European Management Review, (2021)
DOI: 10.1111/emre.12452

From Big Data to Rich Theory: Integrating Critical Discourse Analysis with Structural Topic Modeling

Ana M. Aranda,[1] Kathrin Sele,[2,3] Helen Etchanchu,[4] Jonne Y. Guyt[5] and Eero Vaara[6]

[1] Católica Lisbon School of Business and Economics, Lisbon, Portugal
[2] Aalto University School of Business, Espoo, Finland
[3] School of Business and Economics, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands
[4] Montpellier Business School, University of Montpellier, Montpellier, France
[5] Amsterdam Business School, University of Amsterdam, Amsterdam, the Netherlands
[6] Saïd Business School, University of Oxford, Oxford, UK

A growing interest in the study of discourses has spread in management research, but so far, it has mostly relied on in-depth qualitative analyses of textual material. With the increasing availability of large textual data, several challenges arise. This paper offers a mixed-methods approach to integrate critical discourse analysis with structural topic modeling to turn these challenges into valuable opportunities. We argue that combining both approaches overcomes their limitations and provides great potential for exploring phenomena that matter in our mediatized society. Based on an explanatory sequential mixed-methods design, we develop a stepwise model that provides practical and theoretical guidance to conduct a critical analysis of large textual data. Our illustrative example focuses on the discursive legitimation struggles around the tobacco industry. We demonstrate how an integrated mixed-methods approach allows capturing the breadth and depth of discourses used by different actors in the tobacco debates.

Keywords: critical discourse analysis; structural topic modeling; mixed methods

Introduction

Over the past three decades, we have seen an increasing interest in the role that language and discourse play in organizing and managing (Phillips and Oswick, 2012). Management scholars have employed a variety of discursive approaches to study topics such as strategy (Knights and Morgan, 1991; Vaara *et al*., 2004; Mantere and Vaara, 2008), organizational change (Heracleous and Barrett, 2001; Sonenshein, 2010), institutions (Green *et al*., 2009; Maguire and Hardy, 2009), or leadership (Fairhurst and Uhl-Bien, 2012).

Critical discourse analysis (hereafter CDA) has been a particularly influential approach in management studies (Phillips *et al*., 2008; Vaara and Tienari, 2008; Chouliaraki and Fairclough, 2010). CDA does not only address questions about complex and challenging social and organizational phenomena but entails an in-built critical stance, and is thus problem-oriented and eclectic (Fairclough, 2013; Wodak and Meyer, 2016). It considers discourses and discourse users in light of the broader historical and sociopolitical context, and sees texts as 'sites of struggles' (Wodak, 2001, p. 11) that are inextricably entwined with material social elements (Mumby, 2011). Existing work mobilizing CDA, and the broader assumption of critical studies that organizations are political sites 'accomplished in conditions of struggles and domination' (Deetz, 1996, p. 202), has, for example, looked at how power, legitimacy, identity, or inequality are (re)constructed in and through discursive struggles (Vaara *et al*., 2005; Lefsrud and Meyer, 2012; Barros, 2014; Vaara, 2014; Samdanis and Lee, 2019).

In today's mediatized and digitized society, sense is made and reality constructed in and through discourses. Accordingly, many social and organizational phenomena

Correspondence: Kathrin Sele, Aalto University School of Business, Ekonominaukio 102150 Espoo. E-mail kathrin.sele@aalto.fi

are increasingly pervasive and discursively interwoven, leading to what Wodak (2001) calls 'discursive swarming.' Consequently, management scholars in general, and CDA scholars in particular, are often confronted with an overwhelmingly large and unstructured amount of data, which despite its advantages creates some challenges. One key challenge for scholars applying qualitative methods relates to the manual processing and analysis of large text corpora in a systematic and reproducible manner (Wodak and Meyer, 2016). Not only does this often lead to the necessity of premature sampling and early selection of focal texts (Phillips *et al*., 2008), but also to difficulties in integrating context (Leitch and Palmer, 2010) or operationalizing intertextuality (Farrelly, 2020) and interdiscursivity (Reisigl and Wodak, 2016).

In response to this challenge, some CDA scholars have started to adopt quantitative methods originating from machine learning and computational linguistics to make sense of big data (Kobayashi *et al*., 2018). In particular, topic modeling (Blei *et al*., 2003; DiMaggio *et al*., 2013) has been identified as a promising text-mining tool for discovering latent semantic structures and meaning-making in textual data—a primary goal of CDA. One recent advancement in topic modeling has been the introduction of structural topic models (hereafter STM), which allow the incorporation of metadata such as time or actors (Hannigan *et al*., 2019) that can help identify topic prevalence and content (Roberts *et al*., 2016). This is important for CDA researchers because, by doing so, they cannot only explore content *within* a given document (as in the case of grounded theory or content analysis) but may also theorize about the relationship *between* texts, discourses, and context (Wodak and Meyer, 2016). Therefore, we argue that STM is well suited to aid CDA scholars to link the content of texts to the material context in which they are produced (DiMaggio *et al*., 2013; Grimmer and Stewart, 2013).

Topic modeling has become popular among management scholars (Hannigan *et al*., 2019; Schmiedel *et al*., 2018) and has been combined with qualitative approaches (Kaplan and Vakili, 2015; Croidieu and Kim, 2018). However, we still lack knowledge on how to best integrate CDA and STM in empirical analysis. This is not only a practical problem but deals with the paradigmatic differences in research traditions and their epistemological and methodological assumptions. For instance, one might argue that the critical and constructivist orientation of CDA is at odds with the neutral and objectivist orientation of topic modeling. While we agree that there are apparent differences and that an 'anything goes' approach is not productive (Hassard, 1988; Scherer, 1998), we consider integrating CDA and STM not only possible but potentially very fruitful. Indeed, boundaries can be bridges if crossing them entails explicitly attending and dealing with actual differences and tensions (Gioia and Pitre, 1990; Deetz, 1996). Accordingly, researchers have to 'make their paradigmatic assumptions explicit' (Lewis and Grimes, 1999, p. 686) and 'move beyond reproduction of the differences that divide us to an appreciation of why we are divided' (Morgan, 1983, p. 382). In our case, this means that bringing STM into CDA cannot only be approached from a methodological perspective, but has to respect the epistemological foundations of CDA, including its critical orientation and contextualized nature (Creswell *et al*., 2003).

Inscribing ourselves in a mixed-methods approach (Molina-Azorin *et al*., 2017), we see untapped potential in combining these two approaches for advancing the field of management by seizing the opportunities and unlocking the potential that large textual data offer. We argue that if scholars are confronted with big data, STM may serve as a basis for the traditionally in-depth and focused approach in CDA, as its crucial idea is to inductively derive an understanding of key topics that can be aggregated to form discourses in a transparent and replicable manner (DiMaggio *et al*., 2013; Chandra *et al*., 2016). We contend that CDA has always been meant as a program that could and should be linked with different theories and methods (Fairclough, 2003; Forchtner and Wodak, 2018). Linking CDA with STM thus seems a promising extension of traditional CDA and can be seen as a response to Wodak and Meyer's (2016) call for multi-methodical approaches to analyze, understand, and explain complex and vast phenomena. Hence, our guiding research question is: *How can we integrate CDA with STM to advance management research?*

The remainder of the paper is structured as follows. We first introduce CDA and STM, focusing on their main characteristics and how each approach has been mobilized in management studies. We continue with a reflection on the challenges of integrating CDA and STM on an epistemological and methodological level, before advocating for an explanatory transformative mixed-methods design (Creswell, 2009). This enables us to ensure the critical stance of CDA (Hardy and Phillips, 2004) while addressing the limitations of both methods (Brookes and McEnery, 2019; Jacobs and Tschötschel, 2019). We then outline a stepwise model designed explicitly to integrate CDA and STM, and to enable scholars to derive broader empirical patterns and deeper theoretical insights. To illustrate our model, we conduct a critical analysis of discursive legitimation processes within the US tobacco industry. We demonstrate how the proposed stepwise model yields a deep understanding of the key tobacco discourses between 1986 and 2016. We conclude with a critical reflection on how STM helps CDA scholars make use of large textual

datasets in a way that opens up new opportunities and avenues for research, and how CDA may offer STM users a much needed theoretic-methodological perspective to guide the interpretation of results.

CDA and STM: An overview and recent advances

Critical discourse analysis

There are many definitions of and approaches to discourse analysis (Van Dijk, 2011; Tannen *et al*., 2015). Potter and Wetherell (1987) have famously stated that one could have two discourse analysis textbooks with virtually no overlap. In management research, several discursive approaches have become popular in recent decades (Grant *et al*., 2004; Phillips and Oswick, 2012); notably, interpretive discourse analysis (Heracleous, 2006), narrative analysis (Vaara *et al*., 2016), rhetorical analysis (Sillince and Suddaby, 2008; Heracleous *et al*., 2020), and CDA (Phillips *et al*., 2008; Vaara, 2010). We focus on CDA as an increasingly popular approach that has guided scholars across the humanities and social sciences in examining the discursive construction of social phenomena from a critical perspective (Wodak and Meyer, 2016).

