Amsterdam
University
Press

COMPUTATIONAL COMMUNICATION RESEARCH 5.1 (2023) 1–26
HTTPS://DOI.ORG/10.5117/CCR2023.1.13.BERN

Topic Model Validation Methods and their Impact on Model Selection and Evaluation

Jana Bernhard
Department of Communication Science, University of Vienna

Martin Teuffenbach
Faculty of Computer Science, University of Vienna

Hajo G. Boomgaarden
Department of Communication Science, University of Vienna

Abstract
Topic Modeling is currently one of the most widely employed unsupervised text-as-data techniques in the field of communication science. While researchers increasingly recognize the importance of validating topic models and given the prevalence of discussions of inadequate validation practices in the literature, there is limited understanding of the consequences of employing different validation strategies when evaluating topic models. This study applies two different methods for topic modeling to the same text corpus. It uses four validation strategies to assess how the choice of validation method affects the final model selection and evaluation. Our findings indicate that different approaches and methods lead to different model choices and evaluations, which is problematic. This might lead to unwanted results in case the choice of model has a decisive impact on findings and, consequently, on theory development and practical implications.

Keywords: Topic Model, Evaluation, Computational Communication Methods, Validation

Introduction

Topic Modeling (TM) has evolved to become one of the most used computational methods in communication science (Chen et al., 2023). While its versatility allows researchers to apply this methodology to diverse, oftentimes rather descriptive research questions, recent publications have called for computational methods, including topic modeling, to go further into the direction of testing and developing theories (see for example: Bonikowski & Nelson, 2022). Regardless of the goal of the research, a thorough evaluation or validation of the model chosen for the analysis is indispensable (Maier et al., 2018).

Validating computational text-analysis-methods, and especially topic models is not trivial (Grimmer et al., 2022), as the process of applying a topic model leaves a large number of researchers' degree of freedom (Denny & Spirling, 2018; Maier et al., 2018). There is no agreement on what kind of validation steps should be included (e.g. Ying et al., 2022) or how all of these steps are to be reported (Reiss et al., 2022). This lack of standardization (Hoyle et al., 2021) makes the scientific application and interpretation of TM difficult, to say the least.

While these difficulties in validation, or the lack of validation in general, have been discussed in recent literature (Baden et al., 2022), and different prescriptive pieces have been published (most notably Grimmer et al., 2022; Maier et al., 2018), we believe that the consequences as well as the extent of this lacking roadmap to topic modeling are not yet discussed enough in the community. We contribute to "the dialogue about the norms and expectations of using topic modeling and other computational text analysis methods properly at this relatively early stage of adopting the methodology" (Chen et al., 2023, p. 2), by assessing the impact of topic modeling evaluation methods on the subsequent model selection when conducting substantive research. Our aim is to investigate whether, if researcher A decides to employ a given validation method, they would run a different TM specification than researcher B, who relies on a different validation method. Furthermore, we consider whether such differences are different for different TM algorithms. Thus, we showcase how a researcher's choice of a particular validation method over another one can, instead of lending credibility to their results, severely influence and potentially bias the results. Our contribution calls for more careful reflection on how validation methods may lead researchers to consider different TM specifications and hence for the dependency of TM approaches on what validations researchers prefer to employ. In addition, we present a four-step recommendation plan in the later sections of this paper, offering guidance to researchers on planning their model selection effectively.

Theory and Related Work

While different topic modeling methods have different underlying assumptions, approaches, and needs, all these techniques employ machine learning to extract previously unknown patterns in large text corpora, which are then interpreted by researchers as topics (Boyd-Graber et al., 2017). Studies on topic modeling differentiate between four steps in the process of applying a topic model: first, the pre-processing of text data, second, choosing hyperparameters, third, model selection, and fourth, model validation (Chen et al., 2023; Grimmer et al., 2022; Maier et al., 2018). Steps three and four are somewhat intertwined in praxis, as the selection of one specific model over other alternative models is often based on the same validation methods that are used to validate the final model. Thus the evaluation of multiple, possible models is – or at least can be – done in the same way as the validation of the final model used to address substantive research questions. All preprocessing and hyperparameter choices as well as model selection introduce on the one hand complexity and researcher degrees of freedom, and on the other hand potentially have an impact on the results (for pre-processing and hyperparameter setting see Denny and Spirling, (2018), Maier et al., (2018) and Tolochko et al., (2022) and for model selection see Grimmer et al., 2022).

We argue that, given the multitude of choices and associated researcher degrees of freedom, it is vital in topic modeling to rely on different validation approaches to come to an informed model selection. In such a scenario we would argue that actual validation work is done in step three, which then would yield a choice of the most valid model to be selected, whereas step four then merely evaluates the overall quality of the validation outcomes against an ideal case. Hence, this study primarily focuses on step three - we assess how different validation approaches may or may not lead to different conclusions which model to select eventually. Thus, the decision on selecting which topic model is used for possible substantive analysis is often based on assessing which model looks useful ("face validity") or which models get better scores at various statistical measures ("statistical validity"). Yet we have little systematic knowledge as to whether and how different validation approaches would converge towards the same model selection.

While these scholars give us some indicators of what to focus on when discussing the validity of the results of applying TM, the validation of the models themselves is inherently challenging due to the characteristics and properties of the method as well as its usage. Methodologically speaking, topic models are applied to find useful text classifications based on the topicality or themes of each document. However, what is useful depends on the research problem in question and is strongly dependent on what the model is used for. There are a number of different topic modeling methods, which propose different functions to cluster text. These objective functions are formulated to identify an optimal partitioning, which is determined by a predefined similarity metric, such as the cosine similarity between sentence embeddings. Whether or not this is useful is for the researchers to decide, thus separating the mathematical, formalized "optimal" model from a "model that can answer my research question". This discrepancy between what is mathematically optimal and what is optimal for research introduces additional complexity (and degrees of freedom) to the social scientist since "model fit" essentially depends on an ad-hoc decision and should be thoroughly investigated and justified, which further complicates validation efforts. While these ex-ante decisions are important (Chen et al., 2023; Gentzkow et al., 2019), other researchers have emphasized the importance of post-hoc tests to ensure validity. The application of topic models to a diverse range of text corpora and research questions requires an individual approach to validation, given the specificity of each case (Barberá et al., 2021).

