Unsupervised models considered dangerous? Discussing the practices of social scientists with topic models for automated text analysis.

Matti Nelimarkka*
Aalto University & University of Helsinki
matti.nelimarkka@aalto.fi

October 15, 2017

1 Introduction

There has been an increasing interest towards automated classification of large scale text data in social sciences. For example automated classification has been applied to

- examine the party alignment based on politicians' speeches (Yu et al., 2008)
- explore what type of policies are discussed (Burscher et al., 2015b) and how they are discussed (Levy and Franklin, 2013; Zhang and Counts, 2015; Burscher et al., 2015a)
- measuring aspects like deliberation (Nelimarkka and Ahonen, 0) and politeness (Danescu-Niculescu-Mizil et al., 2013) from text.
- exploring how political agenda emerges and changes (Wilkerson et al., 2013)

These can be broadly be categorized to three families of methods: approaches where predetermined keywords are used as to signal classes, approaches where a priori-classes are known (supervised) and approaches where a priori-classes do not exist (unsupervised) (Purhonen and Toikka, 2016).

There have been only few papers extending the social science oriented methodological discussion on these methods. There are general reviews of these reviews the methods and approaches at a general level and demonstrates some areas for applications (including Lucas et al., 2015; Purhonen and Toikka, 2016). Beyond these type of summaries, there are works which aim to inform social science relevant approaches further. For example, Grimmer and Stewart's (2013) work suggests that "automated content methods can make possible [- -] the systematic analysis of large-scale text collections without massive funding support." It promotes applications of computational tools for large scale text analysis in social sciences. These tools include methods to label content based on known categories as well as approaches where the goal is to find categories. It also discusses the problems of computational data analysis: the problems of language models, the application of automated data analysis to support (not replace researchers), the need

*Author thanks Kone Foundation for financial support during the research project. Author also thanks the Rajapinta research community for discussions around this work and CSC – IT Center for Science for computational resources required to conduct the analysis.

to choose the computational methods based on the research problems, and the need to validate the findings. However, Hopkins and King (2010) argue, that much of the literature has been focused on computer science relevant topics and have not yet extensively address social science approaches.

While Hopkins and King's (2010) work is already over five years old and improvements are made to adapt computational methods for social sciences (e.g., Grimmer and Stewart, 2013), I argue there is lack on the questions of validity and reliability of these methods. I contribute to the methodological discussion further by examining the practices of automated content analysis methods. I examine the unsupervised methods for textual data analysis and focus on topic models in detail. Topic models are mixed membership models, where each document (e.g., a tweet, a Facebook post, one paragraph of text, one policy paper) "belongs" to several topics, each with some portition. These probabilities are computed in an iterative process (that is to say: chosen initially at random a distribution for each word and change those to improve the model fitness). Nowadays, topic models apply latent Dirichlet allocation (LDA) as the distribution to choose the distributions from (a good and non-mathematical presentation, see Blei, 2012).

I first examine the use of unsupervised methods in social sciences before topic models. This highlights that as such, applications of these approaches are not novel. Rather, they have been applied and critically examined already before. I then present findings from an empirical study where I show the impact of different number of topics have for the analysis and findings. In the next section, I engage the difficulty of validation and present observations from that. Based on these two observations, I conclude this work.

2 Social sciences and unsupervised models

There are similarities between explanatory factor analysis (EFA) and topic models: both aim to produce some groups of data based on applications of statistics. Similarly, in both methods the researcher gives an interpretation for the the outcome of data analysis. In factor analysis, it is the meaning of the factors; similarly in topic models it is the meaning of topics. Therefore, I will first discuss EFA and then briefly discuss further the processes how topic models are used.