The roots of CDA can be traced back to applied linguistics from where it has developed into closely related but distinct approaches including Fairclough's original critical work (1989, 2016) as well as socio-cognitive (Van Dijk, 2016), discourse-historical (Reisigl and Wodak, 2016) and multimodal social actor (Van Leeuwen, 2016) approaches. What unites these approaches is that they consider language a social practice and see discourse as both socially conditioned and constitutive (Fairclough and Wodak, 1997). In this view, discourses do not only reflect reality but are the very means of constructing and reproducing it. In particular, CDA aims at revealing taken-for-granted assumptions in society or ideologies, that is, fundamentally different assumptions, values, and worldviews, shared by people and reflected in discourses (Fairclough, 1989, 2003; Van Dijk, 1998; Forchtner and Wodak, 2018). As CDA is concerned with the relationship between textual production and the broader social structures and material context (Fairclough, 2005), scholars engage as much in 'text or discourse immanent critique' as they do in 'socio-diagnostic critique,' which focuses on detecting problematic aspects in discursive practices (Reisigl and Wodak, 2016).

The main intention of CDA is to abductively understand how language is used to exercise power and how it relates intertextually and interdiscursively when forming systems or technologies of power (Wodak and Meyer, 2016). To uncover often subtle and hidden dynamics, discourses are considered in their historical and socio-political context, irrespective of the analytical categories (e.g., social actors, argumentative features, legitimation strategies) used. In line with the clear critical perspective guiding the research efforts and the motivation to detect and name problematic aspects, CDA naturally includes a social commitment on behalf of the researcher and involves taking a stance toward the phenomena under investigation.

Management scholars have mobilized CDA both as a theoretical framework and as a methodological approach to study power relations, legitimation processes, and identity politics as discursive struggles (Phillips *et al*., 2008; Vaara and Tienari, 2008; Chouliaraki and Fairclough, 2010; Vaara *et al*., 2019). Empirically, management scholars have relied on different analytical CDA frameworks. Focusing on identity-related questions and studying decisions in their historical context (Reisigl and Wodak, 2016), Vaara *et al*. (2005) examined the choice of a common language in a cross-border merger and how it shaped future power dynamics within the newly created organization. In a recent study on inequalities in the UK artistic labor market, Samdanis and Lee (2019) analyzed how different discourses are used by social enterprises and social activists borrowing Fairclough's (1992) framework for CDA. Building on Van Leeuwen's (2007) legitimation strategies, several scholars have studied legitimacy struggles and their corresponding legitimation processes. For example, Vaara *et al*. (2006) studied a merger in the paper and pulp sector, and Lefsrud and Meyer (2012) focused on the discursive construction of climate change. Barros (2014) in turn analyzed how the Brazilian oil company, Petrobas, used different discursive strategies to question the media's legitimacy and create credibility for itself during a period of turmoil.

CDA has been characteristically qualitative. However, recently, we observe that discursive organizational and societal struggles have become more interwoven and more dispersed due to their increased mediatization and digitization. Furthermore, scholars are more than ever confronted with large and complex textual data (Törnberg and Törnberg, 2016). Accordingly, we see a need to consider how to ensure the linking of specific texts to their broader context (Leitch and Palmer, 2010) while at the same time accounting for aspects of intertextuality and interdiscursivity (Farrelly, 2020). In particular, identifying discourses in vast textual material remains a challenge that —despite all computational advances—is still often tackled by interpreting specific text excerpts, rather than systematically analyzing all available data (Vaara, 2014). Moreover, with increasingly large datasets, scholars face the challenge of how to validate the existence of discourses empirically. In line with recent discussions and developments in applied linguistics (Forchtner and

Wodak, 2018), we see great potential in combining CDA with STM to meet some of these methodological challenges.

Structural topic modeling

Topic models are an increasingly popular method to analyze large textual data. They have several distinct advantages over other text analysis methods that require manual input and *a priori* decision making (Blei *et al*., 2003; DiMaggio *et al*., 2013; Schmiedel *et al*., 2018). One particularly attractive feature of topic modeling is that it reduces a given text's complexity by associating topics (or 'bags of words') to documents. To find the thematic topics that are latent within a collection of documents, topic models exploit that terms belonging to a specific topic tend to co-occur more regularly than by chance (Blei, 2012). Since the same word may reflect different meanings in different contexts, topic models allow for attributing the same words to several topics (i.e., polysemy) depending on its co-occurrence with other words (DiMaggio *et al*., 2013). Topic models, thus, appreciate that meaning resides in the relationship between words—a feature that is, for example, missing in corpus linguistics (Brookes and McEnery, 2019). Moreover, they are explicit, meaning 'that data are available for the researcher to test his or her interpretations and for other researchers to reproduce the analyses,' and inductive in that they 'permit researchers to discover the structure of the corpus before imposing their priors on the analysis' (DiMaggio *et al*., 2013, p. 577). As topic models enable the 'analysis of larger corpora than human coders can master, facilitating discovery of unanticipated frames, and distinguishing between different uses of the same term,' DiMaggio *et al*. (2013, p. 593) have advocated them as a useful method for management research.

Recent work has provided some technical guidance on applying topic modeling in organizational research (Banks *et al*., 2018; Schmiedel *et al*., 2018). The most essential topic model is the so-called Latent Dirichlet Allocation (hereafter LDA) model pioneered by Blei *et al*. (2003), which identifies a predefined number of topics and their respective prevalence in each document of a corpus or corpora (i.e., a collection of texts). Here, though, we employ STM, a method that generalizes the LDA model by incorporating metadata (i.e., additional contextual or structural information about a document) into the model (Roberts *et al*., 2016, 2019). The inclusion of observed metadata enables researchers to explore the relationship between topic prevalence and selected covariates in a given text.[1] While LDA models can only reveal the latent topics within a given text, STM allows making inferences about how an observed variable of interest affects a particular topic (Roberts *et al*., 2016).

Management scholars have started to use topic models in a wide range of empirical settings. Initially, topic models were employed to study innovation (Toubia and Netzer, 2017), patents (Kaplan and Vakili, 2015), technological novelty (Wilson and Joseph, 2015), or latent knowledge structures in a scientific journal (Antons *et al*., 2016). The growing popularity of topic modeling in adjacent disciplines is reflected in an increasing number of studies that have used basic topic models in various contexts. For instance, Huang *et al*. (2018) used it to identify investment risk factors, Croidieu and Kim (2018) to explore field level legitimation of the US amateur radio operator, Haans (2019) to model the (in) distinctiveness of firms in the Dutch creative industries, and Taeuscher *et al*. (2020) to study optimal distinctiveness in crowdfunding platforms.

Interestingly, most management studies rely on LDA despite the many advantages that STM has to offer (c.f., Doldor *et al*., 2019). One reason for the slow adaptation of STM might be that scholars mainly use topic modeling as a first and relatively isolated analytical step that lays out the different topics. However, in light of the need to engage in substantial interpretation of the results, such an approach is problematic (Schmiedel *et al*., 2018; Hannigan *et al*., 2019). The interpretive part of the analysis is enhanced—or so we argue—when not only followed but guided by established qualitative methods. In line with prior work by Brookes and McEnery (2019), we posit that CDA is a remarkably well-suited approach as it aims at an understanding of texts and discourses within their social and material context. In principle, we recognize that any type of topic model allows for automated discovery of the discourses present in large textual data (Törnberg and Törnberg, 2016). However, we deem STM as more destined to be combined with CDA than LDA because it explicitly allows for establishing relationships between texts and contextual variables that may affect or be affected by the texts (i.e., metadata). In this sense, STM facilitates the critical interpretation of discourses that is at the heart of CDA, while simultaneously providing an insightful, precise, and thorough account of the contextual discourses present in textual data. All this said, several challenges need to be addressed when integrating these methods.

Challenges in integrating CDA and STM

Mixed-methods approaches are considered a viable means to overcome quantitative and qualitative methodological limitations, conduct a rigorous analysis of complex phenomena by exploring different facets of data, and address theoretical questions that have hitherto remained

[1]While most uses of STM exploit the fact that covariates relate to topic prevalence, STM also allows for word use to be correlated with covariates.

out of reach (Molina-Azorin *et al*., 2017). However, while integrating different research traditions has many benefits, it is also notoriously difficult since it involves epistemological and paradigmatic differences (Creswell *et al*., 2003). Integrating CDA and STM is no different. To integrate the different analytical steps of machine- and human-based analyses, we need a critical discussion on paradigm (in)commensurability that sheds light on how to combine these two methods best.

Epistemologically, CDA is based on a position that combines the idea of the social construction of empirical phenomena with an appreciation for the importance of the social and material reality outside the researcher's interpretations (Fairclough, 2005). In contrast, the origins of STM reflect a positivist tradition rooted in machine learning and data sciences. Accordingly, while CDA seeks understanding, STM seeks answers through textual analysis. These different epistemological stances suggest that rigor is not achieved in the same way. While CDA relies on reflexivity, STM calls upon validity and reliability measures. This creates practical challenges: CDA scholars are used to subjective interpretation and abductive reasoning based on careful (re)reading of texts (Vaara, 2010); STM, however, reduces human input as much as possible during the inductive estimation process. While the algorithmic creation of discourses may appear sterile from an interpretative lens, interpretative work's open-ended and emergent nature may be questioned from an estimation-driven approach. However, the role of subjectivity in interpretation is precisely what may allow for bridging the two paradigms (Gioia and Pitre, 1990). Although STM is an automated approach, interpretation is happening in various steps of the STM estimation process (Hannigan *et al*., 2019). Similar to CDA, interpretation is made by the researcher whose agency becomes intertwined with the agency of the STM algorithm. Neither the researcher nor the actual technique or tool performs the analysis in isolation. In CDA, the researcher guides and is guided by particular analytical methods, which can be more or less prescriptive (Wodak and Meyer, 2016). In STM, choices made by the researcher shape the outcomes of the estimation and *vice versa*. This mutually constitutive process enables a dialogue which allows not only to account for their epistemological differences explicitly but avoids both paradigmatic closure and relativism (Hassard, 1988; Scherer, 1998; Lewis and Grimes, 1999).