Ballester and Penner (2022, p. 2) argue that "the three properties that functional topic models should have [are]: robustness, descriptive power and reflection of reality." Validation relates to the latter property. Validity in social science refers to the accuracy and truthfulness of the results and conclusions of a study. It's the extent to which a study measures what it claims to measure and that the results are a true reflection of the reality being studied. Social scientists differentiate between types of validity that can be taken into account. In general, Scharrer and Ramasubramanian (2021, p. 62f) explain *face validity* ("the measure maps on to common understanding of the concept"), *criterion-related validity* ("the measurement relates in a logical manner with another variable outside of your study") and *content validity* ("degree to which the full range of meanings of the concept are being reflected in the measurement"). On manual content analysis Krippendorff (2013, p. 319) differentiates three main categories *face validity* ("being obviously true, sensible, plausible"), *social validity* ("addressing important social issues"), and *empirical validity* ("The degree to which available evidence and established theory support intermediate stages of a research process and its stages”). Regarding the latter, he concludes that this evidence can be based on *content, internal structure* and *relations to other variables.* He further distinguishes these three subcategories to include *sampling, semantic, structural, functional, correlative* and *predictive validity.* This detailed description of different types of validity can function as a guide when thinking about how we validate automated content analysis, such as TM.

The validation of topic models is critical in scenarios in which ground truth labels are not available for the text corpus being analyzed (as arguably true for most TM application scenarios). DiMaggio and colleagues (2013, p. 586), partly relying on Grimmer and colleagues (2011) emphasize that validation should focus on three different points of view:

1. statistical validation: if the model results are consistent with the assumptions of the model
2. semantic or internal validation: whether the model meaningfully discriminates between different senses of the same or similar terms
3. predictive or external validation: attention to particular topics responds in predictable ways to news events

While the first, statistical validation, has a special place due to the mathematical background of TM, the second validation step is closely related to what has been described in general as *criterion-related validity* as well as *content validity* and Krippendorff subsumed in the *internal structure.* The third then connects well to Krippendorffs *relations to other variables.*

Probably the clearest roadmap for TM in communication science was put forward by Maier and colleagues (2018). They describe the following steps in evaluating TM to ensure reliability and validity: 1. Coherence Metrics to identify useful hyperparameter settings and 2. Qualitative judgement of different, but well-performing models (as found in step 1) by experts based on the top words. This leads to the selection of one topic model, which is validated in more depth, by summarizing different statistical values, excluding topics that are not interpretable, reading documents that are related to each topic, and employing hierarchical clustering on the top words, to identify mergeable topics.

The selected topic model or models are then validated in more depth, by summarizing different statistical values, excluding topics that are not interpretable, reading documents that are related to each topic, and employing hierarchical clustering on the top words, to identify mergeable topics. Thus, validation methods are applied in two steps: Model Selection and Model

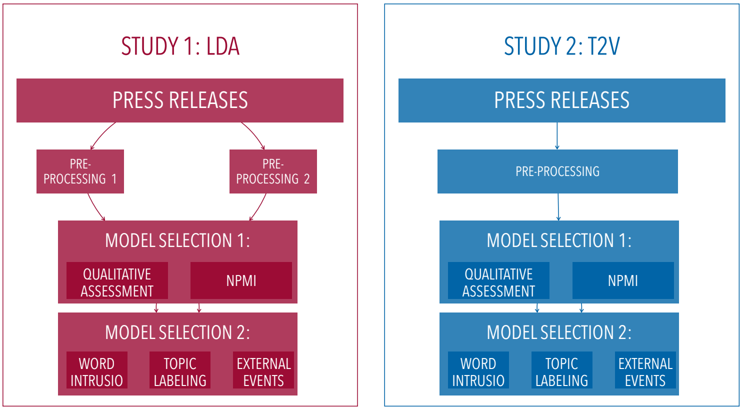
Validation. As it is quite necessary for the research process to decide on one (or at least only very few) topic models to base the substantive results on, this paper aims to showcase that this is often not as straightforward as some scholars have described before. It might be a trivial statement to suggest that each researcher has some influence on the results, however, they should at least aim for as little impact as possible or relatedly for objectivity and transparency throughout all decisions in the research process (Scharrer & Ramasubramanian, 2021). Our goal is to explain how the choice of validation methods can impact the model selection and thus the final results of a research study, that employs TM.

Research Design

We propose an empirical setup to assess how the choice of validation method impacts the model evaluation and selection, and thus, potentially the results of topic models for substantive research questions. Our design, as illustrated in Figure 1, relies on two studies with distinct TM approaches, the results of which in combination would yield insights into our research interest. We apply two topic modeling algorithms, with different pre-processing steps, to one text corpus and then apply different evaluation methods and assess how they would affect model selection. In our first study, we test the impact of validation methods on models generated with LDA, as LDAs are still among the most used methods in the social sciences (Chen et al., 2023; Maier et al., 2018). Second, we use Top2Vec (T2V) (Angelov, 2020), which is an embedding-based model built on a pre-trained neural language model.

As mentioned above, there are many different validation methods and no standards regarding their implementation, not to speak about their potential combination. To choose appropriate approaches for our study, in the first step we looked at prescriptive resources: Maier and colleagues (2018) as well as Ying and colleagues (2022), emphasize the technique of labeling topics by reading the top words or related documents and by relying on automated metrics (topic coherence, mutual information, or hierarchical clustering) regarding internal validity. On external validity, they suggest expert evaluation, manual codings as well as considering external events. Additionally, Ying et al. (2022) refined the intruder method, which was first put forward by Chang and colleagues (2009) to measure semantic coherence in an attempt to create one off-the-shelf validation method that can be used for any topic modeling research question. Grimmer and colleagues (2022) regarding validation without gold-standard data, highlight practices such as

Figure 1: Visualization of the Research Design



intrusion tasks (Chang et al., 2009; Ying et al., 2022), labeling top words for semantic validity, and assessing the correspondence of the model to external events (hypothesis validity). A recent systematic literature review found that the most frequently used validation methods that build on human judgment are: labeling topics based on top words, and human interpretation of topics based on top words and documents, comparing to manual analysis, including theoretical considerations and relating to external events (Bernhard et al., 2022).

