Pääkkönen, Juho; Ylikoski, Petri

# Humanistic interpretation and machine learning

*Published in:*
SYNTHESE

*DOI:*
10.1007/s11229-020-02806-w

Published: 18/11/2021

*Document Version*
Publisher's PDF, also known as Version of record

Synthese
https://doi.org/10.1007/s11229-020-02806-w

OBJECTIVITY IN SOCIAL RESEARCH

# Humanistic interpretation and machine learning

**Juho Pääkkönen**[1,2] · **Petri Ylikoski**[1,3]

Received: 31 October 2019 / Accepted: 22 July 2020


**Abstract**
This paper investigates how unsupervised machine learning methods might make hermeneutic interpretive text analysis more objective in the social sciences. Through a close examination of the uses of topic modeling—a popular unsupervised approach in the social sciences—it argues that the primary way in which unsupervised learning supports interpretation is by allowing interpreters to discover unanticipated information in larger and more diverse corpora and by improving the transparency of the interpretive process. This view highlights that unsupervised modeling does not eliminate the researchers' judgments from the process of producing evidence for social scientific theories. The paper shows this by distinguishing between two prevalent attitudes toward topic modeling, i.e., topic realism and topic instrumentalism. Under neither can modeling provide social scientific evidence without the researchers' interpretive engagement with the original text materials. Thus the unsupervised text analysis cannot improve the objectivity of interpretation by alleviating the problem of underdetermination in interpretive debate. The paper argues that the sense in which unsupervised methods can improve objectivity is by providing researchers with the resources to justify to others that their interpretations are correct. This kind of objectivity seeks to reduce suspicions in collective debate that interpretations are the products of arbitrary processes influenced by the researchers' idiosyncratic decisions or starting points. The paper discusses this view in relation to alternative approaches to formalizing interpretation and identifies several limitations on what unsupervised learning can be expected to achieve in terms of supporting interpretive work.

**Keywords** Humanistic interpretation · Topic modeling · Machine learning · Objectivity · Text analytics · Latent Dirichlet allocation

✉ Juho Pääkkönen
juho.paakkonen@helsinki.fi

1 Sociology, University of Helsinki, Helsinki, Finland

2 Computer Science, Aalto University, Espoo, Finland

3 Institute for Analytical Sociology, Linköping University, Norrköping, Sweden

Published online: 07 September 2020

## 1 Introduction

The objectivity of interpretive text analysis—humanistic interpretation[1]—has been a hot potato in the social sciences since their beginning. The necessity of humanistic interpretation has been generally recognized, but many have retained their suspicions about the sources of bias that could influence the interpretive process. Thus attempts have been made to formalize the interpretive process to make it more transparent and to control some possible biases. These attempts have met with opposition. In particular, formal approaches based on coding have been argued to be limited in terms of replicability and in their ability to account for nuances in textual meaning. At worst, coding procedures have been argued to impose interpretation on text data, distorting their underlying meaning structures, and barring evidence important for forming a well-grounded interpretation (Biernacki 2012a).

The development of machine learning-based tools for text analysis has initiated the most recent debate about the objectivity of humanistic interpretation. While techniques based on supervised machine learning are thought to share the same problems as their coding-based ancestors, unsupervised machine learning seems to promise something new. Instead of imposing interpretation on texts, unsupervised text analysis has been argued to *delay* the moment of subjective engagement with texts until evidence of word patterns has been uncovered (Lee and Martin 2015a; Mohr and Bogdanov 2013). This computer-aided distant reading is not claimed to replace humanistic close reading of texts, but to *support* subsequent interpretive work by providing evidence for interpretations in a manner that is not only much more scalable but also less subject to biases that derive from the interpreters' preconceptions.

In this paper, our aim is to articulate an account of whether and in which sense this unsupervised approach could be said to improve the objectivity of interpretive text analytics. We do so through a close examination of the uses of unsupervised learning in the social sciences. More precisely, we will scrutinize prevalent ways of using topic modeling—a popular unsupervised approach among social scientists—to analyze how it is used to support interpretation. We will distinguish between two attitudes toward the results of topic modeling, namely *topic realism* and *topic instrumentalism*, and argue that under neither is the role of unsupervised learning to eliminate the researchers' judgments from the process of producing evidence from text data. The use of topic modeling always depends on interpretive engagement with the primary materials on the part of the researcher. This sets limits to the scalability of the unsupervised approach and to the use of modeling results as evidence for social scientific claims. Consequently, we argue that the primary ways in which unsupervised learning supports interpretation is by enabling analyses to draw on information that would otherwise be inaccessible and by making the interpretive process more transparent and

[1] The term "humanistic interpretation" is used in the debate we cover to refer to traditional social scientific interpretation that does not employ coding or formalization (Biernacki 2012a, 2014). This paper does not discuss text analysis *in the humanities*, only interpretive text analysis in the social sciences that works according to certain practices often associated with humanistic research.

systematic. Throughout, we illustrate our argument with examples of topic modeling drawn from social scientific literature.[2]

While discussions on the methodology of unsupervised text analytics are motivated by concerns with objectivity, the word rarely occurs in them. As Hacking (2015, p. 20) notes, "objectivity" is an elevator word that is used for a semantic ascent from ground-level questions. In this paper, our focus is very much on the ground-level discussions, and thus we are not interested in how social scientists, in our case cultural sociologists, use the word "objectivity." It is also our view that issues pertaining to substantial methodological problems are best grasped without reading them through the lens of some given philosophical account of objectivity. We agree with Wright (2018, pp. 389–395), who argues that abstract concepts such as objectivity can be useful for clarifying methodological debates by collecting together criteria for comparing different methods. However, to serve this purpose, we hold that accounts of objectivity should be based on a thorough examination of the substantive ground-level issues involved in each of the methodological practices under discussion.

Following this idea, we first investigate ground-level issues in how unsupervised methods can support interpretation and return to the issue of objectivity mostly toward the end of this paper. We argue that, in the context of text analytics, objectivity of interpretation is enhanced when there are resources to justify that one's interpretation is right and it is possible for others to reach the same conclusions given the same evidence. This makes the acceptance of the interpretation less dependent on blind trust in the competence and interpretive fairness of the researcher. This view of objectivity does not amount to mere replicability of the unsupervised analysis or to a mechanical control of the researcher's judgments (cf. Daston and Galison 1992) but rather has parallels with analyses of objectivity as being grounded in interactive debate (Douglas 2004; Longino 1990). In the debate over the meaning of texts, unsupervised learning enhances objectivity by reducing suspicions that particular interpretations are the products of *arbitrary* processes influenced by the researchers' idiosyncratic preferences, theoretical starting points, or methodological decisions. As Hacking (2015, p. 26) puts it, objectivity is not a virtue but the proclaimed absence of this or that vice (Wright 2018, p. 388). However, we also argue there are clear limitations to what machine learning methods, such as topic modeling, can achieve in regard to mitigating such interpretive failings. The main point of this paper is to provide an account of objectivity that is sufficiently well grounded on the social scientific uses of unsupervised learning to help analyze the role and limitations of modeling in these interpretive processes.

This paper thus connects together three literatures. It starts with a recent debate in cultural sociology, then takes a look at discussions around topic modeling methods in the social sciences, and in the end, connects these to philosophical discussions about objectivity. The debate in cultural sociology builds around Richard Biernacki's (2012a, 2014, 2015) spirited defense of traditional humanistic interpretation and Lee and Martin's (2015a, b) challenge to that position. This debate identifies the core issues and provides us a well-articulated present-day account of humanistic interpretation. To have a real-life example of unsupervised machine learning methods, we dip into

[2] While our examples mainly concern modeling of text documents such as newspaper articles, our argument could be extended to cover also materials produced by researchers, such as interviews or open-ended survey responses (e.g., Roberts et al. 2014).

methodological discussions about topic modeling, which is by far the most popular unsupervised machine learning method in the social sciences. As noted, our motivation for looking at the details of topic modeling is to get a grasp of the actual roles this method plays in social scientific research. While machine learning has often been noted to enable interpreters to analyze corpora and draw on information that would otherwise be inaccessible, we have not seen systematic accounts of how modeling in fact supports these processes and how this relates to hopes of improving objectivity. It is our belief that only a good understanding of ground-level details makes it possible to make reliable philosophical observations about "objectivity" and understand the limits of machine learning methods in interpretive research. By engaging with topic modeling as a case, we are able to reinvigorate long-standing issues about objectivity and interpretation in a new and interesting context that has yet to receive the attention of philosophers.[3] As we hope to show in this paper, the introduction of machine learning methods has led to a dislocation of established practices of interpretation in the context of social scientific text analytics. This provides an opportunity to rethink many interesting issues at the core of interpretive research.

The remainder of this manuscript is organized as follows. In the next section, we introduce the debate behind the idea that unsupervised learning makes text analysis more objective by delaying interpretation. Then, in Sects. 3 and 4, we introduce topic modeling as a case of unsupervised text analytics and distinguish between the realist and instrumentalist attitudes. Section 5 argues that neither attitude regards topic modeling results as evidence for social scientific claims in isolation from substantial interpretive engagement with the modeled text materials. When topic modeling can provide evidence for theoretical interpretations, this rests upon a prior interpretive understanding of the contents of the modeled corpus. Lacking such understanding, modeling amounts to interpretive exploration of texts with the help of a computational tool. Sections 6 and 7 then discuss how topic modeling can support interpretive work. This happens by enabling analysts to draw on more comprehensive materials and to uncover information about word patterns that would be inaccessible without the unsupervised method (Sect. 6). Further, the computational procedure facilitates making interpretive decisions explicit and amenable to inspection and criticism by others (Sect. 7). Both advantages apply only under limited circumstances, which we

[3] Treatments of machine learning in the philosophy of science have tended to remain on a general level and to gloss over the details of how different methods are used in practice. One issue in literature concerns the implications of machine learning for the inductive method (e.g., Williamson 2009; Thagard 1990). Another strand focuses on the ethical implications of using black-box algorithms in academic research and beyond (e.g., de Laat 2018; Krishnan 2019). Sullivan (2019) presents a recent discussion which draws on philosophical accounts of modeling and the case of neural networks to argue that model opacity does not pose a problem for gaining understanding from machine learning. While these discussions are important for understanding how earlier methodological work in philosophy relates to machine learning methods, they remain quite unconnected to how these methods are actually used in different domains. In particular, to the best of the authors' knowledge, philosophers of social science have not studied the use of machine learning in social scientific text analytics, despite the growing popularity of these methods. One exception is Stuart (2019), who discusses the possibility of automated discovery in ethnographic research. Albeit interesting in its own right, this question is largely irrelevant for our purposes. As explained below, the extant practice of using machine learning in social scientific text analytics depends inextricably on human interpretation. As such, the pertinent issues in this context do not bear on the conceptual conditions of fully automated discovery.

discuss in the respective sections. Finally, Sect. 8 concludes our discussion by linking our main observation back to ideas about the objectivity of interpretation.

## 2 The debate about the objectivity of interpretation

The role of interpretive text analysis has been long debated in the social sciences. On the one hand, contextually sensitive interpretation has been recognized as necessary for developing nuanced accounts of social phenomena (see, e.g., Rabinow and Sullivan 1979 for a defense). On the other, reconciling the subjective underpinnings of interpretation with the evidentiary standards of the social sciences has proven difficult.[4] Thus, the history of social scientific text analytics has seen recurrent attempts to build objectivity into interpretive methods (Denzin and Lincoln 2011). In this paper, we focus on unsupervised machine learning as one of the most recent of such attempts. To see what the stakes of unsupervised learning are in terms of the objectivity of text analysis, we first have to discuss some alternative approaches.

What we refer to as *humanistic interpretation* is the hermeneutic close reading of text materials—such as newspaper articles, political manifestos, or literary works—with as much sensitivity to contextual detail as possible. The starting point in humanistic interpretation is that the meaning of texts cannot be settled a priori. The aim in humanistic interpretation is to form an understanding of the meaning of the analyzed text in relation to information about what kind of text it is taken to be (e.g., news, fiction, a book review), by whom, where it was published, during what time, surrounded by what kind of events, what kind of conceptions, and so on. This process is necessarily open-ended, not based on a predefined system for processing evidence, and thus essentially driven by the judgments of the researcher. This is not to say that humanistic interpretation completely lacks guiding standards. However, the criteria which guide the interpretive process are typically loose by design—such as coherence, breadth, or analytical sophistication of claims (Biernacki 2014, pp. 338–340)—to retain enough leeway for the analyst to grasp the open-ended and complex nature of textual meanings. Accordingly, while the validity of humanistic interpretation has traditionally been regarded to depend on consensus emerging in collective debate over evidence (Hirsch 1967), no strict rules exist for deciding which particular reading of a text is correct.