This reflection provides the necessary ground to focus on the methodological challenges and, in particular, the question of whether and how these fundamental differences can be addressed. Advocates of mixed-methods have argued that scholars have several ways to engage with differences in a meaningful way (Creswell, 2009). They highlight the importance of expansion or augmentation—both processes carrying the notion of complementarity (Deetz, 1996; Creswell *et al*., 2003). Complementarity does indeed apply to CDA and STM: while the former is based on a critical engagement with the data at hand, the latter does not imply any particular orientation but adopts a neutral stance. Although the approaches are very different, we believe that it is possible to retain the critical stance of CDA when interpreting STM results. This is because, from its inception, CDA has been intended as an approach that can, and should, be combined with other theoretical and methodological perspectives to make meaningful contributions (Fairclough, 2003). For all of these reasons, the foci of these methods can be seen as complementary.

Building on these insights, we propose an explanatory 'transformative strategy' (Creswell, 2009, p. 215) for combining CDA with STM. Such an approach ensures the critical ideological stance inherent to CDA. As

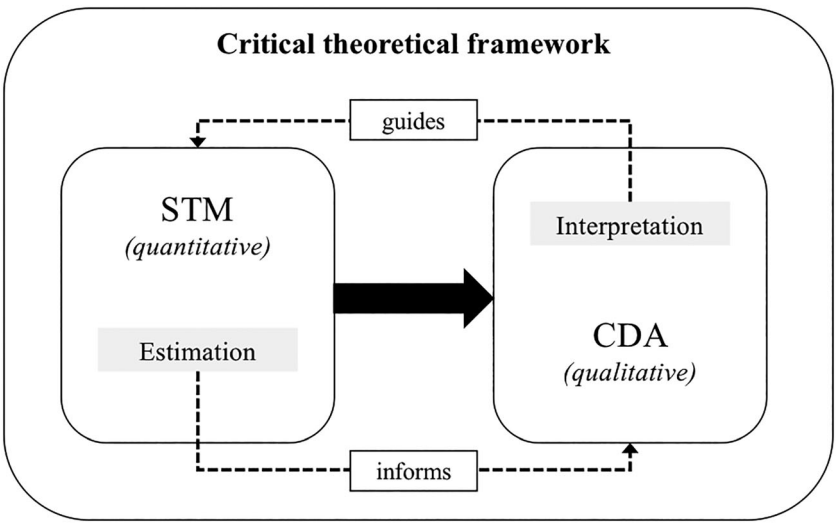

FIGURE 1 Transformative exploratory design

outlined by Mertens (2012, p. 808), transformative mixed methods are particularly suitable to study issues of power and inequities, and the methodological choices are 'made with conscious awareness of contextual and historical factors.' In line with CDA's idea that researchers take a stance, the transformative paradigm considers researchers as agents interested in advancing advocacy issues. Figure 1 illustrates how the mixed-methods design is embedded in a broader theoretical framework that guides and informs theorizing throughout the entire research process. As can be seen, we advocate for a sequential design that moves from mainly quantitative to mainly qualitative analysis. However, the process must be seen as iterative, as several interpretations guide the estimation of the STM while the derived estimates inform CDA. Thus, the two approaches should be seen as interrelated (Creswell *et al*., 2003), as each step draws on both, albeit to different degrees (see Table 1). Next, we outline how to integrate the two methods in our stepwise model practically.

A stepwise model for combining CDA with STM

Our stepwise model aims at capitalizing on the advantages of combining CDA and STM for zooming in and out of large textual data. Figure 2 illustrates how textual data can be analyzed to identify broad discourses based on the automated identification of relations between different topics, followed by a detailed exploration of discourses and their dynamics. The Figure further shows that whereas steps 2–4 are driven by STM and guided by CDA, steps 5–7 are driven by CDA and informed by STM. As an illustration, we use the legitimacy struggles in the US tobacco industry between 1986 and 2016. While tobacco control regulations are now widely enforced, their enactment did not come without decades-long debates between actors defending particular interests and ideologies. Next, we present each of the steps in detail.

Step 1: ***Choose a theoretical focus***

Having a 'critical' research question and an apt empirical phenomena characterized by power struggles and inequalities—as suggested by a transformative design—the first step entails reflecting on the theoretical approach for the textual analysis. We put forward three non-exhaustive and not mutually exclusive alternatives on how to integrate CDA and STM, each of which has a different emphasis: ideology-based, actor-based, or document-based.

(A) ***Ideology-based approach***

Building on core assumptions of CDA, this approach looks at discursive components that add up to significant social issues such as ideologies. The theoretical focus aims to deepen the understanding and role of linguistic

TABLE 1 A stepwise model to integrate CDA with STM

	Steps	*CDA*	*STM*
1	Choose theoretical focus	**Adopt critical stance and choose discourse, actors, or documents as theoretical focus.**	
2	Collect large textual data	*Use critical reflexivity to guide the selection of corpus and metadata.*	**Choose textual data and prepare the corpus. Collect and prepare relevant metadata.**
3	Define and interpret topics	*Evaluate the most interpretable number of topics to guide topic interpretation.*	**Compare several solutions with different numbers of topics. Inspect bags of words that load on each topic and word associations to assign labels to topics.**
4	Identify discourses based on topic relations	*Use interpretation to cluster topics into overarching discourses.*	**Generate network graph to visualize how topics are correlated.**
5	Explore linkages between discourses and context	**Explore textual data to obtain fine-grained understanding of key phenomena (e.g., time, actors, document type), and to identify key moments and trends.**	*Generate an overview of topics over the metadata of interest (e.g., time, actors, document type, etc.) to inform qualitative analysis.*
6	Select a sample to zoom in on	**Select texts for further analysis.**	*Select the texts to be analyzed in detail.*
7	Code selected texts	**Abductively develop theoretical coding typology.**	*Include codes based on metadata and topic model results of texts.*
8	Develop findings and generalizations	**Draw inferences based on zooming in and out of textual data.**	

In bold: driving method.
In italics: supporting method.

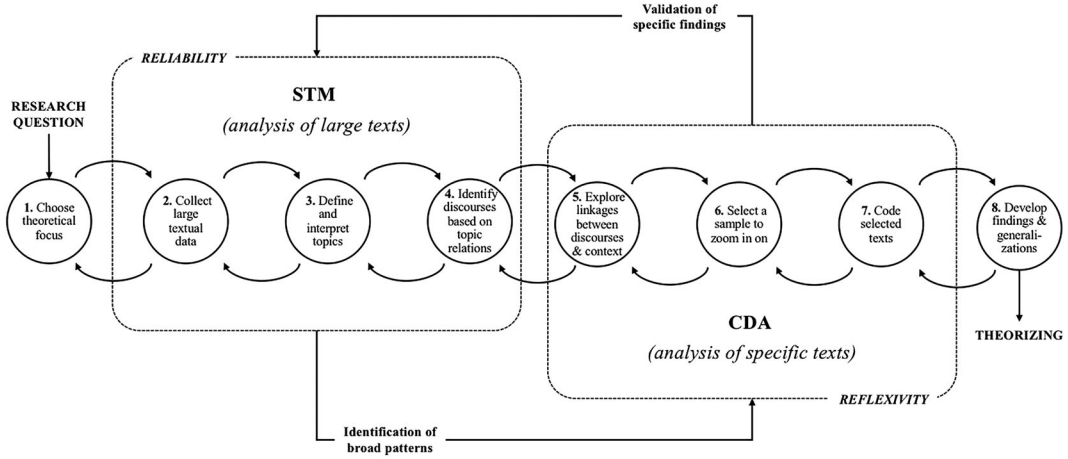

FIGURE 2 A stepwise model to combine STM with CDA

vehicles such as tropes, metaphors, or linguistic processes. Based on closely related topic clusters, broader-level discourses and ideologies can be interpreted (Van Dijk, 1998).

(B) ***Actor-based approach***

Scholars may choose to explore the discursive strategies that actors mobilize to (de)legitimate or re-legitimate specific phenomena (e.g., Suddaby and Greenwood, 2005; Vaara and Tienari, 2008; Vaara, 2014). This approach's theoretical focus is concerned with the actors and their strategies while accounting for the historical and socio-political context (Reisigl and Wodak, 2016; Van Leeuwen, 2016).

(C) ***Document-based approach***

In line with CDA's traditional intertextuality (Fairclough, 1992), specific documents can be seen as the driving force behind the discursive construction of meaning since texts are understood as having an actor-like role in the generation of meaning. This approach makes it possible to explore orders of discourse and genres through time and space (Fairclough, 2003).

Step 2: ***Collect large textual data***

The choice of texts to be included in the corpus is informed by the research question and may include any textual data type. We collected newspaper articles that reflected the public debates around smoking over time using LexisNexis (https://www.lexisnexis.com). We obtained all articles published in *The New York Times* (NYT) between 1986 and 2016 that contained the keywords tobacco,' 'smok!' (i.e., smoking, smokers, etc.), and 'cigarette.' In doing so, we followed research that has identified the NYT as arguably the most influential newspaper in the US (Fiss and Hirsch, 2005). Our initial search resulted in 6,336 articles. In order to validate our sample, we manually skimmed through these articles and discarded those that: had length issues (e.g., were too short or extremely long), were not related to the US, or were unrelated to the tobacco debates (e.g., obituaries). After pre-processing the data, our final sample consisted of 3,688 articles with a mean word count of 701.

