Political Science Research and Methods (2025), 1–19
doi:10.1017/psrm.2025.10008

ORIGINAL ARTICLE

Beyond standardization: a comprehensive review of topic modeling validation methods for computational social science research

Jana Bernhard-Harrer, Randa Ashour, Jakob-Moritz Eberl, Petro Tolochko and Hajo Boomgaarden

Department of Communication, University of Vienna, Vienna, Austria
Corresponding author: Jana Bernhard-Harrer; Email: jana.bernhard@univie.ac.at

(Received 24 November 2023; revised 28 October 2024; accepted 17 January 2025)

Abstract

As the use of computational text analysis in the social sciences has increased, topic modeling has emerged as a popular method for identifying latent themes in textual data. Nevertheless, concerns have been raised regarding the validity of the results produced by this method, given that it is largely automated and inductive in nature, and the lack of clear guidelines for validating topic models has been identified by scholars as an area of concern. In response, we conducted a comprehensive systematic review of 789 studies that employ topic modeling. Our goal is to investigate whether the field is moving toward a common framework for validating these models. The findings of our review indicate a notable absence of standardized validation practices and a lack of convergence toward specific methods of validation. This gap may be attributed to the inherent incompatibility between the inductive, qualitative approach of topic modeling and the deductive, quantitative tradition that favors standardized validation. To address this, we advocate for incorporating qualitative validation approaches, emphasizing transparency and detailed reporting to improve the credibility of findings in computational social science research when using topic modeling.

Keywords: convergence; standardization; text-as-data; topic modeling; validation

1. Introduction

With the maturation of computational text analysis within the social sciences (Bonikowski and Nelson, 2022), topic modeling has become a particularly popular method. Topic modeling is a process for identifying the underlying themes or topics present in a collection of text documents (Boyd-Graber *et al.*, 2017; Grimmer *et al.*, 2022), making it a versatile method that has been applied to various areas, such as journalism studies, political communication, international relations, political science, or migration studies (Roberts *et al.*, 2014; Lucas *et al.*, 2015; Jacobi *et al.*, 2016; Heidenreich *et al.*, 2019; Watanabe and Zhou, 2022). A recent literature review by Chen and colleagues emphasized the importance of topic modeling specifically for communication science, arguing that it is "an effective and innovative tool for many communication researchers" (Chen *et al.*, 2023, p.1), due to the abundance of digitized text data. The popularity of topic models can be linked to their vast applicability and cost-effectiveness. As a bottom-up approach, it can help researchers identify latent structures within large volumes of text (DiMaggio *et al.*, 2013), detect new categories or concepts (Nelson, 2020), and overall infer meanings from text (Grimmer *et al.*, 2022; Chen *et al.*, 2023).

Topic modeling was first developed in the computer sciences but was quickly adopted by scholars in the core social sciences. Nevertheless, there are considerable differences between the computational and the social sciences. As Wallach (2018, p. 4) puts it: "[C]omputer scientists may be interested in finding the needle in the haystack [...], but social scientists are more commonly interested in characterizing the haystack." This fast adoption of a new methodology to a different field required novel methods of validation that would be suitable for social scientific problems. While some could be borrowed from computer science, they would not always fit social scientific applications (Baden *et al.*, 2022). Furthermore, topic modeling methods are being further developed. For example, the introduction of neural networks into topic modeling (Zhao *et al.*, 2021) necessitates developing validation methods to account for new modeling approaches and their uses in the social sciences. Often, however, the validation approaches have lagged behind the widespread use that topic models have enjoyed (e.g., Baden *et al.*, 2022).

Given the largely automated and inductive nature of the process, it is particularly crucial to validate the outcomes and interpretations of topic models to ensure their accuracy and scientific veracity. First and foremost, a lack of validation practices is problematic from a scientific point of view, as missing validation signifies a lack of scientific rigor (Scharrer and Ramasubramanian, 2021). Second, a lack of validation practices complicates the use of topic modeling for theory building (Grimmer *et al.*, 2022; Ying *et al.*, 2022) as well as giving policy recommendations (Baden *et al.*, 2022). Third, neglecting validation gives way to criticism and skepticism around using computational methods in the social sciences more generally (Bonikowski and Nelson, 2022). Some have already voiced such critique (Margolin, 2019) due to a lack of transparency and uncertainty around applications and outcomes (DiMaggio *et al.*, 2013).

During the past years, multiple scholars have raised their concerns about a lack of clear patterns toward validating topic models (see for example Hoyle *et al.*, 2021; Baden *et al.*, 2022). The computational social science community has started to respond to these claims of lacking standardization, with studies providing first road-maps to using topic modeling in the social sciences, more generally (Maier *et al.*, 2018; Chen *et al.*, 2023), as well as first studies discussing topic model evaluation, specifically (Harrando *et al.*, 2021; Ying *et al.*, 2022; Bernhard *et al.*, 2023). However, these studies look only at specific subfields or specific evaluation tasks. In light of this, we argue that a systematic overview of validation methods applied to topic modeling is still lacking. We thus propose a thorough and systematic review of research applying topic models to assess the alleged lack of standardization of validation methods in the field. This study inductively analyses 789 substantive and methodological studies applying topic modeling pertinent to the social sciences. We shed light on the following research question: Is there a convergence toward a gold standard of validation methods for topic modeling? To do so, we will first consider which methods are applied regularly, whether there are changes over time, and whether there are combinations of validation methods that are applied more often than others. Additionally, we take into account potential methodological differences between research published in the different fields of interest (i.e., core social sciences and peripheral social sciences) and analyze them separately as well. Such an overview will make visible the breadth of potential validation methods that exist for topic modeling, thus serving as a benchmark against which researchers can compare their work. Additionally, tracking the frequency with which different validation methods are used over time can help identify emerging trends in research applying topic models. Awareness of widely accepted validation methods can foster consensus within the research community. This could eventually lead to more standardized practices and more efficient resource allocation. Last but not least, understanding the prevalent validation methods can guide the education and training of students and researchers interested in applying topic models to social scientific research questions. This paper should provide researchers with the information needed to navigate the rather complex landscape of topic modeling validation techniques.

2. On validation

Validity in the social sciences is concerned with the accuracy and scientific veracity of measures and by that, as well, of research results and downstream conclusions and recommendations. In simple terms, validity very generally refers to the question of whether measures actually measure what they are designed to measure. Therefore, the quest for validity underpins the very essence of scientific progress, also serving as a cornerstone for the construction of credible knowledge upon which impactful and effective policies and interventions can be built. Various types of validity are considered in social science, including *face validity* (aligning with common understanding), *criterion-related validity* (logical connection with external variables), and *content validity* (representing the full concept's meanings) (Scharrer and Ramasubramanian, 2021). Since validation is important for all methods in social science, many differing terminologies and sub-dimensions have been developed. In the following, we will specifically discuss those concepts central to content analysis overall and topic modeling in particular.

When it comes to manual content analysis, Krippendorff (2013) uses a threefold classification of validity into *face*, *social*, and *empirical* validity. Face validity is understood as a result of being plausible. Social validity, here, refers to a meta-perspective addressing the question of whether a scientific inquiry and measurements connected to it have societal relevance. Of more central concern for the current study, empirical validity is further differentiated into three sub-dimensions. First, there is the sub-dimension *content validity* (see also above), which, here, also includes questions relating to the appropriateness of sampling strategies. Second, he discusses the sub-dimension that he defines as *relations to other variables*, which is similar to the aforementioned criterion-related validity. The third sub-dimension, *construction and use*, relates to the internal structure of measures, which includes taking a look at the structural correspondence between available data or established theory and the modeled relationships, and demonstrating functional correspondence between what a content analysis does and what successful analyses have done. We argue that this detailed taxonomy of different types of validity—although developed for manual content analysis—can also function as a guide when thinking about how to classify validation methods in topic modeling.