Broadly speaking, the difficulties associated with humanistic interpretation in the social sciences can be summed up in three related worries. First, the open-ended hermeneutic process is typically *non-transparent*. It is practically impossible to present in a traceable manner all relevant evidence and decisions involved in an interpretation. This leads humanistic interpretation to presuppose a lot of trust in the skills and background knowledge of the interpreter. She should have extensive knowledge

[4] For instance, the cultural sociologist Jeffrey Alexander et al. (2012, p. 21) express this difficulty as follows: "…while thick descriptions of meaning have often produced virtuoso interpretations, such deep readings tend to finesse the issue of breadth. How representative are these beliefs? Which groups and individuals subscribe to them? Have the best interpreters been selective in their reading of the evidence? Are the authors of spectacular accounts actually making sense of empirical meanings, or are they merely good writers who are skilled at packaging highly idiosyncratic and subjective views?"

of the context but, at the same time, be open-minded to be able to identify novelty. She should also be patient, pedantic, and honest in her procedures, while retaining her interpretive imagination. Thus one can easily doubt the virtuosity of a particular interpretation without denying the general possibility of objective interpretation. For instance, lack of transparency makes it difficult to assess whether evidence for an interpretation has been cherry-picked to confirm the researcher's preconceptions or theoretical expectations (Baker and Levon 2015). Second, humanistic interpretation suffers from *underdetermination* by evidence, in the sense that plausible but diverging interpretations can often easily be articulated on the basis of the same materials without clear criteria to choose between them (Jones 1998). This happens especially when contextual information about the interpreted materials is inaccessible, when the cultural distance between texts and researchers is great, or when researchers work with different samples or interpretive aims. Finally, close reading is *limited in scalability*, being confined to work with small datasets from local settings. This hampers the generalizability of claims (Williams 2000) and can lead to arbitrary sampling when working with large materials (cf. Moretti 2000)—a problem that is arguably becoming increasingly acute due to the recent proliferation of digital sources of textual data (Tangherlini and Leonard 2013). Deep humanistic interpretations with widely varying and comprehensive source materials depend on the sole efforts of virtuoso researchers, whose analyses are difficult to assess or follow by others not as well-versed in the particularities of the approach (Lee and Martin 2015b, p. 406).

The prevalent way to mitigate these problems has been to formalize parts of the interpretive process. While varying approaches to formalization exist (Mohr and Rawlings 2012), the common idea is to specify an explicit system of procedures which the researcher follows to arrive at an interpretation. Most often, this has taken the form of systematic procedures for *coding* text passages into a set of meaningful categories, either provided by an underlying theoretical framework or generated through comparisons between texts (Glaser and Strauss 1967). The aim in coding is not to eliminate subjective interpretation in text analysis but to constrain the researcher to work in accordance with systematic procedures, which will facilitate the evaluation of interpretations by explicitly specifying the steps that produced them.[5]

Despite their popularity among social scientists, coding approaches have also met with opposition. As a recent example, Biernacki (2012a) has argued forcefully that constraining interpretation to predefined procedures does not mitigate the problems of transparency and underdetermination but rather leads to the violation of the open-ended nature of meaning. Biernacki based this claim on his failed attempt to replicate the sampling and coding processes of three seminal studies in cultural sociology (Griswold 1987; Bearman and Stovel 2000; Evans 2002). In each case, Biernacki (2012a, p. 128) found unsystematic application of coding categories to confirm pregiven theoretical viewpoints as well as neglect contextual information that could have compromised the interpretation. He argued that this is a principal problem with any interpretive process that relies on coding: it hides interpretive decisions behind seemingly neutral

[5] That coding retains an element of subjective interpretation is testified to by approaches which emphasize the constructed nature of categorizations (e.g., Charmaz 2006; Braun and Clarke 2013). Also these approaches provide systematic procedures which enable researchers to specify *how* they arrived at their interpretations and *which* text passages support their interpretive claims.

procedures that in reality break evidentiary connections between the interpretation and the primary materials. In other words, according to Biernacki (2012a, pp. 5–8), coding avoids open-endedness in interpretation by fixing an arbitrary system which assumes a priori that the texts are comparable to each other according to some shared criteria. This leads researchers to gloss over nuanced information and allows *imposing interpretation* on texts by forcing them to fit into prespecified categories. Biernacki (2014) concludes that humanistic craft interpretation—despite its challenges—is *epistemologically superior* to coding-based content analysis because the latter misses the causal processes that generate the materials to be analyzed and frustrates the retrieval of nuanced meanings. Further, he argues that humanistic interpretation is also more traceable and replicable, and it preserves the referential ties to contextual information needed to challenge any proposed interpretation.

Perhaps unsurprisingly, Biernacki's claims have spurred a critical debate in cultural sociology.[6] We find Biernacki's criticism of coding instructive but inconclusive. Indeed, many social scientists would undoubtedly dispute the claims made against coding as too strong. Biernacki's attempt to replicate the three coding studies, albeit seminal, does not provide a general demonstration of the interpretive unviability of coding. Further, even if social scientists were to abandon coding in favor of humanistic interpretation, the above discussed issues of transparency, underdetermination, and scalability of interpretation would still remain (Lee and Martin 2015b). This means that attempts to improve objectivity would continue to be motivated, and methodological development would likely continue to aim toward systematizing interpretive processes. Despite these reservations, we believe that Biernacki revealed an interesting tension between the attempt to make interpretation in text analytics *mechanically objective* (Daston and Galison 1992)—in the sense of systematizing interpretive processes to make them replicable and less dependent on individual judgment—and what might be called the *substantial objectivity* of the interpretive product, i.e., the actual contextual meaning of the text. The aim behind systematization through coding was to control the interpretive process and to reduce suspicions that interpretations are the products of a researcher's particular biased judgments. The formalized system for arriving at interpretations seeks to provide the researcher with resources to show that one's interpretation is not arbitrary. Given that the procedures for producing interpretations are accepted, researchers could (in principle) show others that their interpretations are supported by the original text materials. Formal systems thus aim at tackling the problem of objectivity in interpretive work by laying bare the exact procedures which yield the evidence for any particular interpretation. The challenge in doing so, identified by Biernacki, is that formalized procedures simultaneously *constrain interpretation* and prevent it from grasping the nuanced information that is required for well-grounded accounts of textual meaning. The worry is that formalization ends up producing evidential artifacts tied to the particular prespecified system and not rigorous interpretations based on deep understanding of the original materials. Indeed, Biernacki's argumentation can be given a stronger reading, according to which *all* formal approaches to text analysis *in principle* share the same problem (Biernacki 2012b, 2015). It is this

[6] See Lee and Martin (2015a) and the subsequent issue of *American Journal of Cultural Sociology* edited by Jeffrey Alexander (Biernacki 2015; Reed 2015; Spillman 2015; Lee and Martin 2015b).

stronger claim that has become the gist of subsequent critical debate; And it is here that unsupervised machine learning enters the debate, as a solution to the problems associated with coding.

To see how unsupervised methods are thought to improve upon coding, it is useful to contrast them to another branch of machine learning text analytics, namely *supervised* methods. The aim in supervised text analytics is to categorize text documents using a classification scheme prespecified by the researcher. The machine learning algorithm is first provided with a subset of texts which have been manually categorized by the researcher in accordance with the classification scheme. Using this training data set, the algorithm records how features of the text documents—such as words used in them—relate to their manually assigned categories. This process produces a model of the data which, after training, can be used to infer categorizations for other documents in the collection.[7] Thus, the aim in supervised text analytics is to learn and extend as accurately as possible the researcher's *coding* of a subset of texts (Nelson et al. 2018). By contrast, in *unsupervised* learning, text documents are not categorized according to a scheme specified by the researcher. These methods model the data by *counting* features directly in the full collection of texts. For instance, if one text frequently uses political vocabulary and another uses words related to education, they are allocated to different categories without any a priori information about what those categories might mean, simply on the basis of word occurrences.[8] This inductive working of unsupervised methods is thought to give them the advantage over coding-based approaches. In contrast to coding, which constrains interpretation to work in accordance with an a priori scheme, unsupervised methods are argued to introduce a formal procedure for generating evidence that *delays* the moment of interpretation in analysis to a phase after modeling (Mohr and Bogdanov 2013, p. 560; Lee and Martin 2015a, p. 14).

The issue in this debate thus is not whether unsupervised machine learning could do away with humanistic interpretation. The claim of the proponents of machine learning is that unsupervised methods enable researchers to (1) retain the open-endedness of hermeneutic interpretation in text analysis *while* (2) supplementing it with a mechanically objective procedure for distant reading texts (cf. Moretti 2013), which (3) can generate evidence in a manner that alleviates the shortcomings of unaided humanistic close reading.[9] The idea is that unsupervised methods can not only improve the scalability of interpretive research but also make the analysis process more transparent and less subject to underdetermination by enabling researchers to *show rather than tell* their audience how their interpretive claims are supported by the original materials. Central to this is the notion that counting enables researchers to delay interpretation, which enables intersubjective scrutiny over evidence:

[7] Features used for constructing the model need not be words but can also include metadata information about the documents, such as their type or author. However, words are perhaps the most commonly used feature in interpretive text analytics with machine learning, given the aim of understanding the meaning of text *contents*.

[8] See Grimmer and Stewart (2013) and Schwartz and Ungar (2015) for extensive discussions of different supervised and unsupervised methods of text analytics in the social sciences.

[9] This approach has been variously labeled in literature as "hermeneutic modeling" (Mohr and Rawlings 2012) or "semantic modeling" (Ignatow 2015). See Mohr (1998) for an early account of such formal approaches to meaning.

> When it comes to formal analyses, we might say that bad sociologists code, and good sociologists count. The reason is that the former disguises the interpretation and moves it backstage, while the latter delays the interpretation, and then presents the reader with the same data on which to make an interpretation that the researcher herself uses. Even more, the precise outlines of the impoverishment procedure is explicit and easily communicated to others for their critique. And it is this fundamentally shared and open characteristic that we think is most laudable about the formal approach. (Lee and Martin 2015a, p. 24.)

Despite these alleged strengths, it remains unclear how unsupervised methods can in fact support interpretive work and in what sense this could be said to make interpretation more objective. For one, as critics have pointed out (Reed 2015), while counting words as such might not involve interpretive judgments, there is a prior decision to be made about how exactly the words are to be counted. For instance, how should one account for differences in meaning when a particular word has different uses across documents? More generally, unsupervised learning comprises a wide variety of methods for text analysis, with a multitude of diverging modeling approaches within each family of methods. It would be strange to claim that decisions between these methodological approaches should *not* be influenced by researchers' preinterpretive understanding of the relevant documents. In fact, the researcher has to choose which documents to include in the analysis, so it is very difficult to see how unsupervised machine learning methods could completely remove the role of preconceptions. Finally, as Biernacki (2015) has argued, it is unclear whether formal representations can really address the problem of underdetermination. What guarantees that the patterns produced through counting words can limit the range of diverging interpretations to any notable degree? Why should we think that interpreters can any more easily agree over the meaning of word counts than the meaning of the original texts?

The relevance of these questions for the issue of objectivity is also vindicated by parallel discussions about computational methodology in the context of digital humanities. Many digital humanities scholars have been decidedly critical about the idea that computational methods could mitigate human interpretive involvement in text analysis (Ramsay 2005; Clement 2013; Gibson and Ermus 2019). Instead, this debate has focused on articulating how these "reading machines" (Ramsay 2011) relate to more traditional close reading of texts (Bonfiglioli and Nanni 2016). The literature studies scholar Earhart (2015) has argued that tools for distant reading texts might introduce their own kinds of biases into the reading process, in addition to the "cultural biases" that traditionally influence the work of the literary critic. However, as methodological decisions may always introduce biases into the interpretive process, the crucial issue is not *whether* methods influence or constrain interpretation but what are the possibilities for recognizing and controlling biases in interpretive processes. In the context of the history of ideas, Betti and van den Berg (2016) have argued that all interpretive processes—hermeneutic interpretation included—are always influenced by the conceptual frameworks that researchers employ in analysis. In their view, the problem with established interpretive methodology is that this influence is often implicit and not recognized, and thus the preconceptions that guide analysis cannot be subjected to rigorous empirical scrutiny. In this sense, computational methods might be superior

to traditional close reading—and by extension, perhaps also to coding-based analyses—because their particular biases might be easier to detect than those embedded in established interpretive practices (Earhart 2015). Thus, the emerging debate about computational methods for text analysis provides a space for critically assessing the biases of interpretive practices of humanistic scholarship more broadly. In a similar vein, the recent critical attention given to formal methods in the social sciences provides an opportunity to reassess interpretive methodology more generally. As in the digital humanities context, the issue is not the necessity of humanistic interpretation in formal analyses, but a better understanding of interpretive practices and of how the involvement of formal methods could improve them.