2.1 Explanatory Factor Analysis

I have chosen to discuss about EFA as it is familiar to (quantitatively oriented) social scientists. The mathematical foundations for analysis were developed in the early 20th century, but as it was computationally heavy and became more mainstream in the 1960s (Metsämuuronen, 2006). The computational capabilities in 1960's increased EFA's utility for researchers (e.g., Kaiser, 1960). Nowadays EFA is part of commonly used statistical software tools, like SPSS and R. The process of EFA is rather simple thanks to the software tools: data is loaded, correlation within the data is computed, the factor loadings are estimated, the matrix is rotated to increase its interability and finally, factor loadings are computed to each data item. The outcomes are then examined based on eigenvalues and the factor loadings to examine the overall quality of them. The rotation is commonly done with an orthogonal VARIMAX rotation, but other rotations to e.g., account for autocorrelation (PROMOX) have been developed (e.g., Metsämuuronen, 2006, or any other method book of your choosing).

For me, the benefit of EFA relates to its history – the critical discourse has already emerged to examine how social scientists traditionally apply EFA (e.g., Fabrigar et al., 1999; Russell, 2002; Bandalos and Boehm-Kaufman, 2010). Fabrigar et al. (1999) argue there are five decisions

which researchers make before starting the EFA analysis:

1. choosing the data and variables for the analysis,
2. deciding EFA is suitable method for analysis,
3. choosing the method,
4. number of factors, and
5. rotation method for the data.

Each of these decisions can have implications on the outcomes of EFA and can therefore be critized. Analysis on the problems of EFA highlight issues with the sample sizes and number of variables (Fabrigar et al., 1999; Russell, 2002) and choosing the the number of factors (Bandalos and Boehm-Kaufman, 2010, 79–83). Furthermore, the choice of EFA as data analysis method has raised concerns (Bandalos and Boehm-Kaufman, 2010) as well as the reporting of reliability and stability of factors (Fabrigar et al., 1999). The problems are extensive: Fabrigar et al.'s (1999) analysis showed that about 20% of articles in psychology journals had some issues with EFA.

To summarize, while EFA is a valuable method for data analysis, the practices how it is applied are questioned. The critical texts I have briefly reviewed (Fabrigar et al., 1999; Russell, 2002; Bandalos and Boehm-Kaufman, 2010) do not even address the problems of interpretation of the findings. Rather, they focus only on the statistical processes of extracting the factors from data and even with that focus, give rather pessimistic view of the method.

In my view, this in part relates to the unsupervised characteristic of EFA. Unlike in hypothesis driven research, the expectations are not formulated but rather the data 'guides' the analysis. This can lead to issues on choosing the method, rotation and number of factors as results are not clearly wrong. Compare this to more traditional analysis methods, like regressions, where the errors (residuals) can be analyzed. Similarly in regression analysis, applying logistic regression in case where ordinary least square regression should have been used clearly gives a wrong answer. However, in unsupervised methods the intuition is much weaker guide for analysis. I agree with Watts (2011): the problem is that human can rationalize every outcome (even for hypothesis-driven research). This makes data analysis extremely complicated.

2.2 Topic models

As this paper focuses on topic models as an exemplar of modern unsupervised methods, I wills briefly examine the current practices for conducting analysis with topic models. I have used R-package `topicmodels` (Hornik and Grün, 2011), however tools for analysis also include `STM` (Roberts et al., 2013), `Gensim` (Rehurek and Sojka, 2010) and `Mallet` (McCallum, 2002), among others. Each tool may require somewhat different process as they provide different support throughout the process. The process normally includes following steps

1. Transforming the data into computational format. This step includes removing commas, periods and other markings like that, making the letters lowercase, and stemming or lemming the content to base forms. The last part is conducted to ensure different forms of the word are correctly measured as instances of the same word. For example, cats and cat are both changed to cat to make sentences like: "I like a cat" and "I like all cats" have same meaning. Overall the aim is to transform the data into bag-of-words type of structure for further analysis.

2. Creating the document-term matrix. This step focuses on quantifying the textual data into format where each document is represented by a row in a matrix and each unique word is presented in a column. This matrix is naturally sparse: for each document there are many columns which have the value 0 to indicate the word is not present in this topic. To reduce the number of columns in the document-term matrix, both common and rare words can be removed from the document-term matrix.

