“What Topic model Should I choose?” - measurement error induced by model choice in automated text analysis

- Snorre Ralund, Hjalmar Bang Carlsen, Robert Klemmensen, David Drejer Lassen

Abstract

The dream of unsupervised automated text analysis is that we become able to identify valid and reliable topics at scale without polluting data with arbitrary human judgements. In this paper we show that current criteria for labeling and accepting topic models include arbitrary judgements and that they lead to nontrivial measurement errors. Model labeling and choice relies on three unwarranted assumptions 1) semantic coherence between the most predictive documents within topic cluster and the rest of the documents 2) insignificant variation between topic models classified as the same topic 3) explicit performance measures are not necessary.

Using data from Facebook 46.000+ posts from Danish politicians we compare simulations of 1600 topic models to a labels created by a pre-defined dictionary of 15 topics and find 8969 matches. To investigate the varying performance of possible choices we narrow this down to a set of 2960 equally plausible choices. Our results strongly suggest that topic models are not 1) semantically coherent and 2) do not treat all topics – or the topic articulation of different actors – equally, leading to differential bias in the results. To illustrate the consequences for research in political communication we simulate 1600 end-to-end research projects(model labeling, model acceptance and conclusions) of opportunism in political communication. We show that political opportunism can be assigned to all political parties depending on choice of model. We end by pointing at alternative research strategies for automated text analysis.

Introduction

The amount of textual data available to social scientists have exploded in recent decades and has led to the development of more and less automated methods that can extract different features from texts. Topic models have played a particularly important role in many branches of the social sciences including political communication. Topic models have been used to analyze the distorting effects of electoral incentives on Congressional representation (Grimmer, 2013); the dynamics of democratic agenda setting in Congress speeches (Quinn et al., 2010); the substantive traits of trade bills that were lobbied and those that were not (Kim, 2017); the political discourse on cryptomarkets (Munksgaard & Demant, 2016) and the cultural carrying capacities of NGO’s in their political communication (Bail, 2016). The rich array of empirical applications of topic models clearly demonstrate the need and usefulness of automated text analysis. But as many proponents of topic modeling have stressed over and over again, we have to run iterative runs of our models on the data in order to obtain substantive model fits and we have to validate results in a variety of ways.

We argue that current criteria for model selection are flawed in ways that can lead to nontrivial measurement errors. This is due to a lack of representativeness between the sample used to label topics and the underlying distribution of words that the topic represent. We will call that the *problem of semantic coherence*. Topic modeling procedures *assume* semantic coherence when they choose and label their topic model by interpreting the most predictive words and documents (Grimmer & Stewart, 2013; Quinn et al, 2010). This naturally makes the procedure vulnerable to misclassification and it can lead to situations where the analyst chooses a model that does not have high precision. These problems could of course be inconsequential if the lack of precision did not introduce systematic biases in topic propositions and in the mapping of topics onto other variables of interest (time, actors aso). Unfortunately, it does just this. Our results strongly suggest that topic models do not treat all topics – or the topic articulation of different actors – equally, leading to systematic biases in the result. We contribute to critical analysis of topic model research (Schmidt, 2012; Lancichinetti et al, 2015) by 1) focusing on the consequences of model choice for social scientific research results and 2) setting up explicit criteria for how researchers can and can't use topic models for measurement and statistical inference.

In this paper we illustrate the problems of model selection by investigating how opportunistic Danish politicians are in their Facebook updates across topics. Our definition of opportunism is that actors priorities topics after the amount of positive response they get from their audience, a variant of what Jacobs & Shapiro(2000) refer to as pandering or political responsiveness. We show all political parties can be identified as opportunistic and that all parties can be identified as non-opportunistic depending on which topic model we choose to believe. On the basis of these results, we argue that strict criteria for model validation and that studying the distribution of model candidates using simulations can help to us make valid decisions across models.

2. The choice and classification of topic models

Topic models are increasingly used in the social science. They have been praised for being able to scale texts with minimum costs and assumptions (Quinn et al. 2010). The intuition behind topic modeling is that texts are about multiple topics. More formally, a topic is a distribution of words and a document is a distribution of topics (Blei 2012). A distinction commonly drawn in the literature is between topic models and supervised machine learning. While supervised methods require the researcher to know before hand what categories are in the data, topic models inductively derive topic candidates from statistical patterns within data. Topic models thereby simultaneously identifies topics in the data and make it possible to measure the content of those topics. Where other text methods require a lot of validation work before scaling, topic models requires it after scaling. In the following we run through the different parts of the topic modeling procedure: classification, validation and reliability testing.