Literature on validation in automated content analysis is particularly concerned with quality assurances regarding human annotation as the "gold standard" or "ground truth" that is used in dictionary (lexicon-based) approaches and supervised machine learning approaches (Song *et al.*, 2020; Grimmer *et al.*, 2022; Birkenmaier *et al.*, 2024). As an inductive approach, topic modeling, however, cannot rely on such "ground truth" measurements to be compared against. When DiMaggio and colleagues (2013) write about different perspectives on topic model validation, specifically, they refer to what they call *semantic or internal validity*—defined as whether the model meaningfully discriminates between different meanings of the same or similar terms (i.e., similar to content validity and validity on internal structure) as well as *external validity*, which is similar to previous ideas of criterion-related validation. With topic modeling, furthermore, being a statistical approach to content extraction (see also Laver *et al.*, 2003; Lowe, 2008), importantly DiMaggio and colleagues put a novel emphasis on the critical role of *statistical validity*, which assesses if the model specification inherent to the specific topic modeling approach is appropriate for the data at hand.

3. On validation of topic models

As described above, topic modeling approaches are inductive, and most are unsupervised, which means that the data generation process and, with that, model outputs cannot be well assumed prior to analysis. This makes their validation less straightforward than that of supervised methods in computational content analysis (Grimmer *et al.*, 2022). To make matters even more complicated, previous studies have shown that specific decisions relating to pre-processing (Denny and Spirling, 2018; Tolochko *et al.*, 2024), vocabulary choice (Maier *et al.*, 2020), as well as model selection (Bernhard *et al.*, 2023), can lead to tremendous changes in the model results. Validation is thus important for

both ex-ante (i.e., to decide which topic modeling algorithm should be applied) and ex-post (i.e., to evaluate the model's performance in relation to its designated task) evaluation (Gentzkow *et al.*, 2019). However, the degrees of freedom in pre-processing and hyperparameter settings that researchers tend to have, combined with the fact that topic models learn and assign documents in one step, place particularly high importance on the post-hoc validation of the models in connection to their results.

In a first hands-on user guide, Maier and colleagues (2018) provide an overview of using and evaluating LDA topic models in communication research. Importantly, they also discuss a wide range of validation approaches, including coherence metrics, qualitative expert judgment in the first step of model selection, as well as statistical validation, interpretability checks, document-topic relationships, and hierarchical clustering for mergeable topics on model validation post hoc. Notwithstanding these first efforts, Baden and colleagues (2022) have recently criticized the sustained emphasis on technological advancements over validation concerns in computational text analysis methods, including topic modeling, when it comes to the field of computational social science as a whole. While the first user guides to topic model validation may exist, it is unclear whether or to what extent researchers follow them. Moreover, concerning the increasing use of computational methods, and in particular topic modeling, in theory-driven research, researchers have criticized that computational social science studies suffer from a lack of social scientific contextualization (Baden *et al.*, 2022; Bonikowski and Nelson, 2022). Neglecting validation in the face of technological advancements makes it challenging to evaluate the methodological soundness of topic modeling studies, build theories, or make policy recommendations based on their model outputs (Baden *et al.*, 2022). Thus, especially studies posing substantive research questions, which measure social constructs, are at risk of misinterpreting the results they get from topic models. However, methodological research building on and further developing topic modeling approaches also requires proper validation to ensure (among other things) generalizability and comparability of the methods.

Due to the inherent abstraction in computational analyses, the call by researchers for establishing well-defined and universally accepted validation standards becomes even more self-evident when studying more latent social science concepts (Jacobs and Wallach, 2021). In addition to the lamented insufficient emphasis and reporting on validation practices, full stop, researchers also criticized the absence of agreed-upon methods or benchmarks applied across studies utilizing topic modeling more generally. The "validation gap" (Baden *et al.*, 2022) for topic models is thus argued to be accompanied by a "standardization gap" (Hoyle *et al.*, 2021). There are also no standardized forms for reporting the method or its validation (Reiss *et al.*, 2022), thus introducing a "reporting gap" as well.

In the past decade, many validation methods and metrics have been proposed and put to use. At first glance, however, persistent criticisms would suggest that no convergence was achieved and that this abundance of possibilities has made it even more challenging to develop points of comparison between studies and to judge the quality of the models and subsequent model outcomes. Given the widespread use of topic models and the plethora of proposed validation methods, it is high time for a systematic overview of applied validation strategies.

Of general interest to the current investigation would be to understand if, in the two decades of topic modeling application, researchers have started to come up with a set recipe for validating topic models, or at least whether specific validation techniques or combinations of validation techniques tend to be more favored over others when using topic modeling depending on the designated task. In other words, has the field started to converge to a common validation framework of topic models, and are there first signs that universally accepted standards of topic modeling validation are beginning to develop?

Methodology

To address our research question, we conducted a comprehensive systematic literature review encompassing all studies using topic modeling pertinent to the social sciences. We recognize that the

computational social sciences are still a young field. The core social sciences (e.g., communication science, political science, and sociology) have only recently adopted topic modeling. At the same time, it also has a rich history in other disciplines (e.g., computer sciences and linguistics), which also apply topic modeling to research questions relevant to studying human-generated text in the periphery of social sciences. Thus, we deliberately broadened our systematic review scope beyond the confines of social science journals. We argue that this inclusion of studies beyond the field of the core social sciences adds interesting information to the study of topic model validation, as this will allow us to take into account differences between the core social sciences and peripheral social sciences, as well as between substantive and methodological studies.

4.1. Sampling

To create our sample, we searched for relevant studies in four scientific databases (Web of Science, Mass Media Complete, ACM Digital Library, and EBSCOhost (Communication & Mass Media Complete, Humanities Source Ultimate, SocINDEX)) using a search string, which looked for different spellings of "Topic Model" as well as "Topic Modelling" in the title, abstract, or keywords of a research study. Additionally, we specified that "valid*" needs to be found at least once in the full text.[1] We did not limit our search regarding the publication date so that we could give an overview since the introduction of topic models in social research. Initially, this search yielded 1,556 studies. In the first step, we coded the entire sample to assess which studies would be relevant for our review.

Regarding formal characteristics, studies were excluded if they were (I) not accessible to the coders, (II) duplicates of already selected studies, or (III) texts like extended abstracts, posters, presentation slides, panel descriptions, or studies without empirical analyses. Furthermore, studies were excluded if (IV) not at least one of the text corpora used in the studies was based on human-generated speech or (V) if the term "valid*" was mentioned in the paper but did not refer to the topic model or its output. After reviewing, 789 studies met our criteria and were included in the sample for further analysis. The earliest study was published in January 2004, and the most recent one was published in March 2022, which coincides with the month of data collection.[2]

4.2. Description of the dataset

Figure 1 shows how the number of topic modeling studies pertinent to the domain of the social sciences has increased over the past two decades. The number of studies with substantive research questions has risen steadily since 2011. However, the number of studies focusing on methodological advancement peaked in 2017 before dropping and plateauing for the upcoming years. Overall, more studies were published in conference proceedings (54%), than in journals (46%). Top publication outlets suggest that a considerable number of studies pertinent to this investigation are (still) published in computational science journals and conferences (see Table 3 in Appendix C).

4.3. Codebook and manual classification

The identified studies were then coded by two of the authors. We first annotated whether the studies had substantive research questions from the social sciences and/or focused on methodological advancements in topic modeling. Note that these two categories are not mutually exclusive. However, our main variables of interest were the validation methods mentioned and applied by the original study authors. Here we employed an inductive coding scheme where we added columns each

[1]The initial search yielded 8,878 studies without the "valid*" keyword. This sample was too large for manual inductive coding, so we limited our search to studies explicitly mentioning validation. However, this does not imply that all studies lacking the "valid*" term do not engage in validation. These could include theoretical studies, studies that reference topic models or studies that are not explicitly using this terminology when discussing validation approaches.

[2]Please see Appendix C and osf.io for our sample and replication data.

