HIIG Discussion Paper Series

Discussion Paper No. 2016-05

Topic modeling for media and communication research: A short primer

August 2016

Cornelius Puschmann
cornelius.puschmann@hiig.de
Alexander von Humboldt Institute for Internet and Society

Tatjana Scheffler
tatjana.scheffler@uni-potsdam.de
University of Potsdam

Abstract

A variety of powerful tools for the automated and semi-automated analysis of textual data are increasingly at the disposal of media and communication researchers. Among the assemblage of methods, the school of techniques known as topic modeling has recently attracted particular interest. What utility does one popular type of topic model, *latent dirichlet allocation* (LDA), have for media and communication research? This paper illustrates some distinct strengths and weaknesses of LDA. We first briefly introduce its conceptual foundations, along with a selection of studies from the social sciences that apply it to different types of content, from newspapers and scientific publications to literary texts and social media. We then present a case study of news coverage of the Syrian civil war. After describing our data, we turn to two facets of the results in particular: the relation of terms and topics and the proportions of topics in documents, aggregated into months. We make the case for contrastive (rather than descriptive) uses of topic modeling that build broader analyses on the initial output of a model, rather than concluding with a list of terms.

Keywords

methods, content analysis, topic modeling, LDA, news

Contents

1. Introduction

Recent years have seen a surge in the interest in automated content analysis across the social sciences (Grimmer & Stewart, 2013; Lewis, Zamith, & Hermida, 2013; Zamith & Lewis, 2015). With the proliferation of digital textual content, the ability to derive meaning from large volumes of data from news sources and social media has become pivotal to promising new areas of scholarly inquiry. One focus of methodological innovation has been the intersection of human coding and supervised machine learning (Scharkow, 2013). But what utility do unsupervised and inductive computational techniques, such as latent dirichlet allocation (LDA), have for media and communication research? Due to their data-driven approach to topicality and meaning, the set of procedures referred to as topic modeling has sometimes been compared to grounded theory and inductive reasoning about content, with all the analytical baggage that this entails (Mohr & Bogdanov, 2013; Murthy, 2015). Yet topic modeling has in recent years also attracted attention among scholars from a wide range of disciplines, such as political science, digital humanities, health communication, and the sociology of science, in addition to being widely used in commercial applications (e.g. Grimmer & Stewart, 2013; Jockers & Mimno, 2013; Paul & Dredze, 2014; Hall, Jurafsky & Manning, 2008, see Mohr & Bogdanov, 2013, for a helpful overview). Its proponents consequently see great potential for topic modeling in media and communication studies, while some scholars have voiced skepticism and argue that its capabilities overlap with simpler, established procedures, such as co-word analysis (Boumans & Trilling, 2016; Leydesdorff & Nerghes, 2015). Yet others suggest solutions that may conveniently bridge topic modeling and manual approaches through combining expert judgment and crowdsourcing with machine learning in techniques such as labeled LDA and sub-corpus modeling (Ramage, Dumais & Liebling, 2010; Tangherlini & Leonard, 2013).

In this paper, we illustrate some distinct strengths and weaknesses of LDA. We first very briefly introduce its conceptual foundations, along with a select set of studies from the social sciences that apply it to different types of content, from newspapers and scientific publications to literary texts and social media. We then present a case study of news coverage of the Syrian civil war that

illustrates applications of LDA in media research, while also pointing out its limitations. After describing our data and methods, we turn to two facets of the results in particular: the relationship of terms and topics to each other and the proportions of topics in documents, aggregated by month. We close with a summary of the approach's strengths and weaknesses, and, based on our example, make the case for contrastive (rather than descriptive) uses of topic modeling that build broader analyses on the initial output of the model, rather than to conclude with a list of terms.

2. A brief description of latent dirichlet allocation (LDA)

Topic modeling describes a family of computational techniques for analyzing textual content that is increasingly popular in both industry research and across different academic fields (see Blei, 2012, Mohr & Bogdanov, 2013, for two relatively non-technical introductions). Their origins lie at the intersection of computational linguistics and mathematical techniques for the efficient processing of large matrices. The foundation of topic modeling is the bag-of-words approach (sometimes bag-of-features approach) to written data, in which words are considered to be features that possess distinct distributions in a text or collection of texts. Its forerunner, Latent semantic analysis (LSA; Deerwester et al, 1990; Landauer & Dumais, 1997) can among other things be used to determine the similarity of documents or authors based on shared word co-occurrence. Latent dirichlet allocation (LDA; Blei, Ng & Jordan, 2003) and correlated topic models (CTM; Blei & Lafferty, 2007), proposed by David Blei and colleagues, are among the most prominent forms of topic modeling derived from LSA and pLSA (probabilistic LSA). Among the commercial applications of topic modeling are information retrieval tasks and recommender systems (Wang & Blei, 2011; Zhao et al, 2011). LDA is a generative model in which the occurrence of individual words in a document is explained by topics that generate the words. Each document is characterized by a multinomial combination of topics and each topic has a probability of generating certain words, with high-frequency function words such as 'the' or 'and' occurring in roughly even distribution, while words that probabilistically characterize a topic exhibit a skewed distribution among topics. Like other

computational approaches, LDA is quite scalable, but in-depth qualitative knowledge of the data under analysis is still crucial, particularly to determine whether the association of terms with topics is the result of a sampling error or some other interference. LDA partly curtails the performance issues of earlier approaches by applying computationally efficient sampling methods which mitigate some of the problems of high dimensionality that characterize large document-term matrices.