2.1 Classification

The discovery and classification part of topic modeling involves an interpretation of the topics the model returns, and a decision of whether to accept them as meaningful and relevant or not. This is done by sampling the words and documents that are most predictive of a topic. It is this step that ensures the *substantive fit*, the interpretive coherence and theoretical usefulness, of the model. This is arguably the most important part of the topic modeling procedure, because we cannot rely on *statistical fit* as the main and only principle of model selection, due to the fact model fit is based on transformation of documents into word frequencies, which by itself does not represent the entity of interest, namely the meaning or topicality of the original document (Grimmer and Stewart 2013: 286). It is therefore mainly the substantive fit of the topic model that leads researchers to proceed with a set of topics to be validated by other means. This procedure has a set of related problems. 1) It unwarrantedly assumes semantic coherence throughout the topic, which leads to need for 2) methods to handle and minimize classification errors which is not implemented in current practice. By interpreting a topic only from its most predictive words and documents, the risk of misrepresenting what the topic contains at lower levels is large (less predictive words, less predictive documents). This becomes evident if we read across topics and interpret them using only the most paradigmatic cases instead of a random sample. Thereby we risk bias by seeing the large difference in data and losing sight of the overlap between topic clusters. The most predictive words and documents from each topic should, by definition, carry the most distinct topical traits. Formulated in classic clustering terms, one is biased towards assuming large between-topic variation and small within-topic variation. This is not a problem for initial exploration, indeed, it might be a virtue, but it is a problem for cluster categorization. In a situation where we intentionally do not sample randomly, it seems obvious to ask what this does to one's measurement.

2.2 Validity and Reliability

To be fair, proponents of topic models are well aware of the problems of validity, and seminal papers on topic models present various ways of validating topic models such as semantic validity, predictive validity and concurrent validity among the most important. But as Grimmer and Stewart (2013) argue these validation methods are, unlike in supervised machine learning, not made directly on the validity of the models categorization of documents. As a consequence, we do not have any information of the potential biases introduced by the chosen model. In the context of big data, just a little systematic noise will have significant effect on your results (McFarland & McFarland 2015).

Reliability in content analysis refers to the way categorization devices are "free of influences by circumstances that are extraneous to processes of observation, description, or measurement" (Krippendorff, 2009: 350). This concept of reliability contains more that replicability as implied by

Quinn et al (2010). Quinn et al's (2010) conception of reliability lead them to state that topic models can be regard as "100 % reliable, completely replicable" (Quinn et. al 2010: 216). However, it is has been argued that the setting of hyperparameters, number of clusters (Grimmer & Stewart, 2013) and priors (Wallach et al, 2009), have significant effects on the performance of topic models - choices that are exogenous to the process of observation, description and measurement. The choices of hyperparameters are however, more or less, explicated and therefore demand the researcher attention and justification. The reliability problem we address in the following is the consequence of the variation between seemingly similar topics, something not directly accessible to researcher given that the topic can share the same or similar predictive words and documents and yet have very different underlying documents. The instability and fluctuations in and across models is a well-known property of topic models (Roberts et. al 2016). These issues can't be solved alone by optimizing statistical fit, as we argue above, because statistical fit does not ensure substantive fit (Chang et al. 2009). The big question is then what this variation across similar topic model candidates does to measurement error and ultimately to research results.

3. Simulation segment.

In order to test the consequence of choosing among equally plausible model candidates we simulate more than 1600 possible topic models. We then use the qualified models to test the same hypothesis, and show the variance of the results based on the choice of different plausible model candidates. We specifically address the way a model substantively fit is determined (as described above). We mimic this by having a set of dictionaries with topical words describing 15 different political topics (environment, economy, health, etc.); we can then determine the overlap between each dictionary and the most predictive words of a topic candidate from the topic models. Using the same dictionary logic, we can assess the degree to which the most predictive documents includes the topic we expect. By training 1600 different topic models on the same data, we can find topic candidates that would have been accepted with some plausibility, - i.e. has a known topicality of the top words, and has the same topic represented in the most predictive documents.

Because there is no information that differentiates the different topic candidate, we can treat this as a random choice, and investigate the consequences and large variances in results obtained from choosing one specific model. In this test design we assume that supplementary methods for choosing topic models (the validation methods described above and optimizing after statistical fit) does not systematically change the topic model choice situation.

3.1 Dataset

Using a dataset consisting of 400.000+ Facebook posts from Danish politicians from 2008 to 2016, we assess the potential consequences of model choice using conventional methods for a research project. As it has been demonstrated elsewhere (Tang et al., 2014), topic models are not particularly well suited for social media data, because the length of the documents make the inference of topic proportions too dependent on the prior. To make our dataset more suited to the model, we filtered out posts with less than 200 characters, and with more than 2000 (footnote: One problem with using this type of probabilistic inference as measurement is that comparing documents of different size is similar to comparing the p-value of a study with 100 participants to a study with a million participants.). Furthermore, we manipulated the dataset to include a set of predefined topics that we have defined using the dictionary method of classification (Grimmer & Stewart, 2013) (Dictionary includes 701 curated topical words of 15 different topics, see supplementary material). We reduce the dataset to documents only consisting of one of these topics, to ensure good conditions for the topic model to identify the predefined categories (and possible labels of a topic discovered by the model).

After these reductions we are left with a dataset of 46209 posts which we labeled according to our dictionary. We rely on the assumption that dictionary labels will have a biased but relatively stable relationship with the true labels, allowing us to compare the relative performance of different models. For each plausible topic choice and label (i.e. topics with dictionary words in the top 10 most predictive words and corresponding dictionary labels in the 5 most predictive documents), we can evaluate the performance of each model. It is, however, important to note that the main point of the following investigations is not to evaluate the performance directly, but show the high variance in performance of equally plausible choices of meaningful topics and labels.