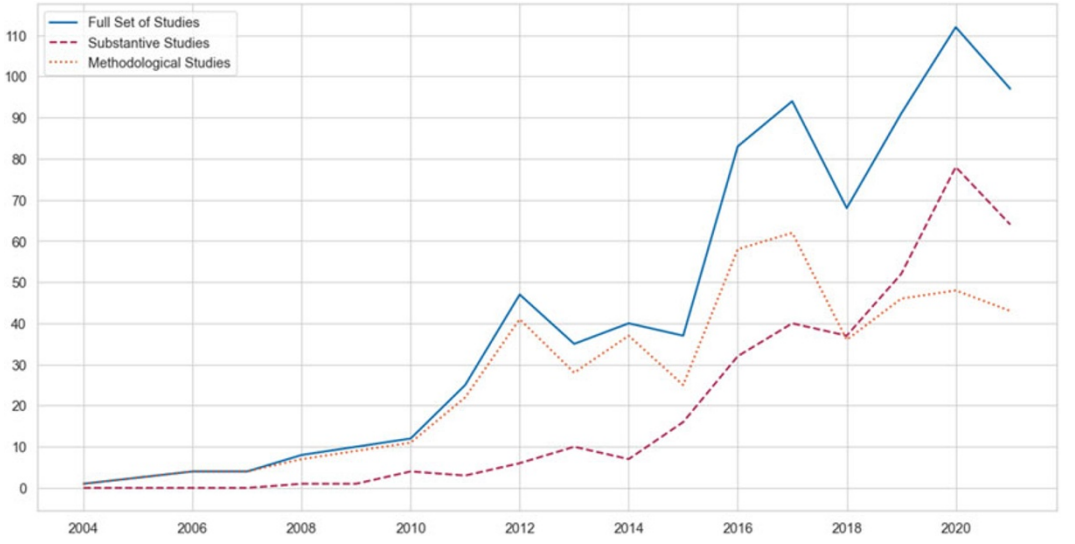

Figure 1. Number of Studies with a substantive or methodological goal in our sample over time.
Note: As only a quarter of the year 2022 is included in the sample, we did not include it in the graph, as it would have resulted in a misleading trend.

time a method was mentioned for the first time. Given the absence of a comprehensive systematic review on topic modeling validation methods to date, the lack of standardization in reporting validation methods, and the ongoing debates and developments in this area, we believe that an open and inductive approach to coding validation methods is needed. Rather than relying on a potentially incomplete pre-defined list of validation methods, our methodology involved coding any approaches mentioned by the authors of our analyzed studies. We included those approaches that the authors deemed relevant to validating their topic models and topic model outputs.[3]

This approach allowed us to capture the breadth of practices and strategies in this dynamic and evolving field but has two key implications. Firstly, we might have incorporated methods that some researchers might not consider adequate for topic model validation, yet were communicated by the study authors as a means to validate their approach. Secondly, our analysis focused solely on the validation methods explicitly articulated in the final manuscripts, potentially overlooking crucial, albeit unreported, steps.

In the initial phase of inductive coding, we identified a total of 445 distinct methods that were mentioned as pertaining to topic model validation. Subsequently, a meticulous refinement process was implemented, addressing instances where different terminologies denoted identical procedures. We, furthermore, amalgamated closely related approaches.[4] This consolidation led to a streamlined set of 138 validation methods. Recognizing the need for a more manageable framework, we further pruned the list by excluding validation methods that accounted for less than 1% of our sample, resulting in a more practical and focused compilation of methods. The remaining 32 validation methods were then grouped into 8 overarching categories based on their shared methodological perspective: *Model Comparison*, *Internal Qualitative Inspection*, *External Qualitative Inspection*, *Error Rate Analysis*, *Distinctiveness of Top Words*, *Information Theory Metrics*, *Similarity and Distance Metrics*,

[3]The studies obtained from the EBSCOhost database have been added later in response to a reviewer comment during peer review. Thus, these 59 studies were coded deductively based on the categories identified during the inductive coding of 730 studies sampled through the other databases used for the initial submission of the manuscript.

[4]For example, we aggregated methods that are variations of each other, such as micro- and macro Precision, or different k-fold splits.

and *Downstream Tasks*. These eight categories can also be understood in relation to the theoretical perspectives on validation introduced earlier in this study. While *Model Comparison* can be connected to Krippendorff's (2013) understanding of internal validity, *Internal Qualitative Inspection* is understood in the sense of Krippendorff's understanding of face and context (semantic) validity or DiMaggio and colleagues (2013) internal validity. *External Qualitative Inspection* can be related to internal and relational validity (for Krippendorff) and external validity (for DiMaggio et al.). The four groups comprising statistical measures *Error Rate Analysis, Distinctiveness of Top Words, Information Theory Metrics as well as Similarity and Distance Metrics* are difficult to characterize in Krippendorff's assessment, as he was focusing on manual analysis, however, they can be connected to what Dimaggio and colleague's (2013) understand as statistical validity. The last category *Downstream Tasks* is again related to Krippendorff's relational validity.

These eight overarching categories synthesize the numerous, often highly specific approaches commonly associated with validation. They comprehensively encompass previously suggested types of validation, providing a broader and overarching classification of the diverse methods for validating topic models that are pertinent to social research and are widely used in practice. Employing our inductive approach, we successfully grouped all the methods used into these eight overarching categories. It should be noted that intercoder, as well as intracoder reliability, was assessed using Krippendorff's alpha, revealing satisfactory levels of agreement. Please refer to Table 1 in Appendix A for detailed information. For an overview of the mentioned validation methods corresponding to each of the categories above, please see Table 2 in the Appendix.

5. Results

Our comprehensive review uncovered a diverse array of approaches to topic model validation. Subsequently, we will delve into these approaches in greater detail, emphasizing their development over time. We will present our findings according to the frequency with which each category was mentioned in our sample (see Figure 2 for a visualization).[5]

Most studies (61.2%) in our sample mention at least one validation method pertaining to the overarching validation category of *Comparing Models*. This category encompasses instances where authors emphasize having executed various types of topic models or specified their topic models differently to determine the most suitable approach for the specific task. The decision to either use a single model or compare multiple models should be regarded as a precursory step preceding all subsequent methods of validating topic modeling. Of course, the basis for comparing topic model outputs must derive from one of the other seven overarching categories. As methodological advances in topic modeling tend to require comparisons to pre-existing modeling techniques, this category is mentioned more often in methodological studies (77.9%) as compared to substantive studies (37.1%).[6]

In the second place, we find *Internal Qualitative Inspection* (54.2%), aiming at internal, face, and content validity. This category entails the application of qualitative methods to evaluate the quality and relevance of topics generated by a topic modeling algorithm, relying solely on the model's output. Common practices within internal qualitative inspection include assessing the plausibility of topics, which heavily relies on the concept of face validity. Substantive studies rely much more on these methods (76.7%) than studies with a methodological focus (38.3%).

A bit under half of the studies (44.5%) in our sample use different kinds of *Error Rate Analysis*, which is based on the assessment and quantification of the performance of the model by comparing its predictions against a ground truth (oftentimes manual annotations based on a deductive coding schema). These metrics help in understanding the model's effectiveness and its ability to make

[5]For a visualization of these categories, which presents them separately for substantive and methodological papers, see Figure 4 in Appendix C.

[6]Studies can have both a substantive and a methodological focus. Thus, we give the prevalence of each category in percentage and not frequency, as the numbers, in this case, would not add up to the number of studies in our sample.