LDA encodes a number of very specific assumptions about meaning. To provide one contrast with content analysis noted by Grün and Hornik in their description of the *topicmodels* package for R (2011, p.1-2, emphasis mine): "In mixed-membership models documents are not assumed to belong to *single* topics, but to simultaneously belong to *several topics* and the topic distributions vary over documents." Furthermore, as Blei and Lafferty explain, LDA assumes that the words of each document arise from a mixture of topics, where each topic is a multinomial over a fixed word vocabulary. The topics are shared by all documents in the collection, but the topic proportions vary stochastically across documents (2007, p. 18). In practice, this means that conducting an analysis based on just two topics may result in a very coarse classification, in which the topic shares are equal across documents, or in a coarse classification of documents, where one set of documents is mostly in topic A, and another set mostly in topic B. By contrast, a large number of topics will result in a strongly skewed distribution with many topics only occurring in a small number of documents and heuristically appearing very similar to each other. Terms are likewise associated with topics, by way of the log-likelihood of occurrence in them. Some words are not very distinct for any single topic, but occur across topics, making a stop-word list containing high-frequency function words particularly useful.

In contrast to most forms of content analysis common in social science, topic modeling induces functional categories purely from structural features. This can create an awkward situation in which the expert knowledge of the researcher seemingly has no place, because a topic model generates its analysis from word distributions alone. Firstly, the selection of features is crucial: words are not necessarily to meaning what bricks are to a wall, i.e. in the case of phrasal expressions or idioms, even if the bag of words evokes such imagery. Secondly, the approach implicitly assumes both that all topics are similar in the

sense of possessing a unique feature distribution imprint, and that world knowledge, while important, is non-essential to distinguish topics from each other. The number of topics in LDA must be predetermined. Depending on the document type and the planned analysis, the number of topics can be very large (several hundreds). While LDA is relatively robust to changes in the number of topics (Stevens et al., 2012), many topics learned this way are either duplicates of each other or irrelevant for classification or further analysis. Human judgment is needed to distinguish between them. Relatedly, topic models itself are rarely directly interpretable. In particular, LDA does not output interpretable labels of the learned topics. Usually, topics are represented by the top n (e.g., top 10) most probable terms generated by each topic. However, these terms are heavily dependent on preprocessing such as stopword removal and can repeat across topics. Only the human analyst can make sense of the topics that have been learned. Finally, the discovered latent topics may or may not correspond to what the human analyst would call 'topics'. Depending on preprocessing and aggregation, 'topics' may also reflect authorship (through style or specialized vocabulary), genre, or other orthogonal categories, and must be manually evaluated. To summarize, the meaning of topics in LDA is deeply relational and sometimes runs counter to human intuition. Topics mean something *in relation to other topics*, rather that in relation to world knowledge, thus complicating their interpretation and setting them apart from how other popular units of analysis in media and communication research, such as frames, are frequently conceptualized.

3. Applications of LDA in media and communication research

In this section we briefly present a set of six studies that utilize LDA. We focus on empirical studies which answer research questions relevant to media and communication research, rather than analyses that are largely methodological or technical in nature.

Hall, Jurafsky and Manning (2008) apply an LDA-related approach, Dynamic topic modelling (Blei & Lafferty, 2006) to a corpus of 12,500 papers from the ACL Anthology to describe significant intellectual developments in the history

of computational linguistics. Their article centers on a time series analysis of 43 topics that exemplify the field's dynamics over time, with some areas of investigation declining while others gradually establish themselves. The authors also invert their approach to investigate the topical diversity of papers presented at three conferences in computational linguistics to establish whether an originally more narrow workshop has over time become more thematically diverse in relation to two larger conferences. Both the field of computational linguistics and the three conferences, both of which are characterized by their topic shares, are essentially metadata that can be linked to the topics by aggregation or annotation, again highlighting the relational potential of topic modeling.

Bonilla and Grimmer (2013) fit a customized LDA to assess media attention in the U.S. news media in the wake of terror attacks. In doing so, the authors are able to investigate how newspapers and nightly newscasts divide their attention among topics both before and after a terror alert. After having operationalized attention shifts in this way, the authors are able to show that terror alerts are widely reported following a terror attack, and are able to relate this to shifts in the attention paid to other topics. In their approach, Bonilla and Grimmer (2013) proceed to the logical next step of testing a hypothesis on the relational results produced by topic modeling.