However, it should also be noted that the practice and aims of computational text analysis are often different in the digital humanities than in social research. For instance, Betti and van den Berg (2016) argue that to guard against the implicit influence of preconceptions, computational modeling of texts should always be driven by an explicit conceptual model prespecified by the researchers. This idea is markedly different from that of delaying interpretation with unsupervised learning. In Betti and van den Berg's view, computational modeling should effectively be used as a method for applying the researchers' conceptions to text data. By contrast, the role of unsupervised learning in social research is closer to data modeling (Suppes 1962). In social scientific text analytics, unsupervised models are used for producing regimented representations of complex text corpora that help researchers identify which textual patterns have evidential value for answering substantive theoretical questions. In such use, the role of modeling cannot be to straightforwardly apply a chosen conceptual framework to the texts. Instead, the philosophically interesting questions in this context concern the different roles given to formal modeling and interpretation in the process of generating evidence for theories. Further, while modelers in the social sciences are typically interested in studying social phenomena for which text data provide evidence, digital humanities often investigates textual patterns in themselves. Such work focuses, for instance, on tracking the prevalence of certain keywords over time (e.g., Ramsey and Pence 2016) or on predicting the authorship of unfamiliar texts not included in the literary canon (Bonfiglioli and Nanni 2016). In such cases, issues concerning the relationship of computational results to the substantive phenomena studied are not similarly pressing. Finally, as Kaltenbrunner's (2015) study of a large digital database in literary studies shows, the predominant challenges posed by digital text data in the humanities often concern issues such as how to develop a database ontology which adequately reflects the respective field's expert understanding of relevant issues (Buckner et al. 2011). These issues are orthogonal to the aim of mitigating biases in social scientific textual interpretation, which we focus on below.

We contend that the best way to address the issue of objectivity in unsupervised learning is to examine how these methods are used in social scientific research to support interpretations. With this intent, the next sections analyze the use of the most popular unsupervised machine learning method—topic modeling—in social scientific text analysis.[10] On the basis of this case, we argue that unsupervised learning cannot be

[10] Topic modeling has also been extensively applied and discussed in the digital humanities (see, for instance, the *Journal of Digital Humanities* 2012 volume 2, issue 1 dedicated to topic modeling). Indeed,

taken to alleviate underdetermination in interpretation. If there is a sense in which the unsupervised approach can improve objectivity, then this primarily happens through improving scalability and transparency.

## 3 Topic modeling

Topic modeling is a family of unsupervised methods for discovering hidden thematic structure in text collections (Blei 2012a). Topic models were initially developed for information retrieval in large unstructured collections. During the past decade, the method has become increasingly popular in the social sciences, with applications to substantive research problems such as cross-national comparison of research agendas in academic journals (Marshall 2013), tracing prevalent framings in art discourse (DiMaggio et al. 2013; Roose et al. 2018), and examining topical similarity between the communications of parties and political movements (Stier et al. 2017).[11]

The most common variant of topic modeling methods in the social sciences is known as latent Dirichlet allocation (LDA).[12] This method works by taking a collection of text documents as input and modeling them with a generative Bayesian process which tries to predict words occurring in the documents. To do so, LDA assumes that the documents are made up of a fixed number of *topics*, which they each exhibit to a varying degree. Technically, topics are probability distributions over the vocabulary of all words occurring in the documents, but intuitively, they are often described as lists of words which frequently occur together in the modeled documents (Blei 2012b). The generative process of LDA works as follows: for each word in each document, a topic is first randomly sampled from the document's respective topic distribution; then, a word is randomly sampled from the word distribution of that topic; and this routine is repeated until all words in all documents in the collection have been predicted. The generative process begins with a random initialization for both topic and word distributions, but iteratively tunes these parameters to maximize predictive accuracy through repeated rounds of Bayesian inference. The resulting distributions of document–topic and topic–word probabilities are finally returned as a model of the document collection.[13]

Footnote 10 continued
many topic modeling contributions to cultural sociology—the case we focus on—share an important lineage with fields such as literary history (e.g., Tangherlini and Leonard 2013; Jockers and Mimno 2013). Cultural sociology can be regarded as a borderline field, methodologically rooted in both sociological and humanistic text analytics. This is also the reason why issues of the objectivity of interpretation are particularly well visible in this context (Alexander et al. 2012). While we recognize this dual lineage, below we focus on discussing cases of topic modeling from social research because the debate between formalization and humanistic interpretation was framed within this context.

[11] For a use of topic modeling in philosophy of science, see Malaterre et al. (2019).

[12] LDA topic modeling was originally introduced by Blei et al. (2003). See Blei (2012a) for an accessible discussion of the probabilistic family of models to which LDA belongs. Mohr and Bogdanov (2013), Maier et al. (2018), Grimmer and Stewart (2013), and Blei (2012b) provide introductions to topic modeling in the contexts of cultural sociology, communication research, political science, and digital humanities, respectively. Isoaho et al. (2019) includes a good discussion of different topic modeling variants and their use in qualitative political research.

[13] The name "latent Dirichlet allocation" derives from this generative process, which uses the Dirichlet distribution to discover topics, regarded as latent variables influencing word distributions in the data. LDA

**Table 1** Topics and their five most probable words

| Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 |
|---|---|---|---|---|
| Government | Percent | Police | I | People |
| Soviet | Million | Court | Bush | Two |
| United | Year | State | President | Officials |
| States | New | Two | New | Air |
| President | Billion | Case | House | City |

**Table 2** Prevalence of topics in documents

| Document | Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 |
|---|---|---|---|---|---|
| $d_1$ | 0.051 | 0.051 | 0.617 | 0.144 | 0.137 |
| $d_2$ | 0.349 | 0.196 | 0.249 | 0.131 | 0.075 |
| $d_3$ | 0.056 | 0.075 | 0.498 | 0.230 | 0.141 |
| … | | | | | |
| $d_D$ | 0.657 | 0.069 | 0.099 | 0.058 | 0.117 |

Tables 1 and 2 below depict what typical LDA results look like.[14] To use LDA for interpretive research, the lists of words which constitute topics will have to be assigned an interpretation by the researcher. Typically, this happens by examining the most probable words in each topic and reading the documents with the highest probability of exhibiting the given topic. As Blei (2012b) notes, these results are useful for grasping the thematic content of a document collection because sets of frequently cooccurring substantive terms that "tend to be about the same subject." Further, the document–topic distribution can be used to draw additional inferences about the thematic structure of the corpus, such as the prevalence of words belonging to different topics at different times.[15] To facilitate the interpretation and display of topic models, various visualization packages have been developed, providing users with additional information and options for exploring the results (e.g., Sievert and Shirley 2014).

LDA topic modeling has been extensively applied in the social sciences and is generally regarded as providing an accessible, fully replicable, and transparent method for representing large collections of documents (Nelson 2017).[16] The method can handle

Footnote 13 continued
is based on what is known as the "bag-of-words" assumption, according to which the order of words in the documents makes no difference for the generative process. Thus, intuitively, LDA draws on Dirichlet distributions to "allocate" each word in the documents, represented as bags of words, to latent topic variables. See Blei (2012a) for a more extensive discussion.

[14] These figures were produced by running LDA with five topics on the Associated Press example dataset with the R package "topicmodels" (Grün and Hornik 2011). The words and topics in the tables were selected for display solely for illustrative purposes.

[15] Later variants of topic modeling facilitate examining the dependency of topics on each other and on metadata variables (Blei and Lafferty 2006; Roberts et al. 2019; see Isoaho et al. 2019 for discussion).

[16] While LDA involves random processes, modeling can be made replicable by controlling the pseudo-random number generator responsible for initializing the document–topic and topic–word distributions. Further methods also exist for evaluating result robustness upon changes in initialization (Roberts et al. 2016).

massive amounts of text data, while also being applicable to relatively small datasets (e.g., thousands of newspaper articles; Roose et al. 2018). Finally, the computational procedure underlying LDA is well understood, and clear documentation exists for different implementations of the algorithm (e.g., Grün and Hornik 2011). Thus, the method enables systematic exploration of the influence of different modeling choices on the results. Moreover, these can be easily replicated and expanded upon by the wider researcher community due to the availability of accessible software packages which do not presuppose extensive technical understanding.[17]

Replicability and transparency are crucial features, given that topic modeling—like most unsupervised methods—depends on a number of steps that must be performed by the researcher as modeling proceeds. First, the texts must be *preprocessed* to make them amenable for computational analysis (Denny and Spirling 2018; Schofield and Mimno 2016). Second, the researcher has to set three modeling parameters. The first two of these are Bayesian prior parameters, which control the shape of the Dirichlet distributions used in modeling (cf. Wallach et al. 2009), and the third is the number of topics to be generated. The question of how to select the value of this parameter has been extensively discussed in literature. While computational methods exist for selecting the optimal number of topics in terms of predictive accuracy (Greene et al. 2014), experimental evidence suggests that predictively optimal topics can be difficult to give a coherent semantic interpretation (Chang et al. 2009). Consequently, social scientific use tends to emphasize result *interpretability* as the relevant model selection criterion (e.g., Mohr and Bogdanov 2013). While computational methods for assessing the interpretability of topics have been developed (Mimno et al. 2011), the "ground truth" in model selection remains the researcher's assessment of how well the modeling results address the given research problem at hand (DiMaggio et al. 2013). Because of these decisions, the importance of *validating* the modeling results has been heavily emphasized in literature (Grimmer and Stewart 2013).

The LDA procedure is unsupervised in that it does not draw on any researcher-provided semantic categories in constructing topics.[18] The patterns generated by the method are based on statistical associations between words alone and, as such, do not draw on interpretive involvement on the part of the researcher (see, however, Schmidt 2012). However, the multitude of decisions that the modelers have to make in curating data, selecting modeling parameters, and validating their models already shows that it is far from clear what objectivity could amount to in relation to these methods.[19] In addition, as we argue below, social scientific uses of LDA involve different attitudes

[17] As we discussed below, this is both an advantage and a potential source of problems.

[18] Also a supervised variant of LDA exists. It uses the topic structure of a set of preclassified documents to infer classification for further documents based on topics identified in them (Blei and McAuliffe 2007).

[19] This range of decisions involved might raise a suspicion that topic modeling is not a good case for investigating the issue of objectivity. Indeed, some unsupervised methods for text clustering seemingly involve fewer choices than LDA. For instance, density-based clustering methods group texts solely on the basis of their distance in vector space and, as such, do not require that the researcher specify the number of clusters to be identified. However, this does not mean that these methods would not involve decisions that the researchers cannot make in a data-driven manner. In general, all unsupervised methods for text analytics depend on preprocessing, model selection, and validation that cannot be built into the methods in any fully automated fashion. Topic modeling is a suitable case for investigating objectivity in this context because the method has been associated with high hopes of enabling data-driven analyses in social research. However,

toward the method. To argue that unsupervised learning does not alleviate underdetermination, in the next section, we first distinguish between two different attitudes which have implications for how the modeling results support interpretation.

## 4 Topic realism and topic instrumentalism

A central challenge in computational text analysis geared towards interpretation is to articulate how the modeling results relate to meaningful phenomena that social scientists are interested in investigating (see Marres 2017)—that is, issue framings and agendas and the phenomena underlying them, such as discourses or relations of power. Social scientists who apply unsupervised methods are interested in producing models of text data that could help them identify the textual patterns that have evidential value for substantial theoretical questions. As seen in the previous section, this task is as complicated as any kind of data modeling, involving a host of decisions to be made on the part of the researcher. However, the context of interpretive text analysis also involves the additional difficulty that analysts will have to coordinate their interpretive reading of texts with the formal modeling process. Ignatow (2015, pp. 106–108) has noted that, in practice, computational text analysts tend to subscribe to an assumption that textual meanings are "real social entities with qualities and causal powers" which can be analyzed using formal methods. However, analysts are also acutely aware that texts are open to multiple interpretations and that, in formal analysis too, meanings are construed in part by the researcher (e.g., Mohr and Rawlings 2012). Consequently, the role of modeling in interpretive research is not clear, and approaches in computational text analytics involve varying combinations of both formal modeling and interpretive reading.

This tension between the faith in independently existing evidentially relevant patterns in data and the necessity of interpretive engagement is particularly visible in topic modeling, and this is what makes it a philosophically interesting case for discussing objectivity. In social scientific uses of topic modeling, two attitudes towards the role of formal modeling in interpretation can be distinguished, hereafter called *topic realism* and *topic instrumentalism*. Topic realism is the view that the modeling process can capture and produce representations of various social scientific theoretical constructs—such as frames, discourses, or narratives—as they *actually exist* in the modeled text materials. That is, for topic realism, there is a match between the mechanically objective modeling procedure (the LDA generative process) and the defining features of a theoretical concept, which enables the method to *operationalize* and *measure* that concept in the texts.