3. Running the analysis. This step focuses on mechanical application of the algorithm and LDA distributions to determine the word loadings per topic and therefore, the topic loadings per document. When doing the mechanical application of the algorithm, the researchers is required to choose the number of topics (k). The algorithm produces exactly this number of topics. Therefore, the algorithm assigns each document to every topic (in some proportion) and assigns each word similarly to every topics. This can be considered similar to the execution of the EFA in other data formats.

4. Validating the results. This part is not computational, but includes validation and sensibility results of these methods.

This description should give the reader a general sense of approaches for these methods. I will not in this work address the internal workings of the LDA further to keep the work accessible for non-statistic oriented scholars. Similarly, as several other works provide further description of the methods, I have not aimed to replicate these efforts. Those interested on applications of automated (unsupervised) text analysis are advised to review Lucas et al. (2015) and Blei (2012) and other publications cited above for further illustrations.

3 Potential challenges in topic model process

I will now address two critical observations about the topic model process based on the reflections on EFA. The first one addresses the difficulty of choosing the number of topics and second one on validation strategies for these topics.

Challenge 1: Choosing the number of topics (k)

As discussed above, the topic model analysis requires a number of topics (k) as an input to the algorithm. In social sciences the common approach has been to examine few alternative ks and then choose from those the one which has most meaningful one for analysis (as a single example, see Levy and Franklin, 2013). Grimmer and Stewart (2013) even goes as far as to say "[d]etermining the number of clusters is one of the most difficult questions in unsupervised learning." They continue to discuss the difference between substantive fit, goal of the process for social scientists, and statistical fit, the common approach in the communities to examine model fit to the data. Due to the need for substantive fit, social scientists have been keen to continue handpick the number of topics manually based on seeing the words most heavily presenting certain topics.

I am concerned on this practice. To illustrate its similarities on EFA, this would mean that researchers examine some potential number of factors and choose the number which for them give best outcomes. This is why in EFA, there are numerous alternative methods available to determine the number of factors. To return to Watts's (2011) thinking, the challenge emerges as an human can rationalize almost any outcome once the result is shown. In this case, how we ensure the overall validity and reliability of the process?

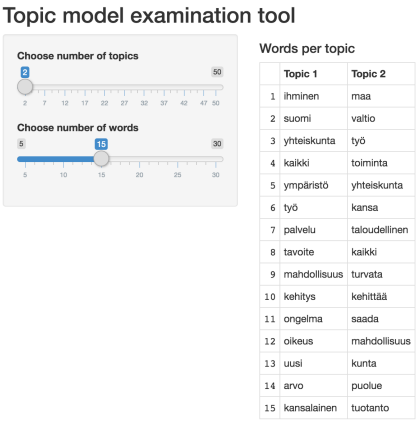


Figure 1: Experimental user interface

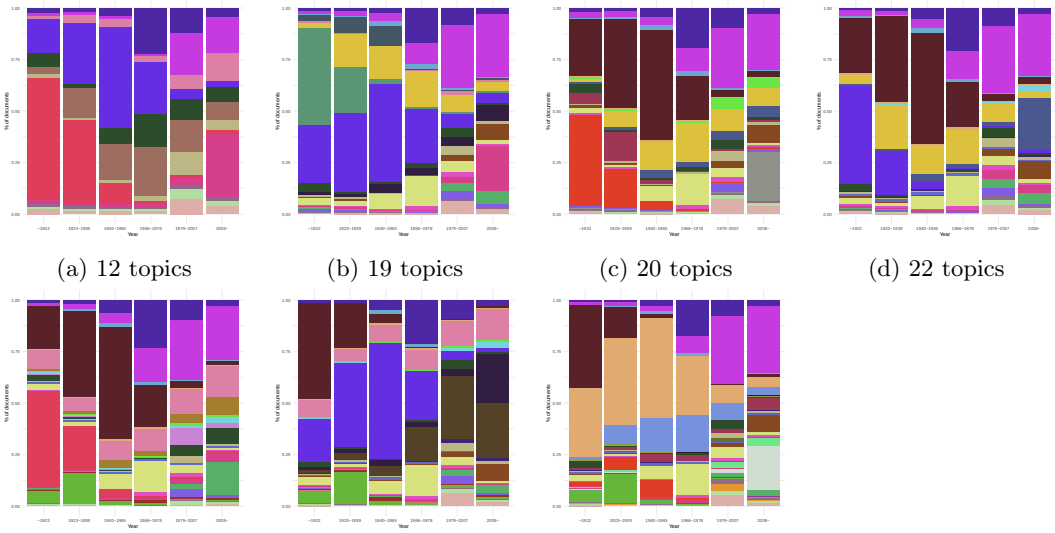
To further address this challenge, I created an experimental setup to examine how researchers determine the number of topics. I asked five researchers, all familiar with topic model process, to determine the number of topics which, for them, produced best substantive fit. Researchers changed the number of topics using a slider from an user interface (Figure 1) and could see the words and topics change. I asked them to change the number of topics and examine them until they were happy.