A similar form of abstracting away from the description of topics to their relation to metadata is taken by Ghosh and Guha (2013) in a public health study of obesity-related tweets. The aim is to geographically map messages related to obesity and investigate possible correlates, in this case, the location of fast food restaurants across the United States. Tweets are sampled through a series of search queries, and LDA is applied both to distinguish obesity-related tweets from false positive (tweets matching a search query but unrelated to obesity) and for a qualitative differentiation of aggregated topics, to which the authors refer to as *themes* (e.g. childhood obesity and schools, obesity prevention, food habits). The authors use bigrams for this purpose, an interesting variation of the standard bag of words (unigram) procedure.

Yang, Torget and Mihalcea (2011) study historical shifts in news writing, while Koltsova and Koltcov (2013) apply LDA to Russian LiveJournal blog posts to assess political discourse. Both studies report topic keywords only, though Koltsova and Koltcov (2013) aggregate topics into themes, similar to the

approach taken by Ghosh and Guha (2013). Both forgo a further analysis beyond a description of the topics, focusing instead on a qualitative interpretation of the model output.

Finally, Paul and Dredze (2014), present topic modeling as a single component of a larger ensemble in their study of health-related discourse in social media. The authors propose a specialized Ailment Topic Aspect Model (ATAM), pointing into the direction of more narrowly focused approaches that take the specificity of text type (social media vs. news writing) and subject matter (health vs. politics) into account, rather than proposing a type of model universally applicable to all types of data. The ATAM distinguishes principally between non-ailment topics (television, family, transportation, music) and ailments (influenza, sleep issues, cancer) in an approach that deliberately blends supervised and dictionary-based techniques, with dictionaries developed by relying in part on crowdsourcing.

What distinguishes the more elaborate studies in this list from less comprehensive ones is that they (a) rely on topic modeling to sample from a larger body of data, (b) find subtopics in a broadly defined theme, informed by domain knowledge, and (c) relate topic membership to external variables, contained in or aggregated from available metadata, such as time, geography, or authorship.

4. Data and methods

We compiled a corpus of 2,083 items from the World News section of the British newspaper *The Guardian* published between January 2011 and September 2015 and matching the search query "Syria", using the GuardianR package for R (Bastos & Puschmann, 2014). Photo galleries and items from the 'as it happened' subsection (which contains live blog coverage of events) were omitted. We aggregated the documents on a monthly basis to allow us to make observations on the chronology of the Syrian civil war, resulting in 57 documents, one for each month. We then used the *tm* package (Feinerer, Hornik, & Meyer, 2008) to remove numbers and punctuation (preserving intra-word dashes), trim excess whitespace, and transform the terms to lower case. We furthermore omitted a list of 82 stop words, a combination of high frequency English function words, and terms used for captions and other

functional purposes in *The Guardian* (*gmt, bst, photograph*) as well as the search term, *Syria*. After this preprocessing step, we first calculated a document term matrix (DTM) from the corpus, and then removed sparse terms from this DTM, i.e. terms only occurring in 25% or less of all documents (similar to Ghosh and Guha's approach). The result is a trimmed matrix with 7,610 types and 253,889 tokens. We then applied LDA to this DTM (k = 8). Following the suggestions made by Grün and Hornik (2011), we first modeled with k = 30, but later reduced this number to k = 8. The recommended approach to find the optimal k is to proceed stepwise and compare the perplexity of the different *k*-models (Pleplé, 2013). Another strategy is to apply hierarchical clustering based on Euclidean distance to examine topic similarity. We use this approach for illustrative purposes, but without setting a fixed threshold value on which similar topics are to be merged, though tree height should in theory be usable for this purpose.

5. Results

As has already been noted, in the LDA philosophy topics are distributions of terms over documents, or perhaps more plainly put, the likelihoods of terms occurring in documents. In uniformly distributed data, all terms are equally likely to occur, but this intuitively is not the case within written texts. Table 1 shows terms associated with the eight topics calculated by LDA.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
people	forces	assad	air	chemical	arab	assad	government
refugees	gaddafi	attack	forces	international	egypt	government	israel
	government	damascus	iraq	military	forces	homs	morsi
	libya	government	iraqi	president	killed	opposition	people
	libyan	opposition	isis	russia	military	people	reports
	people	rebels	islamic	russian	people	regime	
	security	regime	military	use	protesters	security	
	tripoli	weapons	state	weapons	security		
			strikes				

Table 1: Terms associated with eight topics in the corpus (log likelihood >= 0.005).