The topic realist attitude is illustrated by the study of DiMaggio et al. (2013), who used LDA to measure how the framing of public support for arts evolved in major US news outlets during 1986–1997. They argue that the topics produced by LDA, by uncovering cooccurrence patterns between substantive terms in documents, may be taken to "index" discursive environments which correspond to the communication

Footnote 19 continued
it has been applied extensively enough to provide ample material for analyzing how these hopes are squared with the various researcher-dependent decisions necessarily involved in all unsupervised modeling.

research concept of frames (2013, p. 593). As a concept, frame denotes "semantic contexts" that associate certain substantive notions or ideas with an issue of interest (2013, p. 578). For instance, arts funding can be framed negatively by discussing it in connection to politically controversial events or positively by mentioning it in "connection with happy news about enjoyable or edifying exhibitions and performances" (2013, p. 574). To measure such framings, DiMaggio et al. used LDA to generate 12 topics from a dataset of 7958 news articles, 3 of which they identified as particularly interesting "conflict" frames. Through examining the prevalence of different frames over time, they showed that the percentage of words belonging to the conflict frames sharply increased during 1988–1989, in contrast to noncontroversial framings (2013, p. 594).[20]

This approach contrasts with what we call topic instrumentalism. Under topic instrumentalism, modeling is not taken to measure theoretical constructs but instead to provide information about word patterns, which can be *usefully employed to guide subsequent interpretation of the primary text materials*. This view is explicitly propounded by, for instance, Törnberg and Törnberg (2016, p. 407), who "choose not to collapse topic modeling with any theoretical concept, but rather to use it as it was initially intended: as a tool for inductive empirical categorization." Similarly, Jacobs and Tschötschel (2019, p. 3) argue that equating LDA results with theoretical concepts is misleading, since in reality, "topics are clusters of words that reappear across texts," the interpretation of which "as themes, frames, issues, or other latent concepts (such as discourses) depends on the methodological and theoretical choices made by the analyst," In this vein, Törnberg and Törnberg (2016) used LDA to guide their critical discourse analysis of the connections between Islamophobia and antifeminism on the Swedish discussion forum Flashback.[21] They ran topic modeling with 20 topics on a corpus of 12,796 posts, sampled using keywords connected either to feminism or Islam. Their aim in doing so was to identify posts where Islam and feminism are discussed together. Modeling arguably could provide them with just such information, given that the topic word lists are representations of words which frequently occur together in the documents. Thus, through examining topic word lists, they identified 17 topics as including vocabulary about both Islam and feminism and proceeded to conduct a critical discourse analysis of the most representative documents in the topics. As one of the main results of this analysis, they argued that gender inequality is used as a discursive strategy on Flashback to criticize Islam (Törnberg and Törnberg 2016, pp. 417–418).

Topic realism and topic instrumentalism form a continuum of attitudes towards topic modeling, not a contrast between considered philosophical positions. For instance, Mohr and Bogdanov (2013, p. 547) seem to assume the topic realist attitude in suggesting that there exists an actual number of topics in a text corpus and a corresponding

[20] For other examples of measuring frames with topic models, see Fligstein et al. (2017) and Schnable (2018). For measurement of other concepts than frames, see for instance, the study of issue homophily by Schmidt-Petri et al. (2018), word polysemy in DiMaggio et al. (2013), and literary themes in Jockers and Mimno (2013).

[21] See also Light and Cunningham (2016), Miller (2013), Baier and Gengnagel (2018), and Roose et al. (2018) for other examples of topic modeling which accord with topic instrumentalism.

**Table 3** Topic realism and topic instrumentalism

| | Topic realism | Topic instrumentalism |
|---|---|---|
| Role of modeling | Measures social scientific theoretical constructs as they actually exist in text materials | Provides an organizing representation for subsequent interpretation |
| Role of interpretation | Model evaluation, validation | Model evaluation, in-depth analysis of primary materials |

*correct* model which properly captures the underlying thematic structure of the texts.[22] Simultaneously, however, topic realist modeling often holds that, while LDA can measure theoretical constructs, the meanings represented by those constructs can occur at different levels of "granularity" (e.g., Greene et al., 2014, p. 498). Accordingly, DiMaggio et al. (2013, pp. 582–583) argue that the decision over different modeling parameters is ultimately tied to the researchers' analytical interest and an interpretive assessment of which level of granularity in representation most clearly depicts the intended constructs in the corpus.[23] Nevertheless, even though they at times emphasize that topic modeling is an exploratory approach which provides a "lens" for viewing text collections (DiMaggio et al. 2013, p. 582), their overall aim is to use modeling for measuring theoretical constructs as they actually occur in the corpus. Thus it is possible that researchers flip-flop between realism and instrumentalism, even within the same paper. It seems that social scientists express topic realist attitudes when they are optimistic about the prospects of topic modeling doing some of the interpretive heavy lifting, but when the implicit interpretive steps are challenged or when the multiplicity of alternative models is raised, they tend to fall back to a more modest instrumentalist attitude that highlights the exploratory value of the method.

Both topic realism and topic instrumentalism can be argued to be compatible with the idea that unsupervised learning allows delaying interpretation (see, e.g., DiMaggio et al. 2013, p. 577; Törnberg and Törnberg 2016, p. 405 for statements to this effect). The relevant difference between the attitudes concerns the respective roles that are given to the modeling procedure and subsequent interpretive reading, not the idea of delaying interpretation. Crucially, while in topic realism the primary role of interpretation is to evaluate and validate the measurements of the modeling process, in topic instrumentalism, the substantive part of the analysis is performed through interpretive reading of the text materials. Despite this difference, presented in Table 3, both attitudes are often associated with optimism about the prospects of topic modeling in making interpretation more objective.

In the next section, we argue that under neither topic realism nor topic instrumentalism can the modeling process be taken to eliminate the researchers' judgments from the process of producing evidence for social scientific theories. This implies

[22] See also Isoaho et al. (2019, p. 7), who argue that incorrectly setting preprocessing and modeling parameters may lead to the model mistakenly representing actually distinct topics as single entities.

[23] This seems to be the predominant approach taken in social scientific topic modeling. Even studies which draw on computational metrics for evaluating modeling parameter optimality usually combine them with the researchers' final assessment of result interpretability (e.g., Bail et al. 2017).

that topic modeling cannot be understood as improving the objectivity of humanistic interpretation by mitigating the problem of underdetermination. Further, the necessity of interpretive engagement with the original text materials sets limitations to the scalability of the unsupervised approach.

## 5 Evidential role of unsupervised learning

Recall that the main advantage of unsupervised methods over coding, as argued by Lee and Martin (2015a, p. 24), is that counting enables researchers to bring into debate the *same data* which they themselves used as the basis for their interpretation. Provided with shared data to be used as evidence, debate over divergent interpretations no longer has to be based on the authority of the interpreters' hermeneutic readings. Coding failed in this respect, according to Biernacki, because systematization of the interpretive process in fact did not prevent researchers from imposing their particular a priori viewpoints on the materials. Furthermore, the connections to original data were lost as coding products replaced the original texts in further analyses. By delaying interpretation in analysis altogether, unsupervised learning could arguably overcome these problems. Thus, unsupervised learning forces researchers to articulate how their interpretations are consistent with evidence uncovered by modeling. However, as noted in the previous section, the pertinent question in this context is *what is the evidential role of modeling in the interpretation?* The distinction between topic realism and topic instrumentalism allows us to treat this question systematically. According to topic realists, modeling results have an evidential role because they provide the information relevant for deciding whether the interpretive claims made are correct. Topic instrumentalists disagree.

To see the difference between these two positions, let us start with the instrumentalist attitude. In Törnberg and Törnberg's (2016) study of Islamophobia and feminism, formal modeling did not have an evidential role, all work was done by discourse analysis. It was the authors' interpretive reading of the text materials that yielded, for instance, their claim about gender inequality as an anti-Islamist discursive strategy. The role of topic modeling was to point to subsets of data where Islam and feminism were discussed together, but these results in themselves did not support the interpretive claims. Of course, the model results did provide information about *which* documents the authors should read and *what kind of content* they could expect to find there. The topics provided guidance about where discourses about feminism and Islam could be expected to be present in the materials. However, it was only through subsequent interpretive reading that Törnberg and Törnberg identified and analyzed the different discourses present within that subset of documents.

More generally, the topic instrumentalist attitude allows the formal model to guide the interpretative reading of the original materials, but it does not have an evidential role. On this view, the patterns uncovered by topic modeling do not represent anything substantive over and above word occurrence regularities. Another way to put this is that, under topic instrumentalism, the researchers do not take the model results to provide evidence of theoretical constructs. The upshot is that, upon disagreeing over interpretations, researchers will have to turn to evaluating each others' readings of

the text materials to defend their claims. In topic instrumentalism, modeling in itself cannot show any particular interpretation to be better than another. It can only indicate where the evidence relevant for evaluating interpretations could be found.[24]

Topic realism is different in this respect. In topic realist measurement, the aim is to confirm that modeling results *qua* modeling results provide evidence of certain theoretical constructs. As modelers cannot prima facie know for certain that their chosen model correctly captures the intended phenomenon, they need additional ways of assessing the modeling results. This is achieved via validating the model. As Nelson (2017, p. 28) argues, the underlying aim of validation in computational modeling is to ensure that "the identified patterns are not an artifact of a specific algorithm, or are based on a biased interpretation of the output and deep reading." Validation seeks to check against a trusted "ground truth" that the modeling results can be interpreted as providing evidence of the intended constructs. Thus, also DiMaggio et al. argue (2013, p. 596) that, upon successful validation, the modeling results can be "used reflexively as evidence about the state of the world."

This does not imply that modeling is the only source of evidence. In this vein, DiMaggio et al. (2013) argue that the kind of evidence produced by modeling essentially depends on the kind of validation used. For instance, to confirm that topics capture word *polysemy*—or relationality of word meaning with respect to context—it is necessary to *semantically validate* the modeling results by checking that occurrences of the same word in different topics in fact correspond to uses of the word in different senses (DiMaggio et al. 2013, pp. 586–590). This involves close reading of documents from different topics and interpreting the meanings they give to certain words of interest. In addition, to validate that topics operationalize certain kinds of frames, it is necessary to confirm that they behave in the materials as expected. One way of doing so is through close reading the materials. This was the approach taken, for instance, by Fligstein et al. (2017, p. 890), whose use of topic modeling results for measuring how the 2008 financial crisis was framed in economic regulators' meetings depended heavily on the authors' close interpretation and familiarity with the modeled meeting transcripts. Another way is via *external validation*, where topics are compared against information about known external events. For instance, DiMaggio et al. (2013, pp. 593–596) checked that the prevalence of their conflict frames increased in the materials as expected during times of known controversies about arts funding.

This variety of approaches indicates that there is no single obvious ground truth against which the results of unsupervised modeling can be compared. Validation always requires "a certain amount of creativity" (Nelson 2017, p. 31) and skill in combining different methods available for evaluating interpretations (Grimmer and Stewart 2013, p. 271). Most pertinently, validation depends on the *researchers' understanding of how the measured construct should behave in the modeled collection of texts*. In this

[24] Naturally, researchers can challenge each others' modeling choices and use a different model or method to guide their readings. For instance, Törnberg and Törnberg's claims could be challenged by selecting a different number of topics and consequently arguing that, in the most representative documents of those topics, gender inequality was not used as a discursive strategy or was used differently. The important thing is that the evidence relevant for challenging the original claims is produced by interpretive reading. Thus, under topic instrumentalism, modeling does not do away with the need to engage in debate over interpretive readings, although this debate now also has to take into account modeling decisions.

vein, DiMaggio et al. (2013, p. 603) emphasize that topic modeling requires "domain expertise on the part of the analyst" and that any "effort to apply topic modeling to a corpus to answer interpretive questions must include a subject-area specialist on the team." Similarly, Mohr and Bogdanov (2013, p. 560) argue that topic modeling results should always be evaluated by a "well informed observer (a subject-area specialist) who understands the discursive context of the corpus." We take this to mean that, for topic realist measurement, an *interpretive background understanding* of the modeled corpus is always required for using the modeling results as evidence for substantial social scientific claims.

Consider, for instance, semantic validation of polysemy. To validate that topics capture context-dependent meanings of a word, researchers need an understanding of what the word means across documents representative of different contexts (that is, topics). This involves not only examining how the word is used across documents but also investigating what constitutes differences between the varying contexts of word use. That is, to understand the meaning of different uses of the word, researchers need to know what kind of text it is that they are reading, written and read by whom, in what kind of outlet, and so on. This calls for hermeneutic interpretation of the modeled materials. This point is also illustrated by external validation of frames. Here, ground truth seems to reside in brute correspondence between topic prevalences and events external to the text collection. For instance, given a topic that is taken to operationalize controversial framings of arts funding, the measurement of that frame's prevalence in the materials is validated if the prevalence increases at times of known public controversies about arts funding. Also this comparison is essentially based on a prior assessment of how the topic should behave in relation to the known events given that the topic is interpreted as a frame. Again, an understanding of the "discursive context of the corpus" is needed.

This necessity of interpretive background understanding has two implications for the evidence that we can take topic realist measurement to provide for interpretation. First, because the use of modeling results as evidence rests on an interpretative understanding of the corpus, there can be substantive disagreement over interpretations by researchers with divergent background knowledge, expertise, or theoretical aims. In this vein, user studies suggest that topic models are interpreted differently by users with different levels of expertise (Lee et al. 2017). Likewise, we should not assume a priori that researchers with divergent theoretical starting points or different understandings of a corpus will come to the same interpretive conclusions with the same model. This implies that topic realism has no stronger claim to resolving underdetermination than topic instrumentalism. Also, in topic realist measurement, disagreement over diverging interpretations will ultimately have to turn on arguments about the researchers' understanding of the text materials. The model has evidential relevance only against the backdrop of a consensual agreement over the meaning of a well-understood corpus.