Evaluation Methods

In line with the research presented above (DiMaggio et al., 2013; Maier et al., 2018) we chose different points of view of validity: statistical (NPMI) as well as face validity (qualitative judgment of top words and reading documents) to choose useful hyperparameters as well as internal (word intrusion and topic modeling) and external validation (relation to external events) to further evaluate our models. As the steps on internal and external validation are quite labor intensive, we used the information from statistical and face validity to choose two models which were evaluated in more detail. Of course, these different validation approaches provide different perspectives on model validity, yet as a baseline, we would argue that ideally, TM should show high validity on all accounts. If TM is used for theory building and refinement, it appears important to draw on models that do not compromise on certain types of validity but rather converge on high validity in different areas and through different approaches.

Mutual Information

As a first step, adhering to the advice of Maier and colleagues (2018), we took a statistical indicator to determine which hyperparameter settings would lead to the most “useful” topic models. We initially computed various coherence metrics (Röder et al., 2015) for all models to systematically evaluate the hyperparameter settings. Adhering to the recommendation of Hoyle and colleagues (2021) as well as Grimmer and colleagues (2022), we ultimately relied on the Normalized Pointwise Mutual Information (NPMI) metric as well as human judgment (qualitatively checking top words and documents) for deciding which are useful models. The NPMI score is high if the top N words have a high joint co-occurrence probability, i.e. the words often co-occur in the corpus. This is an intuition similar to what most statistical topic models (e.g. LDA) make use of, where topics are generated based on word co-occurrence patterns. Neural-topic models on the other hand rely on text representations generated by neural-network-based models (e.g. transformers). These embedding models are optimized to find semantically meaningful representations of texts. Therefore, we expect statistical models to perform better when compared to neural models in terms of automated topic coherence metrics. Thus, we do not compare the NPMI scores between topic modeling methods but only within one method.

Word Intrusion

We first implement a word intrusion task, as put forward by Chang and colleagues (2009). This evaluation method is extremely versatile and straightforward. The method uses the top words that are calculated to be indicative of each topic and postulates that a human should be able to spot a randomly included word, that is not part of these top words. Thus, it is a test of internal validity (as defined by DiMaggio et al., 2013, or a matter of face or semantic validity as defined by Krippendorff, 2013). We took nine top words from each topic and randomly included one of the top ten words from another topic as the intruder. We instructed three student assistants who were not familiar with the details of the research project but were aware that they were evaluating press releases, to mark the intruder word. Each of them completed this task in two days. We then calculate the percentage of correctly identified intruders, thus, this measure can go from 0% to 100%, allowing us to compare models with a different number of topics.

In the topic modeling process, the LDA model assigns a topic probability to each word in the corpus. For the generation of topic top words, we selected the top n-words for each topic, representing the words with the highest topic probability. Due to the LDA model's statistical nature, these top words are in general words that often co-occur in the text corpus. T2V aligns words and documents within a shared latent vector space. The algorithm identifies document clusters within this space, defining them as topics. For selecting top words, we extracted the top n-words from this latent space with the highest similarity to documents within the topics. In other words, we selected words that the model represents as semantically similar to the document clusters. These are words that are used in the same context, which is in general not equivalent to the LDAs word co-occurrence approach. We, thus, expect the T2Vs approach to produce superior results for this evaluation, as word co-occurrence often finds words that are not related to the topics.

Topic Labeling

To include further human oversight (Grimmer et al., 2022), we read 10 documents per topic to assess whether they can be meaningfully interpreted (as suggested by Maier et al., 2018). Meaningful in this case is defined as the documents relate to one, distinct issue of Austrian Politics. This was done by one of the authors with an education in political and communication science so that topical expertise is given. To do so, the topic of each document was paraphrased with one or two words before trying to find one label for the topic encompassing all of the ten documents. To compare the number of meaningful topics, we additional distinguished between three cateories: no label found; label found that would relate to all documents, labels found if the texts in the topic included more than one topic. For example, a topic, with documents on health policy and voluntary work. Thus, in the classification of DiMaggio and colleagues (2013) this task points us to internal validity. For Krippendorff (2013) this task would be in the area of face and content validity. Yet, LDA allows for documents to have multiple topics, while T2V classifies each document into one topic. Thus, to get to the documents of each topic, we only chose documents for which the topic made up more than 50% of each press release. However, we do not expect this difference to substantively impact the evaluation method, thus this kind of evaluation can be used within and between the different topic modeling methods.

External Events

The last method, comparing to external trends (Maier et al., 2018; Ying et al., 2022) aims at comparing the findings of a topic model (e.g. the number of topics in a given timeframe) to some kind of external baseline (e.g. official statistics or the occurrence of specific events). Thus, this evaluation method would be classified as external (DiMaggio et al., 2013) or correlative (Krippendorff, 2013). Often this method is only partly implemented, as it is only possible for topics that can be reasonably expected to be related to quantitatively measurable external events. This can either be done for topics that are of specific interest for the analysis or as many topics as feasible. As an *example*, we show how the topic of unemployment develops over time and compare this development to official unemployment statistics (WKO, 2022). We then calculate a correlation index (Person's r) to assess how close the two developments are. We argue that this is a reasonable comparison, as it can be expected that parties talk more about unemployment when it is high, as this also leads to unemployment being discussed in the news. However, as it could also be that unemployment is discussed more when it is exceptionally low, we do not expect a strong correlation. We thus argue that the differences in the correlation should be focused on, not the strength of the correlation itself. The decision of which topics to compare to which statistic has to be taken, in part, after the topic model has been evaluated as to which topics it includes. For this study, we wanted to be able to compare all four models based on the same topic-statistic correlation. We chose unemployment, over the other connecting topics Health (too ambiguous in the LDA models), Pension (lack of external event), and Feminism (lack of suitable external statistics). Regarding the different TM methods we do not have a strong reason to expect this validation approach to work better or worse for one or the other method.