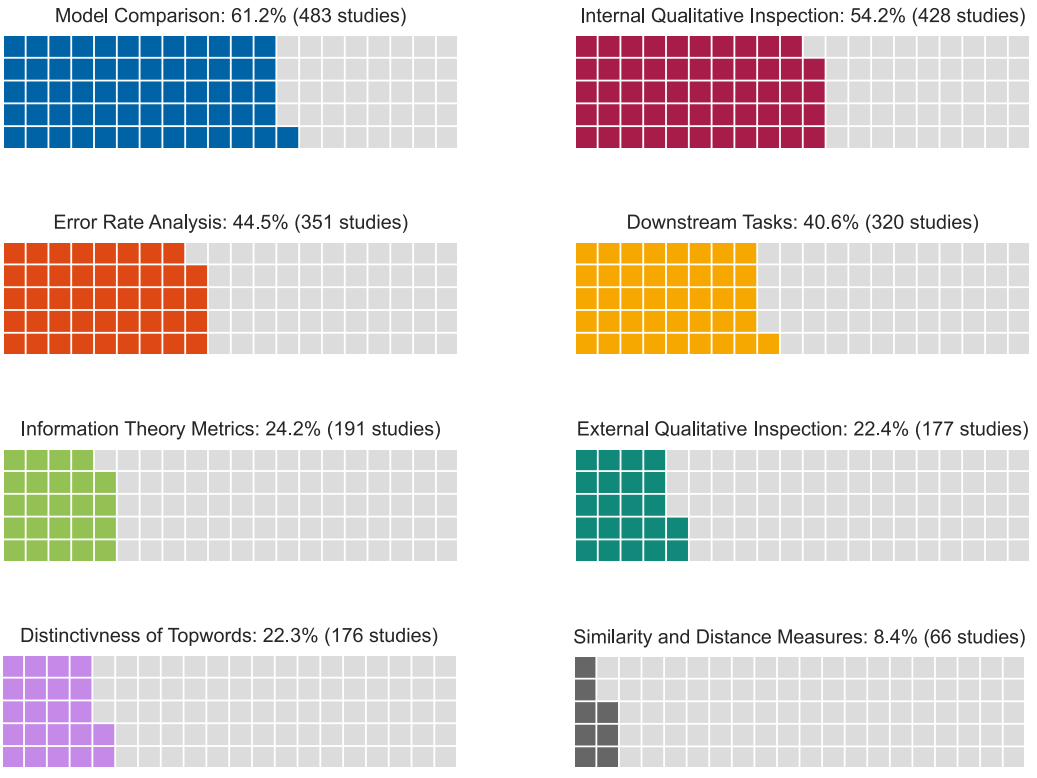

Figure 2. Percentage of studies employing validation methods.

accurate predictions. Metrics of *Error Rate Analysis* include well-known statistics such as calculating Recall, Precision, or the F-Score. While only 22.7% of substantive studies employ some kind of error rate analysis, 60.2% of methodological studies do so.

Still, 40.6% of the studies argue that they evaluate the validity of their model and its output by applying it to a specific *Downstream Task*. *Downstream Tasks* in the context of topic model validation refers to the assessment of the effectiveness and utility of topic models by evaluating their performance in tasks that depend on the output of these models. Instead of concentrating on the internal characteristics or outputs of the topic models, this approach assesses the contribution of the generated topics to the success of subsequent tasks, like, for example, serving as a covariate in a regression analysis. This method is more widespread for substantive research (49.3%), as compared to methodological research (36.3%).

A quarter of studies (24.2%) in our sample rely on validation methods building on statistical validity, using *Information Theory Metrics*. These are statistical measures used to assess the quality, uncertainty, and information content of topic models. These metrics help quantify the differences between probability distributions and evaluate the efficiency and accuracy of the models. Included in this category are the Jensen Shannon as well as Kullback–Leibler Divergence, Perplexity, and Entropy. 28.6% of methodological studies are validated with methods from this group, while only 19.1% of substantive studies apply the same methods.

The category of *External Qualitative Inspection* is applied by 22.4% of studies. *External Qualitative Inspection* involves qualitatively evaluating the meaningfulness and relevance of topics, explicitly leveraging model-external information, such as theoretical assumptions or real-world contexts, including events or dynamics. Some studies also compare topic modeling outputs with inductively

human-annotated subsets of the text corpus. Similar to *Internal Qualitative Inspection*, substantive studies apply these more frequently (32.4%) than methodological studies (15.2%).

Metrics relating to *Distinctiveness of Top Words* are referred to in 22.3% of all studies analyzed. This category comprises methods that evaluate statistical validity based on the uniqueness and quality of high-probability words (i.e., Top Words) within the topics generated by a topic model. Such measures aim to evaluate the meaningfulness and relevance of the top words within each topic. This category is used by substantive and methodological studies at a similar rate (substantive: 22.4% and methodological 22.1%).

Finally, *Similarity and Distance Metrics* are the smallest overarching category, present in 8.4% of studies overall. In the context of topic models these metrics quantify the degree of (dis)similarity between topics, documents, or words, enabling the evaluation of the relationships, overlaps, and distinctiveness within the generated topics. Related metrics are, for example, the Jaccard Coefficient, or Silhouette. Again, this category is distributed similarly across all studies, 7.8% in substantive studies and 9.5% in methodological studies.

Next, we examine how the salience of different categories of validation methods changed over time (see Figure 3). Given the limited number of cases in the initial years of our dataset and to ensure meaningful comparisons, we narrowed our focus to a subset of the data, commencing from 2011 ($n = 750$). Three discernible trends emerge: categories that exhibited consistent use over time, those that experienced a decline in usage, and those that observed an increase. Among those used consistently over time are *Downstream Tasks* and *Similarity and Distance Metrics*. Categories such as *Error Rate Analysis*, *Information Theory Metrics*, and *Model Comparison* have lost popularity over time. For instance, *Model Comparison* was referenced in 80% of studies in 2011, but only in 60% of the studies in 2022. Similarly, *Error Rate Analysis* decreased by nearly 30 percentage points, from being utilized in 60% of studies to only a third in 2022. Lastly, *Information Theory Metrics* appeared in almost half of all studies in 2011, but only in every fifth study in 2022. Conversely, validation methods such as *Internal Qualitative Inspection*, *External Qualitative Inspection*, and *Distinctiveness of Top Words* approaches are experiencing growing prominence. The *Internal Qualitative Inspection* has surged from under 40% to almost 70%. The growth in studies mentioning *External Qualitative Inspection* is even more striking, starting at 10% in 2011 and surging to nearly 25% within a decade, possibly speaking to the increasing adoption of topic modeling, particularly by researchers trained in social sciences rather than computer sciences. The validation methods pertaining to the *Distinctiveness of Top Words*, as well, have witnessed a large increase, initially being employed in fewer than 10% of studies and now featuring in over a third of the studies.

Numerous scholars have underscored the importance of accommodating diverse perspectives validation as a prerequisite for appropriate topic modeling validation (DiMaggio *et al.*, 2013; Krippendorff, 2013; Maier *et al.*, 2018; Chen *et al.*, 2023), thereby emphasizing the necessity of incorporating a synthesis of various validation categories in research studies. Aiming to shed light on the extent to which researchers acknowledge and integrate diverse validation perspectives within their studies, we first look at the average number of validation categories within each study: On average, each study incorporates approximately three overarching validation categories, as illustrated in Figure 4). The average number of individual validation methods used in a single study is four, with no significant disparities observed between substantive and methodological studies.[7] Overall, we find no discernible trends suggesting a growing inclination toward the integration of a higher number of validation perspectives in conjunction within single studies.

Figure 5 illustrates the extent to which different validation categories co-occur in individual studies within our dataset. We acknowledge that we limit our examination to dyadic combinations, given the complexity introduced by the increasing number of possible combinations (e.g., triads). Despite this

[7]Please note that the average number of validation methods should be interpreted carefully, as this is strongly dependent on how we summarized them.