Second, this suggests that topic realist measurement is limited in scalability. When working with large and poorly understood corpora, topic realist modelers face the challenge of coming to grasp the context of the corpus before they can validate their measurements. A related point is that the materials need to be sufficiently homogeneous so as to enable measurement in terms of a certain prespecified concept. For instance,

as Rhody (2012) has argued, topic modeling is poorly suited to measuring "themes" in texts which contain a lot of figurative language, such as poetry. Interestingly, this means that, when data are large or of high variety, topic instrumentalist exploration provides a way to develop a background understanding of the corpus. Topic instrumentalism is not similarly limited by the need to validate the modeling results as evidence of theoretical constructs. Of course, under topic instrumentalism too, models have to be evaluated for their usefulness in guiding subsequent interpretation. For instance, a model which does not capture any terminology related to Islam or feminism would presumably be judged to be a bad guide to studying discursive connections between those phenomena. However, as the aim is not to use modeling results as evidence, models can be treated more as tools for reading texts (Ramsay 2011) than as evidence for deciding between interpretations. A corollary of this is that there seems to be no in principle reason to stick with just one model or method in topic instrumentalism. The analyst can simultaneously employ multiple different methods for representing the corpus, given that they each provide different and possibly useful entry points for interpretive analysis.

To summarize, we argued above that under neither topic realism nor topic instrumentalism can modeling results be taken as evidence for adjudicating between social scientific claims without the researchers' engagement with the primary text materials. In particular with respect to topic realism, we argue that the use of the model as evidence rests on an interpretive background understanding of the modeled corpus. This implies that unsupervised learning cannot be claimed to make interpretation objective in the sense of mitigating underdetermination. Modeling might be able to uncover information about word use patterns, but the use of these data as evidence for interpretations is possible only against the backdrop of a sufficiently shared understanding of the modeled corpus and the modeling decisions made. Unsupervised learning does not make interpretation mechanically objective. In the next two sections, we argue that, despite this, modeling can help researchers to ground their interpretations on more diverse text materials and to draw on information in the texts that would otherwise be inaccessible. Further, modeling can greatly facilitate systematic collective scrutiny over the resulting interpretations. This is the way in which unsupervised learning can improve the objectivity of interpretation.

## 6 Expanding the hermeneutic horizon

As discussed in Sect. 2, Biernacki argued that formalization prevented researchers from grasping nuanced information in texts that is required for well-grounded interpretations (what we referred to as substantial objectivity). However, one of the primary shortcomings of humanistic interpretation is that the method is limited in scalability. Traditional hermeneutic interpretations of large or very diverse materials can be produced only by virtuoso interpreters whose approach is inaccessible to follow or imitate by less experienced researchers. Furthermore, even virtuoso interpreters are forced to work with relatively small samples of materials drawn from local settings. The question of how to select the samples becomes pressing when data are available from many different sources or when the volume becomes so large that forming an understand-

ing of the corpus through unaided interpretation is impossible. Even in cases where the researcher can go through all the material, there is a possibility that the order in which the texts are read affects the interpretation, i.e., the materials read later get lesser evidential weight as the interpretation has already settled.

In relation to these problems, unsupervised learning can support hermeneutic work in two primary ways: First, unsupervised methods provide access to information in the texts that is difficult or impossible to grasp through unaided interpretation; Second, they enable the analysts to deal with comprehensive materials from a wide variety of sources. A model produced by machine learning provides an overview of the materials that is not biased by arbitrary sampling, reading order, or just by reader exhaustion.[25] In these regards, unsupervised learning can help make hermeneutic interpretation more objective through *expanding* it to draw on more comprehensive and diverse materials, and to take into account information in the materials that otherwise would go unnoticed. Let us look at these two ways to *enhance objectivity* more closely.

The key advantage that topic modeling is thought to have over simpler quantitative approaches such as coding or keyword frequency analyses is that this method does not require the researchers to prespecify which words or categories are central for capturing a given phenomenon (e.g., DiMaggio et al. 2013, p. 577). In this sense, topic modeling can be regarded as a method for exploring large corpora that is importantly *independent* from the researchers' conceptions of the modeled contents. It is technically possible to model a corpus in its entirety using a range of parameters before having read a single document or having any idea about their contents. Thus, the method enables the *discovery of unanticipated information* in text data in a manner not possible for coding or keyword-based approaches. Further, computational modeling provides a perspective into the data that is *different* from that of the analysts' viewpoint—such as identifying large-scale word-use patterns that are hard to detect due to the limited scale of close reading or subtle differences in the use of a word across contexts that analysts may fail to notice but perfectly systematic machines are able to pick up. It is important to note that both topic realists and topic instrumentalists can employ these ideas. For instance, Törnberg and Törnberg (2016, p. 404) maintain that data-driven exploration with topic modeling enables identifying "small but systematic patterns and tendencies that may not be visible to the naked eye when restricted to small-n studies." When used for measurement, DiMaggio et al. (2013, p. 593) hold that the unsupervised approach facilitates the "discovery of unanticipated frames, and distinguishing between different uses of the same term."

This also allows us to clarify the sense in which it is possible to say that unsupervised learning delays the interpretation. As argued in the previous section, the evidential role of modeling always rests on an interpretive background understanding of the modeled materials. Nevertheless, topic modeling is often described as enabling researchers to explore text collections "without prior manual analysis or a priori assumptions" (Jacobs and Tschötschel 2019, p. 9). We can now see that such statements should be understood as referring to the computational process of modeling word cooccurrences rather than the process of producing evidence using the model. It is the computational

[25] It should also be noted that the overview provided by machine learning makes possible collaborative work in a manner that is unimaginable in the context of traditional humanistic interpretation. This allows scaling up of close readings of texts. See the next section for more details.

procedure of counting words rather than the broader process of evaluating and validating the results of different models as evidence that is independent from the researchers' understanding of the materials. Although subsequent interpretation draws on the modeling results, the modeling process itself is meaning-agnostic. It does not eliminate the researchers' judgments from the process of producing evidence, but it does produce information about statistical word patterns in a manner that is independent from the researchers' preconceptions of the modeled materials. Further, the meticulous process of uncovering these regularities in even small corpora is very different from the way in which human interpreters ordinarily engage with texts. Consequently, unsupervised modeling can discover information in text data that is surprising and hard or even impossible to detect for human analysts and can thus challenge the researchers' expectations about the materials.

Topic modeling also demonstrates how unsupervised learning can help with the issue of sampling large or diverse materials. The model provides an overview of the corpus which can be used to regiment interpretive engagement with the texts. Typically, topic modelers focus their interpretation on topics that are most interesting from their particular analytical perspective, reading the most representative words and documents of these topics (see Schmidt 2012 for a criticism of this practice). For instance, in this vein, DiMaggio et al. (2013) focused their analysis on the "conflict frames." Likewise, Törnberg and Törnberg (2016) engaged in reading only from topics where Islam and feminism overlapped. To take another example, Roose et al. (2018) used topic modeling to identify what they called "modernist" and "contemporary" themes in 25 years worth of arts magazines. These examples show how modeling enables researchers to identify the subset of materials *within the full collection* that has the greatest evidentiary value with respect to their particular analytical interests. These analyses can be supplemented with metadata information about the documents to test how different kinds of documents relate to the topics of interest (e.g., DiMaggio et al. 2013, pp. 599–601). Modeling thus enables rigorous sampling and provides researchers a way to argue for their particular selection of documents to read from among the full collection. This can lower the threshold for researchers to approach large text collections, as their analyses no longer have to be based on unaided virtuoso reading of vast and complex materials or sampling choices that are difficult to justify substantially.

These examples show how unsupervised learning can both mitigate the scalability limitations of unaided humanistic interpretation and enable access to information about texts that otherwise might have gone unnoticed. These expansions of the hermeneutic horizon are possible because the unsupervised process works on statistical word patterns, which is a very different approach than that taken in traditional interpretive work. In other words, modeling enables researchers to recognize and take into account information in the data that unaided interpretation might have glossed over. Furthermore, modeling can recognize patterns in vast and diverse collections of texts. Thus it can create access points for researchers who otherwise would have no means of grasping the full complexity of the materials. Unsupervised learning can thus potentially improve the *substantial objectivity* of interpretation by providing a more thorough grasp of the information in texts than unaided hermeneutic work can achieve. The idea here is that having more—and novel—evidence enhances the chances of getting the interpreta-

tion right. Naturally, this only happens when the products of unsupervised learning are additions to traditional humanistic interpretation but not, for example, when the use of machine learning methods leads to dramatically reduced interpretative engagement with the source materials.

## 7 Objectivity through transparency

As argued in Sect. 5, social scientists who employ unsupervised text analytics must work under an unavoidable "interpretive uncertainty" concerning the meaning of their modeling results (DiMaggio 2015, p. 2). This does not imply that no systematic methods exist for evaluating interpretations. Validation is one such way of circumscribing interpretive text analytics. However, given that validation also depends on interpretation, further ways are needed to probe the *interpretive process itself*. Unsupervised methods can importantly support collective scrutiny of the interpretive work that is needed to turn modeling results into evidence through allowing for increased transparency with respect to many interpretive decisions. As Lee and Martin (2015a, p. 24) note, this "explicit and easily communicated" nature of formal modeling is a key advantage over unaided interpretation, where explication of the hermeneutic process presents a challenge.

Topic modeling provides ample examples of how unsupervised learning facilitates collective scrutiny over interpretations. First, as already noted in the previous section, the modeling results provide an explicit criterion for sampling documents from the topics for further interpretive analysis. Through specifying which documents are most likely to contain words from any given topic, modeling guards against cherry-picking text passages in support of particular a priori viewpoints or preconceptions (Törnberg and Törnberg 2016, pp. 408–409; Jacobs and Tschötschel 2019, p. 14). Through this information included in the document–topic distribution (Table 2), the model provides exact quantification of how representative each topic is in relation to the full collection of texts. This helps evaluate the generalizability of claims and to contextualize them in relation to corpus-level information about the materials. Modeling thus helps make explicit the scope of the interpretive claims and the data that support any particular claim.

Importantly, all representations produced by topic modeling—such as the topic word lists, the prevalence of topics in the materials, and possible further analyses performed on these—can be easily visualized and shared. In this vein, for instance, Light and Cunningham (2016) used network representations of correlations between topics at different times to track the thematic evolution of Nobel Peace Prize speeches. Similarly, Baier and Gengnagel (2018) employed principal component analysis to construct a two-dimensional representation of topic correlations, which they then used to analyze discursive similarities between European Research Council (ERC) grant applications. Thus, topic modeling facilitates displaying different corpus-level representations of the materials, which can then be supplemented by close readings of texts selected on the basis of explicit criteria (Jacobs and Tschötschel 2019, p. 14). This kind of computational scaffolding for interpretive work improves transparency not only by helping researchers be more explicit about how their particular interpretations relate

to the materials but also by making important parts of the analysis process replicable. If different researchers have the exact same data and use the same software implementations of topic modeling with the same parameter settings and text preprocessing steps, they can reproduce each other's modeling results with perfect accuracy.[26] Replicability alleviates the worry associated with coding processes and unaided humanistic interpretation that the categorization of documents is driven by the preconceptions of the researcher or influenced by the order in which the texts are read. Through replicating the modeling procedure, others can formulate their own interpretations of the same materials and compare these against the original claims. This process can help researchers locate the documents that bear most strongly on the differences between their interpretations, given particular modeling results. Further, it can help researchers contest each other's categorization of the documents by making explicit the modeling decisions that were made in producing them.[27] This is a clear advantage over manual coding approaches, where systematically explicating and reproducing the text categorizations of others is a cumbersome and sometimes impossible task (Biernacki 2012a).

The improved explicitness and communicability of text categorization does not imply that the entire interpretive process become more transparent. After all, as argued above, the modeling process is always assessed through the researchers' interpretive engagement with the primary text materials. However, modeling can facilitate collective scrutiny of how researchers develop their interpretive understanding of the modeled materials by dividing textual interpretation into more easily evaluated tasks. Focusing interpretation to work with clearly delineated parts of the materials facilitates systematic comparison between potentially divergent interpretations. For example, teams of experts can work independently from each other to evaluate topics to assess the extent to which their interpretations differ (Maier et al. 2018). Some studies have even crowdsourced the process of topic evaluation to broaden the range of interpretive viewpoints (Stier et al. 2017). Such teamwork is greatly supported by the modeling results which help researchers coordinate the collective process of reading texts. For instance, on the basis of the document–topic distribution, they can decide to focus their reading on the ten most representative documents from each topic and then collectively settle for an order in which to analyze them. To be sure, such coordination is not impossible in coding-based analyses, either. However, the point is that the modeling results can be easily shared between a group of researchers and enable them

[26] Moreover, the details about modeling are easy to share in code repositories on platforms such as GitHub. For an example, see the replication code shared by Laura Nelson as companion to her topic modeling analysis of the literature produced by feminist movements in the US during the nineteenth and early-twentieth centuries (2017): https://github.com/lknelson/computational-grounded-theory. By running the program code with the data provided in this repository, using the software specifications indicated in the readme file, researchers can exactly reproduce the modeling results that ground Nelson's analysis.