Next, we cleaned the data by filtering out stop-words that carry no thematic meaning (e.g., 'you,' 'and,' 'I') and may hamper the estimation, as well as words that occur so infrequently that they are unlikely to be representative for a specific topic. Following standard practice (Hannigan *et al*., 2019), we stemmed words (reduced 'company' and 'companies' to their stem 'compan') throughout the text corpus. Alternative decisions to make when cleaning the data, which we have not used, are whether to lemmatize the words or whether only to use certain parts of the documents (e.g., nouns/verbs, see Hannigan *et al*., 2019). For these steps, we chose the opensource statistical software R and its 'stm' package (Roberts *et al*., 2019), which offers several functions to import/manipulate textual data.[2] Besides preparing the text corpus, the package allows processing additional metadata for each document. Such metadata can be any type of information deemed relevant for exploring relationships between external factors and textual content (e.g., authorship, outlet, date, tone, length, etc.). The critical perspective of CDA provides researchers with the necessary theoretical guidance to identify the appropriate metadata to be included in the STM.

[2]For convenience, we refer the reader to the R package 'stminsights' (Schwemmer, 2018). Most R packages come with a tutorial, which is also available for STM. For creating the text corpus, the reader is referred to https://cran.r-project.org/web/packages/corpus/vignettes/corpus.html. For estimating the STM, the following vignette is very useful: https://cran.r-project.org/web/packages/stm/vignettes/stmVignette.pdf. For additional resources on the 'stm' package, see www.structuraltopicmodel.com

For our case, we use two pieces of meta-information, namely the publication year and the main actors present within each article.[3] We grouped the actors into three interest groups: (1) Tobacco Industry, (2) Government, and (3) Anti-Smoking groups (Aranda and Simons, 2018). Tobacco Industry contains actors that are pro-smoking and whose interests are aligned with those of the companies (e.g., smokers, tobacco growers, etc.). Government includes federal and state authorities. Anti-Smoking groups comprise individuals, health organizations, and other societal members (e.g., the WHO, medical schools, various NGOs). By including metadata on time and actors, we can trace how topics evolve and show which actors are associated with specific topics.[4]

Step 3: ***Define and interpret topics***

Once the corpus is built, and before running the STM, an essential step is to decide how many topics are to be chosen. There is statistically no way to determine how many topics are needed to explain a given text corpus best. It is, therefore, customary to do a grid search over a feasible range of topics (Roberts *et al*., 2019). The range of the grid search depends on the corpus, particularly on the number of documents that are to be studied and the research question, which determines the required level of detail. Generally, the larger the number of documents and the higher the level of detail required, the more topics are needed to capture the corpus' thematic content adequately. In previous work, the number of topics ranges from 12 to 200 (DiMaggio *et al*., 2013; Puranam *et al*., 2017). While it is important to note that there is no single rule for deciding the 'right' number of topics, diagnostic statistics such as the semantic coherence of topics and the held-out likelihood may help select the number of topics that best suits the needs of the researcher.[5] In specific, the statistics provided can be used to identify: (1) statistical saturation in terms of number of topics; and (2) superior solutions among neighboring solutions (Kuhn, 2018). Inspection of Figure 3 informs us that, as expected, the model captures more nuances as we add more topics (k). Specifically, we observe that both semantic coherence and held-out likelihood improve significantly up until somewhere between 40–50 topics, but observe decreasing improvements for these criteria beyond that number of topics, whereas more than 100 topics seems to be undesirable. While not definitive, these results suggests that solutions with 100 or less topics identify the most discriminating topics, that is, those characterized by distinct top identifying words (see also Schmiedel *et al*., 2018). Moreover, the statistics are of particular importance to identify solutions that clearly outperform neighboring solutions. In general, we observe a relatively monotonous increase in these statistics, without observing many outliers, implying that there is no solution clearly outperforming neighboring solutions.

At this point, the quantitative assessment should be cross-checked with a qualitative one. In CDA, the decision on the number of topics is intimately linked to the actual topic interpretation. Once the STM is estimated, the researcher needs to make sense of the bags of words associated with each topic for the retained solution. The word associations with a specific topic are an essential source of information to label the topics. There are various word association indices (e.g., probability, FREX, lift, and score), and several statistics and visualization tools that can prove useful in this step (Roberts *et al*., 2019).[6] However, it is often incredibly insightful to read a set of representative documents in each topic. Such a manual and interpretational assessment allows for creating an in-depth knowledge of the empirical material and guides the interpretation of topics resulting in meaningful themes that provide a contextual understanding of the topics from a critical stance. Moreover, topics are best interpreted with the research question and the theoretical focus in mind (step 1). As usual, in inductive research, emergent insights may lead to a refinement of the initial focus.

In our example, we inductively derived detailed and distinct labels for each of the topics to capture their richness. We first did so for a wide range of topics (20, 25, 30, 40, 50, 100), to determine the correct trade-off between detail and abstraction, after which we focused on the different solutions around the 40–50 topics optimal point—which was considered the right level of abstraction, but also were identified using Figure 3 and the statistics discussed above—to find the model that offered the best depiction of the debates' key themes. In specific, we inspected several solutions (i.e., 40, 43, 46, 48, 49, 50) that performed well from a statistical perspective, using the measures identified earlier. While new themes emerged as we increased the number of topics, we seemingly reached theoretical saturation at 43

[3]There are several methods to extract actors in a given corpus (Pinto *et al*., 2016). To deal with possible endogeneity concerns, we have compared the STM solutions including and excluding actors as covariates; the results are substantively similar.

[4]Using publication date as a covariate to explain topic prevalence, STM allows for time-correlated topics.

[5]In our case, we use the two most common diagnostic statistics, semantic coherence and held-out likelihood (Wallach *et al*., 2009; Mimno *et al*., 2011), to evaluate the outcome of models with different numbers of topics (see Figure 3). The semantic coherence of topics indicates the degree to which the top words in each topic co-occur. This has been found to correlate strongly with the human judgment of topic quality (Mimno *et al*., 2011). The held-out likelihood summarizes how well the estimated model predicts the occurrence of topics in held-out data, that is,, in data that was dropped from the actual model estimation (Wallach *et al*., 2009). Another diagnostic statistic that can be used is exclusiveness, which measures the degree to which the top words in topics are restricted to the focal topic (Schmiedel *et al*., 2018), but it is not readily shown using the STM plotting function.

[6]To aid interpretation, one can use the R package 'stminsights' to create visualizations.

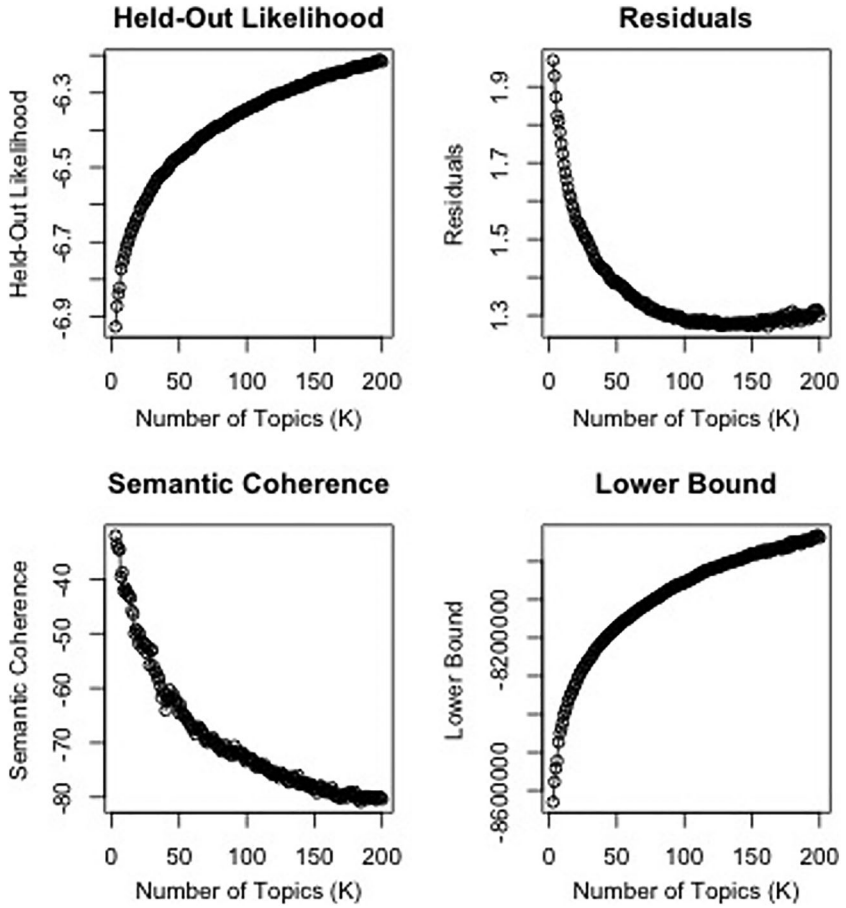

FIGURE 3 Diagnostic values by number of topics

topics. Indeed, for the 43-topic solution, we could assign an exact thematic meaning to almost all topics, which indicates that the model identifies the main discourses in the debate—not requiring the additional complexity associated with interpreting a larger number of topics. Since the qualitative interpretation results were coherent with the results of the previous quantitative analyses, we settled for an STM with 43 topics. Table 2 gives an overview of the thematic meaning of each topic.