The experimental results diverged, meaning that different researchers found different number of topics highest in substantive fit. The number of topics were 12, 19, 22, 24 and 30. As their range of alternatives was from 10 to 30, we can summarize five researchers' substantive fit covered the whole range rather well. Therefore, the substantive fit can be almost everything, depending on who was conducting the analysis. The researcher chosen the lowest number explained to us that he aimed at reducing the number of topics, but observed that the 10 topics model had too much overlap and therefore, opted to 12 topics model to reduce overlap. Instead, the researcher who considered 30 topics model as the best argued that he increased the number of topics and observed that 'junk' did not increase while doing that. He therefore opted to the highest number of topics as it still increased the ability of the models to produce insights.

I also conducted the analysis of statistical fit using harmonic mean loglikelyhood measurement (Griffiths and Steyvers, 2004; Wallach et al., 2009). Using this approach, I observed that 20 topics was of best fit. There have been discussion on variety of metrics to evaluate fitness on topic models, including perplexity (e.g., Blei et al., 2003), loglikelihood (e.g., Griffiths and Steyvers, 2004) and Chib-estimator (Wallach et al., 2009). Therefore, the jury is clearly still out the statistical fitness evaluations. I have chosen harmonic loglikelyhood measurement as it is well integrated with the `topicmodels` package.

The question therefore is: which of these six different alternatives is the "right" one for analysis for the data. To answer this, I will illustrate the effect different ks have in the data analysis. The dataset consists of Finnish political party programs from early 20th century to 21st century.

Figure 2 shows the distribution of topics per time period, following Mickelsson's (2015) model. In that Figure, a single color demonstrates the share of all documents which belong to that topic. The topics have been mapped over time, therefore between Figures 2a–2g colors have approximately the same meaning (same words loaded to them). According to him e fifth period (1979-2007) related to major changes in the political agenda and emergence of new political topics, such as ecology and feminism. In my view, the 12 topics model (Figure 2a) does not fully



(a) 12 topics (b) 19 topics (c) 20 topics (d) 22 topics

(e) 24 topics (f) 26 topics (g) 30 topics

Figure 2: Examples of Finnish party data as seen through different topic models

demonstrate it. There is some increase on topics which has been previously non-existing, but this is rather minor. Instead, the 20 topics model (Figure 2c) and 30 topics model (Figure 2g) illustrate the change by allocating more novel topics to analysis.

Figure 2 shows also that the different models create rather different distributions per year. For example 12 and 24 topics models are dominantly pink in the years before 1940's while in 26 topics model it is more dark brown. While the topic loadings in these structures are rather different, the qualitative differences behind these are less critical. Some topic models seem to have initialized more to focus on e.g., ideological language (communism, capitalism and socialism) while others have highlighted party specific topics (like topics explicitly mentioning parties by name). This may relate to the stochastic process of initializing the topic models and as such as partly an artifact. (However, based on this observation, having an explicit seed for random process might be something worth of considering.)