[27] In this sense, modeling processes which force researchers to explicitly make a number of decisions fare better in terms of transparency than methods where many decisions are made "under the hood." The multitude of decisions to be made in many topic modeling implementations, for instance, place the burden on the researchers to openly articulate the assumptions that underlie their parameter and preprocessing choices. As Biernacki (2015) argues, formal methods which seemingly produce results without any researcher involvement are merely rhetorical tools that hide the fact that texts are always analyzed from a particular perspective, based on a particular approach to counting.

to easily specify explicit criteria that guide the reading process, such as document representativeness. Further, these processes can be scrutinized through comparing the interpretations of different researchers not only against each other but also against quantitative metrics of topic semantic coherence (e.g., Bail et al. 2017). Finally, the interpretations can be assessed by subjecting them to robustness testing upon changes in modeling parameter values or preprocessing steps (e.g., DiMaggio et al. 2013; Fligstein et al. 2017; Denny and Spirling 2018). Each of these procedures enabled by topic modeling can be explicitly articulated and communicated to others to facilitate scrutiny of the interpretive process. Thus, unsupervised modeling provides a range of possibilities for systematically coordinating collective debate over interpretations, which improves over coding-based approaches that typically base their evaluation on simple intercoder reliability assessments.[28]

Based on the above observations, it should be clear that the way in which unsupervised learning improves transparency is *relative to* interpretive processes which traditionally have relied on unaided human judgment. In comparison with unaided interpretation, modeling enables researchers to explicate more precisely how their interpretive judgments are grounded in the materials and how exactly the categorization that guides their analysis was arrived at. Further, through forcing researchers to explicate their decisions about how to model materials, unsupervised methods can help them become more conscious about the points at which their preconceptions might come to play in the analysis. To give one more example, DiMaggio et al. (2013, online supplement pp. 2–4) explained that, to measure frames, they initially produced 5 different models with the number of topics varying between 8 and 12. The 12-topic solution was selected through interpretative assessment of how clearly the different models depicted "government arts support as problematic" (online supplement p. 2).[29] It is precisely in this sense that we understand the claim made by Lee and Martin (2015a, p. 14) that unsupervised modeling can "shift the burden of proof dramatically" in interpretive debate. Through laying out explicitly the criteria which modelers use to ground their interpretive claims, the burden is shifted to others to challenge those criteria.

However, it is important to note that whether this will turn out to be helpful for subsequent interpretive debate depends on the participants' understanding of the tech-

[28] It has also been suggested that a rigorous way to evaluate the results of unsupervised learning is to compare them with the researchers' independent coding of a subset of the modeled materials (e.g., Grimmer and Stewart 2013; Isoaho et al. 2019). In view of our arguments in this section, we maintain that this approach can facilitate transparency only to the extent that the coding procedures adopted can be made explicit and replicable. Sanctioning the results of unsupervised learning with coding that itself is not circumscribed by the modeling results must find a way to independently answer Biernacki's criticism.

[29] They also compared this model with other solutions using as a metric the discrepancy between the topic word distributions and how words in the materials were actually distributed in relation to different news outlets (DiMaggio et al. 2013, pp. 597–599; see Mimno and Blei 2011 for further explanation of the technique). LDA is based on the assumption that all words allocated to a given topic are drawn independently from the same underlying distribution (a multinomial parameterized with the respective topic–word distribution) (Blei et al. 2003). This assumption is violated if an external variable—such as the news outlet which published the article—influences how words are in fact distributed in the materials. Using this criterion to compare different solutions, DiMaggio et al. found the 12-topic model to be superior in terms of capturing thematic content independently of news outlet, which further supported their model selection (online supplement, pp. 3–4).

nical aspects of topic modeling. Indeed, our intention is not to argue against the often repeated claim that machine learning methods are opaque to many users (Burrell 2016). By talking about transparency in the context of interpretive text analytics, we do not wish to suggest that *unsupervised models* are somehow unproblematically interpretable. Our point is rather that the reasons underlying human interpretive judgments are often notoriously difficult to explain, and in an important sense, the workings of unsupervised modeling can be easier to communicate to others and subject to collective assessment. This improved explicitness does not automatically dissolve disagreements, but it can facilitate collective scrutiny over interpretations and thus improve the transparency of *the interpretive process*. However, this also implies that the sense of transparency we associate with unsupervised methods cannot be reduced to merely displaying one's analysis code to others. The transparency of the interpretive process also necessitates the existence of a sufficient degree of *shared understanding* about how modeling works. Machine learning processes that most participants in the interpretive debate do not understand will not be useful for increasing interpretive transparency.

This emphasis on the transparency of the interpretive process does not imply that complex or opaque models could not produce reliable data from text materials. As philosophers have argued in the epistemology of simulation literature (Winsberg 2019), modelers can rely on complex and epistemically opaque (Humphreys 2009) computational processes for data generation, given that their reliability and performance have been sufficiently tested by the relevant expert communities (Barberousse and Vorms 2014; Symons and Alvarado 2019). Similarly, data modeling and exploration techniques (e.g., instrumentalist uses of topic modeling to obtain information about word patterns) can be relied on to produce optimal outcomes even when their users do not comprehensively understand the statistical details of the method. As we have seen above, social scientists employ unsupervised learning to produce models of text data (Suppes 1962) that help them identify which word patterns have evidential value for answering substantial theoretical questions. Although the evidential role of modeling depends on interpretive engagement, what is important for the reliability of computational pattern discovery is that the unsupervised methods be demonstrated to work as intended, using well-understood test datasets or for instance simulated data. In this sense, the situation is similar to that of observational instruments. As Bogen and Woodward (1988, pp. 331–333) argue, the reliability of scientific instruments can be established through independent tests, although their users might not have a systematic explanation for their operation. Our claim about transparency thus is not that social scientists should become experts in statistical learning to rely on unsupervised methods. What we mean is that participants in the interpretive debate will have to share a sufficient understanding of what is implied by different modeling decisions (e.g., preprocessing, choice of the number of topics) to be able to effectively criticize each other's claims. This understanding is analogous to knowing how to operate a microscope—a craft skill that researchers can have without being able to systematically explain how the instrument works.

As Symons and Alvarado (2019) have argued in the context of simulations, mere craft skill of individual modelers will not be sufficient for modelers to be able to evaluate and trust each other's analyses (cf. Hubig and Kaminski 2017). Modelers might

have different understandings of their modeling procedures, which can at worst lead to breakdowns in communication. Thus there needs to be a common recognition of some established practices that the modelers can refer to in explaining their decisions to each other. This is what we mean by the claim that interpreters need to have a sufficiently *shared* understanding of modeling processes. For instance, there is some evidence that using word stemmers as a preprocessing step in topic modeling with English language texts does not significantly improve the statistical quality of topics under a number of metrics (Schofield and Mimno 2016). Indeed, in the social sciences, word lemmatization is often recognized to be a more sophisticated—albeit computationally more demanding—method for vocabulary reduction in English (e.g., Grimmer and Stewart 2013). To understand the relevance of each other's preprocessing decisions, topic modelers need to be aware of such recommended practices. As Symons and Alvarado (2019, p. 54) have argued, developing established practices that can ground trust in the results of computational models in any given substantive field is a long and gradual process. In the case of unsupervised learning in the social sciences, this process has only very recently begun, and the best practices are in many contexts still being negotiated.

Finally, another important precondition of transparency is the access to the modeled materials (cf. Lee and Martin 2015a, pp. 22–23; DiMaggio et al. 2013, p. 577). For collective scrutiny over interpretations to be possible, other researchers need to be able to access and replicate analyses using *the same materials* that grounded the original claims. Transparency is not a feature intrinsic to computational modeling but depends on the extent to which the authors are able to share and document their materials, modeling decisions, and subsequent interpretive steps in the analysis.[30] When the modeled materials cannot be openly shared or when the analysis process is inadequately documented or otherwise opaque to interpreters in debate, the unsupervised formal modeling becomes suspect.

We maintain that the improved transparency, exemplified above by topic modeling, is a further sense in which unsupervised learning can be taken to make humanistic interpretation more objective. This *objectivity through transparency* is based on the ability to *show* how the researcher has reached her interpretation. Unsupervised methods enable researchers to specify explicitly and exactly to others how and where in the analysis their interpretive engagement with the materials influenced the analysis. This can facilitate collective evaluation of their particular starting points and additionally enables others to replicate the modeling and formulate their own interpretations. The transparency enhances the trustworthiness of the analysis by reducing the space available for speculation about the processes by which the interpretation might have been reached.[31] It makes it easier to identify when the interpretation is a product of a particular researcher's idiosyncratic viewpoint. Unsupervised machine learning thus

[30] See Jockers and Mimno (2013) for a case of topic modeling where parts of the modeled materials came from proprietary sources that prevented opening the dataset.

[31] However, Symons and Alvarado (2016) have argued that the increasing reliance on software brought about by computational methods can significantly compromise the reliability of knowledge claims (Symons and Horner 2014). The extensive software infrastructure and multitude of packages that are required for collecting and modeling even moderately large digital materials hamper the researchers' capability to systematically test for errors in their analyses. In this sense, the situation in computational text analytics is arguably similar to simulation modeling, where software complexity and the necessity of case-specific

does not improve objectivity by eliminating biases or preconceptions from the analysis but reduces worries about their invisible influence. This openness also creates incentives for the researcher to be more reflective about the possible sources of bias as the interpretive choices are more visible to attentive readers. The objectivity afforded by increased transparency is thus based on the ability of others to check how an interpreter arrived at her claims and to see whether their own interpretations agree.

## 8 Do unsupervised machine learning methods enhance the objectivity of interpretive research?

In this paper, we take a close look at the ground-level issues related to the use of unsupervised machine learning methods in support of interpretive research and connect them to concerns about objectivity in interpretive research. We use the most popular unsupervised method in the social sciences—topic modeling—as an example, but we believe our observations have more general relevance. To show this, in this section, we connect our discussion of topic modeling to philosophical discussions about objectivity. Let us begin by summarizing our argument.

In Sects. 4 and 5, two quite different attitudes towards the products of topic modeling are identified: topic realism and topic instrumentalism. According to the former, topic modeling makes it possible to measure and produce representations of social scientific theoretical constructs—such as frames, discourses, or narratives—as they actually exist in the modeled text materials. According to the latter, topics are just statistical patterns of word occurrence that can be usefully employed to guide subsequent interpretation of the primary text materials. Topic instrumentalism thus is not tied to any particular theoretical starting point. In practice, this amounts to the following difference: a topic realist aims to find and validate the optimal topic model, while a topic instrumentalist is happy to work with multiple different models if they provide useful insights about the primary materials. Another important difference between these two stances concerns the evidential role of the topic models. According to topic realists, modeling results have an evidential role in deciding whether the interpretive claims made are correct. Topic instrumentalists disagree; for them, all the relevant evidence comes from the substantive interpretation of the source materials.

In Sect. 6, we identified two ways in which unsupervised learning could be claimed to expand the evidential basis for interpretive research. First, unsupervised methods provide access to information in the texts that is difficult or impossible to grasp through unaided interpretation. Computational modeling provides a perspective into the data that is different from that of the analysts' viewpoint and thus enables *the discovery of unanticipated information* in text data in a manner not possible for coding or keyword-based approaches. Second, unsupervised methods enable the analysts to deal with comprehensive materials from a wide variety of sources, in other words, it enables *scalability*. A model produced by machine learning can provide an overview

Footnote 31 continued
engineering solutions to optimize computation have been argued to lead to the loss of researcher control over epistemically relevant aspects of modeling (Symons and Alvarado 2019; Barberousse and Vorms 2014). How the problems caused by increasing software-dependency in research should be addressed is currently an open question. See Floridi et al. (2015) for a discussion of different types of software malfunction.

of large corpora without choosing search terms or schemes for classifying materials. Furthermore, this overview is not biased by arbitrary sampling, reading order, or just by reader exhaustion. In both of these cases unsupervised machine learning can be taken to improve the objectivity of interpretation by enabling researchers to draw on a more extensive pool of evidence.

In Sect. 7, the role of *transparency* in reducing doubts about the possible arbitrariness of interpretation is analyzed. We argue that unsupervised modeling can facilitate collective scrutiny of interpretations by making the interpretive process more transparent. It enhances the objectivity of interpretation by providing resources for researchers to show how their interpretation is systematically linked to the materials, thus reducing doubts about arbitrary or idiosyncratic influences on interpretation. By facilitating replication and collective evaluation of how texts were sampled and analyzed, modeling makes humanistic interpretation less dependent on blind trust. However, it is important to recognize that transparency is not an intrinsic feature of computational tools but presupposes open research practices that involve access to primary materials, models, and detailed documentation of the research procedures.

These observations help us to identify and clarify the reasons why unsupervised methods need not be subject to Biernacki's criticism of coding and the sense in which they can be said to improve the objectivity of interpretation. The tension uncovered by Biernacki is between attempts at systematizing interpretation to make it mechanically objective and the interpreters' ability to grasp the contextual meaning of texts—or the substantial objectivity of interpretation. We argue that unsupervised learning can enable researchers to simultaneously engage with close reading text materials and grasp nuanced information in texts, while providing a systematic means for scrutinizing interpretations. The approach does not require that researchers specify an interpretive scheme for reading texts, but it *does* depend on the researchers reading of the primary materials. It is significant that these benefits are independent of one's choice between topic realism and topic instrumentalism.