Step 4: ***Identify discourses based on topic relations***

Next, we explore topic relations to derive broader meaning structures (i.e., discourses). For this purpose, it is useful to display the linkages between the different topics in a correlation graph, which captures the network of topic relations and highlights the likelihood of two topics being discussed together. The correlation graph may reveal clusters of topics or identify topics that are distant from one another.

Figure 4 presents the correlation graph for our empirical example. We explored Figure 4 in detail to understand the correlations across topics and identify whether those particular topics that are displayed close to each other could be aggregated into broader discourses. We identify four main discourses through the manual analysis of topic clusters. In the upper part, we see a cluster around *health,* which relates to smoking's health consequences. There is a cluster around *marketing* in the middle*,* which is related to the advertising strategies of the tobacco companies. On the lower left-hand side, the *legal* cluster represents the lawsuits faced by the industry. On the lower right-hand side, there is a *regulatory* cluster that comprises different tobacco control regulations. These clusters provide a comprehensive discursive profile characterizing the tobacco debates.

Step 5: ***Explore linkages between discourses and context***

To explore linkages between discourses and context, we consider the metadata included in the estimation of the STM and qualitatively explore the data to identify key moments and trends of interest. The STM estimated the proportions by which the metadata are linked to each topic. Since the amount of metadata that can be meaningfully accounted for within the STM is practically limited (e.g., actors need to be grouped coherently), CDA can enrich the analysis by providing a more fine-grained understanding. Hence, at this point, we switch from

TABLE 2 Thematic content of topics*

Topic	Issue	Top words	Topic	Issue	Top words
1	Health research	smoke, percent, studi, year, smoker	23	Trial	court, judg, rule, case, appeal
2	Stock market	compani, tobacco, rjr, nabisco, analyst	24	Smoking Bans	smoke, restaur, bar, law, ban
3	Indian reservation taxes	indian, state, tax, reserv, cigarett	25	MSA	settlement, industri, tobacco, state, attorney
4	Tar content	cigarett, smoker, smoke, tar, light	26	Nicotine	regul, tobacco, drug, nicotin, agenc
5	Cessation	nicotin, smoke, smoker, quit, addict	27	Heart disease	studi, research, risk, smoke, heart
6	Business strategies	brand, brown, american, williamson, agenc	28	Branding	brand, marlboro, market, like, promot
7	Product placement	smoke, campaign, tobacco, california, anti	29	Vending machines	tobacco, cigarett, law, state, machin
8	---	build, art, apart, street, new	30	Youth access	smoke, children, school, teen, use
9	Presidential actions	clinton, presid, hous, administr, white	31	Tobacco documents	compani, document, tobacco, industri, research
10	Lung cancer	cancer, smoke, lung, death, diseas	32	Public places	smoke, counti, ban, park, prison
11	Civil cases	tobacco, compani, damag, class, case	33	Lobbying	state, lobbi, tobacco, group, money
12	Senate bill	bill, senat, legisl, tobacco, republican	34	Advertising	advertis, cigarett, tobacco, camel, compani
13	Packaging	tobacco, product, warn, health, use	35	---	cigarett, book, like, man, one
14	Corporate strategies	philip, morri, compani, altria, cigarett	36	Cigarette sales	cigarett, new, state, sell, sale
15	Sponsorship	billboard, player, sign, tobacco, game	37	City bans	citi, new, york, mayor, council
16	Menthol	cigarett, menthol, tobacco, health, product	38	Tobacco farmers	tobacco, farmer, year, carolina, north
17	Cigars	cigar, pipe, smoke, year, store	39	Stories of smokers	smoke, one, year, peopl, say
18	Taxes	state, tax, new, year, budget	40	Tobacco companies	reynold, compani, tobacco, lorillard, cigarett
19	Non-smokers	smoke, ban, nonsmok, smoker, health	41	Lawsuit	smoke, cigarett, tobacco, case, compani
20	Companies' profits	percent, price, billion, share, bond	42	---	public, can, will, peopl, may
21	Legal strategies	lawyer, state, tobacco, lawsuit, suit	43	Academic reports	univers, health, professor, medic, research
22	Cigarette prices	tax, cigarett, increas, pack, price			

* All topics were assigned to a theme based on the 'bags of words' that relate to that specific topic. Although the majority of themes were easy to identify using the most prevalent words, three topics (8, 35, 42) could not be clearly assigned to a single issue. These topic contains words that did not reflect any meaning in particular.

STM to CDA as the driving force of the analytical process.

Figure 5 presents the relation of actors and topics in our example and informs our qualitative interpretation.[7] We see that anti-smoking groups mainly draw on health discourse. The legal discourse is mainly used among the government and the industry. Interestingly, the regulatory and marketing discourses are used somewhat equally by all three actors, albeit from opposite perspectives. While the industry links regulatory and marketing discourses to an economic neoliberal discourse, anti-smoking groups and the government relate it to a discourse that foregrounds public well-being. Thus, the way discourses are used by different actors provides preliminary insights on their different roles in shaping the broader debates.

Moreover, to uncover the evolution of topic proportions over time, we visually examined Figure 6, which shows each marketing-related topic's time dynamics.[8] Specifically, Figure 6 tells us the expected topic proportion of the selected topics: the proportion of each topic *k* in the corpus. This step is essential for identifying key moments of interest, for example, to pinpoint when a specific topic is most or least used. It helps understand how topics evolve and reveals the underlying forces shaping the broader discourses, providing the ground for critical reflection on how the relations between actors evolve in the debates.

Step 6: ***Select a sample to zoom in on***

At this point, exemplary texts need to be selected to engage in a qualitative analysis of specific features. In CDA, it is typical to select texts based on contextual analysis, but usually without a systematic analysis of the broader dataset. By starting with STM, CDA scholars may significantly enlarge the datasets they can consider, allowing sampling to happen at a later stage in the research process. Since we wish to identify broad patterns based on large textual data and intend to explore actors' legitimation strategies over time in detail, we use CDA to interpret texts that reflect structuring moments within the debate. In practice, we focus on the points in time and on those documents most instructive about changes reflected in discursive dynamics.

In our example, and for illustrative purposes, we focus on the marketing discourse. We explored the time dynamics illustrated in Figure 6 to identify critical moments during which the marketing discourse seemed particularly prevalent. We picked topic #16, *menthol in cigarettes*, as particularly impressive given its sharp increase in the later years of analysis. A critical reading of the texts helped us to understand the detailed content of the topic in any year of interest. For example, in 2013 it revolved around banning menthol in cigarettes and flavors in cigars after 'Congress exempted menthol from

[7] In order to calculate the association between topics and actors, we drew on the coefficient estimates of the topic prevalence model within the STM. Actors with a higher (absolute) coefficient are comparatively more likely associated with a particular topic.

[8] All other plots are available from the authors.

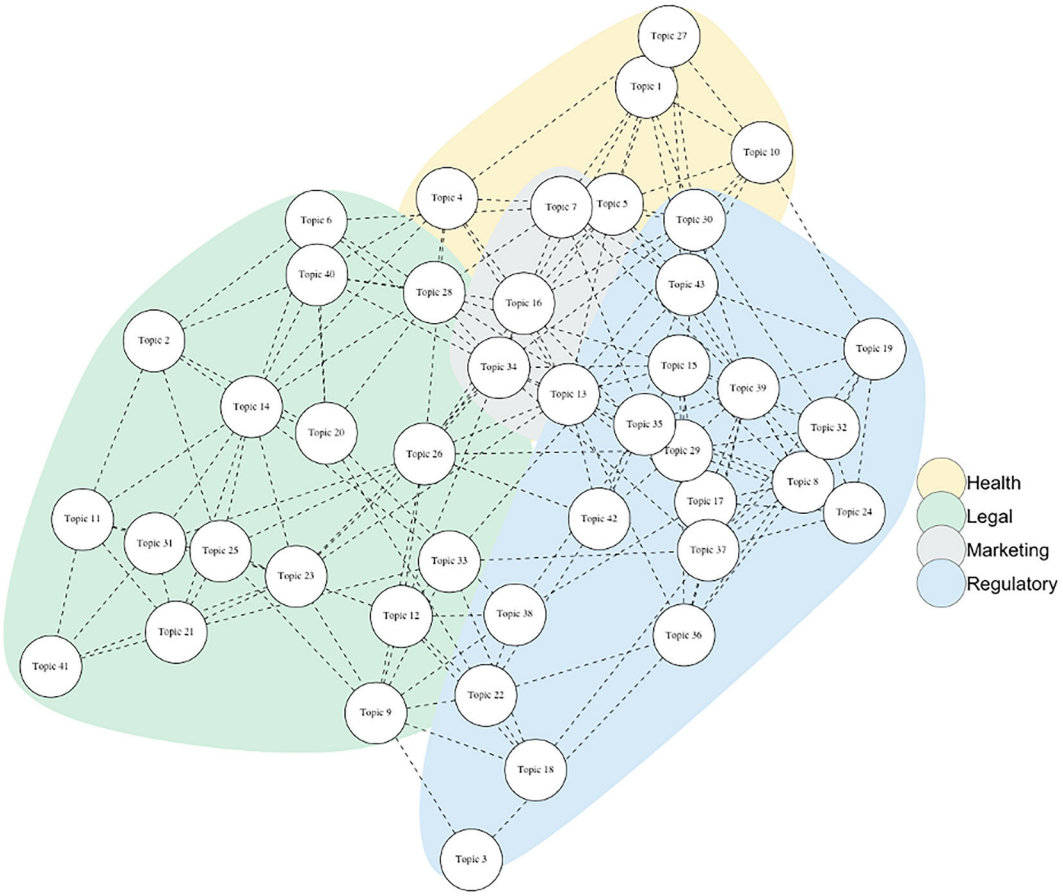

FIGURE 4 Topic relations and associated discourses [Colour figure can be viewed at wileyonlinelibrary.com]

a ban on flavors in cigarettes [in 2009]' and 'the law was silent on flavors in cigars.'