Challenge 2: Validating findings

Validation is seen as a critical part of topic model process. Topic models are not expected to replace close reading of the text material (Grimmer and Stewart, 2013). Indeed, big data always needs context information to provide a meaning for the data (Boyd and Crawford, 2012). However, this far the literature has been sparse to provide exact approaches to work with the data. I will shortly review them and following this, describe some observations from my own work.

Chang et al. (2009) proposed an experimental methodology to examine topic models. Their methodology focused on measuring the coherence of topics. They showed participants six words and five of them were picked from a topic and they were asked to identify the sixth intruder word. They also showed participants few sentences of a topic and again six words, one of which did not belong to that topic. Overall, Chang et al.'s (2009) observation based on these experiments was that experiment participants did not success in the experiments. Therefore, they conclude that the human understanding of a topic and their computational meaning of a topic did not match well.

Similarly, Towne et al. (2016) conducted an experimental setup to measure the coherence of

topics. Their methodology is radically different from Chang et al's approach. Instead of single words, participants were asked to identify a document that did not match to computational analysis. They were provided with three documents, two from a same topic and third from another topic. They observed that 75% of participants successfully identified the wrong document. Participants also reported high level of understanding why this took place and what are the meanings of topics.

The differences between Towne et al.'s (2016) and Chang et al.'s (2009) works give rather different idea of the ability of the methods to produce meaningful results. I believe this is in part related to the data sets used as those can produce rather different topics. Moreover, the execution of the topic model algorithms may be to blame. Nonetheless, for me Chang et al.'s approach is more strict than Towne et al.'s work. For research purposes, focus on coherence at a document level would be sufficient in many cases.

Furthermore, I am doubtful if such formal and extensive settings are required for validation. Running these experiments for each paper which applies topic models require too much efforts. Rather, following Boyd and Crawford's (2012) argumentation, contextual understanding is required to provide meaning for the data. We have applied *big-data-augmented ethnography* and during the (online) data collection ethnographic field work to observe what the data consists of (Laaksonen et al., 2017). The field notes were then examined to inform the computational data analysis. In this manner, we were aware of the contents and no extensive reading after analysis was required to understand the meanings for observations. Rather, the close reading work was already conducted during the ethnographic period.

Another opportunity to apply previous work if it is available on the same topic. For example, the analysis of party programs has benefitted Mickelsson's (2015) work on Finnish political systems. The previous work provided insights and therefore allowed us to 'validate' the topics. This is to say, clear outlines can be detected and addressed based on that. For example, one topic was clustered around words like joga and natural law. While initially surprising and clearly needing further clarification, Mickelsson (2015) indicated it was one of the radical ecology parties in Finland.

The existence of several approaches for validation calls for further analysis of these differences. In EFA this stage focuses more at labeling the factors, not on the internal coherence of each item in the factor. (These are identified based on the factor loadings rather explicitly.) The focus on validation could therefore attribute more on the human interpretation of topics. The work process could focus on examining differences on traditional qualitative coding and categorization to computational results, an experimental approach I am not aware of being used.

In the test setting, 20-30 short texts are selected from a larger corpus. The corpus has already been topic modelled and therefore, the texts can be chosen so that they represent mostly fiveish topics. Following methods are tested to examine the results emerging from them:

- Comparison to traditional qualitative methods. Test participants is asked to organize the texts to groups and continue working on the texts until the number of groups is the same as the the number of topics.
- To simulate our big-data-augmented ethnography, participants are asked to read introduction material on a topic. Following this, participants are given set of words and are asked to produce number of topics clusters of these words.
- Chang et al.'s (2009) approach to identify an outlier word.
- Towne et al.'s (2016) approach to identify an outlier document.

After all these experiments, the success and error rates are measured and each of them is compared and sorted. The goal is to identify the benefits and challenges of each method for validation discussed above.