In order to ensure that our model has enough data we tested its performance and the learning curve saturating around 40.000 samples (see appendix), we can therefore conclude that the size of the dataset is a fair test ground.

3.2 Simulating a set of plausible topic choices

Until now we have referenced topic models very broadly. For the purpose of this experiment we have chosen Latent Dirichlet Allocation (LDA). The specific problem of labeling without knowing the underlying reliability of topic distribution is shared by the whole family of models, but more complex developments of the basic topic model such as the structural topic model (Robert et al, 2013) might behave differently in relation to our tests of differential bias.

Latent Dirichlet Allocation is a probabilistic model, and the high dimensional space (N words * M documents) it operates on combined with the optimization algorithms used makes the solutions non-deterministic and therefore also suboptimal (footnote: Roberts et al. (2016) has investigated this, but only to obtain some sense of the variability of the measurement, and not to question the model itself.). Running the model many times, and with different parameter settings we can create a population of possible topic models. We search a limited space of possible solutions varying the number of topics, K, from 15 to 30, to imitate actual practice, where simulation of many solutions is costly computationally and time consuming in the model inspection and interpretation step (*footnote: To be fair, model selection using a larger grid search is also done accompanied by evaluations of the model using statistical measures. Substantially, a model is only accepted if it can be interpreted, and it is well known that these measures do not necessarily correspond to how humans understand and interpret topics (Chang et al. 2009)). For each K we simulate 100 different models, amounting to 1600 different topic models, and 36000 different topic candidates.

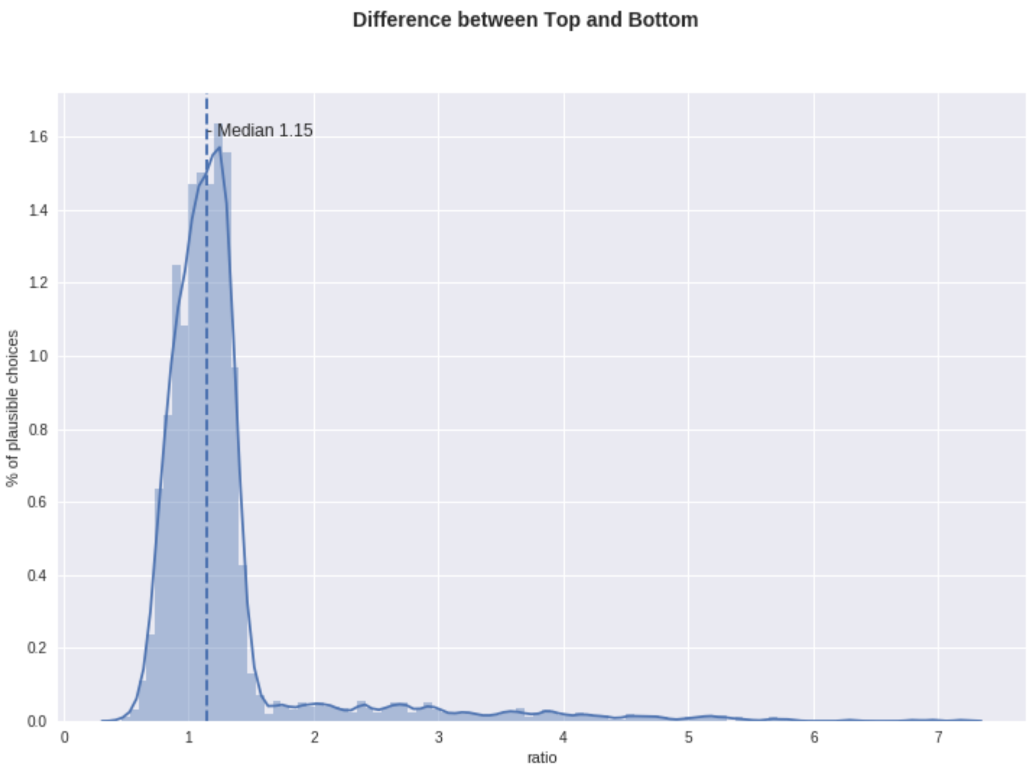
3.3 Matching Topic Candidate to Dictionary Label

For each topic candidate we simulate the labeling process done by a researcher. This is done with the assumption that one tries to create an abstraction that fits the most predictive words and the most predictive documents. By curating a dictionary of words expressive of well-known topical categories like education, crime, health, environment, we assume that given a set of these words, the researcher label it a given topic. If the most predictive documents also contains these words, then we have what we call a plausible choice of topic and label. Across the 36000 topic candidates, 2960 were matched as a plausible choice.

After all plausible choices have been found and labeled, we can investigate the effects of accepting one models educational topic over another plausible choice. As already noted the idea is not to assess the quality of each topic directly, but rather to bring forward the large variance between essentially arbitrary choices.