These four validation methods correspond to different kinds of validity, as described above. While mutual information relates to statistical validation, the intention behind the metric can be seen as relating to internal validity as well. This connects it to the task of word intrusion, and topic labeling, which in the classification of DiMaggio (2013) all relate to internal validity, while the comparison to external events, would be external validity. If we take into account the more detailed description of Krippendorff, we can see some differences between the three internal validation methods, as they could relate more to semantic (word intrusion) or content (topic labeling), however, of course, both still go into the direction of internal validity. In sum, we would expect the results of approaches one to three (mutual information, word intrusion, and topic labeling) to converge more and clearly show which model a researcher should prefer since they arguably relate to the same types of validity. Approach four (external events) may, somewhat in contrast, diverge more from the pattern, as it aims at measuring a different kind of validity. Yet ideally models are valid on all accounts.

Case Description

As a case for this setup we analyze which topics parties in Austria have talked about in the past 15 years. To do so we aim to find the most useful text classification put forward by the method of topic modeling. We define "usefulness" as the number of topics that can a) be meaningfully interpreted by humans and b) are theoretically sensible for the context of Austrian politics between 2004 and 2020. For this analysis, we use 218.471 press releases that have been sent out by the five parties currently in the Austrian Parliament (SPOE, OEVP, FPOE, GRUE, NEOS). [1]

Topic Modeling Methods

Study 1: Latent Dirichlet Allocation(LDA)

The Latent Dirichlet Allocation (LDA) model (Blei et al., 2003) is one of the most widely used topic modeling methods (Bernhard et al., 2022). It is a statistical model that simultaneously estimates a document-topic and a topic-word distribution. With those two functions, one can estimate the membership probability for each document to the topics, as well as the most descriptive words for each topic. The classical LDA model requires the number of topics k to be specified beforehand. The resulting distributions can be adjusted with two parameters, usually denoted as α and β. The parameter α is the prior concentration parameter representing document-topic density. Hence this parameter controls how many topics are assumed to be in a document. High α results in more topics per document. β represents the topic-word density prior, which influences how many words are ascribed to each topic. As with most statistical models, LDA requires pre-processing of the data. Pre-processing has a strong influence on the results (Denny & Spirling, 2018). Therefore we decided to follow best practice conventions (Maier et al., 2018) and performed the following steps:

[1] Replication material for this study can be found on https://doi.org/10.17605/OSF.IO/PYFDT.

1. Removal of punctuation and digits, lowercase all characters
2. Stemming and tokenization
3. Remove the most frequent and least frequent words.

We performed the pre-processing with two settings, once with removing all words that appear in $\geq 95\%$ or $\leq 0.5\%$ of the documents and once with $90\%/1\%$. We evaluated several values of k (4 to 50) and α (0.1 to 1). β was set to $\frac{1}{k}$ (*symmetric prior*, for more information on the parametrization of the LDA model, see Maier et al. (2018)). For every parameter-setting, we performed three runs and averaged the topic coherence score.

For all settings and parametrizations, we validated the models with the Normalized pointwise mutual information (NPMI). This score returns a value between 0 and 1, the higher the better. The upper Figure 2 depicts the achieved results for our two settings. As expected, the LDA model produced topics with nearly perfect NPMI scores (over 0.96 for all parameter settings for all $k \geq 10$). Additionally, we found that this coherence score (1) improved with the number of topics and (2) hardly varied for different parametrizations (less than 0.025 for $k \geq 10$).

Due to the high time consumption of manual validation, we decided to pick only two models for further analysis, one for each pre-processing setting. We chose models with different k so that we get an overview of how k impacts the results (6, 14, 30, 40, 50). We manually inspected ten top words as well as five documents related to each topic. We then chose two models which have the most interpretable topics for further human-based validation.

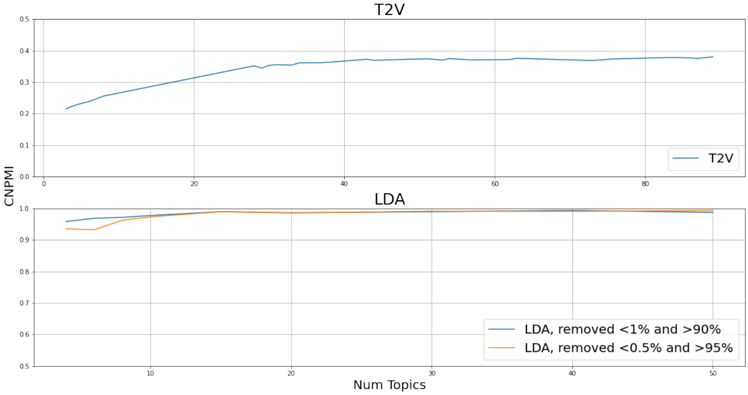
Study 2: Top2Vec (T2V)

The T2V (Angelov, 2020) model is a neural-network-based topic model. In contrast to statistical models, neural language models utilize context-aware embeddings instead of word frequencies. Therefore, these models do not require extensive pre-processing of the input texts. To find topics, T2V embeds the input corpus with a pre-trained embedding model and clusters them. The resulting clusters are interpreted as topics. Next, the vocabulary of the corpus is embedded in the same vector space. For each cluster of documents (i.e. topic), the closest word embeddings based on Euclidean distance in the embedded space are computed and used as topic representatives.

T2V utilizes a density-based clustering algorithm, namely HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise)

(Campello et al., 2013), combined with an additional dimensionality reduction algorithm. The results of T2V can be adjusted with the initialization of the HDBSCAN algorithm. Note that, unlike LDA, this algorithm does not use a pre-defined number of clusters (i.e. topics) k. The framework supports a variety of embedding models. For our experiments, we decided to use an SBERT (Reimers & Gurevych, 2019) model that is pretrained on a multilingual-text corpus (*distiluse-base-multilingual-cased*), which is a state-of-the-art transformer model. As T2V requires no further pre-processing, to achieve various numbers of output topics k we adjusted the min-cluster-size parameter of the HDBSCAN algorithm (see Campello et al., 2013). After several runs, we again evaluated the NPMI metric (see Figure 2). Similar to the LDA model, the metric increased with the number of topics. However, the results were significantly worse than for the previous model (between 0.22 and 0.38 compared to 0.91 to 0.99). Again, we picked two models for further validation. To do so, we again manually assessed the quality of models with different k (4, 19, 30, 54, 63).

Figure 2: NPMI coherence scores for LDA and T2V



For a summary of the models parameterization and the preprocessing of textual data please refer to Tables 3 and 4 in the Appendix.