Step 7: ***Code selected texts***

The next step of our model focuses on the actual coding of texts and their interpretation thereof. As is usual in CDA, this analytical step is abductive, meaning that theorizing happens iteratively with an increasingly focused analysis (Wodak and Meyer, 2016). The actual coding depends on the research question and the theoretical focus (step 1). In our example, we chose the actor-based approach, which allowed us to analyze legitimation strategies within the previously identified discourses. The coding typology naturally includes codes that come from the metadata (i.e., actor, time) and codes derived based on the close reading of the text. Accordingly, the interpretation builds on quantitative and qualitative results (Creswell, 2009).

Extant research has established various discursive legitimation strategies that provide varying typologies to choose the starting point for CDA. In our exemplary case, we used Van Leeuwen's (2007) framework of authorization, rationalization, moralization, and mythopoiesis (see Table 3). We examined the selected texts, coded for a specific period, actor, and discourse based on the STM results. Within the texts, we focused on paragraphs that revealed the actors' discursive legitimation strategies to shape the tobacco debate in their interest. For illustration purposes, we conduct the analysis for topic #16 (*menthol in cigarettes*) in 2013.

Table 3 illustrates the strategies of (de)legitimation in the discursive struggles surrounding tobacco. The results show how government actors did not yet take a stance, postponing the announcement of actions into the future, a strategy that is potentially due to the change in leadership in the tobacco unit of the U.S. Food and Drug Administration. In contrast, anti-smoking groups advocated for regulating menthol in tobacco products using all four (de)legitimation strategies. Interestingly, the tobacco industry sought to legitimate smoking of menthol cigarettes, arguing against regulations by referencing a lack of scientific insights and using rationalization. It becomes clear that while the government was positively associated with the menthol issue over time, in 2013, the topic was pushed onto the government's agenda by anti-smoking groups and the tobacco industry. As the illustrative analysis above indicates, CDA is very useful in refining and broadening

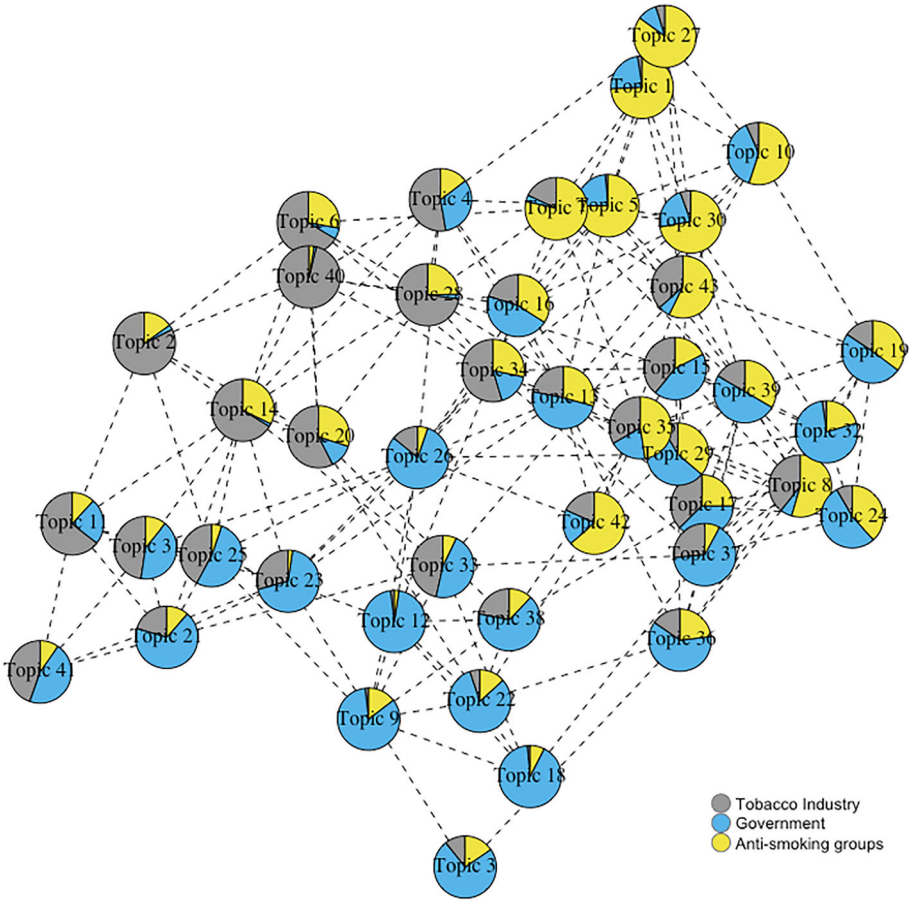

FIGURE 5 Topic relations and associated actors [Colour figure can be viewed at wileyonlinelibrary.com]

theoretically relevant interpretations of STM results, which on their own would not have revealed these complex actor/topic/time constellations.

Step 8: ***Develop findings and generalizations***

To further strengthen the analysis with an additional layer, we propose a last step, which brings together insights from both methods. We integrate the actors' legitimation strategies (see Table 3) and time dynamics (see Figure 6). After analyzing the sharper increase of the menthol issue (topic #16) at the end of the period of study, we enhanced our interpretation of the articles with the respective metadata. Specifically, we explored associations between actors and articles or between actors and topics, which provided us with an indication of the discursive struggles' peculiarities at a specific point in time. For example, we uncovered that the texts discussing menthol had a strong relationship with topic #30 (youth access).

We explored the connection between the menthol issue (marketing) and youth access (regulatory) (see Figure 7). Closer reading suggests that this connection reflects a debate about whether and how menthol cigarettes lure young people. Specifically, the articles associated with the increasing discussions about youth in and after 2013 showed a debate on smoking among children and teens. Within this debate, anti-smoking groups referenced statistics supporting the increase in youth smoking (rationalization) and projected that e-cigarettes would eventually lead young people to smoke regular cigarettes (mythopoesis). They further linked menthol and youth using moral evaluations, while the tobacco industry tried to dissociate these issues using rationalization and mythopoesis to legitimate smoking and protect the developing e-cigarette market.

Discussion and conclusion

This article introduces a stepwise model to combine CDA with STM that allows management scholars to confront the ever-increasing amount of textual data in our mediatized society. Building on a transformative explanatory mixed-methods research design (Creswell, 2009), we suggest that CDA, a qualitative discursive approach, can be enhanced with STM, a sophisticated topic modeling application. As shown by our illustrative example, if the two approaches are combined and considered mutually constitutive, researchers can draw inferences that are empirically

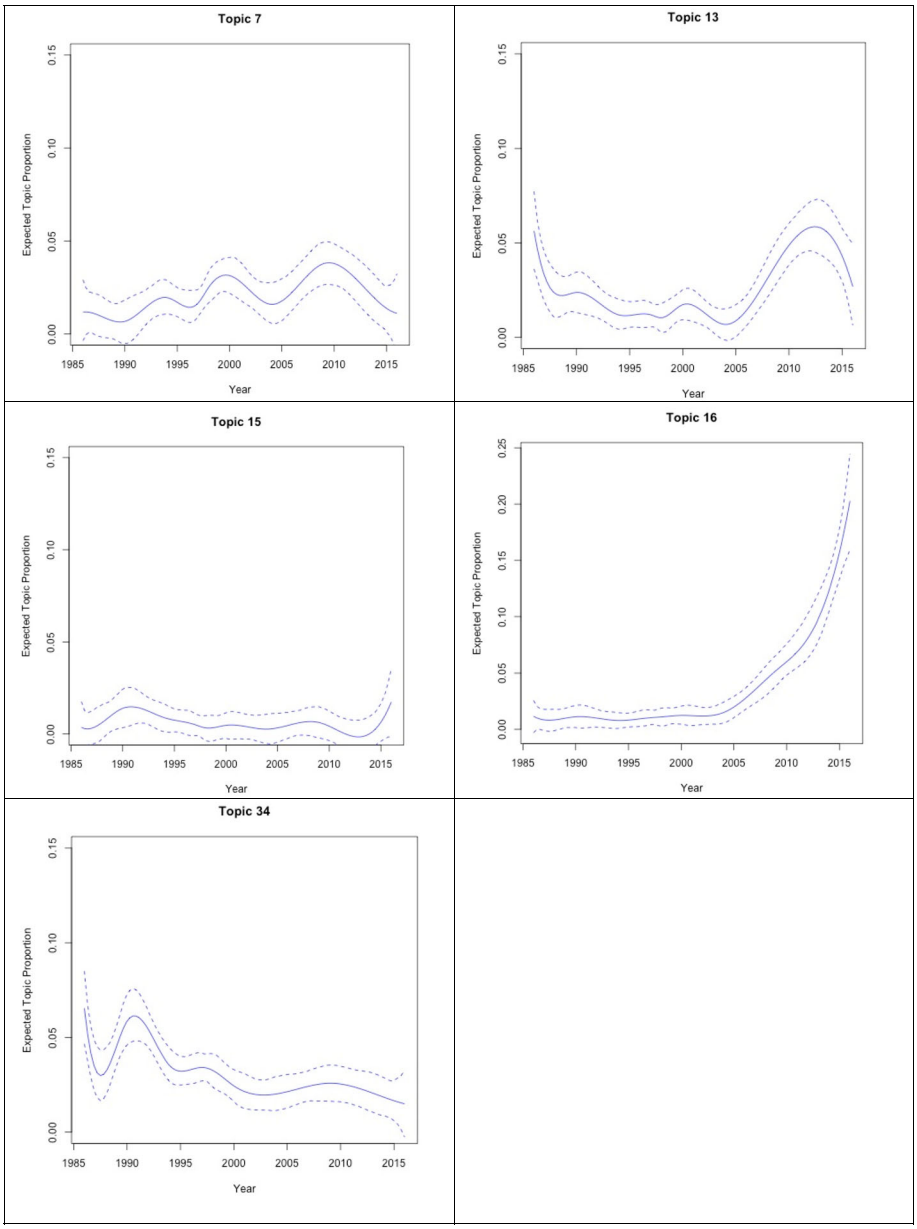

FIGURE 6 Time dynamics (selected marketing topics) [Colour figure can be viewed at wileyonlinelibrary.com]

