hong Kong, November 2–6, 2009.

Lucas, Christopher, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer, and Dustin Tingley. 2015. "Computer-Assisted Text Analysis for Comparative Politics." *Political Analysis* 23: 254–277.

Matthes, Jörg, and Matthias Kohring. 2008. "The Content Analysis of Media Frames: Toward Improving Reliability and Validity." *Journal of Communication* 58 (2): 258–279. doi:10.1111/j.1460-2466.2008.00384.x.

McCallum, Andrew K. 2002. *Mallet: A Machine Learning for Language Toolkit*. Retrieved from http://mallet.cs.umass.edu

Meyer, David, Kurt Hornik, and Ingo Feinerer. 2008. "Text Mining Infrastructure in R." *Journal of Statistical Software* 25 (5): 1–54.

Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54 (1): 209–228. doi:10.1111/j.1540-5907.2009.00427.x.

Ramage, Daniel, Evan Rosen, Jason Chuang, Christopher D. Manning, and Daniel A. McFarland. 2009. "Topic Modeling for the Social Sciences." *In Nips Workshop on Applications for Topic Models: Text and beyond* 5: 1–4.

Ramirez, Eduardo H., Ramon Brena, Davide Magatti, and Fabio Stella. 2012. "Topic Model Validation." *Neurocomputing* 76 (1): 125–133.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana K. Gadarian, B. Albertson, and David G. Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58(4): 1064–1082. doi: 10.1111/ajps.12103.

Carina Jacobi (author to whom correspondence should be addressed), Department of Communication, University of Vienna, Austria; E-mail: carina.jacobi@univie.ac.at

Wouter van Atteveldt, Department of Communication Science, VU University Amsterdam, The Netherlands; E-mail: wouter@vanatteveldt.com. http://vanatteveldt.com

Kasper Welbers, Department of Communication Science, VU University Amsterdam, The Netherlands; E-mail: k.welbers@vu.nl