Results: Study 1: LDA

Word Intrusion

Three student assistants completed the word intrusion task for both LDA models (see Table 1 for detailed results). We found that the LDA40 TM performed on average a bit better (one-quarter of intruder words correctly identified), but the detailed results of the student assistants differ from each other, which suggests that this estimate is unstable. The LDA50 TM performed worse (one-fifth of intruder words were correctly identified). However, both scores are not good overall, which would suggest that both LDA models are not sufficiently well suited to be used for substantive research.

Table 1: Results of Word Intrusion Task for LDA Models

	TA1	TA2	TA3	mean
LDA40	17.5% (7/40)	32.5% (13/40)	22.5% (9/40)	24.2%
LDA50	18.0% (9/50)	20.0% (10/50)	16.0% (8/50)	18.0%

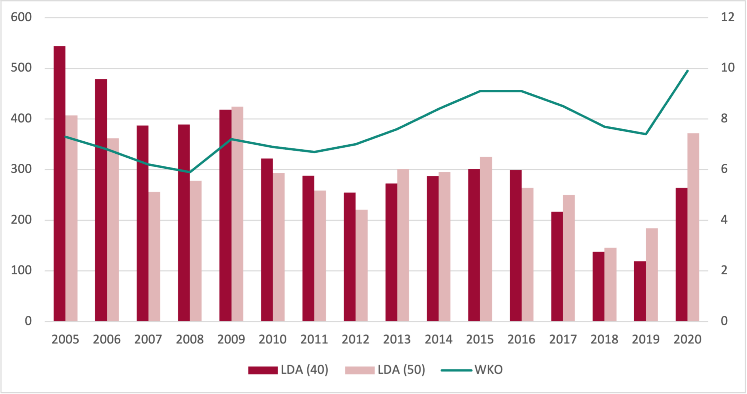
Reading Documents

We found that in the first LDA model (40 topics), 13 topics (32.5%) revolved around a meaningfully interpretable topic. Additionally, 16 topics (40%) could be interpreted, even though they included two topics that were connected but not the same. Only 11 topics (27.5%) could not be interpreted at all. Similarly, the second LDA model (50 topics) included 15 meaningful topics (30%), however, only 13 topics confounded two connected topics (26%). This led to 22 topics (44%) that could not be interpreted. This validation method would suggest that the first LDA model is more suitable for substantive analysis, yet still with about a third of the models representing nonsense.

External Events

Figure 3 shows the development of the *Unemployment*-topic for both models, as well as the official monthly unemployment statistics. We see that although there is some parallel movement in the development of unemployment salience in press releases and the statistics, they do not correlate strongly (LDA40: $r(14) = -0.35, p = .181$ and LDA50: $r(14) = 0.13, p = .63$). More worryingly, however, is that one of these correlations is positive, while the other is negative, however, neither of the correlations is significant. This would suggest that neither of the models adequately captures the topic of *Unemployment*.

Figure 3: Number of press releases on the topic of Unemployment in both LDA models versus the official unemployment statistic for Austria



Results: Study 2: T2V

Word Intrusion

Three student assistants completed the word intrusion task for both T2V models (see Table 2 for detailed results). For the top words of the T2V30 TM, 84% of intruders were found consistently by the student assistants, while 75% of intruders were found for the T2V30 TM. Both scores are indicative of models that have coherently clustered documents into topics, but it is clear that, based on these results, the TM with 30 topics would be preferred for further substantive research.

Table 2: Results of Word Intrusion Task for T2V Models

	TA1	TA2	TA3	mean
T2V30	83.3% (25/30)	86.6% (26/30)	83.3% (25/30)	84.4%
T2V63	74.6% (47/63)	76.2% (48/63)	74.6% (47/50)	75.1%

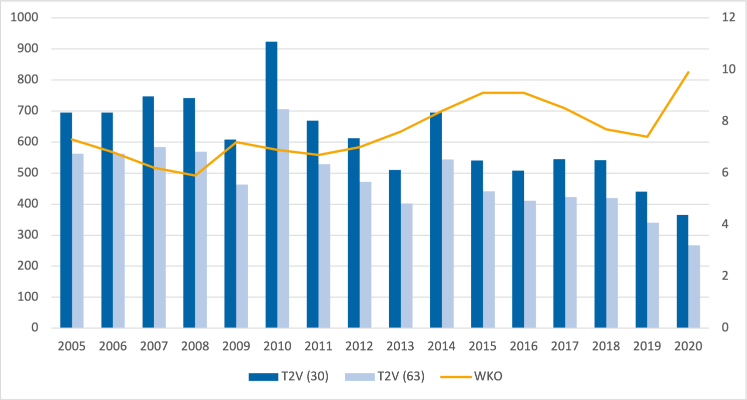
Reading Documents

Regarding the first T2V Model (30 topics) we found 26 useful and meaningful topics (86.7%) and only two which confounded two topics, as well as two that could not be interpreted (6.7% each). The second T2V Model (63 topics) included 57 meaningful and useful topics (90.5%) and only three topics that confounded two topics and another three that could not be interpreted (4.8% each). Thus, both models seem to cluster the press releases in the most meaningful ways, which would allow for further substantive research.

External Events

Again, we want to see how the number of press releases in the *Unemployment*-Topics relates to the official statistics. Figure 4 shows some parallel development, and this time we see a stronger correlation (T2V30: $r(14) = 0.68, p = 0.04$ and T2V63: $r(14) = 0.67, p = .005$). Thus, both models seem to adequately capture this topic.

Figure 4: Number of press releases on the topic of Unemployment in both T2V models versus the official unemployment statistic for Austria



Final Thoughts on Model Evaluation

These validation steps reveal different things about the different topic models. First, on LDA: Our results show that both models score similarly on NPMI, the word intrusion task, and the comparison to external events. However, had we relied on statistical validity only and not taken into account any human validation approach, we would have been confident with our models and would have used them for substantive research. Had we relied on the word intrusion task or the comparison to the Unemployment Statistics we might have concluded that both LDA models are insufficient for further substantive research. The two LDA models seemed to be successful when looking at the topic labeling method, where the LDA40 model performed slightly better. Thus, depending on which validation method we would put our trust in, we would have come to different conclusions about our models. This has a clear impact on possible substantial results, as the two models give us vastly different topics (see Table 5 in the appendix). Thus, even though the word intrusion task and the topic labeling both aim at face or internal validation they point us in different directions.