Table 1 shows the list of terms for each topic that occur with a threshold log likelihood of >= 0.005. Note that this is distinguished from raw term frequency, but a measure of distinctness that a term has for a topic. The word *people* is both relatively frequent in the corpus and distinguishes Topics 1, 2, 6, 7 and 8 from Topics 3–5, while the term *refugees* is only distinct for Topic 1. Similarly,

place and person names (*libya, tripoli, assad, gaddafi, morsi*) identify individual topics with great precision and are relatively infrequent. Topic term lists are a hallmark of analyses that apply topic modeling, yet they do very little other than provide a convenient heuristic to validate a topic. Topic keyword lists can both be quite sparse, as in Topic 1, where the term *refugees* in combination with *people* is sufficient to identify the topic, or overly specific, as in Topic 5, where terms such as *use* and *international* reliably distinguish the topic without semantically describing it. The term *people* identifies five out of eight topics because of its co-occurrence with other terms, which in turn are only associated with a single topic (*refugees, gaddafi, arab, homs, morsi*). The finding that person and place names are highly discriminant for topics in news writing echoes the finding of Koltsova and Koltcov (2013, p. 217), who identify a number of such terms in their data, along with specialized lexis, such as *president, opposition, minister* and *state*. This highlights that topic models are much more sensitive towards frequency than humans: An item with reasonably high frequency is likely to appear non-distinctive to a human, even though it reliably predicts a topic in co-occurrence with other, less frequent terms.

How can the similarity of topics by means of the terms that distinguish them be assessed? One way of making an informed decision about topic validity is the use of a distance measure. Figure 1 plots Euclidean distance among the eight topics in the data on the basis of term distribution (the beta statistic, in Grün & Hornik's terminology). On visual inspection we see that Topic 2 and Topic 5 branch off early, followed by a subdivision of Topics 1 and 4, Topics 6 and 7, and Topics 3 and 8. Applying this technique to a much larger number of topics quickly shows the communalities of tightly clustered topics. Techniques such as pruning may be used to reduce the number of topics, or the perplexity of a test and training set of documents can be relied upon (Pleplé, 2013). However, it is worth pointing out that perplexity is not a good reflection of human judgment, as the words most frequent within a topic may distinguish but not describe a topic very well. For example, the association of Topic 1 and Topic 4 raises the question of why mentioning of Islamic State/ISIS and refugees co-occurs and what the implications of this topical proximity may be (We stress that this is very much an example for *the kind of issue* that may be relevant for research, rather than this particular distinction identifying an issue of actual relevance).

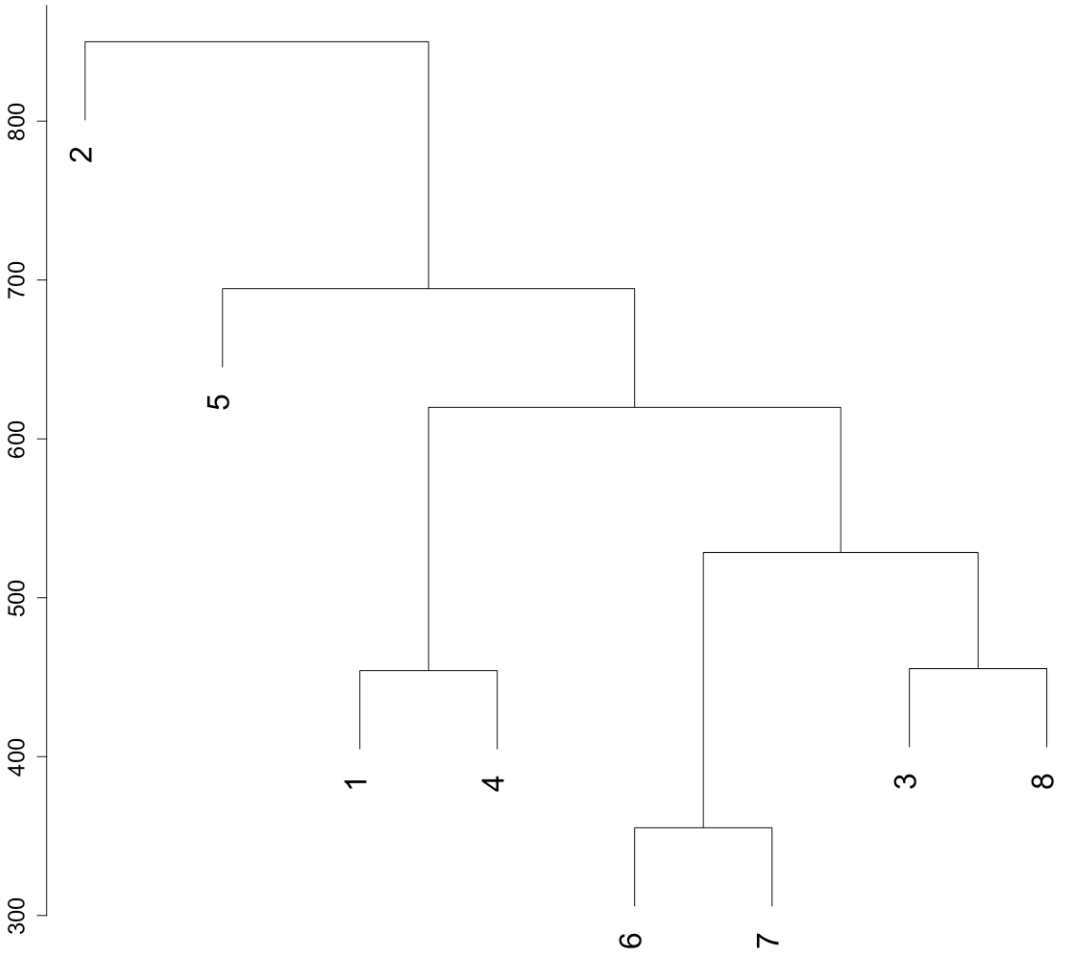

Figure 1: Hierarchical cluster dendrogram showing the degree of similarity among topics. Basis is Euclidean distance, calculated from the log likelihood scores of terms (the beta statistic) within topics (Ward's method).