3.4 Tip of the iceberg assumption

As already noted, when labeling a topic using the most representative examples, we have to assume semantic coherence all the way down to the less representative examples. If this assumption does not hold, we risk overestimation of the topic. In figure 1 we investigate the distance between top 10 % most predictive documents and the bottom 90 % of each accepted topic. As all documents have some probability (although small) of belonging to any given topic, we follow Quinn et. al (2010:213) and simplify the task by only looking at the most prevalent topic of each document (i.e. convert the soft assignment to a hard-assignment). For each plausible choice of topic we then compare the label distribution according to our dictionary of the top 10 % to the bottom 90 %. We answer the question: How much more likely is the believed label in the top 10 % than in the bottom 90 %, normalized by the expected difference, as expressed by topic proportions in the documents of the top and the bottom?

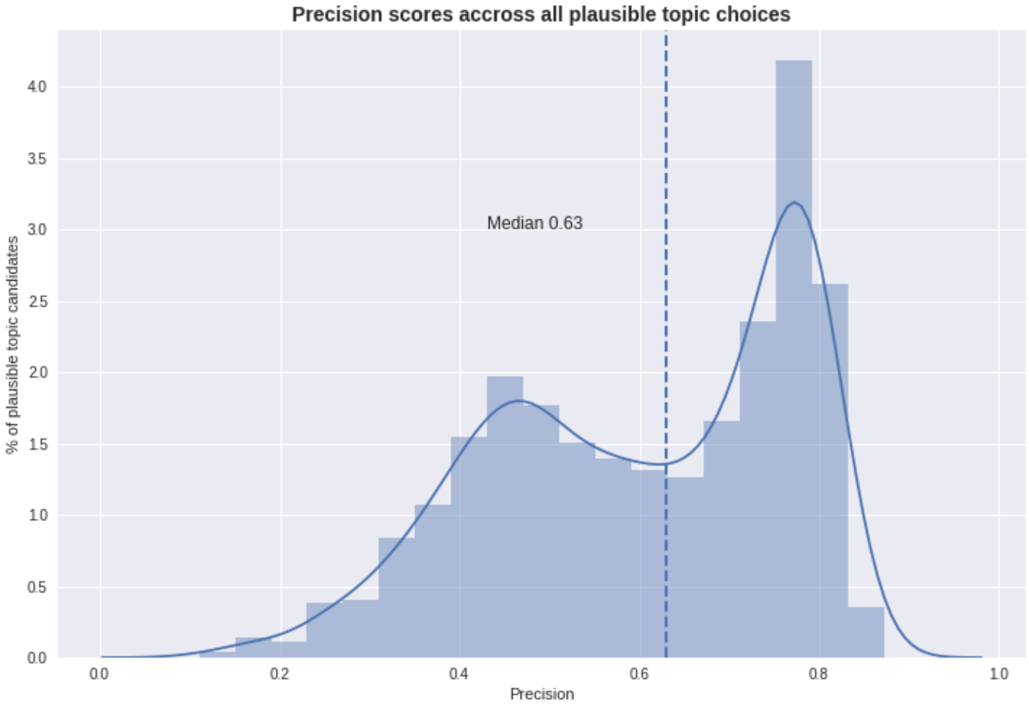


[Figure 1 Comparison of Top and Bottom. The figure expresses the distribution of ratios between the fraction of the believed label in the top 10 % documents of each topic compared to the fraction in the bottom 90% normalized by the topic proportions of the top and bottom].

Figure 1 clearly show there is no 1:1 relation between the top and the bottom, this suggests that is it not possible to assume semantic coherence. Looking only at the topic candidates where all the most predictive documents had the same dictionary label (seemingly coherent), and even looking further down to the top 10 % documents, it is evident that the top is not representative of the whole topic. We can see that it is fairly common (almost 10 percent) that the believed label is more than twice as prevalent in the top 10 % as the bottom 90 %. The median of 1.15 ratio, points to a 15 percent overestimation, and only 13 % of plausible topic choices fall in a reasonable 0.95 to 1.05 difference.