Second, on T2V: Our results show that the models get similar results for NPMI, reading documents, and external events. Both models score very high in the validation task that is based on reading documents, which will lead us to believe that we have great models that can be used for substantive research if we were to rely on this approach only. Our models do show differences in their performance of word intrusions and how they compare to external events. This suggests that as researchers we would have to make a trade-off between internal and external validity. Both models seem to work similarly well when compared to unemployment statistics, suggesting that they might be used to some extent, for substantive research on this topic. T2V30 performs better on the word intrusion task, which could sway researchers to choose this model when only considering this validation method. Thus, again, depending on which validation method we choose, we would either assume that both models are equally good, or that the T2V30 is slightly better. For this method, the impact on results is smaller, as the topics are more stable (see Table 5 in the Appendix and Rodriguez and Spirling, 2022).

For both TM methods, we thus see a divergence in terms of how different validation approaches may lead to different conclusions about the substantial usefulness of a particular model. If, as argued above, an ideal scenario would show the strong validity of a model in all accounts, this is something that we do not clearly see in any of the scenarios above. So, in the absence of a standardized approach to topic model validation (which validation approaches to apply, how many of them) our results demonstrate a situation in which different validations, were they used exclusively, would point researchers to use different models for substantive research. This problem, however, appears to be less strong for T2V, since here we see a stronger convergence of different approaches. Thus it seems that in our scenario, the T2V method showed more robust models than the still widely used LDA.

Discussion

Validity in the social sciences refers to the accuracy and truthfulness of the results and conclusions of a study and is often defined as the extent to which a study measures what it claims to measure. Especially when talking about computational methods, which utilize algorithms that work as a black box or are applied by researchers without a background in computational science, validation is often performed post-hoc on the models' results. The consequence is that each validation strategy depends on the research question, text corpus, and (maybe) the theory behind the analysis. This setup is less than ideal and the reliance on models to interpret results in light of theory has been named as one of the causes of the replication crisis in psychology (see for example: Wiggins & Christopherson, 2019).

In more straightforward statistical models (like regression models), certain criteria evaluate how well the model fits the data and if these criteria are met — conditional on theoretical expectations — one can be sure that the output is correct. Topic models, however, do not have a method that defines the "correctness" of the model. Regardless of the post-hoc model validation, there are several "useful" models (as demonstrated above). But this means that none of these models can be shown to adhere to a single theory. Ultimately the choice of the model would determine which "theory" we are testing (without our explicit knowledge). Thus, every judgment is dependent on a, more or less, arbitrary model selection, and is therefore post-hoc and not suited for theory building. In a quantitative setting, even if we build on gold-standard data (see e.g., Song et al., 2020) and have a good model fit, researchers have to rely on existing theories for interpreting the results. As argued earlier, in the absence of clear guidelines, topic modeling is not yet a standardized methodology (but first steps are provided by Denny and Sperling (2018) and Maier et al. (2018)).

Topic Model Selection is a crucial step in the topic modeling process, which is often brushed over or presented as being very straightforward (taking the model with the best scores regarding different statistical values). However, as evident in this study, it is not that easy. Above we have shown that the application of different validation methods exclusively would lead researchers to put their faith in different models that at points show vastly different substantial results. Our study also showed that this is more problematic for LDA models, as compared to T2V models. It appears that T2V would show a somewhat better convergence of different validation approaches and therefore might be the preferred modeling method to yield TMs with higher overall validity. Where do we go from here? Planning the validation of the topic model should start before the application of the topic model. Here are some steps to take when using topic models to research communication scientific phenomena.

Where do we go from here? Planning the validation of the topic model should start before the application of the topic model. Here are some steps to take when using topic models to research communication scientific phenomena.

1. Before starting the application of the method, in the first step researchers should consider several questions that may inform validation steps: What would a good model look like? Although this might seem like a somewhat obvious suggestion at first glance, it is not trivial at all to formulate a short description of a) what topics a good model would include or not include, b) how many topics you would expect at a minimum and maximum, or c) which patterns would you expect to find. These decisions have to come from knowledge of the text corpus, text context, and theoretical considerations. For deductive research, this process is close to hypothesis generation, however, not about the model outcome, but the model itself. This description should be saved, so that it can be used for the upcoming steps.
2. A second step researchers could ask themselves: If these data were created by manual content analysis, that you did not conduct yourself (for example in secondary data analysis), how would you go about checking the quality of the data and the validity of the topic classification? Content Analysis has been applied in communication science for a long time, and as a scientific community, we have found ways of thinking about the validity (Krippendorff, 2013). We can and should use this knowledge to build validation strategies for automated analysis, and topic models. We thus suggest using the validation classifications we already have from manual content analysis (Krippendorff, 2013; Scharrer & Ramasubramanian, 2021), or from prescriptive publications such as (DiMaggio et al., 2013) and looking at different kinds of validation and how they relate to the description made at the first step. Which validation angle is helpful to gain insights into the description we formulated? We suggest deliberately taking into account as many different kinds of validation as possible so that it can be assessed whether a model is good on more than one account. The goal of this second step is to come up with a list of kinds of validations that should be applied to all potential models.
3. Third, researchers have to decide which evaluation method they want

to apply. For this, researchers should use the overview of which validation methods correspond with which validation angle for step two. This gives them a rich list of possible validation steps to take. The researchers can then shorten this list by assessing which methods are feasible (in terms of e.g. time or funding), have been applied by researchers with comparable projects (e.g. Maier et al., 2018) or proposed for a specific topic modeling method (e.g. Zhao et al., 2021). We want to highlight, however, as many have before us, the importance of including human-in-the-loop validation methods.

4. Researchers then have to decide on several validation methods, which they want to apply to their model at the a) model selection and b) model validation stage, to avoid arguing circularly. Additionally, researchers should be transparent about why they chose specific methods and disregarded others, and clear about which benchmarks they set for which validation method so that a model can either pass or fail a specific step in the validation process. At this stage, researchers also need to take into account the possibility, of different methods not converging, and pointing at different models, as could be seen in our example. In this case, researchers need to decide which validation method to prioritize.