A second view on the relation of topics, now taking the time-aggregated documents into account is assessing document similarity through the topic distributions in them. Figure 2 shows one technique for assessing the relationship of topics and documents by means of visualizing the correlations of topic distributions with documents. The map illustrates the career of a topic over time, showing the unsurprising time-based ordering of issues that the topics identify (in contrast to, for example, the genre-based ordering of topics in literary corpora). Topic 1 (*people, refugees*) features prominently since Fall 2014, after having already occurred in 2013 and in early 2011, before the beginning of the Syrian civil war in February 2011. Topic 2 (associated with many terms related to Libya) is prominent from March to October 2011. Topic 4 combines reference to Iraq and Islamic State, while Topic 6 (terms associated with Egypt

and the Arab Spring) is also prominent early in the campaign and more scattered than Topic 2. The same applies to Topic 7, which is prominent from late 2011 to the middle of 2012. It is noteworthy how abruptly Topic 2 drops from the agenda to be replaced by Topic 6 in November 2011, or how the refugee crises is 'blotted out' by reporting on Islamic State and Egypt in June and July 2014.

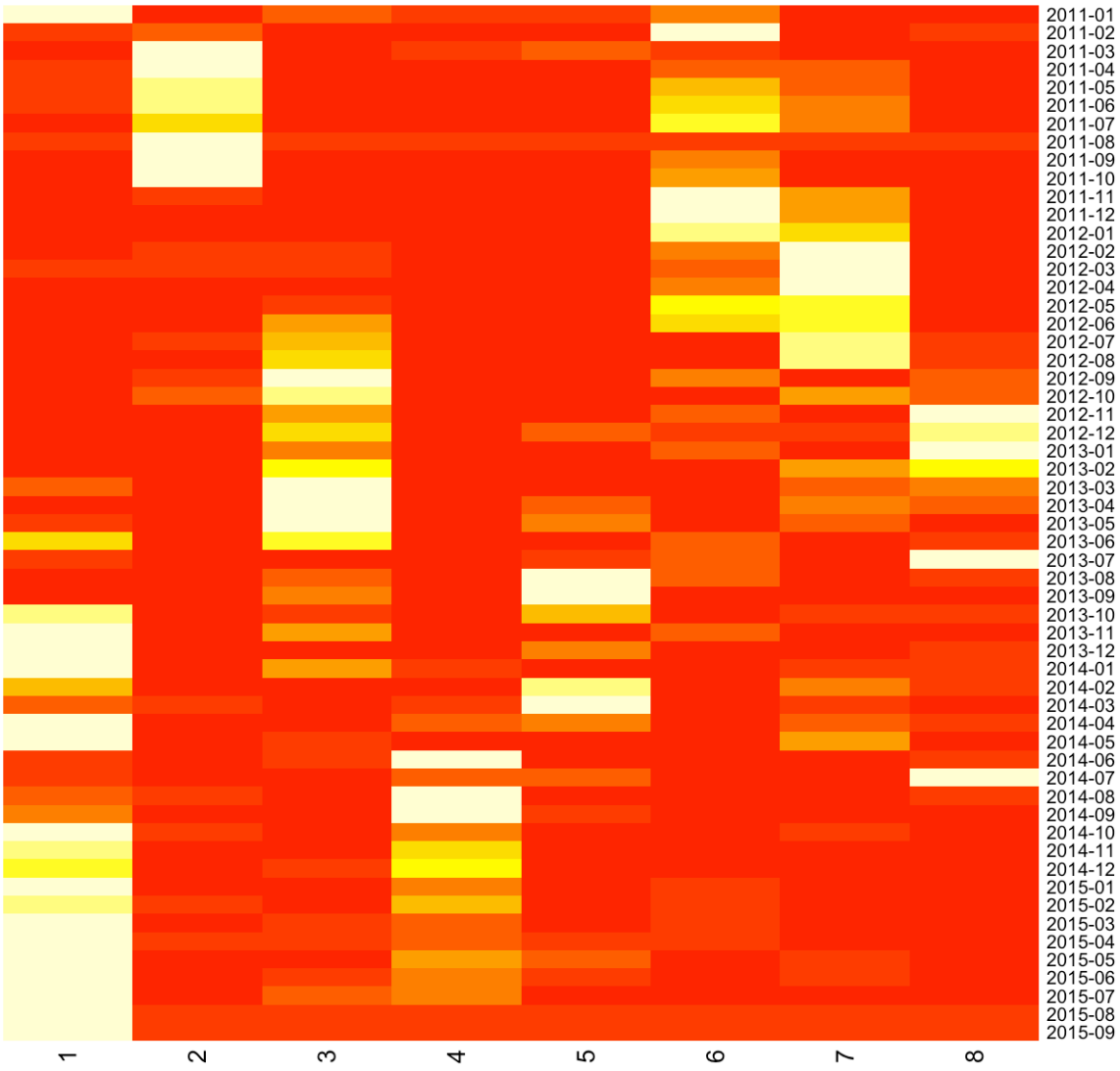

Figure 2: Heatmap showing the association of the eight topics (x axis) and documents (y axis) based on the proportion to which each document is associated with each topic (the gamma statistic).