validated and theoretically grounded. Whereas STM helps reveal the occurrence of topics and discourses within large datasets, CDA and its analytical techniques enable the researcher to gain in-depth insights. The combination of both approaches yields a fine-grained understanding of broad but theoretically relevant patterns. Accordingly, our model enhances Hannigan *et al.*'s (2019) notion of 'rendering' by iteratively zooming in and out of textual data. By doing so, our study makes two important contributions to management research.

First, our model broadens the repertoire of how we can approach and analyze the role of discourses in a wide range of organizational phenomena (Phillips and Oswick, 2012). Studying organizations through a discursive lens has gained popularity, and CDA has allowed for many interesting insights on how we understand management processes and practices (Phillips *et al.*, 2008; Chouliaraki and Fairclough, 2010; Vaara, 2010). Interested in the overarching relationship between language and power, CDA has traditionally relied on qualitative methods based on in-depth critical analysis of texts in their broader context (Fairclough, 2013; Wodak and Meyer, 2016). However, with the increasing availability of texts, manual analysis has become more difficult, impractical, or in some instances, even unmanageable, complicating the identification of discourses and their dynamics. Further, accounting for context, intertextuality, and interdiscursivity—a central

TABLE 3 Legitimation strategies (based on Van Leeuwen, 2007 and Vaara, 2014)

Legitimation strategies	*Definition*	*Empirical example*	*Actors/Adjunct topics and discourses*
Authorization	Legitimation by reference to personal authority (e.g. experts or role models) and to impersonal authority (e.g. tradition, custom and law)	*Legitimating ban on menthol cigarettes:* The White House should free the F.D.A. to ban menthol flavoring in cigarettes to protect the health and save the lives of Americans, especially minority Americans. The European health ministers agreed last month to ban menthol cigarettes to curb youth smoking.	Anti-smoking groups referencing European health ministers as expert authorities
Rationalization	Legitimation by reference to the goals and uses of institutionalized social action, and to the knowledge society has constructed to endow them with cognitive validity (e.g. through science or definitions)	*Legitimating and delegitimating ban on menthol cigarettes:* The FDA released a scientific review on Tuesday that found that mint flavoring made it easier to start smoking and harder to quit [which] pleased smoking opponents. […] Lorillard, the biggest manufacturer of menthol cigarettes in the US, said in a statement that 'the best science demonstrates that menthol cigarettes have the same health effects as nonmenthol cigarettes and should be treated no differently.'	Anti-smoking groups as well as the tobacco industry are both referencing scientific knowledge to rationalize their arguments.
Moralization	Legitimation by reference to value systems (e.g. through evaluative adjectives, abstractions, analogies)	*Legitimating ban of flavors in cigars* Nothing is more popular than a chocolate-flavored little cigar. They are displayed just above the Hershey bars along with their colorful cigarillo cousins – white grape, strawberry, pineapple. […] Smoking opponents contend that the agency's delay [in banning flavors in cigars] is threatening recent progress in reducing smoking among young people.	Anti-smoking groups create an analogy between flavored cigars and candy and use evaluative adjectives.
Mythopoiesis	Legitimation conveyed through narratives that project the future by contrasting it to the present (e.g. imaginaries, dramatization)	*Legitimating and delegitimating (often flavored) e-cigarettes ban* The share of middle and high school students who use e-cigarettes doubled in 2012 from the previous year. […] 'This is really taking off among kids' said Dr. Frieden, director of the CDC. Producers promote them as a health alternative to smoking, but researchers say their health effects are not yet clear. […] Murray Kessler, the chief executive of Lorillard, said that the rise in youth usage was "unacceptable" and added that the company was 'looking forward to a regulatory framework that restricts youth access" but does not "stifle what may the most significant harm reduction opportunity that has ever been made available to smokers.'	Anti-smoking groups dramatize the use of (flavored) e-cigarettes among young people projecting a high increase. The tobacco industry's claims sustain that flavors should be regulated only for young people, for it foresees that regulating them broadly will deter smokers from switching to an allegedly less harmful product.

feature of CDA—in a meaningful way is challenged due to difficulties in moving between analytical levels (Leitch and Palmer, 2010).

Our article explicitly addresses these difficulties by proposing to enhance CDA with STM. Although there may be other related mixed methods, we argue that STM offers a particularly useful automated method to explore and analyze large textual data, as it allows us to discover patterns and meaning structures in a way that is in line with CDA's epistemological assumptions. In practice, STM can provide not only an important methodological component for CDA scholars focusing on how specific discourses and strategies are used in and around organizations, but also enable new research questions around complex longitudinal and multiactor processes, which would be more difficult to address without this methodological integration. While our paper's main contribution is that it elucidates how CDA can benefit

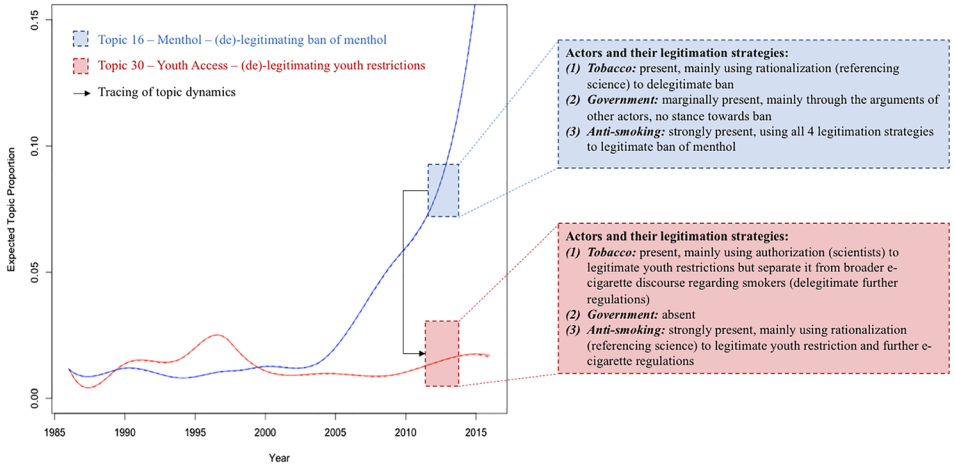

FIGURE 7 Tracing actors' legitimation strategies through time and across topics [Colour figure can be viewed at wileyonlinelibrary.com]

from STM in the field of management and beyond, STM's approach of seeing discourses as composed of interrelated topics may also help other discursive approaches, such as interpretive discourse analysis (Heracleous, 2006), and rhetorical analysis (Sillince and Suddaby, 2008; Heracleous *et al*., 2020). To combine these approaches scholars need to consider their integration from a methodological and epistemological perspective, similar to our present analysis, an aspect that is often missing when simply transferring quantitative approaches into traditionally qualitative fields.

Second, and in line with our call for a mixed-methods approach, we maintain that CDA helps make theoretical sense of STM results. Despite the many advantages of using STM when analyzing large textual datasets, this method must always be combined with higher-level theoretical reasoning and researcher driven interpretation. This issue has become apparent in recent advances, which warn about the dangers and illusions of pure empiricism. Indeed, prior work has argued that different approaches can significantly add to topic modeling's standard applications (e.g., Baumer *et al*., 2017; Croidieu and Kim, 2018; Hannigan *et al*., 2019). We offer essential insights into this line of work by adding two crucial aspects of STM in which CDA can be beneficial. Firstly, the critical perspective of CDA and its strong consideration for context can help select the inclusion of relevant factors. Second, as STM requires a great deal of interpretation, CDA can provide critically oriented and theoretically grounded guidance once the topics are derived. Moreover, CDA allows exploring theoretical relations within and between discourses, for which STM lacks the analytical depth.

All of this said, there are two critical points to bear in mind. First, STM is not a substitute for the in-depth analysis that CDA scholars typically engage in. We see it as a handle that supports researchers when confronted with large, complex, and unstructured textual data. This means that collecting more data is not always needed; sometimes, a single text can be just as informative as an extensive collection (see Vaara and Tienari, 2008). Therefore, we urge scholars to critically assess whether and why 'more is better' in their particular setting. Second, and related to the first point, bringing together methodological approaches rooted in different research paradigms requires a critical discussion about when and how this is possible (Deetz, 1996).

To conclude, although our model offers stepwise guidance on how to proceed with combining both approaches, we maintain that not all steps need always to be taken and that usually, the analysis progresses iteratively rather than in a linear manner. These are essential points to bear in mind with the kind of inductive reasoning and analysis that CDA and STM are based on. In a nutshell, it is only by 'letting the data speak' that researchers can make the most out of the combined potential of CDA and STM. Nevertheless, it helps to pin down key steps, questions and challenges associated with them. We thus hope future research will apply the stepwise model offered in this paper to study phenomena that matter.