This leaves us in an interesting place with respect to objectivity. The benefits of unsupervised methods are commonly associated with the introduction of a mechanical procedure that delays the moment of interpretation in the analysis. However, as argued above, the use of modeling results as evidence always rests on interpretive engagement with text materials, and thus improvements in objectivity cannot amount to a mechanical elimination of the researchers' judgments from the process. In line with Biernacki's (2015) criticism, unsupervised learning does not mitigate underdetermination between competing interpretations but instead turns the issue to concern the interpretation of modeling results against different background understandings of the modeled materials and the modeling process. The important point is that this also dissolves the seeming conflict between mechanically objective formal approaches and substantially objective humanistic interpretation. This insight serves to demonstrate the importance of paying attention to ground-level issues before applying this or that concept of objectivity to analyze novel methods or practices.

Instead of mechanical objectivity, our investigation of topic modeling shows that the aims of formal approaches to textual interpretation can be more productively conceptualized in terms of what Douglas (2004) has called "interactive" objectivity—or objectivity that is based on active discussion and social interaction among a community

of researchers. The key markers of this kind of objectivity are transparency of procedures and openness to criticism. As Douglas (2004, pp. 463–464) argues, the hope behind calls for transparency is that, "by keeping scientific discourse open to scrutiny, the most idiosyncratic biases and blinders can be eliminated." Similarly, as argued above, the transparency afforded by unsupervised learning amounts to increased facility in explicating one's interpretive process to others. This is why we maintain that unsupervised learning improves objectivity by *providing argumentative resources* to convince others that one's interpretation is not an artifact of idiosyncratic processes. Importantly, for interpretative debate to be interactively objective, it is not necessary that agreement be reached between divergent interpretations. Instead, the key criterion of interactive objectivity is that interpretations should become subject to criticism from others and that this criticism can potentially have a transformative effect on them (Longino 1990, p. 76).[32]

Through distinguishing between interactive and mechanical objectivity, we can see that Biernacki's tension between formalization and nuanced humanistic interpretation need not pose a pressing problem for unsupervised learning. As discussed by Daston and Galison (1992), the aim in improving the mechanical objectivity of knowledge-producing processes is to mitigate the influence of individual idiosyncrasies on their outcomes. This requires that the results of these processes can be reliably reproduced by different individuals (Douglas 2004, pp. 461–462; see also Megill 1994; Fine 1998). Our discussion of topic modeling demonstrates that, given a fixed parameter specification, the technical process of modeling can indeed be said to be objective in this sense. However, as argued above, this sense does not apply when the researchers' interpretive engagement with the modeled materials is taken into account. The objectivity of unsupervised interpretive text analysis depends on the whole interpretive process and not just on certain analytical subroutines (which in themselves might be mechanically objective). This implies that it is a mistake to reduce objectivity in this context to the idea that modeling delays interpretation by automating parts of the analysis process. The relevant question to ask about objectivity is not whether some formal method *eliminates* the need for interpretation in some parts of the analysis process but whether the method *facilitates* access to important information in text materials and *enables* systematic collective scrutiny of interpretations. As argued above, topic modeling can do both through improving interactive objectivity and permitting access to information in data that unaided interpreters could not access. There seems to be no principled reason why improved facility in intersubjective criticism should contradict the interpreters' ability to grasp nuanced information in texts. Given Biernacki's argument that coding frustrates the retrieval of nuanced information in texts and hides interpretive decisions behind seemingly mechanical procedures, unsupervised learning indeed seems to fare better in terms of improving objectivity.

At the same time, the increased facility of intersubjective scrutiny and access to more comprehensive information give unsupervised machine learning methods some

[32] Douglas (2004) identified agreement in discussion as a further condition of objectivity, labeled as "concordant objectivity." Both concordant and interactive objectivity are often discussed under the label of "intersubjective" objectivity (e.g., Risjord 2014, p. 23; Janack 2002; Crasnow 2006). Here we rely on Douglas' notion of interactivity because it helps highlight that unsupervised learning can improve objectivity of interpretation even in cases where agreement is not reached.

positive advantages over unaided humanistic interpretation. The transparency and scalability provided by unsupervised learning makes it easier for a wider range of interpreters to approach and work with the increasing volumes of text materials that are presently available. Modeling can help researchers draw on a wider evidential base to back up their interpretations, which improves the substantial objectivity of their claims. It also helps researchers evaluate the generalizability of their claims in relation to the modeled collection of texts, provides a means to justify and assess sampling, and makes rigorous interpretation depend less on virtuoso skill and more on clear and accessible documentation. Ideally, such benefits can lead to researchers gaining additional resources to argue for their interpretations and to the debate between divergent claims to become more driven by evidence (Lee and Martin 2015a). Again, we emphasize that this advantage stems from the increased scope and facility of communicating interpretive judgments that unsupervised methods potentially enable rather than the mechanical elimination of judgments from certain parts of the analysis process.

Finally, however, it is also very important to recognize that these advantages only apply under limited circumstances. As ween above, the theoretical interpretation of modeling results presupposes a background understanding of the modeled materials. Further, the materials must be sufficiently homogeneous so as to enable measurement in terms of certain prespecified concepts. These criteria set limitations to scalability when modeling is applied to corpora that are not well understood. Likewise, for modeling to improve transparency, the modelers need to be able to share the materials on which they based their interpretations (DiMaggio et al. 2013, p. 577). Finally, the modeling results are only useful for subsequent collective scrutiny if others are sufficiently familiar with the particular methods employed. This sets technical limitations to using machine learning methods in domain areas where they are not widely understood as well as requirements with respect to documenting the modeling process.

Given these limitations, the unsupervised approach can also work as a distraction to the interpretive process. Many have expressed skepticism concerning the import of large-scale computational analyses with methods that have not originally been developed for social scientific purposes and the operation of which might not be sufficiently well understood (boyd and Crawford 2012; Halford and Savage 2017; Törnberg and Törnberg 2018). The time invested in learning complex novel methods is easily time away from substantial interpretive research. Concurrently, there are worries that computational methodology might steer research to directions that are peripheral from the theoretical concerns of the social sciences (McFarland et al. 2016; Halavais 2015). Further, computational models might work to produce a false sense of objectivity for audiences not well versed in their technical specifics. In practice, computational algorithms are opaque black boxes for many users (Burrell 2016; Gillespie 2014). The objectivity and impartiality associated with algorithmic methods in many contexts present the danger of computational methods coming to play a rhetorical role justifying analyses which do not fulfill the criteria of rigorous interpretative research (Elish and boyd 2018). Increased attention to the specifics of modeling decisions can lead to glossing over issues related to how the actual interpretive work was carried out.

## 9 Conclusions

In this paper, we focus on issues that are the substance and motivation for anxiety about objectivity in interpretive research rather than analyze how social scientists discuss objectivity or how they use the word "objectivity." We do not develop a new theory or a definition of objectivity, but we are happy if our discussion provides some materials for philosophers who are developing such things. Our focus on the details of interpretive practices enables us to pose questions about objectivity and interpretation in a novel context. While philosophical accounts of objectivity can help elucidate methodological practices across contexts (Wright 2018), applications of abstract concepts to novel domains should pay heed to their particular characteristics. For us, the main philosophical challenge is to increase the clarity about the ground-level issues not to develop metatheory about them. That would be a topic for another paper, anyway.

Our investigation of the uses of topic modeling shows that objectivity in interpretive text analysis with unsupervised learning does not amount to a mechanical elimination of the researchers' judgments from the analysis process. Instead, unsupervised methods can improve the objectivity of interpretation in two primary ways. First, they enable analysts to draw on information that would otherwise be inaccessible and to do so with more comprehensive datasets. Second, they can facilitate collective scrutiny of interpretations by potentially making the interpretive process more transparent. This analysis helps see how unsupervised methods can potentially overcome some of the shortcomings of both unaided humanistic interpretation and coding-based approaches. However, unsupervised methods also have important limitations which show that objectivity is not an intrinsic feature of computational methods but depends on how they are employed within the broader context of interpretive practice.

Our paper does not deal with all the interesting issues related to machine learning methods in interpretative research, but we hope to demonstrate that they raise interesting philosophical questions and enable rethinking of the nature of interpretive research. For instance, more work remains to be done in understanding how model evaluation practices become embedded in the extant methodological traditions of different social scientific fields. It is not clear whether the potential transparency benefits of unsupervised methods are realized in practice and what requirements do the methods set for the participants' expertise. The development of computational methods in the social sciences and humanities will continue, and these developments are going to challenge established views about many things. There is a unique opportunity for present-day philosophers to observe and participate in these debates. It is our hope that this paper raises awareness of this opportunity among the philosophers of social sciences.

**Acknowledgements** Open access funding provided by the University of Helsinki including Helsinki University Central Hospital. This paper is based on ideas presented in the Objectivity in Social Research workshop at the University of Bergen in May 2019 and the ENPOSS conference at the National Technical University of Athens in August 2019. We thank the audiences of these events for their valuable comments. The first author of this study thanks the KONE Foundation (project: "Algorithmic Systems, Power and Interaction") for funding supporting this work.

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflicts of interest.



## References

Alexander, J., Jacobs, R., & Smith, P. (2012). Introduction: Cultural sociology today. In J. Alexander & P. Smith (Eds.), *The Oxford handbook of cultural sociology* (pp. 3–24). Oxford: Oxford University Press.

Baier, C., & Gengnagel, V. (2018). Academic autonomy beyond the nation-state. *Österreichische Zeitschrift für Soziologie, 43*(1), 65–92.

Bail, C., Brown, T., & Mann, M. (2017). Channeling hearts and minds: Advocacy organizations, cognitive-emotional currents, and public conversation. *American Sociological Review, 82*(6), 1188–1213.

Baker, P., & Levon, E. (2015). Picking the right cherries? A comparison of corpus-based and qualitative analyses of news articles about masculinity. *Discourse & Communication, 9*(2), 221–236.

Barberousse, A., & Vorms, M. (2014). About the warrants of computer-based empirical knowledge. *Synthese, 191*(15), 3595–3620.

Bearman, P., & Stovel, K. (2000). Becoming a nazi: A model for narrative networks. *Poetics, 27*(2), 69–90.

Betti, A., & van den Berg, H. (2016). Towards a computational history of ideas. In *CEUR workshop proceedings* 1681.

Biernacki, R. (2012a). *Reinventing evidence in social inquiry*. London: Palgrave MacMillan.

Biernacki, R. (2012b). Rationalization processes inside cultural sociology. In J. Alexander & P. Smith (Eds.), *The Oxford handbook of cultural sociology* (pp. 46–69). Oxford: Oxford University Press.

Biernacki, R. (2014). Humanist interpretation versus coding text samples. *Qualitative Sociology, 37,* 173–188.

Biernacki, R. (2015). How to do things with historical texts. *American Journal of Cultural Sociology, 3,* 311–352.

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

Blei, D. (2012b). Topic modeling and digital humanities. *Journal of Digital Humanities, 2*(1), 8–11.

Blei, D., & Lafferty, J. (2006). Correlated topic models. *Advances in Neural Information Processing Systems, 18,* 147–154.

Blei, D., & McAuliffe, J. (2007). Supervised topic models. In J. C. Platt, D. Koller, Y. Singer, & S. T. Roweis (Eds.), *Proceedings of the 20th international conference on neural information processing systems NIPS '07* (pp. 121–128).

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research, 3*(3), 993–1022.

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *The Philosophical Review, 97*(3), 303–352.

Bonfiglioli, R., & Nanni, F. (2016). From close to distant and back: How to read with the help of machines. In F. Gaducci & M. Tavosanis (Eds.), *Proceedings of the third international conference on the history and philosophy of computing HaPoC '15* (pp. 87–100). Berlin: Springer.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662–679.

Braun, V., & Clarke, V. (2013). *Successful qualitative research: A practical guide*. Thousand Oaks: SAGE Publishing.