Finally, Figure 3 again shows the gamma statistic returned by the LDA, to which both Blei and Lafferty refer to topic assignments (of terms), called beta by Grün and Hornik, and topic proportions (within documents), called gamma by Grün and Hornik. Each time slice is associated with the eight topics in different proportions, from phases in which a number of topics co-occur to

those in which a single topic dominates. Visual inspection reveals a growth and decline of particular topics (as already shown in Figure 2), along with a resurgence of others. Topic 2, describing the conflict in Libya and the protest and civil war following the ousting of Muammar Gaddafi captures a large share of initial coverage, as does Topic 7, making reference to the siege of Homs and clashes of the Syrian army and rebel forces, which declines sharply after 2012. Topic 3, referring to the Aleppo siege, has a similar career of gradual growth and decline. A watershed role is afforded to Topic 5, which describes the use of chemical weapons by the Syrian regime, which spikes in September 2013 to abate soon after. Islamic State captures a large topical share starting in summer 2014, while the refugee crisis dominates the present picture.

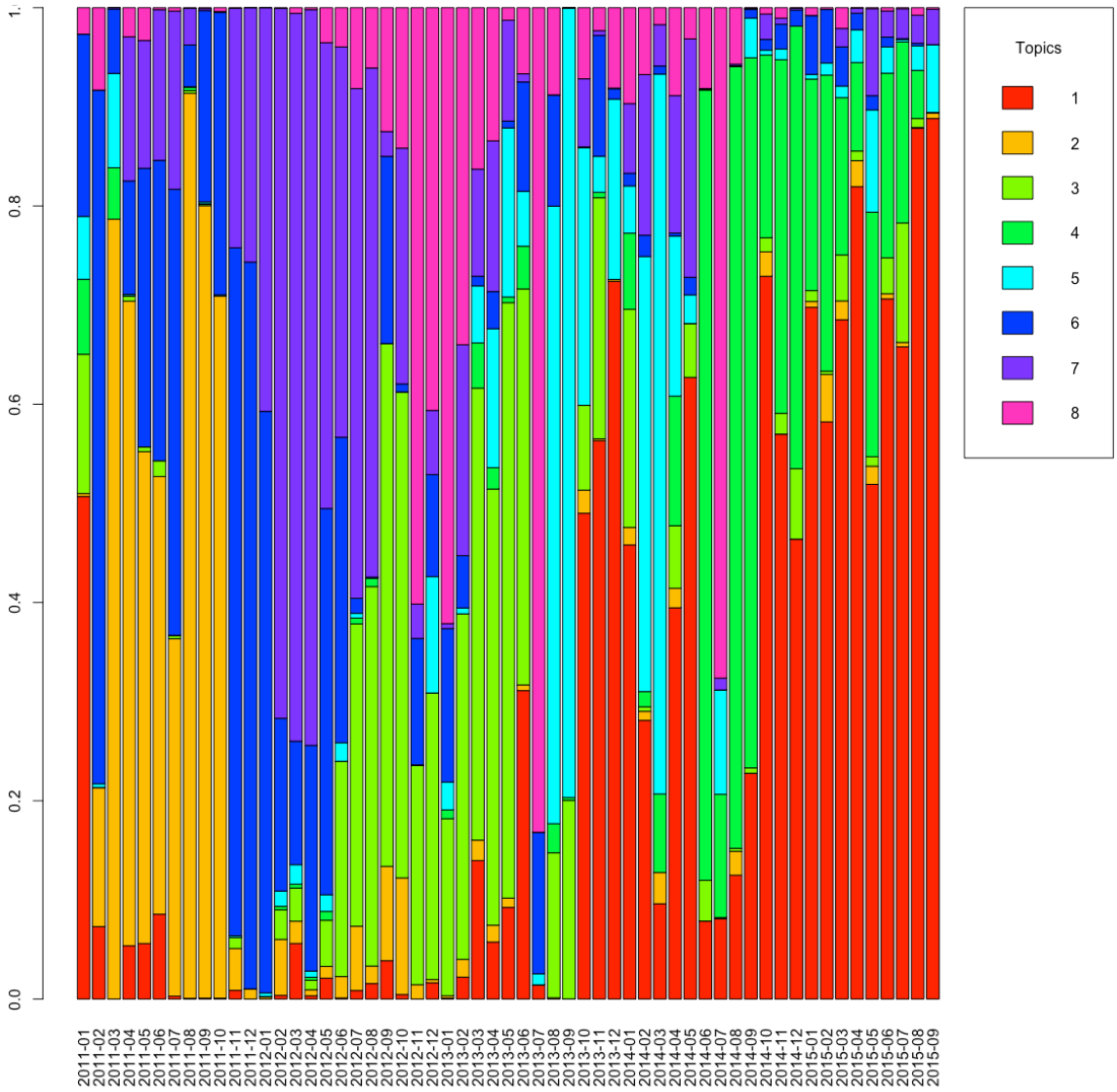

Figure 3: Normalized topic scores across news items. Based on the proportion to which each document 'belongs' to each topic (the gamma statistic).

6. Strenghts and limitations

Apart from purely descriptive results, which can easily be achieved by other means, this brief characterization highlights that there are secondary benefits to topic modeling. LDA topics identify certain categories well, but may miss others that would be obvious to a human coder. The association of topics with words and documents, combined with the ability to aggregate documents on the metadata level (time, geography, authorship) means that some topics that may have gone unnoticed otherwise, for example by occurring only in a small subsection of the corpus, or which are not initially assumed to be related to the issue under study, can be discovered. Secondly, the proportional perspective on topic shares within documents means that topical diversity can be more effectively assessed than human judgment allows. This is perhaps an even more relevant finding for comparative or contrastive analyses. Assuming that a topic model is not entirely wrong – which it will be only if some aspect of the sampling has gone awry – knowledge of the exact proportion of topics appears useful.

Topic modeling cannot qualitatively do things better than an expert coder, both because it lacks human competence and because it was not designed for that purpose in the first place. The uses that David Blei and colleagues have in mind are very much geared towards the discovery of topics in large and unfamiliar collections. If the researcher is intimately familiar with a small dataset, the model is unlikely to reveal anything strikingly new. The greatest advantage of domain knowledge is likely to be the ability to discover errors in sampling (the selection of items), preprocessing (the accidental removal of crucial information) and processing itself (determining the ideal number of topics). When all these aspects are considered and checked, topic modeling can be useful for its ability to discriminate, rather than describe, topics.