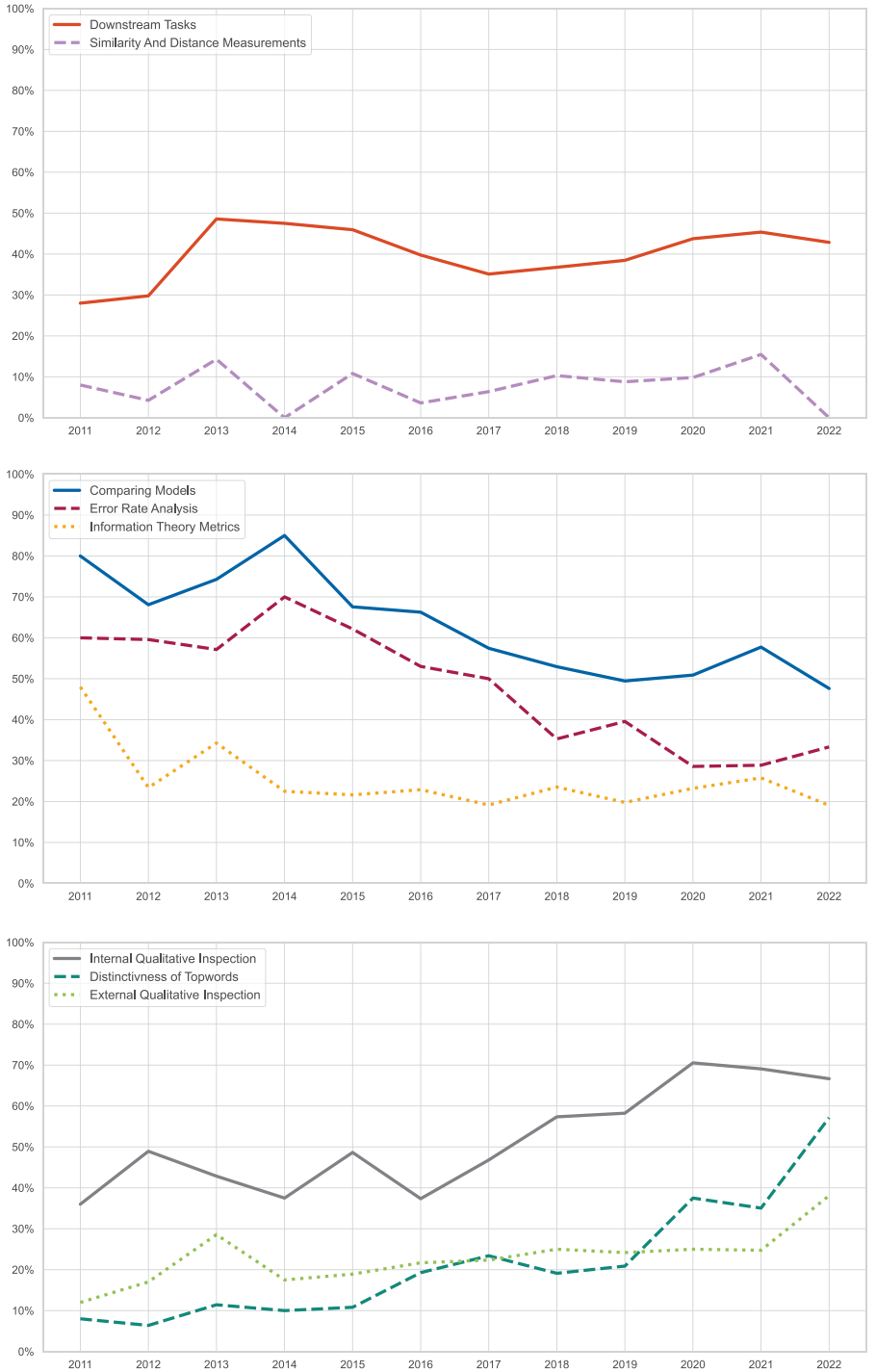

Figure 3. Changes in the application of validation categories over time.

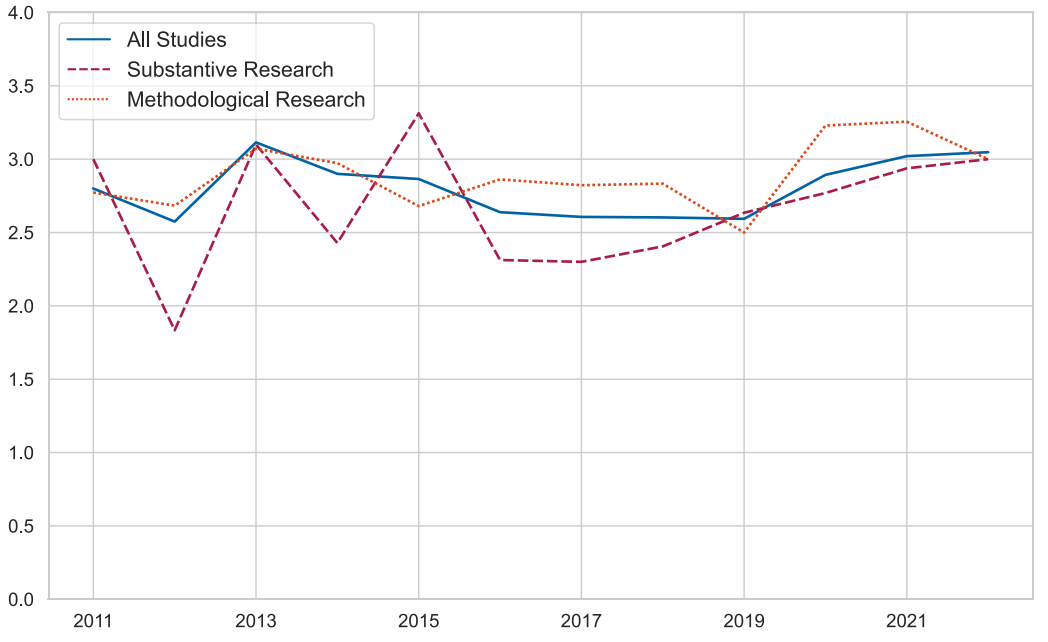

Figure 4. Average number of validation categories used per study over time.

	Model Comparison	Internal Qualitative Inspection	Error Rate Analysis	Downstream Task	Information Theory Metrics	External Qualitative Inspection	Distinctivness of Topwords	Similarity & Distance Measures
Model Comparison	3,7	48,1	81,0	55,9	73,1	41,2	66,1	59,1
Internal Qualitative Inspection	42,6	9,1	34,9	64,7	52,3	76,8	62,1	54,5
Error Rate Analysis	58,6	28,6	4,3	37,8	43,5	29,4	39,0	45,5
Downstream Task	36,8	48,1	34,4	3,8	34,7	49,2	41,2	43,9
Information Theory Metrics	29,0	23,5	23,9	20,9	4,7	20,9	26,0	37,9
External Qualitative Inspection	15,0	31,6	14,8	27,2	19,2	3,4	27,1	25,8
Distinctivness of Topwords	24,1	25,6	19,6	22,8	23,8	27,1	0,6	30,3
Similaity & Distance Measures	8,0	8,4	8,5	9,1	13,0	9,6	11,3	1,5

Figure 5. Dual co-occurrences of two validation categories in percent.
Note: The diagonal marks the share of studies that include only validation methods from one validation category.

limitation, the comparison reveals intriguing insights. For instance, 81.0% of the studies mentioning validation methods related to *Error Rate Analysis* compare these metrics by running different *Model*

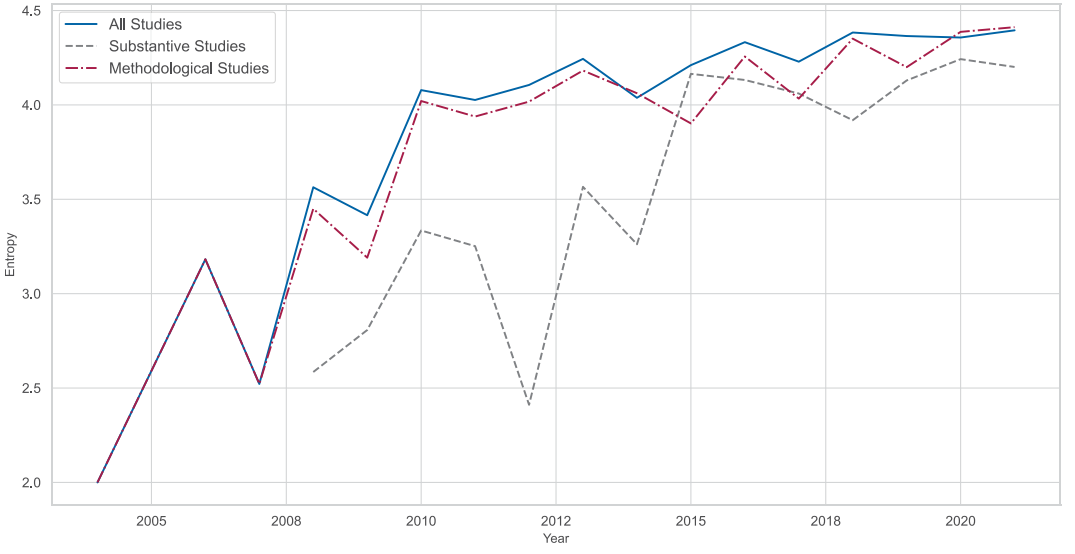

Figure 6. Information entropy for binary validation categories over time.