Buckner, C., Niepert, M., & Allen, C. (2011). From encyclopedia to ontology: Toward dynamic representation of the discipline of philosophy. *Synthese, 182*(2), 205–233.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*. https://doi.org/10.1177/2053951715622512.
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,* 288–296.
Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. Thousand Oaks: SAGE Publishing.
Clement, T. (2013). Text analysis, data mining, and visualizations in literary scholarship. In K. Price & R. Siemens (Eds.), *Literary studies in the digital age.* Retrieved May 4, 2020 from https://dlsanthology.mla.hcommons.org/text-analysis-data-mining-and-visualizations-in-literary-scholarship/. MLACommons.
Crasnow, S. (2006). Feminist anthropology and sociology: Issues for social science. In S. Turner & M. Risjord (Eds.), *Philosophy of anthropology and sociology* (pp. 827–861). Amsterdam: Elsevier.
Daston, L., & Galison, P. (1992). The image of objectivity. *Representations, 40,* 81–128.
de Laat, P. B. (2018). Algorithmic decision-making based on machine learning from big data: Can transparency restore accountability? *Philosophy & Technology*, *31*(4), 525–541.
Denny, M., & 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.
Denzin, N., & Lincoln, Y. (2011). Introduction: The discipline and practice of qualitative research. In N. Denzin & Y. Lincoln (Eds.), *The SAGE handbook of qualitative research* (4th ed., pp. 1–19). Thousand Oaks: SAGE Publications.
DiMaggio, P. (2015). Adapting computational text analysis to social science (and vice versa). *Big Data & Society*. https://doi.org/10.1177/2053951715602908.
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.
Douglas, H. (2004). The irreducible complexity of objectivity. *Synthese, 138,* 453–473.
Earhart, A. (2015). Data and the fragmented text: Tools, visualization, and datamining or is bigger better? In A. Earhart (Ed.), *Traces of the old, uses of the new: The emergence of digital literary studies*. Michigan: Michigan Publishing. https://doi.org/10.3998/etlc.13455322.0001.001.
Elish, M., & Boyd, D. (2018). Situating methods in the magic of big data and AI. *Communication Monographs, 85*(1), 57–80.
Evans, J. (2002). *Playing God? Human genetic engineering and the rationalization of public bioethical debate*. Chicago: University of Chicago Press.
Fine, A. (1998). The viewpoint of no-one in particular. *Proceedings and Addresses of the American Philosophical Association, 72*(2), 7–20.
Fligstein, N., Brundage, J., & Schultz, M. (2017). Seeing like the fed: Culture, cognition, and framing in the failure to anticipate the financial crisis of 2008. *American Sociological Review, 82*(5), 879–909.
Floridi, L., Fresco, N., & Primiero, G. (2015). On malfunctioning software. *Synthese, 192*(4), 1199–1220.
Gibson, A., & Ermus, C. (2019). The history of science and the science of history: Computational methods, algorithms, and the future of the field. *Isis, 110*(3), 555–566.
Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P. Boczkowski, & K. Foot (Eds.), *Media technologies: Essays on communication, materiality, and society* (pp. 167–193). Cambridge: The MIT Press.
Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago: Aldine.
Greene, D., O'Callaghan, D., & Cunningham, P. (2014). How many topics? Stability analysis for topic models. In T. Calders, F. Esposito, E. Hüllermeier, & R. Meo (Eds.), *Machine learning and knowledge discovery in databases. ECML PKDD 2014 Proceedings, part I* (pp. 498–513). Berlin: Springer.
Grimmer, J., & Stewart, B. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(3), 267–297.
Griswold, W. (1987). The fabrication of meaning: Literary interpretation in the United States, Great Britain, and the West Indies. *American Journal of Sociology, 92*(5), 1077–1117.
Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v040.i13.
Hacking, I. (2015). Let's not talk about objectivity. In F. Padovani, A. Richardson, & J. Tsou (Eds.), *Objectivity in science: New perspectives from science and technology studies* (pp. 19–33). Berlin: Springer.

Halavais, A. (2015). Bigger sociological imaginations: Framing big social data theory and methods. *Information, Communication & Society, 18*(5), 583–594.
Halford, S., & Savage, M. (2017). Speaking sociologically with big data: Symphonic social science and the future for big data research. *Sociology, 51*(6), 1132–1148.
Hirsch, E. D. (1967). *Validity in interpretation*. London: Yale University Press.
Hubig, C., & Kaminski, A. (2017). Outlines of a pragmatic theory of truth and error in computer simulation. In M. Resch, A. Kaminski, & P. Gehring (Eds.), *The science and art of simulation I* (pp. 121–136). Berlin: Springer.
Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese, 169*(3), 615–626.
Ignatow, G. (2015). Theoretical foundations for digital text analysis. *Journal for the Theory of Social Behaviour, 46*(1), 104–120.
Isoaho, K., Gritsenko, D., & Mäkelä, E. (2019). Topic modeling and text analysis for qualitative policy research. *Policy Studies Journal*. https://doi.org/10.1111/psj.12343.
Jacobs, T., & Tschötschel, R. (2019). Topic models meet discourse analysis: A quantitative tool for a qualitative approach. *International Journal of Social Research Methodology*. https://doi.org/10.1080/13645579.2019.1576317.
Janack, M. (2002). Dilemmas of objectivity. *Social Epistemology, 16*(3), 267–281.
Jockers, M., & Mimno, D. (2013). Significant themes in 19th-century literature. *Poetics, 41*(6), 750–769.
Jones, T. (1998). Interpretive social science and the "native's point of view": A closer look. *Philosophy of the Social Sciences, 28*(1), 32–68.
Kaltenbrunner, W. (2015). Scholarly labour and digital collaboration in literary studies. *Social Epistemology, 29*(2), 207–233.
Krishnan, M. (2019). Against interpretability: A critical examination of the interpretability problem in machine learning. *Philosophy & Technology*. https://doi.org/10.1007/s13347-019-00372-9.
Lee, M., & Martin, J. L. (2015a). Coding, counting and cultural cartography. *American Journal of Cultural Sociology, 3*(1), 1–33.
Lee, M., & Martin, J. L. (2015b). Response to Biernacki, Reed, and Spillman. *American Journal of Cultural Sociology, 3*(3), 380–415.
Lee, T., Smith, A., Seppi, K., Elmqvist, N., Boyd-Graber, J., & Findlater, L. (2017). The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies, 105,* 28–42.
Light, R., & Cunningham, J. (2016). Oracles of peace: Topic modeling, cultural opportunity, and the Nobel peace prize, 1902–2012. *Mobilization: An International Quarterly, 21*(1), 43–64.
Longino, H. (1990). *Science as social knowledge*. Princeton: Princeton University Press.
Maier, D., Waldherr, A., Mitner, P., Wiedemann, G., Niekler, A., Keinert, A., et al. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures, 12*(2–3), 93–118.
Malaterre, C., Jean-François, C., & Pulizzotto, D. (2019). What is this thing called philosophy of science? A computational topic-modeling perspective, 1934–2015. *HOPOS: The Journal of the International Society for the History of Philosophy of Science, 9*(2), 215–249.
Marres, N. (2017). Do we need new methods? In N. Marres (Ed.), *Digital sociology: The reinvention of social research* (pp. 78–115). Cambridge: Polity.
Marshall, E. (2013). Defining population problems: Using topic models for cross-national comparison of disciplinary development. *Poetics, 41*(6), 701–724.
McFarland, D., Lewis, K., & Goldberg, A. (2016). Sociology in the era of big data: The ascent of forensic social science. *The American Sociologist, 47*(1), 12–35.
Megill, A. (1994). Introduction: Four senses of objectivity. In A. Megill (Ed.), *Rethinking objectivity* (pp. 1–20). Durham: Duke University Press.
Miller, I. (2013). Rebellion, crime and violence in Qing China, 1722–1911: A topic modeling approach. *Poetics, 41*(6), 626–649.
Mimno, D., & Blei, D. (2011). Bayesian checking for topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 227–237). Association for Computational Linguistics.
Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 262–272). Association for Computational Linguistics.

Mohr, J. (1998). Measuring meaning structures. *Annual Review of Sociology, 24,* 345–370.
Mohr, J., & Bogdanov, P. (2013). Introduction—Topic models: What they are and why they matter. *Poetics, 41*(6), 545–569.
Mohr, J., & Rawlings, C. (2012). Four ways to measure culture: Social science, hermeneutics, and the cultural turn. In J. Alexander & P. Smith (Eds.), *The Oxford handbook of cultural sociology* (pp. 70–113). Oxford: Oxford University Press.
Moretti, F. (2000). The slaughterhouse of literature. *Modern Language Quarterly, 61*(1), 207–227.
Moretti, F. (2013). *Distant reading*. New York: Verso Books.
Nelson, L. (2017). Computational grounded theory: A methodological framework. *Sociological Methods & Research*. https://doi.org/10.1177/0049124117729703.
Nelson, L., Burk, D., Knudsen, M., & McCall, L. (2018). The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. *Sociological Methods & Research*. https://doi.org/10.1177/0049124118769114.
Rabinow, P., & Sullivan, W. (1979). *Interpretive social science: A reader*. Berkeley: University of California Press.
Ramsay, S. (2005). In praise of pattern. *TEXT Technology: The Journal of Computer Text Processing, 14*(2), 177–190.
Ramsay, S. (2011). *Reading machines: Toward an algorithmic criticism*. Champaign: University of Illinois Press.
Ramsey, G., & Pence, C. (2016). evoText: A new tool for analyzing the biological sciences. *Studies in History and Philosophy of Science Part C, 57,* 83–87.
Reed, I. A. (2015). Counting, interpreting and their potential interrelation in the human sciences. *American Journal of Cultural Sociology, 3*(3), 353–364.
Rhody, L. (2012). Topic modeling and figurative language. *Journal of Digital Humanities, 2*(1), 19–35.
Risjord, M. (2014). *Philosophy of social science: A contemporary introduction*. Abingdon: Routledge.
Roberts, M., Stewart, B., & Dustin, T. (2016). Navigating the local models of big data: The case of topic models. In R. M. Alvarez (Ed.), *Computational social science: Discovery and prediction* (pp. 49–97). Cambridge: Cambridge University Press.
Roberts, M., Stewart, B., & Dustin, T. (2019). stm: R package for structural topic models. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v091.i02.
Roberts, M., Stewart, B., Dustin, T., Lucas, C., Leder-Luis, J., Gadarian, S., et al. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58*(4), 1064–1082.
Roose, H., Roose, W., & Daenekindt, S. (2018). Trends in contemporary art discourse: Using topic models to analyze 25 years of professional art criticism. *Cultural Sociology, 12*(3), 303–324.
Schmidt, B. (2012). Words alone: Dismantling topic models in the humanities. *Journal of Digital Humanities, 2*(1), 49–65.
Schmidt-Petri, H., Adam, S., Reber, U., Häussler, T., Maier, D., Miltner, P., et al. (2018). Homophily and prestige: An assessment of their relative strength to explain link formation in the online climate change debate. *Social Networks, 55,* 47–54.
Schnable, A. (2018). What religion affords grassroots NGOs: Frames, networks, modes of action. *Journal for the Scientific Study of Religion, 55*(2), 216–232.
Schofield, A., & Mimno, D. (2016). Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics, 4,* 287–300.
Schwartz, A., & Ungar, L. (2015). Data-driven content analysis of social media a systematic overview of automated methods. *The ANNALS of the American Academy of Political and Social Science, 659,* 78–94.
Sievert & Shirley. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63–70). Association for Computational Linguistics.
Spillman, L. (2015). Ghosts of straw men: A reply to Lee and Martin. *American Journal of Cultural Sociology, 3*(3), 365–379.
Stier, S., Posch, L., Bleier, A., & Strohmaier, M. (2017). When populists become popular: Comparing Facebook use by the right-wing movement Pegida and German political parties. *Information, Communication & Society, 20*(9), 1365–1388.
Stuart, M. (2019). The role of imagination in social scientific discovery: Why machine discoverers will need imagination algorithms. In M. Addis, P. Lane, P. Sozou, & F. Gobet (Eds.), *Scientific discovery in the social sciences* (pp. 49–66). Berlin: Springer.

Sullivan, E. (2019). Understanding from machine learning models. *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axz035.

Suppes, P. (1962). Models of data. In E. Nagel, P. Suppes, & A. Tarski (Eds.), *Logic, methodology and philosophy of science: Proceedings of the 1960 international congress* (pp. 252–261). Stanford: Stanford University Press.

Symons, J., & Alvarado, R. (2016). Can we trust big data? Applying philosophy of science to software. *Big Data & Society*. https://doi.org/10.1177/2053951716664747.

Symons, J., & Alvarado, R. (2019). Epistemic entitlements and the practice of computer simulation. *Minds and Machines, 29*(1), 37–60.

Symons, J., & Horner, J. (2014). Software intensive science. *Philosophy & Technology, 27*(3), 461–477.

Tangherlini, T., & Leonard, P. (2013). Trawling in the sea of the great unread: Sub-corpus topic modeling and humanities research. *Poetics, 41*(6), 725–749.

Thagard, P. (1990). Philosophy and machine learning. *Canadian Journal of Philosophy, 20*(2), 261–276.

Törnberg, A., & Törnberg, P. (2016). Combining CDA and topic modeling: Analyzing discursive connections between Islamophobia and anti-feminism on an online forum. *Discourse & Society, 27*(4), 401–422.

Törnberg, P., & Törnberg, A. (2018). The limits of computation: A philosophical critique of contemporary big data research. *Big Data & Society*. https://doi.org/10.1177/2053951718811843.

Wallach, H., Mimno, D., & McCallum, A. (2009). Rethinking LDA: Why priors matter. In *Proceedings of the 22nd international conference on neural information processing systems* (pp. 1973–1981). Curran Associates Inc.

Williams, M. (2000). Interpretivism and generalisation. *Sociology, 34*(2), 209–224.

Williamson, H. (2009). The philosophy of science and its relation to machine learning. In M. M. Gaber (Ed.), *Scientific data mining and knowledge discovery: Principles and foundations* (pp. 77–89). Berlin: Springer.

Winsberg, E. (2019). Computer simulations in science. In E. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2019 Edition). Retrieved May 4, 2020 from https://plato.stanford.edu/archives/win2019/entries/simulations-science.

Wright, J. (2018). Rescuing objectivity: A contextualist proposal. *Philosophy of the Social Sciences, 48*(4), 385–406.