When following all the steps in the list, researchers end up with a description of what a good model would look like, which kinds of validation correspond to this description, and which validation methods can be applied to assess these kinds of validation. The researcher also has a set of decisions that were taken for or against a method, as well as benchmarks for them. This can increase the transparency of the decisions taken by the researcher. We recognize that this process is extremely resource-intensive. However, it is important to recognize the impact validation strategies have on model selection when discussing findings that were obtained through topic modeling. As discussed above, we believe it is vital to come to better-informed model selections through the application of different validation approaches in step three, selecting the best-performing alternative. The same validations can then be used to judge, in step four, to what degree the best-performing model can actually be considered a valid representation of the text.

Our study is not without limitations. The first is, that we cannot solve the problem we describe, only give recommendations and demonstrate its implications. Second, we rely on only one text corpus in one language in our demonstration. We thus want to encourage further research in this area, including different corpora and languages. Third, we showcase the impact of three widely used validation methods, however, there are many more, which were not included (e.g. Bernhard et al., 2022). Fourth, we also had to rely on a pre-selection of topic models, which is based on statistical and face validity, to reduce the number of models that we assessed in-depth.

We see this as yet another indication that we need to shift our attention toward measurement validity (Baden et al., 2022) before we can talk about generating new theories with topic modeling. Indeed, recently, scholars have highlighted how topic models can be used in a qualitative research setting (Isoaho et al., 2021). This allows researchers to put the unsupervised and inductive nature of the method to use. The validation of computational methods, and all methods in general, is an important step in the research process. Continuing our efforts in researching and revising the process of validating is needed if we want to use computational methods to build communication scientific theory.

Conclusion

In conclusion, the validation of topic model selection in communication science research is a crucial step to ensure the accuracy and reliability of study results and any theoretical or practical recommendations derived from them. Computational methods, such as topic modeling, present unique challenges due to their algorithmic nature and reliance on post-hoc validation strategies. We showcased that the choice of validation method has an impact on the selection of the final topic model, which in turn impacts the results. Thus, we argue that topic models offer valuable insights and facilitate exploratory analyses, but their use for theory-building remains problematic. To add to the literature on the validation of computational methods (Baden et al., 2022; Chen et al., 2023), we have proposed an approach to coming up with a validation strategy for topic model selection, emphasizing the importance of formulating a clear description of an ideal model and aligning validation strategies with existing content analysis methodologies. By transparently documenting the decision-making process and benchmarks, researchers can enhance the credibility and replicability of their findings. Additionally, a shift towards measurement validity is essential before topic modeling can become a reliable tool for theory generation. As we continue to explore computational methods' potential, refining and standardizing validation processes will be paramount in advancing communication scientific theory.

References

Angelov, D. (2020, August). Top2Vec: Distributed Representations of Topics. https://doi.org/10.48550/arXiv.2008.09470

Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. G. (2022). Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. *Communication Methods and Measures*, *16*(1), 1–18. https://doi.org/https://doi.org/10.1080/19312458.2021.2015574

Ballester, O., & Penner, O. (2022). Robustness, replicability and scalability in topic modelling. *Journal of Informetrics*, *16*(1), 101224. https://doi.org/https://doi.org/10.1016/j.joi.2021.101224

Barberá, P., Boydstun, A. E., Linn, S., McMahon, R., & Nagler, J. (2021). Automated text classification of news articles: A practical guide. *Political Analysis*, *29*(1), 19–42. https://doi.org/https://doi.org/10.1017/pan.2020.8

Bernhard, J., Ashour, R., & Boomgaarden, H. G. (2022, June). Towards Validity Standards of Topic Models in Computational Social Science [Jahrestagung der Fachgruppe Methoden der DGPuK 2022].

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning*, *3*, 993–1022. https://doi.org/10.5555/944919.944937

Bonikowski, B., & Nelson, L. K. (2022). From Ends to Means: The Promise of Computational Text Analysis for Theoretically Driven Sociological Research. *Sociological Methods & Research*, *51*(4), 1469–1483. https://doi.org/10.1177/00491241221123088

Boyd-Graber, J., Hu, Y., & Mimno, D. (2017). Applications of Topic Models. *Foundations and Trends in Information Retrieval*, *11*(2-3), 143–296. https://doi.org/10.1561/1500000030

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 160–172). Springer Berlin Heidelberg.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., & Blei, D. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems*, *22*. Retrieved October 4, 2022, from https://papers.nips.cc/paper/2009/hash/f92586a25bb3145facd64ab20fd554ff-Abstract.html

Chen, Y., Peng, Z., Kim, S.-H., & Choi, C. W. (2023). What We Can Do and Cannot Do with Topic Modeling: A Systematic Review. *Communication Methods and Measures*, *0*(0), 1–20. https://doi.org/10.1080/19312458.2023.2167965

Denny, M. J., & Spirling, A. (2018). Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. *Political Analysis*, *26*(2), 168–189. https://doi.org/10.1017/pan.2017.44

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, *41*(6), 570–606. https://doi.org/10.1016/j.poetic.2013.08.004

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535–74. https://doi.org/10.1257/jel.20181020

Grimmer, J., & King, G. (2011). General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences*, *108*(7), 2643–2650.

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Hoyle, A., Goel, P., Hian-Cheong, A., Peskov, D., Boyd-Graber, J., & Resnik, P. (2021). Is Automated Topic Model Evaluation Broken? The Incoherence of Coherence. *Advances in Neural Information Processing Systems*, *34*, 2018–2033. https://doi.org/https://doi.org/10.48550/arXiv.2107.02173

Isoaho, K., Gritsenko, D., & Mäkelä, E. (2021). Topic Modeling and Text Analysis for Qualitative Policy Research. *Policy Studies Journal*, *49*(1), 300–324. https://doi.org/10.1111/psj.12343

Krippendorff, K. (2013). *Content analysis: An introduction to its methodology* (3. ed.). Sage. https://ubdata.univie.ac.at/AC08977231

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. *Communication Methods and Measures*, *12*(2-3), 93–118. https://doi.org/https://doi.org/10.1080/19312458.2018.1430754

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. https://doi.org/10.48550/arXiv.1908.10084

Reiss, M. V., Kobilke, L., & Stoll, A. (2022, June). Reporting Supervised Text Analysis for Communication Science [DGPuK Jahrestagung der FG Methoden].