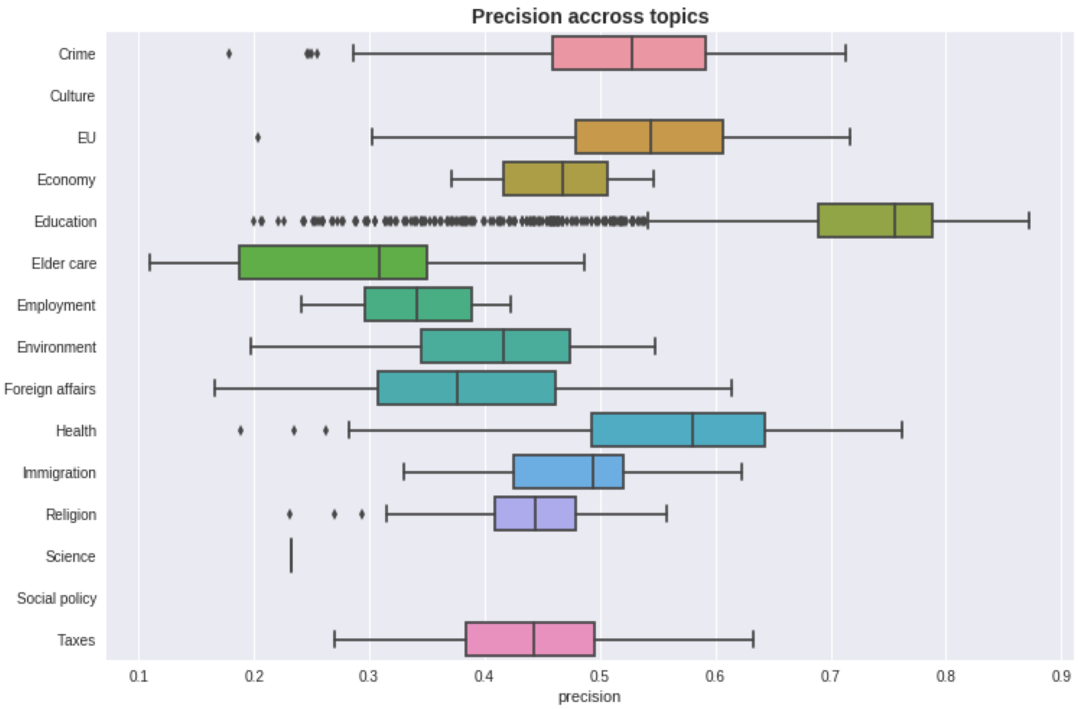
3.5 Arbitrary Choice affecting precision and differential bias

We now turn to the precision of topics in the estimated models. One of the main points of this article is that using topic models as measurement models without explicitly assessing the reliability can lead to major biases in results. Where machine learning traditionally was concerned mostly with performance measures assessing the accuracy of models, there is a growing awareness in the field (*fotnote: Under the theme of Fair Accountable and Transparent Machine Learning) about how models can have differential outcomes on different groups (e.g. Eisenstein et al 2014, Bolukbasi et. al 2016). Each plausible choice of topic is believed to measure one label, but what is the precision in relation to our corresponding dictionary labels. Again we do a hard-assignment of each document to the most prevalent topic, and then we compare the fraction of true labels of each document actually corresponding to the believed label of the topic, using the precision measure (footnote: Precision is defined as the ratio of true positives to the sum of true and false positives).



[Figure 2. Histogram of Precision scores across all plausible topics.]

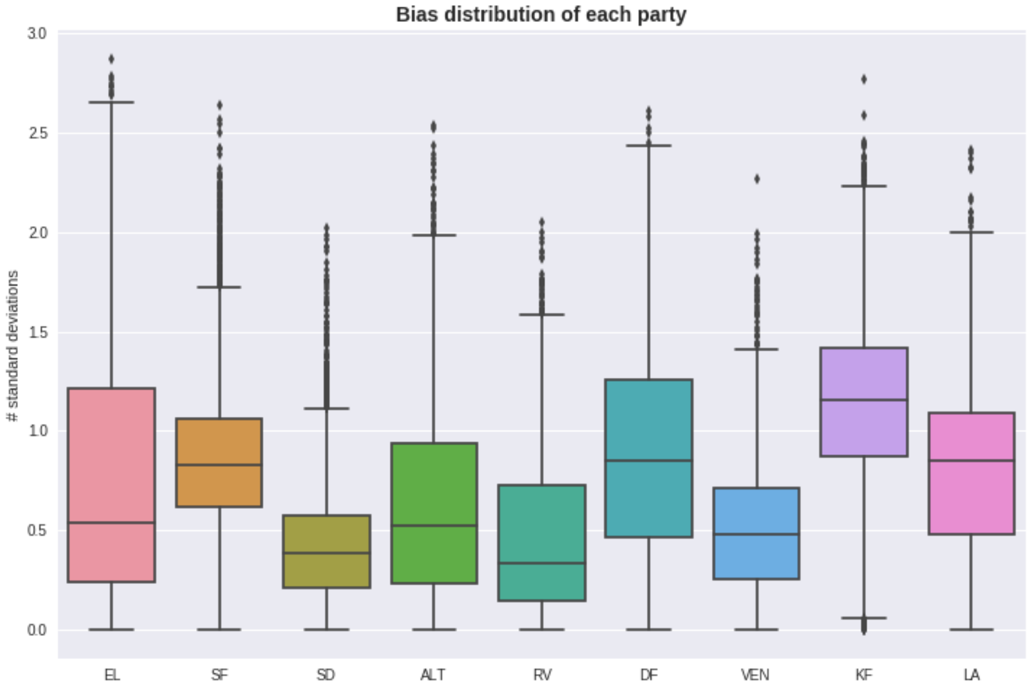
Figure 2 both shows that some models are able to find fairly good reconstructions of our planted topics, but also that the plausible models have high variance in actual performance. The essentially arbitrary choice of model can impact model precision dramatically, suggesting that we need to explicitly investigate performance of the model, and not assume a optimal fit.



[Figure 3 Boxplots of precision scores across our predefined dictionary topics.]