Comparisons. Note that the diagonal represents the number of studies that pertain to a method within the specific category mentioned in the column without any reference to also having used validation methods from other overarching categories.

As expected, there are notable overlaps between methods relying on quantitative metrics, and studies using these categories often employ approaches related to *Model Comparison*. Conversely, studies relying on *External Qualitative Inspection* heavily leverage on *Internal Qualitative Inspection* (76.8%), as well. At the same time, *Internal Qualitative Inspection* emerges as the overarching category of validation methods most frequently used in isolation, without reference to methods from other categories (9.1%).

While we have demonstrated the evolution of the importance of certain overarching categories of validation methods over time and the frequent combination of various perspectives in topic modeling validation, it is intriguing to measure the possibility of convergence in these combinations. Specifically, we seek to understand whether, as time progresses, there is a tendency within the field to increasingly agree on specific combinations of validation categories rather than others. In order to assess this possibility, we calculate *Information Entropy*, a metric from information theory that estimates the diversity within a community, and plot the values over time (see Figure 6). Information entropy is a measure of "surprise" of seeing another data point. In other words, when reading a random topic modeling paper, we would expect that the most "agreed upon" dyadic combination of validation methods is most likely used. If researchers using topic models would converge on some set(s) of validation approaches that are more "standard" than others, we would see the value of *Information Entropy* decreasing over time.

However, the observed values remain relatively consistent over the last decade, suggesting that the diversity of dyadic combinations of validation perspectives employed in studies has not significantly decreased over time. Consequently, no observable convergence of the field toward a specific combination of validation categories is evident. A similar pattern emerges when considering all 32 validation methods instead of the overarching categories (see Figure 8 in Appendix C). For an additional robustness check, we further categorized all relevant studies into studies from the core social sciences and the social science periphery to see whether convergence is happening in either of the

fields. Again, we could not find any indication of convergence. Moreover, we also conducted the analysis using a weighted sample, where the weight for each paper was calculated by dividing the number of citations by the number of years since its publication. The corresponding graphs (Figures 11 , 12, and 13) are available in Appendix E. However, the results show no significant changes from the main analysis.[8]

Despite the widespread adoption of various validation techniques, often spanning distinct categories and examining different facets of model validity, our investigation uncovers a distinct absence of dominant or widely accepted combinations. Curiously, no overarching trends or consensus practices regarding the amalgamation of these diverse validation methods become apparent, highlighting the current lack of standardization of topic modeling validation approaches and no clear signs of convergence in topic model validation more generally.

6. Discussion

Validation is an important part of scientific research. Thus, the critique by many scholars that computational methods, especially topic modeling, lack validation has set into motion a number of projects aimed at enhancing our understanding of topic model validation. While some scholars have focused on presenting a road map (Maier *et al.*, 2018), others have highlighted the process in general (Chen *et al.*, 2023) to see what has been done in the field. However, the notion of missing standards has been a point of discussion at many times (e.g., Maier *et al.*, 2018; Baden *et al.*, 2022). What was missing so far from the literature was a systematic review that approaches validation in a long-term and inductive perspective, which allows us to account for the plethora of possible ways to validate topic models. This is especially important as it allows us to test claims of missing standards while taking into account how topic model validation has changed in the computational social sciences and may have converged in standard routines. We find that in 20 years of applying topic modeling as a method in the computational social sciences, there have neither been clear signals pointing toward standardization of validation practices nor first signs of convergence toward specific validation methods over time. These findings hold true when looking at overarching validation categories as well as when focusing on specific validation methods. Moreover, while some validation approaches more strongly rely on a combined use with other approaches, we could not find any emerging convergence in terms of dyadic combinations of validation perspectives.

However, as evident from this literature review, this lack of convergence is not necessarily indicative of a lack of trying. Perhaps, this inability to converge on a strict set of quantifiable validation criteria comes from an inapt approach to the problem. Unlike classical statistical methods (e.g., a regression analysis), topic models, like many noted before, are an inherently inductive approach to data analysis. Instead of imposing preconceived notions or hypotheses onto the data, topic modeling attempts to uncover hidden thematic structures within texts. This inductive nature means that topic modeling is by nature exploratory, aiming to reveal latent patterns and topics that might or might not be apparent through deductive analysis alone. Classical statistical methods are validated based on how well they fit a predefined hypothesis. The core idea of topic models, by contrast, lies in their ability to uncover hidden structures and emerging themes in the data—a process that, almost by definition, is unsuitable for a deductive validation paradigm that firmly rests on standardization for the sake of comparability.

This distinction becomes even more apparent when considering the assumptions one must make when dealing with classical methods versus topic models (or any other unsupervised machine learning algorithm). In classical deductive analysis, there is a foundational assumption that a singular "real" data-generating process does exist that researchers aim to approximate as closely as possible.

[8]It is noteworthy that there was actually a trend toward divergence in validation methods, particularly for studies from the core social sciences in 2021, however, we want to caution against over-interpreting this result, as it is based on only 27 studies.

Under this assumption, validation techniques can be designed to measure the model's ability to capture this singular underlying truth. For example, classical validation methods in regression, such as cross-validation and goodness-of-fit tests, are tailored to assess how well the model aligns with this assumed reality. Under this assumption, however, topic models are an objectively wrong way to analyze the data. Different topic modeling validation solutions can be valuable for distinct purposes. For instance, one solution might be optimal for summarizing documents, while the other is for analyzing latent themes over time. Also, depending on the corpus at hand and, importantly, *independent of the hypothesis*, a different number of topics may be appropriate for the data. The usefulness of a topic model is deeply dependent on the specific research question, dataset, and expected outcomes of the analysis. This diversity in utility renders the notion of a single, universal "real" data-generating process inappropriate.

Thus, we argue that traditional validation methods rooted in the assumption of a singular "truth" cannot accommodate the multifaceted nature of topic modeling solutions. The appropriateness of a topic model is contingent upon its application context, making it impractical (and, potentially, impossible) to devise a one-size-fits-all validation framework. We suggest that this mismatch makes it incredibly difficult to find standard approaches, or at least a convergence toward a few select methods or even categories of validation methods.

It is also essential to clarify that we are not making an ontological argument about topic modeling validation techniques. This would require us to establish a universal standard for what constitutes a "valid" topic across all contexts—something we explicitly argue against. Our argument is epistemological: we approach topic modeling validation from a perspective of how researchers can ascertain that a model achieves its intended purpose based on its application. We propose that the choice of validation method should be contingent upon the particular utility, or "usefulness," that the topic model provides to the researcher (whether classification, discovery, interpretability, or another goal), irrespective of any essentialist definition of a "topic." By approaching this question from an epistemological position, we argue that model utility should guide the validation criteria, allowing for diverse methods that best serve the research aims. Our perspective acknowledges that validation methods are not absolute but are instead contextually guided.

Naturally, this is not to say that we should forgo topic modeling validation altogether. It does mean, however, that we need to change our way of thinking about it. Especially, when utilizing topic models without "ground-truth" data, instead of understanding validation as hitting a specific cut-off point for F1-Scores, we might focus our attention toward more qualitative interpretations of validation. Humphreys and colleagues (2021), for example, define validity for qualitative research after Kirk and Miller (1986, p. 20) as "the degree to which the finding is interpreted in the correct way." In their opinion, this could mean heightened transparency throughout the whole process (Dienlin *et al.*, 2021) as well as including "thick descriptions" (Humphreys *et al.*, 2021, p. 857) of how the interpretation came to be, triangulating as well as different perspectives, especially also from people with lived expertise. Integrating the methodologies from qualitative research into computational methods, such as topic modeling, can help us face the inevitable need to validate what we find.