4 Discussion and Conclusions

Social sciences has applied unsupervised approaches to data analysis for already since 1960s. However, the experiences of explanatory factor analysis (EFA) have shown that unsupervised approaches are rather often, almost 20% of cases, poorly applied (Fabrigar et al., 1999; Russell, 2002; Bandalos and Boehm-Kaufman, 2010). Given this, the researchers focusing on "Text as Data" – the approach to understand text through computational methods – ought to focus on methodological aspects and investigate those (Grimmer and Stewart, 2013). This work further examined the methodological challenges and pitfalls of topic models, a prominent approach to group text documents to different topics (in an unsupervised manner).

My empirical work examined the subjective nature of unsupervised methods as the first challenge. The experimental setup showed that the substantive fit is subjective. Five researchers, examining the same topic models, recommended number of topics from 12 to 30. The different numbers of topics lead to different outcomes in the analysis phase. We showed that the change of political issues in the 1960s is more clearly shown with higher number of topics compared with the 12 topics model.

To reduce the subjectivity of the substantive fit, researchers can instead apply statistical fit measurements. Researchers warn statistical fit approach can produce topics difficult to conceptualize and therefore recommend substantive fit instead (Grimmer and Stewart, 2013). However, my experiment shows that statistical fit can reduce the potential bias caused by the subjective analysis. This, for me, is more valuable than the subjective fit of the topics. I will further examine this through the narrative of algorithmic power in Section 4.1.

My second observation related to questions of validation. While validation is commonly seen as a critical stage of unsupervised analysis (Grimmer and Stewart, 2013), an explicit process for this has not been to my knowledge presented. I proposed four different methods to evaluate validity and suggested these should be experimentally studied. Most significantly, I wish to highlight methods where context data is brought in the data analysis process, such as the big-data-augmented ethnography (Laaksonen et al., 2017). This observation still lacked empirical analysis, but I will engage further discussion of the unsupervised research process in Section 4.2.

4.1 The power of algorithms and unsupervised methods

Algorithmic objectivity

Social scientists have been recently interested on the power of algorithms (among others, Kitchin, 2017; Gillespie, 2017; Beer, 2017). Given the significant role of algorithms, algorithmic systems and algorithmic decision making, Hallinan and Striphas (2016) ask if algorithms shape the culture Gillespie (2012) seeks to understand the impact of algorithms to relevance and, Bucher (2016) explores the algorithmization of journalistic processes. As this literature is ever increasing, the examples were chosen to illustrate social scientists consider that algorithms (and the creator of algorithms) have power which they impose to citizens using digital systems.

This work is known as critical algorithm studies; its main message is to consider the social role technical systems have today. The literacy is rather critical about algorithms and algorithmic systems and see them as a vehicle of power today. While I think their assessment of

the situation is sometimes extreme, I agree with their attempts to focus on digital systems and power. However, their focus has been on digital systems and everyday society.

As computational methods are applied in the social sciences, the questions asked in the critical algorithm studies become relevant. Researchers can impact the results and clarity of the outcomes when using unsupervised approaches. For example, the choice of the parameters or the random seed of initial locations impacts the findings. In this work, I aimed to make some of the impacts of these choices more transparent.

The challenge can be that the algorithmic process often seems to hide the subjective biases and decisions. These are not addressed in works which have applied topic model analysis, but rather they have focused on the notation of substantive fit and experimented only few alternative number of topics. To provide an alternative for this approach; the analysis of party programs presented above examined total of 299 different values of k, from 2 to 300, to ensure the space for opportunities is covered.

More generally, as digital methods are applied, we need to think what has been the origin and purpose of these methods. What type of assumptions are encoded into the methods and what has been the goals of the developers. To illustrate, the mixed-membership model of topic models assumes that each document belongs to several topics. Topic model emerges from computational data analysis community, which is seen in the bag-of-words approach and thus, missing any context and only focusing the presence of words. Similarly, the discussion of suitable metrics of fitness (Griffiths and Steyvers, 2004; Wallach et al., 2009) have been focused on traditional methods in data analysis. While I have recommended the use of these metrics, the question for social sciences is: is some metrics and concepts we apply to traditional qualitative work more suitable for research process.