Aggregating across topics in Figure 3, we see that the models perform very differently across topics (some topics being absent altogether), suggesting that there is a high chance of producing strong artificial effects in a research setting. The figure furthermore shows that a topic model's performance on *one* topic, does not ensure that it performs well on *other* topics as assumed by Grimmer and Stewart's seminal text (2013:290).

A fair criticism of this result is that the variance across topics could reflect a poor dictionary design. However this points back to the same problems researchers have when labeling more than one topic in topic modelling - i.e. that some labels might not be as good and precise as others. We have to assume that the model can find each topic with equal precision and that they are somehow equally coherent and identifiable. This puts even heavier weight on the labeling process, that cannot be treated as just a casual process of choosing an abstract category that seem to fit the top words presented.



[Figure 4. Boxplot of standardized deviations from the average precision score of a plausible topic across political parties. EL: Unity List, SF: Socialist People's Party, SD: Social Democratic Party; ALT: ALTERNATIVE; RV: Social Liberal Party; DF: Danish People's Party; Ven: Libral Party, KF: Conservative Party; LA: Liberal Alliance]

Finally for each plausible choice of topics we assess a separate precision score for each party. We then transform each precision score to a standardized distance from the precision norm of the plausible topic. The distribution for each party is reported in figure 4. It shows that the bias is distributed very unevenly across party lines suggesting that comparisons across background characteristics can be highly problematic. In the supplementary material we investigate how the number of training examples for each topic and party could be part of the cause.

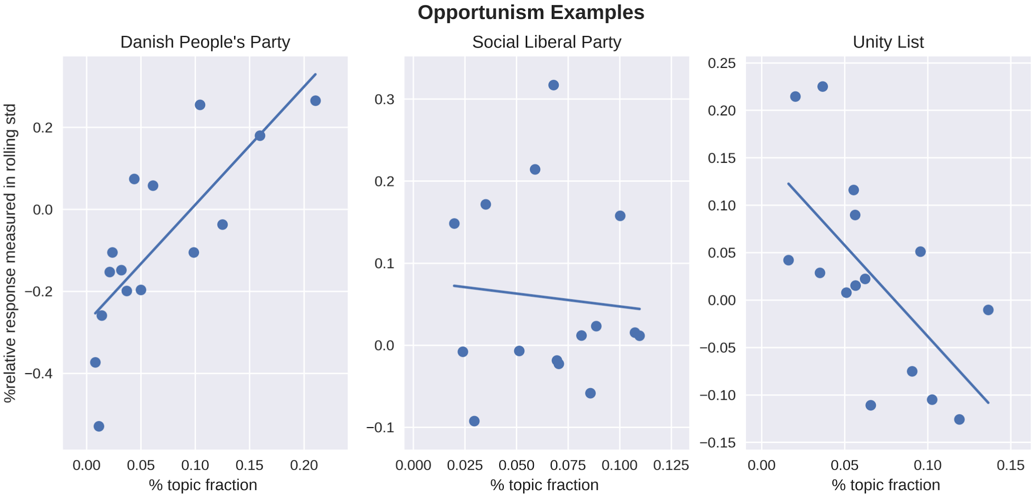
In conclusion, we can say that due to systematic differential bias across both topics and background variables, we can expect a complex set of artificial effects to be mistaken for real statistical evidence of an hypothesis.

3.6 Simulating the end-to-end effect of model choice on research conclusions

Until now we have demonstrated the high variance between what we assumed to be interchangeable model choices, by researchers lacking an explicit test of model performance. In this section we want to explicitly show that the bias introduced and the choice of model can not be treated as random noise, with no impact on our research. By asking a substantive question to be answered by each plausible model we can show how the conclusions can take many different and contradictory forms.

Research question

Social media data allows us to study the interactions between politicians and their followers. Consider the following research question: To what degree do politicians change their political focus as a function of the response they receive? To answer this question we constructed a simple test, quantifying the relation between how much a politician writes about a topic, and the relative popularity this topic holds with his or hers followers measured by 'likes'. Because Facebook is a dynamic platform with users joining (and to a lesser extend leaving) over time, and the number of followers a politician can change over time, we construct a measure of relative response, normalizing the number of likes a post receives by the mean and standard deviation in a sliding window around the post. Now we can ask: Relative to what response the politician gets at one point in time, does a post give him more likes or less likes? For each politician and for each party we can now aggregate this relative measure of response across topics, and compare this to how much the party posts about it. Illustrated in Figure 4, we can now observe and quantify a positive or negative correlation, expressing the degree of correspondence between each party's' priorities and their followers.