Importantly, Humphreys and colleagues (2021, p. 857) note that "it is important to recognize that differences in methods call for different kinds of validity- and credibility-enhancing research practices." In a similar vein, Barberá and colleagues (2021, p. 40) assess that "every research question and every text-as-data enterprise is unique," and thus also call for validation decisions to be adapted to each individual research project. This mirrors our findings that many validation methods can be applied to assess some aspect of validity. The most fatal error would be to neglect validation altogether, as any model requires careful scrutiny to ensure its relevance. While no single method is universally applicable, combining approaches appropriate to the specific research design will yield more valid results. It is difficult to quantify the value of each possible validation method, as this depends on many factors, not only the research question and design but also how the validation is implemented.

While some methods are generally helpful for most topic modeling applications (for example, labeling the topics based on close readings of documents), others are only suitable for some studies (for example, Error Rate Analysis if there is a gold standard). The eight categories outlined in this review represent valid topic modeling validation methods, each with particular strengths and weaknesses to consider:

Model comparison effectively reduces uncertainty as it provides a benchmark against which the performance of one model can be assessed in relation to another. However, its usefulness depends entirely on the quality of the compared models and the reference point, or benchmark, chosen. Thus, this category cannot stand alone and must always be used in conjunction with another method for meaningful analysis.

Internal Qualitative Inspection is critical, particularly in unsupervised learning. It allows human expertise to guide validation through qualitative judgments. Yet, the strength of this approach varies depending on the reviewers' familiarity with qualitative research methods. For this reason, readers, reviewers, and editors must have a foundational understanding of how qualitative analysis works to judge its value.

Error Rate Analysis offers concrete metrics for evaluating models by comparing them to a gold standard. Its effectiveness depends on the quality and relevance of that gold standard, and there is often debate over whether and how a helpful gold standard can be obtained.

Downstream Task validation is helpful because it tests how well a model performs in real-world applications. However, this approach merely postpones the evaluation problem, as the downstream tasks must be assessed for validity. Additionally, determining which models to evaluate and how to evaluate their performance in downstream tasks can be resource-intensive, adding complexity to the validation process.

Information Theory Metrics provide valuable insights by measuring performance in probabilistic spaces. These metrics help narrow down the parameters useful for a given topic model. However, they are purely quantitative and lack human interpretation, which we argue is eventually needed for social science research. Without a qualitative understanding of how different parameter settings affect the model, these metrics can be challenging to interpret in isolation, as they can not distinguish between interpretable and uninterpretable topics.

External Qualitative Inspection is beneficial when there is an external comparison point, offering a form of face validity. However, this is not always feasible, primarily if no suitable external reference exists. Additionally, comparisons between topics based on correlations with external events can be challenging and subject to interpretation, limiting the robustness of this approach.

Distinctiveness of Topwords is a valuable set of metrics that uses the information we get from the topwords of each topic in our model, thus working in linguistic spaces. While these methods can highlight which words are distinct across topics, they also have the limitation of being purely metric-driven, which can hinder their usefulness in contexts that require deeper human interpretation.

Similarity and Distance Measures are metrics that work in geometric spaces. They offer a way to measure the similarity between different topics with a model. However, like the other metric-based approaches, they can fall short when providing meaningful insights into the differences between models without human interpretation.

We argue that similar to how topic modeling usefulness is dependent on the use case, so should the validation procedures be. Striving for a standardized set of validation criteria, applicable universally across diverse use cases, is perhaps a misguided attempt. Depending on the task of a particular topic model in a particular research design, practitioners should choose the validation method that is best aligned with the task. This emphasis on context-specific validation methods fosters transparency in the research process. By explicitly reporting how the model was validated and detailing for what specific reasons this validation was chosen, researchers provide readers with a deeper understanding of the methodology's appropriateness and relevance to the research objectives. This clarity not only

enhances the credibility of the findings but also allows readers to assess the validity of the model's outcomes within the context of the study.

What we can recommend at the end of our review is that computational social scientists should familiarize themselves with qualitative validation criteria and more openly embrace the inductive nature of topic modeling. This might mean realizing that topic modeling is not an ideal method for measurement (e.g., Grimmer *et al.*, 2022), and thus requires putting extra effort into explaining, why and which validation method is chosen, what it tells us, and why this should bolster the credibility of our findings and interpretation. In this, it is important that researchers do not fall into the trap of arguing ex-post, based on the results, that a hypothesis can be accepted or rejected, if they use the same logic to validate their topic model (i.e., Validation via Downstream Tasks). Instead, as it is more appropriate for qualitative research, we should use the outcome to formulate possible hypotheses, which can be tested in a subsequent step, with a different method. We realize that readers might expect clear guidelines or blueprints on how to validate their topic models at this place. However, the lack of convergence and lack of clear patterns as to which methods are actually applied by researchers suggest that there is no one way to validate topic models. We stress that it is the responsibility of each researcher to decide what kind of validation their particular research question warrants. Elsewhere, Bernhard and colleagues (2023) have given an outline of possible questions to ask yourself when validating topic models, which can help as a guide for deciding on a validation strategy and Maier and colleagues (2018) specifically propose a guideline on validating LDA models.

Again, it is important to note, that we do not argue against convergence or standardization of validation strategies per se. Convergence, in the sense of shared methodologies, holds value and is important to the advancements in the field. However, we believe that striving for methodological convergence must not overshadow the need for tailored, task-specific validation in the realm of topic modeling. Unlike in classical quantitative research methods where standardized approaches often prove effective, the diverse nature of textual data and the multifaceted objectives of topic modeling necessitate a more nuanced and adaptable approach to validation.

Overall, scientific progress can only be made if we choose the right method for the research question. This includes taking into account the strengths and weaknesses of a method and catering toward the former. Topic modeling is a computational text analysis method, which helps us *inductively* find patterns in large amounts of text. With the recent surge of digitally available text, it is a great method for identifying underlying themes and thus, gaining insights into rich data. Transparent research and detailed reporting on the application of the method and interpretation of the findings can be an important first step toward credible findings. Applying fitting validation methods, ideally around different validation perspectives can reinforce the validity of a research study using topic modeling.

Our review has certain limitations that need to be considered. First, we only focus on studies using topic modeling that explicitly mentioned validation, which means that we may have overlooked studies that referred to validation only implicitly. It also means that our study cannot speak to the validation gap (i.e., the lack of validation) specifically, rather than the standardization gap in topic modeling research in the social sciences. Second, although we conducted an inductive coding process, it was not feasible to analyze all of the coded validation methods. Many methods that researchers referred to as "validation" were only present in very few studies and could, therefore, not be considered. The lack of standardized reporting (i.e., reporting gap) also may have led to an underidentification or at least to some fuzziness, when it came to categorizing some of the methods. Third, it would have been interesting to separate the analyses further between the subfields in the core social sciences (e.g., political science vs. communication science), however, this was not possible, as there were not enough relevant studies per year and subfield in our sample. And finally, while there is a desire for recommendations on the best validation method for a topic model, this is not something we can do in this study. We argue that, while standardization is important and researchers should

not be overwhelmed with the choice of validation methods, every researcher has to do the work, to derive which validation methods would be the most applicable to a specific topic modeling use case. We do hope that our study can help get this process started.

7. Conclusion

In conclusion, our systematic literature review spanning two decades of topic modeling in the computational social sciences reveals a notable absence of standardized validation practices and a lack of convergence toward specific validation methods over time. This discrepancy may be attributed to the inherent inductive and qualitative nature of topic modeling, which does not align with the deductive, quantitative traditions that typically seek standardization. Building on this, we propose a shift in how we perceive validation of topic modeling, particularly in the absence of "ground-truth" data. We advocate for a more qualitative approach to validation, emphasizing the importance of correctly interpreting findings as well as transparency drawing from qualitative research methodologies. Acknowledging the uniqueness of each research question, we recommend that computational social scientists adapt their validation criteria to suit the specific context of their research projects and transparently motivate this choice. Ultimately, our review underscores the importance of selecting the right methods for the research question, understanding the strengths and weaknesses of these methods, and fostering transparency and detailed reporting to enhance the credibility of findings when employing topic modeling in the computational social sciences.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psrm.2025.10008. To obtain replication material for this article, see https://doi.org/10.7910/DVN/N67BDI.