4.2 Rethinking the research process

My second discussion relates to the research process with unsupervised methods. Kitchin's (2014) characterization of the modern approach for research is data intensive science. In this process, the hypothesis setup and data exploration are not separated stages, but rather go hand in hand during the research process. Based on my experience on the several research projects, this description in my view is well suited to the practices. However, the process is also rather unclear; there is no clear consensus on the stages of analysis. This may make it hard to apply and educate unsupervised methods.

The big-data-augmented ethnography takes a step to address this (Laaksonen et al., 2017) and presents a four stage model for conducting the research (see Figure 3). The research starts with ethnographic field work which informs the data collection strategy. In data analysis phase both the ethnographic field notes and the collected big data is analyzed. The analysis of big data analysis is informed through field notes, and the analysis of the big data informs further and field note analysis informed through data analysis. Finally, there is a synthesis of the the findings.

In similar manner, Muller et al. (2016) examine how traditional grounded theory methods can be applied together with unsupervised models. They discuss sequential models, where either grounded theory or unsupervised models are used first, followed by the other methods. They also propose there is an opportunity to conduct integrative work, where both methods are applied together; e.g., constant comparisons are used.

However, even while Laaksonen et al.'s (2017) and Muller et al.'s (2016) proposals aim to address the challenges of computational data analysis, they do not present a process without an

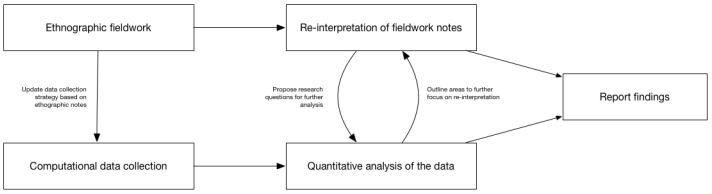


Figure 3: Big-data-augmented ethnography according to Laaksonen et al. (2017)

qualitative phase. It may be that this is the case: computational analysis must be supported through qualitative research stage. However, I believe this is not always the case: the categories can be supported also via e.g., previous literature, as was the case for example above.

As I think the research needs for topic models, they are rather similar to grounded theory. In grounded theory, the aim is to examine the qualitative data through coding and then conduct theory-building based on these codes. Through out the data analysis, memos are made to ensure the transparency of the process. Similar steps of memoing can also support the unsupervised data analysis methods.

I acknowledge that this far, this description and idea this far is rather vague. The question of developing this further relates more to overall need for such methodology – of which I am unsure. The boom of papers using unsupervised computational methods has been increasing, but the question is if the research community would benefit from more structured approach to this. The challenge with such approach could be the various misuses of the terminology, which has been partly the case of grounded theory methods this far.

References

Bandalos, D. and Boehm-Kaufman, M. (2010). *Four common misconceptions in Explanatory Factor Analysis*. Routledge.

Beer, D. (2017). The social power of algorithms. *Information, Communication & Society*, 20(1):1–13.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Boyd, D. and Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, 15(5):662–679.

Bucher, T. (2016). 'Machines dont have instincts: Articulating the computational in journalism. *New Media & Society*.

Burscher, B., Vliegenthart, R., and de Vreese, C. H. (2015a). Frames Beyond Words: Applying Cluster and Sentiment Analysis to News Coverage of the Nuclear Power Issue. *Social Science Computer Review*, (1991):1–16.

Burscher, B., Vliegenthart, R., and De Vreese, C. H. (2015b). Using Supervised Machine Learning to Code Policy Issues: Can Classifiers Generalize across Contexts? *The ANNALS of the American Academy of Political and Social Science*, 659(1):122–131.

Chang, J., Gerrish, S., Wang, C., and Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems 22*, pages 288–296.

Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., and Potts, C. (2013). A computational approach to politeness with application to social factors. *The 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., and Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3):272–299.