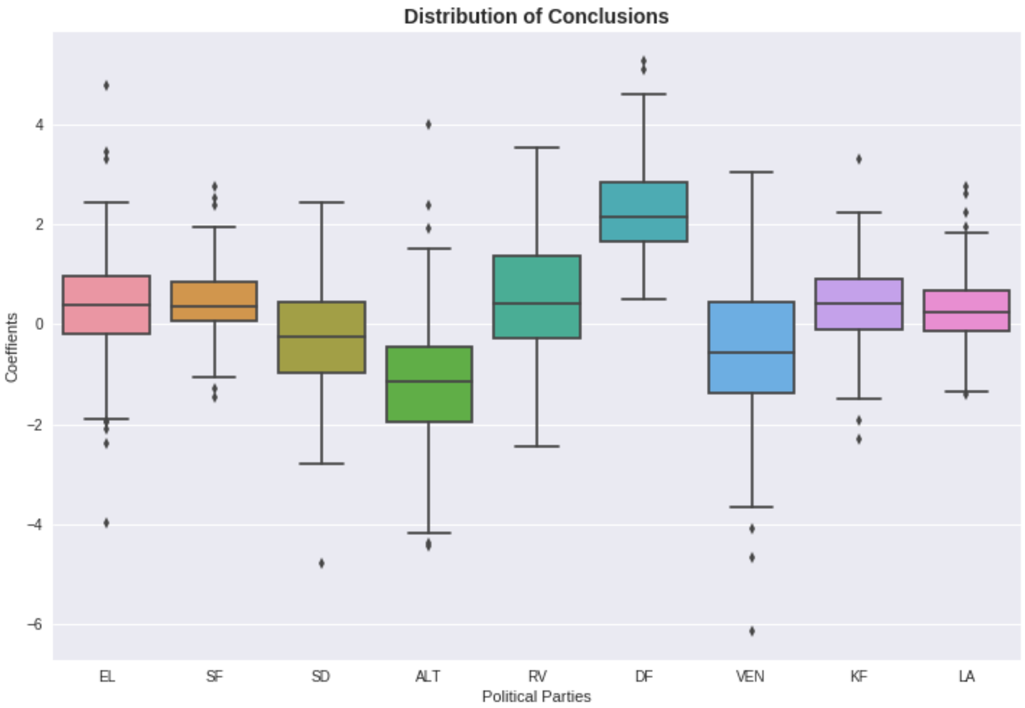
[Figure 5. Correspondence between the politicians topical priorities and the followers responses. The figure show three examples of correspondence between the fraction of posts written by a topic by a party, and the relative response the topics receives, using our dictionary labels as classification scheme, and a rolling normalization of likes as measure of response.]

If we were to accept the topic categories and classification of our dictionary we could now conclude (looking at figure 5) that the rightwing populist party (Danish People's Party) have a very high degree of positive correspondence with their followers, the social-liberal party (Radikale Venstre) has a very weak and spurious relation, and finally that the most leftwing socialist party (Enhedslisten) has a negative relationship between their own topical priorities and their voters response.

However one might rightly conclude that the method used can be criticized, and that the results are very much dependent on the choice and definition of categories (since it defines the space of possible movements). Instead we could use a bottom up approach, letting clusters in the data speak for themselves. For this purpose topic modelling does actually seem like a suitable candidate. Using 217 plausibly accepted topic models (with at least 5 plausibly accepted topics), we investigate what different conclusions a researcher might draw about which party is acting in accordance with their followers.

For each model, we fit a simple linear model between the fraction of posts containing each topic and the relative response of the topic, for each party. This is our measure of correspondence between the politicians' and followers' topical priorities.

According to all the simulated research projects what is the distribution of conclusion?



[Figure 6. Distribution of Research Conclusions based on 217 plausibly accepted topic models.]

From Figure 6 you can conclude that: of all the plausible models, you can pick any conclusion about each party as the distribution ranges from highly coherent to very discordant for almost any party. This is obviously deeply problematic. However if one were to look at the overall distribution of all possible models one might actually find a more credible answer. We note here that the only party which is always in concordance with its followers priorities is actually the danish right wing populist party (Danish People's Party). As we discuss below however, further research is needed to take this as a final conclusion.

4. Discussion and Conclusion

It is important to note that we do not question the use of topic modelling generally. As a method of discovery, clustering methods and in particularly topic modelling are very powerful tools (e.g. Grimmer and King 2011). Furthermore topic modelling techniques in a supervised setting where performance measures are involved can be very useful (e.g. Mcauliffe & Blei 2008).We are by no means the first to show the problematic properties of topic modelling. Various model issues has be illustrated by proponents of topic modeling, but the necessary conclusions for using topic models as a measuring device, as opposed to exploration device, has not been taken.

The implication of our results point to 2 research strategies. The first is to only use topic modeling as one, of potentially many, ways of exploring categories of interest in a large corpus and then make coding scheme out of these categories and follow supervised machine learning procedures to scale and measure features in text. Our results show the current practice of model selection can introduce large biases to one's result, supporting Lowe and Benoit (2013) argument that one needs to explicitly evaluate unsupervised models on a manually coded test set to interrogate the possible biases of the model. After creating a manually annotated dataset, it is natural to switch to supervised learning algorithms that explicitly optimize performance as suggested by Hillard et. al (2008). We therefore return to Hillard et. al's (2008) original conclusion that only a supervised setup can ensure reliability in topic classification (Hillard et. al 2008:33). The supervised learning setup furthermore enables bias detection and more importantly bias correction (see Hopkins and King 2010; Edwards and Storky 2016). Finally it gives the researcher the freedom to define the precise object to study, and directly optimize models for that purpose.