Acknowledgements. We gratefully acknowledge the insightful and constructive feedback provided by the reviewers and editors, which significantly improved the clarity, rigor, and overall quality of this paper.

Funding statement. This work was supported by the Austrian Federal Ministry for Education, Science and Research as part of the funding program Digital and Social Transformation in University Education.

Competing interests. The authors declare no competing interests.

Data availability statement. Replication data and code associated with this study are available in the Harvard Dataverse.

Ethical standards. This research complies with all applicable ethical guidelines. Approval was granted by the Ethics Committee of the Department of Communication at the University of Vienna.

Author contributions. Conceptualization: JBH, HB; Methodology: JBH, PT; Data Curation: JBH, RA, JME; Data Visualization: JBH; Writing—Original Draft: JBH, JME; Writing—Review & Editing: All authors. All authors have read and approved the final manuscript.

References

Baden C, Dolinsky A, Lind F, Pipal C, Schoonvelde M, Guy S and van der Velden MACG (2022) Integrated standards and context-sensitive recommendations for the validation of multilingual computational text analysis. Technical Report Deliverable 6.2. OPTED.

Baden C, Pipal C, Schoonvelde M and van der Velden MACG (2022) Three gaps in computational text analysis methods for social sciences: a research agenda. *Communication Methods and Measures* **16**, 1–18.

Barberá P, Boydstun AE, Linn S, McMahon R and Nagler J (2021) Automated text classification of news articles: a practical guide. *Political Analysis* **29**, 19–42.

Bernhard J, Teuffenbach M and Boomgaarden HG (2023) Topic model validation methods and their impact on model selection and evaluation. *Computational Communication Research* **5**, 1–26.

Birkenmaier L, Lechner CM and Wagner C (2024) The search for solid ground in text as data: a systematic review of validation practices and practical recommendations for validation. *Communication Methods & Measures* **18**, 249–277. doi:10.1080/19312458.2023.2285765.

Bonikowski B and Nelson LK (2022) From ends to means: the promise of computational text analysis for theoretically driven sociological research. *Sociological Methods & Research* **51**, 1469–1483.

Boyd-Graber J, Yuening H and Mimno D (2017) Applications of topic models. *Foundations and Trends in Information Retrieval* **11**, 143–296.

Chen Y, Peng Z, Kim S-H and Choi CW (2023) What we can do and cannot do with topic modeling: a systematic review. *Communication Methods and Measures* **17**, 1–20. doi:10.1080/19312458.2023.2167965.

Denny MJ and Spirling A (2018) Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it. *Political Analysis* **26**, 168–189.

Dienlin T, Johannes N, Bowman ND, Masur PK, Engesser S, Kümpel AS, Lukito J, Bier LM, Zhang R, Johnson BK, Huskey R, Schneider FM, Breuer J, Parry DA, Vermeulen I, Fisher JT, Banks J, Weber R, Ellis DA, Smits T, Ivory JD, Trepte S, McEwan B, Rinke EM, Neubaum G, Winter S, Carpenter CJ, Krämer N, Utz S, Unkel J, Wang X, Davidson BI, Kim N, Won AS, Domahidi E, Lewis NA and de Vreese C (2021) An agenda for open science in communication. *Journal of Communication* **71**, 1–26. doi:10.1093/joc/jqz052.

DiMaggio P, Nag M and 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**, 570–606.

Gentzkow M, Kelly B and Taddy M (2019) Text as data. *Journal of Economic Literature* **57**, 535–74.

Grimmer J, Roberts ME and Stewart BM (2022) *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Oxford: Princeton University Press.

Harrando I, Lisena P and Troncy R (2021) Apples to apples: a systematic evaluation of topic models. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*. Held Online. INCOMA Ltd, pp. 483–493. Available at https://aclanthology.org/2021.ranlp-1.55 (accessed September 14, 2022).

Heidenreich T, Lind F, Eberl J-M and Boomgaarden HG (2019) Media framing dynamics of the 'European refugee crisis': a comparative topic modelling approach. *Journal of Refugee Studies* **32**, i172–i182.

Hoyle A, Goel P, Hian-Cheong A, Peskov D, Boyd-Graber J and Resnik P (2021) Is automated topic model evaluation broken? The incoherence of coherence. Ranzato, M, Beygelzimer, A, Dauphin, Y, Liang, P. S., Wortman Vaughan, J (Eds.), *Advances in Neural Information Processing Systems*. Vol. **34**. Curran Associates, Inc, pp. 2018–2033.

Humphreys L, Lewis NA Sender K and Won AS (2021) Integrating qualitative methods and open science: five principles for more trustworthy research. *Journal of Communication* **71**, 855–874.

Jacobi C, van Atteveldt W and Welbers K (2016) Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism* **4**, 89–106. doi:10.1080/21670811.2015.1093271.

Jacobs AZ and Wallach H (2021) Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT'21. New York, NY: Association for Computing Machinery, pp. 375–385.

Kirk J and Miller ML (1986) *Reliability and Validity in Qualitative Research*. Vol. **1**. Newbury Park: Sage.

Krippendorff K (2013) *Content analysis: an introduction to its methodology*. 3rd ed. Los Angeles [u.a.]: Sage.

Laver M, Benoit K and Garry J (2003) Extracting policy positions from political texts using words as data *American Political Science Review* **97**, 311–331.

Lowe W (2008) Understanding wordscores *Political Analysis* **16**, 356–371.

Lucas C, Nielsen RA, Roberts ME, Stewart BM, Storer A and Tingley D (2015) Computer-assisted text analysis for comparative politics. *Political Analysis* **23**, 254–277.

Maier D, Niekler A, Wiedemann G and Stoltenberg D (2020) How document sampling and vocabulary pruning affect the results of topic models *Computational Communication Research* **2**, 139–152, https://computationalcommunication.org/ccr/article/view/32.

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

Margolin DB (2019) Computational Contributions: A Symbiotic Approach to Integrating Big, Observational Data Studies into the Communication Field *Communication Methods and Measures* **13**, 229–247.

Nelson LK (2020) Computational Grounded Theory: A Methodological Framework *Sociological Methods & Research* **49**, 3–42.

Reiss MV, Kobilke L and Stoll A (2022) Reporting Supervised Text Analysis for Communication Science. *Conference Presentation DGPuK Jahrestagung der FG Methoden*. München, Germany.

Roberts ME, Stewart BM, Tingley D, Lucas C, Leder-Luis J, Gadarian SK, Albertson B and Rand DG (2014) Structural Topic Models for Open-Ended Survey Responses *American Journal of Political Science* **58**, 1064–1082.

Scharrer E and S Ramasubramanian (2021) *Quantitative research methods in communication: the power of numbers for social justice*. *Routledge Social Justice Communication Activism Series*, Routledge, New York.

Song H, Tolochko P, Eberl J-M, Eisele O, Greussing E, Heidenreich T, Lind F, Galyga S and Boomgaarden HG (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**, 550–572.

Tolochko P, Balluff P, Bernhard J, Galyga S, Lebernegg NS and Boomgaarden HG (2024) What's in a name? The effect of named entities on topic modelling interpretability *Communication Methods and Measures* 1–22.

Wallach H (2018) Computational social science computer science + social data *Communications of the ACM* **61**, 42–44.

Watanabe K and Zhou Y (2022) Theory-Driven Analysis of Large Corpora: Semisupervised Topic Classification of the UN Speeches *Social Science Computer Review* **40**, 346–366.
Ying L, Montgomery JM and Stewart BM (2022) Topics, Concepts, and Measurement: A Crowdsourced Procedure for Validating Topics as Measures *Political Analysis* **30**, 570–589.
Zhao H, Phung D, Huynh V, Jin Y, Lan D and Buntine W (2021) Topic modelling meets deep neural networks: A survey.

Cite this article: Bernhard-Harrer J, Ashour R, Eberl J-M, Tolochko P and Boomgaarden H (2025) Beyond standardization: a comprehensive review of topic modeling validation methods for computational social science research. *Political Science Research and Methods*, 1–19. https://doi.org/10.1017/psrm.2025.10008