Gillespie, T. (2012). The relevance of algorithms. In *Media Technologies: Essays on Communication, Materiality, and Society*, pages 167–194.

Gillespie, T. (2017). Algorithmically recognizable: Santorum's Google problem, and Google's Santorum problem. *Information, Communication & Society*, 20(1):63–80.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Supplement 1):5228–5235.

Grimmer, J. and 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.

Hallinan, B. and Striphas, T. (2016). Recommended for you: The Netflix Prize and the production of algorithmic culture. *New Media & Society*, 18(1):117–137.

Hopkins, D. and King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247.

Hornik, K. and Grün, B. (2011). topicmodels: An r package for fitting topic models. *Journal of Statistical Software*, 40(13):1–30.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and psychological measurement*, 20(1):141–151.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1):1–12.

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society*, 20(1):14–29.

Laaksonen, S.-M., Nelimarkka, M., Tuokko, M., Marttila, M., Kekkonen, A., and Villi, M. (2017). Working the fields of big data: Using big-data-augmented online ethnography to study candidate–candidate interaction at election time. *Journal of Information Technology & Politics*, 14(1):1–22.

Levy, K. E. C. and Franklin, M. (2013). Driving Regulation: Using Topic Models to Examine Political Contention in the U.S. Trucking Industry. *Social Science Computer Review*, 32:182–194.

Lucas, C., Nielsen, R. a., Roberts, M. E., Stewart, B. M., Storer, A., and Tingley, D. (2015). Computer-Assisted Text Analysis for Comparative Politics. *Political Analysis*, pages 1–24.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Metsämuuronen, J. (2006). *Tutkimuksen tekemisen perusteet ihmistieteissä*. International Methelp ky, Helsinki, Finland.

Mickelsson, R. (2015). *Suomen puolueet: Vapauden ajasta maailmantuskaan*. Vastapaino, Tampere.

Muller, M., Guha, S., Baumer, E. P., Mimno, D., and Shami, N. S. (2016). Machine Learning and Grounded Theory Method. In *Proceedings of the 19th International Conference on Supporting Group Work - GROUP '16*, pages 3–8, New York, New York, USA. ACM Press.

Nelimarkka, M. and Ahonen, P. (0). Measuring deliberation using machine learning.

Purhonen, S. and Toikka, A. (2016). "Big datan" haaste ja uudet laskennalliset teksiaineistojen analyysimenetelmät. *Sosiologia*, (1):6–26.

Rehurek, R. and Sojka, P. (2010). Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

Roberts, M. E., Stewart, B. M., Tingley, D., Airoldi, E. M., et al. (2013). The structural topic model and applied social science. In *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*.

Russell, D. W. (2002). In Search of Underlying Dimensions: The Use (and Abuse) of Factor Analysis in Personality and Social Psychology Bulletin. *Personality and Social Psychology Bulletin*, 28(12):1629–1646.

Towne, W. B., Rosé, C. P., and Herbsleb, J. (2016). Measuring Similarity Similarly: LDA and Human Perception. *ACM Transactions on Intelligent Systems and Technology ACM Reference Format ACM Trans. Intell. Syst. Technol*, 7(2):1–25.

Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, number 4, pages 1–8, New York, New York, USA. ACM Press.

Watts, D. J. (2011). *Everything is obvious: Once you know the answer*. Crown Business.

Wilkerson, J., Smith, D., Stramp, N., and Dashiell, J. (2013). Tracing the Flow of Policy Ideas in Legislatures: A Computational Approach. *2013 Annual Meetings of the Comparative Agendas Project, Antwerp*, 00(0):1–19.

Yu, B., Kaufmann, S., and Diermeier, D. (2008). Classifying Party Affiliation from Political Speech. *Journal of Information Technology & Politics*, 5(1):33–48.

Zhang, A. X. and Counts, S. (2015). Modeling Ideology and Predicting Policy Change with Social Media. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, pages 2603–2612, New York, New York, USA. ACM Press.