The above research strategy is based on the assumption that the text features of interest are possible to manually code (and scale through machine learning) in a way that does not introduce systematic biases. An expert researcher is surely better than any given topic model at correctly interpreting the topical content and level of abstraction of a single document. But it is likely that with the task of deciding among many possible sets of categories amongst a large set of documents any given categorization schema chosen by the researcher can introduce biased results (for a similar argument see Grimmer and Stewart 2013). Each choice can have both non trivial and non obvious effects on measurements performance and consequently on the final results. In the above case of quantifying opportunism by the priority given to popular topics, we show that each categorization scheme proposed by a given topic model can lead to very different results. A supervised approach would have to assume that it is possible to construct a stable set of categories in which both the author and the reader of the content operate (B), when in fact the actors sense of popular content works along a variety of dimensions hardly possible to capture with a coding schema that is in any way practical.

In effect this critique leads to a possible second research strategy. The necessity of using unsupervised methods, however with at least one important extension: one should not use *one* model alone, but instead investigate the distribution of possible conclusions different models could create. This necessarily means discarding semantics of the topics (footnote: Semantics could be kept by curating dictionaries similar to this project) and exhaustively searching the space of possible models, ensuring that the bias of one is cancelled by others, taking only the aggregated results as valid measurement. The assumption being that the dimensions along which actors actually operate would be captured by some of the models and the rest would be more or less randomly distributed noise. Furthermore great efforts should be used to investigate the properties of the unsupervised framework to ensure that potential results are not driven by artificial properties of the model used (see for instance the important work by Dubossarsky et. al (2017) on frequency effects in the estimation of semantic change).
This is of course a very radical step in that it treats language in a purely statistical manner where we lose the potential, set forth by supervised machine learning approaches and normal topic models alike, of combining qualitative and quantitative insights in the study of political communication.

5. References

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

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems* (pp. 4349–4357).

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems* (pp. 288–296).

Dubossarsky, H., Grossman, E., & Weinshall, D. (2017). Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1147–1156). Copenhagen: Association for Computational Linguistics.

Edwards, H., & Storkey, A. (2015). Censoring Representations with an Adversary. *ArXiv:1511.05897 [Cs, Stat]*.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., … Lempitsky, V. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, *17*(59), 1–35.

Grimmer, J. (2013). Appropriators not position takers: The distorting effects of electoral incentives on congressional representation. *American Journal of Political Science*, *57*(3), 624–642.

Hillard, D., Purpura, S., & Wilkerson, J. (2008). Computer-assisted topic classification for mixed-methods social science research. *Journal of Information Technology & Politics*, *4*(4), 31–46.

Hopkins, D. J., & King, G. (2010). A Method of Automated Nonparametric Content Analysis for Social Science. *American Journal of Political Science*, *54*(1), 229–247. https://doi.org/10.1111/j.1540-5907.2009.00428.x

Jacobs, L. R., & Shapiro, R. Y. (2000). *Politicians don't pander: Political manipulation and the loss of democratic responsiveness*. University of Chicago Press.

Kim, I. S. (2017). Political cleavages within industry: firm-level lobbying for trade liberalization. *American Political Science Review*, *111*(1), 1–20.

Krippendorff, K. (2009). Testing the reliability of content analysis data. *The Content Analysis Reader*, 350–357.

Lancichinetti, A., Sirer, M. I., Wang, J. X., Acuna, D., Körding, K., & Amaral, L. A. N. (2015). High-Reproducibility and High-Accuracy Method for Automated Topic Classification. *Physical Review X*, *5*(1), 011007. https://doi.org/10.1103/PhysRevX.5.011007

Lowe, W., & Benoit, K. (2013). Validating estimates of latent traits from textual data using human judgment as a benchmark. *Political Analysis*, *21*(3), 298–313.

Mcauliffe, J. D., & Blei, D. M. (2008). Supervised topic models. In *Advances in neural information processing systems* (pp. 121–128).

McFarland, D. A., & McFarland, H. R. (2015). Big Data and the danger of being precisely inaccurate. *Big Data & Society*, *2*(2), 2053951715602495. https://doi.org/10.1177/2053951715602495

Munksgaard, R., & Demant, J. (2016). Mixing politics and crime – The prevalence and decline of political discourse on the cryptomarket. *International Journal of Drug Policy*, *35*(Supplement C), 77–83. https://doi.org/10.1016/j.drugpo.2016.04.021

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, *54*(1), 209–228.

Roberts, M. E., Stewart, B. M., & Tingley, D. (2016). Navigating the local modes of big data. In R. M. Alvarez (Ed.), *Computational Social Science* (p. 51).

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

Schmidt, B. M. (2012). Words alone: Dismantling topic models in the humanities. *Journal of Digital Humanities*, *2*(1), 49–65.

Tang, J., Meng, Z., Nguyen, X., Mei, Q., & Zhang, M. (2014). Understanding the limiting factors of topic modeling via posterior contraction analysis. In *International Conference on Machine Learning* (pp. 190–198).

Wallach, H. M., Mimno, D. M., & McCallum, A. (2009). Rethinking LDA: Why priors matter. In *Advances in neural information processing systems* (pp. 1973–1981).