Topic models are also subject to conceptual limitations that apply with regard to the units that they take into account, and how these units are treated. As Grün and Hornik point out in their documentation of the *topicmodels* package, using only raw term frequencies assumes that the order in which words occur in a document is negligible. Another limitation unaccounted for are changes in meaning over time. LDA essentially assumed word meaning to be held constant, allowing the topics that generate the occurrence of words to

fluctuate. Word meaning changes slowly, but longitudinal analyses that rely on centuries of data may pose problems.

Finally, it is vital to note the particular choices made in *our* analysis, in contrast to the characteristics of topic modeling more generally. Particularly the aggregation of multiple documents by month represents such a choice. This results in very large documents which may well flatten topical distinctions visible in more granular analyses. Secondly, the elimination of sparse terms also risks missing signals that are strong at a particular point in time, but do not occur across a sufficient number of documents.

7. Conclusion

The aim of this article has been to describe topic modeling, specifically latent dirichlet allocation (LDA) and its utility for media and communication research, based on a set of prior studies and a step-by-step description of its application to an example corpus.

It should have become evident that topic modeling is prone to somewhat superficial uses that add little in the way of insightful analysis, particularly when topic terms only are provided (Schmidt, 2012). Often, however, this is a consequence of how the results of topic modeling are analyzed, rather than of the quality of these results as such. Once a degree of trust to a set of topics can be established, the actual work of examining topic proportions according to one or more independent variables can begin. Time, geography, author identity and publication outlet are among the examples in the studies we have presented. Both the relation of these variables to topics and the degree of diversity of topics within documents can be of interest, as can be the similarities and differences in the association between documents and topics. Here the potential interactions with traditional content analysis and supervised machine learning become apparent.

The ability to aggregate documents on an analytically interesting variable, or to derive such a variable from available document metadata, is the key to unlocking the potential of topic modeling for innovative analyses in media and communication research. Beyond exploration, this squarely places the brunt of work on compiling a suitable corpus in which influencing factors such as genre differences or language change are successfully mitigated. Multilingual topic

modeling, in which topics are aligned based on similar terms, is such an example, as are diachronic studies, and studies that introduce further variables that may explain topic fluctuation, as time does in our Syria corpus. As Grimmer and Stuart point out, computational methods complement humans, rather than replacing their judgment and expertise with that of an algorithm (2013, p. 281).