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399–408. https://doi.org/10.1145/2684822.2685324

Rodriguez, P. L., & Spirling, A. (2022). Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research. *The Journal of Politics*, *84*(1), 101–115. https://doi.org/10.1086/715162

Scharrer, E., & Ramasubramanian, S. (2021). *Quantitative research methods in communication : The power of numbers for social justice*. Routledge,

Song, H., Tolochko, P., Eberl, J.-M., Eisele, O., Greussing, E., Heidenreich, T., Lind, F., Galyga, S., & Boomgaarden, H. G. (2020). In validations we trust? The impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis. *Political Communication*, *37*(4), 550–572. https://doi.org/https://doi.org/10.1080/10584609.2020.1723752

Tolochko, P., Balluff, P., Bernhard, J., Galyga, S., Lebernegg, N., & Boomgaarden, H. G. (2022). What's in a Name? The Effect of Named Entities on Topic Modelling Interpretability [[Presented at the annual ICA Conference 2022]].

Wiggins, B. J., & Christopherson, C. D. (2019). The replication crisis in psychology: An overview for theoretical and philosophical psychology. *Journal of Theoretical and Philosophical Psychology*, *39*(4), 202.

WKO. (2022, December). *WIRTSCHAFTSLAGE UND PROGNOSE Arbeitslosigkeit* (tech. rep.). https://wko.at/statistik/prognose/arbeitslose.pdf

Ying, L., Montgomery, J. M., & Stewart, B. M. (2022). Topics, Concepts, and Measurement: A Crowdsourced Procedure for Validating Topics as Measures. *Political Analysis*, *30*(4), 570–589. https://doi.org/10.1017/pan.2021.33

Zhao, H., Phung, D., Huynh, V., Jin, Y., Du, L., & Buntine, W. (2021). Topic modelling meets deep neural networks: A survey. *arXiv preprint arXiv:2103.00498*. https://doi.org/https://doi.org/10.48550/arXiv.2103.00498

Appendix

Table 3: Parametrization of models. To implement the LDA model, we utilized pythons gensim library, for T2V the GitHub provided by Angelov, 2020. For more information regarding the parameters please refer to the corresponding implementation

	min_count	**leaf_size**	**min_cluster_size**
T2V30	50	40	850
T2V63	50	40	300
	α	β	
LDA40	0.9	1/40	
LDA50	0.9	1/50	

Table 4: Data statistics. Setting 1 corresponds to 95/0.5 %, setting 2 to 90/1 %. Stopwords removed with pythons $nltk$ library, frequent words with gensims library. For more information please referre to the corresponding documentation

	# of texts	length vocab	avg # tokens	min # of tokens	max # of tokens
tokenized data	24k	68k	207	7	2.9k
setting 1	24k	2.6k	86	2	1.2k
setting 2	24k	1.3k	69	2	858

Table 5: Overview of all topics per model, that could be labeled.

LDA 40	LDA 50	T2V63	T2V30
Agricultural Policy & Climate Change	Attacks on WKSTA & Problems Judicial System	Agricultural Policy	Agricultural Policy
AUGE Union	Care	Agricultural Policy (EU Level)	Alcohol & Smoking Ban
Budget Policy	Corruption & various Political Topics	Alcohol Ban	Antisemitism
Commemoration days	COVID & Government Criticism	Anti-Muslim Racism	Asylum Policy
Criticism of Others 1	Dates 1	Antisemitism	Climate Policy
Criticism of Others 2	Dates 2	Asylum Policy	Congratulations
Danube Island Festival	Discrimination	Austrian Armed Forces	Cultural Policy
Date Announcements	Education Policy	Banks	Democracy
Dates and Announcements	Election Results	Budget Policy	Disputes
Dates and Commemorations SPÖ	Energy & Traffic	Care	Equality for Women
Economic policy	EU Policy	Carinthia	Health Policy
Education policy	Festivals	Christian Trade Union	Islamism
Election Lists & Youth Politics	Financial Policy	Climate Policy	LGBTQIA
Election Results & various Political Topics	Health & Animals	Commitment & Volunteer Fire Brigades	Nuclear Power
Equality Women	Pension Politics	Construction KH Nord	OEVP
EU Politics & Group Members	Promotion (associations & commuters)	Covid Pandemic	Parliamentary Investigations
Health, Development, Nutrition	Public Transport & Rural Areas	Criticism of Others	Pension Policy
Inequalities 1	Rumors & Speculations	Criticism SPOE	Police
Inequalities 2	SPOE 1	Cultural Policy	Press Conference
Names and reports	SPOE 2	Democracy	Reforms
Pension Policy	Suffrage and others	Digitization	Socialist Youth
Renewable Energy & Climate Change	Taxes 1	Drug consumption	Sports Clubs
Rural Area	Taxes 2	Electoral Success	Tax Policy
Scandals 1	Unemployment	European Climate and Energy Policy	Tourism
Scandals 2	Unemployment & Benefits	Floridsdorf Infrastructure	Transport Policy
Security & Social Affairs	Violence against Women	FPOE against all	Unemployment
Unemployment	World trade & Food	Genetic Engineering	Viennese Topics
Violence against women & Antisemitism	Youth Policy	German Language Skills	Youth and Children Policy
Working time & Care		Health Budget	
		Health Policy	
		Housing Communal Building	
		LGBTQIA*	
		New Elections	
		Nuclear Power	
		Obituaries	
		Parking	
		Parliamentary Investigations	
		Pension Policy	
		Police	
		Press Conference 1	
		Press Conference 2	
		Press service 1	
		Press service 2	
		Psychiatry	
		Redesign Mariahilf	
		Reforms	
		Smoking Ban	
		Social Youth	
		Sports Clubs	
		Statistics Austria	
		Tax Policy	
		Tax Policy 2	
		Tourism	
		Transparency	
		Transport Policy	
		Ukraine	
		Unemployment / Labor Policy	
		Vienna	
		Women’s Policy	
		Youth and Children Policy	