8. References

Bastos, M. T., & Puschmann, C. (2014). GuardianR: The Guardian API wrapper. Retrieved from https://cran.r-project.org/web/packages/GuardianR/index.html

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. International Conference on Machine Learning, 113–120. http://doi.org/10.1145/1143844.1143859

Blei, D. M. (2012). Topic modeling and digital humanities. Journal of Digital Humanities, 2(1). Retrieved from http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. The Annals of Applied Statistics, 1(1), 17–35. http://doi.org/10.1214/07-AOAS114

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3, 993-1022. http://doi.org/10.1162/jmlr.2003.3.4-5.993

Bonilla, T., & Grimmer, J. (2013). Elevated threat levels and decreased expectations: How democracy handles terrorist threats. Poetics, 41(6), 650–669. http://doi.org/10.1016/j.poetic.2013.06.003

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit. Digital Journalism, 4(1), 8–23. http://doi.org/10.1080/21670811.2015.1096598

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391–407. http://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. Journal of Statistical Software, 25(5).

Ghosh, D., & Guha, R. (2013). What are we “tweeting” about obesity? Mapping tweets with topic modeling and Geographic Information System. Cartography and Geographic Information Science, 40(2), 90–102. http://doi.org/10.1080/15230406.2013.776210

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political Analysis, 21(3), 267–297. http://doi.org/10.1093/pan/mps028

Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. Journal of Statistical Software, 40(13), 1-30. http://doi.org/10.18637/jss.v040.i13

Hall, D., Jurafsky, D., & Manning, C. D. (2008). Studying the history of ideas using topic models. In M. Lapata & H. T. Ng (Eds.), Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP ’08 (pp. 363–371). Morristown, NJ, USA: Association for Computational Linguistics. http://doi.org/10.3115/1613715.1613763

Jockers, M. L., & Mimno, D. (2013). Significant themes in 19th-century literature. Poetics, 41, 750–769. http://doi.org/10.1016/j.poetic.2013.08.005

Koltsova, O., & Koltcov, S. (2013). Mapping the public agenda with topic modeling: The case of the Russian LiveJournal. Policy & Internet, 5(2), 207–227. http://doi.org/10.1002/1944-2866.POI331

Landauer, T. K., Dutnais, S. T., Anderson, R., Carroll, D., Fbltz, P., Pumas, G., … Streeter, L. (1997). A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 1(2), 211–240. http://doi.org/10.1037/0033-295X.104.2.211

Lewis, S. C., Zamith, R., & Hermida, A. (2013). Content analysis in an era of big data: A hybrid approach to computational and manual methods. Journal of Broadcasting & Electronic Media, 57(1), 34-52. http://doi.org/10.1080/08838151.2012.761702

Leydesdorff, L., & Nerghes, A. (2015). Co-word maps and topic modeling: A comparison from a user’s perspective. Retrieved from http://arxiv.org/abs/1511.03020

Mohr, J. W., & Bogdanov, P. (2013). Introduction—Topic models: What they are and why they matter. Poetics, 41(6), 545-569. http://doi.org/10.1016/j.poetic.2013.10.001

Murthy, D. (2015). Critically enganging with social media research methods. In L. McKie & L. Ryan (Eds.), An End to the Crisis of Empirical Sociology? Trends and Challenges in Social Research (pp. 81–97). London: Routledge.

Paul, M. J., & Dredze, M. (2014). Discovering health topics in social media using topic models. PloS One, 9(8), e103408. http://doi.org/10.1371/journal.pone.0103408

Pleplé, Q. (2013). Perplexity to evaluate topic models. Retrieved from http://qpleple.com/perplexity-to-evaluate-topic-models/

Ramage, D., Dumais, S., & Liebling, D. (2010). Characterizing microblogs with topic models. In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM '10) (pp. 130–137). Washington, DC: AAAI Press.

Scharkow, M. (2013). Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. Quality & Quantity, 47(2), 761–773. http://doi.org/10.1007/s11135-011-9545-7

Schmidt, B. (2012). Words alone: Dismantling topic models in the humanities. Journal of Digital Humanities, 2(1). Retrieved from http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring topic coherence over many models and many topics. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 952–961). Stroudsburg, PA, USA: Association for Computational Linguistics.

Tangherlini, T. R., & Leonard, P. (2013). Trawling in the sea of the great unread: Sub-corpus topic modeling and humanities research. Poetics, 41(6), 725–749. http://doi.org/10.1016/j.poetic.2013.08.002

Wang, C., & Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11 (pp. 448–456). New York, New York, USA: ACM Press. http://doi.org/10.1145/2020408.2020480

Yang, T.-I., Torget, A. J., & Mihalcea, R. (2011). Topic modeling on historical newspapers. In K. Zervanou & P. Lendvai (Eds.), LaTeCH '11 Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (pp. 96–104). Stroudsburg, PA, USA: ACM.

Zamith, R., & Lewis, S. C. (2015). Content analysis and the algorithmic coder: What computational social science means for traditional modes of media analysis. The ANNALS of the American Academy of Political and Social Science, 659(1), 307–318. http://doi.org/10.1177/0002716215570576

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E., Yan, H., & Li, X. (2011). Comparing Twitter and traditional media using topic models. In P.

Serdyukov, P. Braslavski, S. O. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, ... E. Yilmaz (Eds.), Advances in Information Retrieval (pp. 338–349). Heidelberg: Springer. http://doi.org/10.1007/978-3-642-20161-5_34