Acknowledgements

We would like to thank the editor Bill Lee and four anonymous reviewers for their support and insightful comments throughout the process. We would also like to thank Mickaël Buffart, Alice Comi, Meri Jalonen, Christoph Rheinberger, Hanna Timonen and Ruth Wodak who provided feedback on earlier drafts of the paper.

Funding

This research has been supported by the Academy of Finland, grant numbers 315665 and 321362.

References

Antons, D., R. Kleer and T. O. Salge, 2016, "Mapping the topic landscape of JPIM, 1984–2013: In search of hidden structures and development trajectories". *Journal of Product Innovation Management*, **33**: 726–749.

Aranda, A. M. and T. Simons, 2018, "On two sides of the smoke screen: How activist organizations and corporations use protests, campaign contributions, and lobbyists to influence institutional change". In Briscoe F., B. G. King and J. Leitzinger (eds.), *Social movements, stakeholders and non-market strategy*. Bingley: Emerald Publishing, pp. 261–315.

Banks, G. C., H. M. Woznyj, R. S. Wesslen and R. L. Ross, 2018, "A review of best practice recommendations for text analysis in R (and a user-friendly app)". *Journal of Business Psychology*, **33**: 445–459.

Barros, M., 2014, "Tools of legitimacy: The case of the Petrobras corporate blog". *Organization Studies*, **35**: 1211–1,230.

Baumer, E. P., D. Mimno, S. Guha, E. Quan and G. K. Gay, 2017, "Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence?" *Journal of the Association for Information Science and Technology*, **68**: 1397–1410.

Blei, D. M., 2012, "Probabilistic topic models". *Communications of the ACM*, **55**: 77–84.

Blei, D. M., A. Y. Ng and M. I. Jordan, 2003, "Latent dirichlet allocation". *Journal of Machine Learning Research*, **3**: 993–1022.

Brookes, G. and T. McEnery, 2019, "The utility of topic modelling for discourse studies: A critical evaluation". *Discourse Studies*, **21**: 3–21.

Chandra, Y., L. C. Jiang and C.-J. Wang, 2016, "Mining social entrepreneurship strategies using topic modeling". *PLoS One*, **11**: e0151342.

Chouliaraki, L. and N. Fairclough, 2010, "Critical discourse analysis in organizational studies: Towards an integrationist methodology". *Journal of Management Studies*, **47**: 1213–1,218.

Creswell, J. W., 2009. *Research design: Qualitative, quantitative, and mixed methods approaches*. Thousand Oaks, CA: SAGE Publications.

Creswell, J. W., V. L. Plano Clark, M. L. Gutmann and W. E. Hanson, 2003, "Advanced mixed methods research designs". In Tashakkori A. and C. Teddle (eds.), *Handbook of mixed methods in social behavioral research*. Thousand Oaks, CA: SAGE Publications, pp. 209–240.

Croidieu, G. and P. H. Kim, 2018, "Labor of love: Amateurs and lay-expertise legitimation in the early U.S. radio field". *Administrative Science Quarterly*, **63**: 1–42.

Deetz, S., 1996, "Crossroads – Describing differences in approaches to organization science: Rethinking Burrell and Morgan and their legacy". *Organization Science*, **7**: 191–207.

DiMaggio, P., M. Nag and D. Blei, 2013, "Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding". *Poetics*, **41**: 570–606.

Doldor, E., M. Wyatt and J. Silvester, 2019, "Statesmen or cheerleaders? Using topic modeling to examine gendered messages in narrative developmental feedback for leaders". *The Leadership Quarterly*, **30**: 101308.

Fairclough, N., 1989. *Language and power*. London: Longman.

Fairclough, N., 1992. *Discourse and social change*. Cambridge: Polity Press.

Fairclough, N. (Ed), 2003. *Analysing discourse: Textual analysis for social research*. London: Routledge.

Fairclough, N., 2005, "Peripheral vision: Discourse analysis in organization studies: The case for critical realism". *Organization Studies*, **26**: 915–939.

Fairclough, N., 2013. *Critical discourse analysis: The critical study of language*, 2nd ed. London: Routledge.

Fairclough, N., 2016, "A dialectical-relational approach to critical discourse analysis in social research". In Wodak R. and A. D. Meyer (eds.), *Methods of critical discourse studies*. London: SAGE Publications, pp. 86–108.

Fairclough, N. and R. Wodak, 1997, "Critical discourse analysis". In Van Dijk T. A. (ed.), *Discourse as social interaction*. London: SAGE Publications, pp. 258–284.

Fairhurst, G. T. and M. Uhl-Bien, 2012, "Organizational discourse analysis (ODA): Examining leadership as a relational process". *The Leadership Quarterly*, **23**: 1043–1,062.

Farrelly, M., 2020, "Rethinking intertextuality in CDA". *Critical Discourse Studies*, **17**: 359–376.

Fiss, P. C. and P. M. Hirsch, 2005, "The discourse of globalization: Framing and sensemaking of an emerging concept". *American Sociological Review*, **70**: 29–52.

Forchtner, B. and R. Wodak, 2018, "Critical discourse studies: A critical approach to the study of language and communication". In Wodak R. and B. Forchtner (eds.), *The Routledge handbook of language and politics*. Abingdon: Routledge, pp. 135–149.

Gioia, D. A. and E. Pitre, 1990, "Multiparadigm perspectives on theory building". *Academy of Management Review*, **15**: 584–602.

Grant, D., C. Hardy, C. Oswick and L. Putnam (Eds), 2004. *The SAGE handbook of organizational discourse*. London: SAGE Publications.

Green, S. E., Y. Li and N. Nohria, 2009, "Suspended in self-spun webs of significans: A rhetorical model of institutionalization and institutionally embedded agency". *Academy of Management Journal*, **52**: 11–36.

Grimmer, J. and B. M. Stewart, 2013, "Text as data: The promise and pitfalls of automatic content analysis methods for political texts". *Political Analysis*, **21**: 267–297.

Haans, R. F., 2019, "What's the value of being different when everyone is? The effects of distinctiveness on performance in homogeneous versus heterogeneous categories". *Strategic Management Journal*, **40**: 3–27.

Hannigan, T., R. Haans, K. Vakili, H. Tchalian, V. Glaser, M. Wang, S. Kaplan and D. Jennings, 2019, "Topic models in management research: Rendering new theory from textual data". *Academy of Management Annals*, **13**: 586–632.

Hardy, C. and N. Phillips, 2004, "Discourse and power". In Grant D., C. Hardy, C. Oswick and L. L. Putnam (eds.), *The SAGE handbook of organizational discourse*. Thousand Oaks, CA: SAGE Publications, pp. 299–316.

Hassard, J., 1988, "Overcoming hermeticism in organization theory: An alternative to paradigm incommensurability". *Human Relations*, **41**: 247–259.

Heracleous, L., 2006. *Discourse, interpretation, organization*. Cambridge: Cambridge University Press.

Heracleous, L. and M. Barrett, 2001, "Organizational change as discourse: Communicative actions and deep structures in the context of information technology implementation". *Academy of Management Journal*, **44**: 755–778.

Heracleous, L., S. Paroutis and A. Lockett, 2020, "Rhetorical enthymeme: The forgotten trope and its methodological import". *European Management Review*, **17**: 311–326.

Huang, A. H., R. Lehavy, A. Y. Zang and R. Zheng, 2018, "Analyst information discovery and interpretation roles: A topic modeling approach". *Management Science*, **64**: 2833–2855.

Jacobs, T. and R. Tschötschel, 2019, "Topic models meet discourse analysis: A quantitative tool for a qualitative approach". *International Journal of Social Research Methodology*, **22**: 469–485.

Kaplan, S. and K. Vakili, 2015, "The double-edged sword of recombination in breakthrough innovation". *Strategic Management Journal*, **36**: 1435–1,457.

Knights, D. and G. Morgan, 1991, "Corporate strategy, organizations, and subjectivity: A critique". *Organization Studies*, **12**: 251–273.

Kobayashi, V. B., S. T. Mol, H. A. Berkers, G. Kismihók and D. N. Den Hartog, 2018, "Text mining in organizational research". *Organizational Research Methods*, **21**: 733–765.

Kuhn, K. D., 2018, "Using structural topic modeling to identify latent topics and trends in aviation incident reports". *Transportation Research Part C: Emerging Technologies*, **87**: 105–122.

Lefsrud, L. M. and R. E. Meyer, 2012, "Science or science fiction? Professionals' discursive construction of climate change". *Organization Studies*, **33**: 1477–1506.

Leitch, S. and I. Palmer, 2010, "Analysing texts in context: Current practices and new protocols for critical discourse analysis in organization studies". *Journal of Management Studies*, **47**: 1194–1212.

Lewis, M. W. and A. I. Grimes, 1999, "Metatriangulation: Building theory from multiple paradigms". *Academy of Management Review*, **24**: 672–690.

Maguire, S. and C. Hardy, 2009, "Discourse and deinstitutionalization: The decline of DDT". *Academy of Management Journal*, **52**: 148–178.

Mantere, S. and E. Vaara, 2008, "On the problem of participation in strategy: A critical discursive perspective". *Organization Science*, **19**: 341–358.

Mertens, D. M., 2012, "Transformative mixed methods: Addressing inequities". *American Behavioral Scientist*, **56**: 802–813.

Mimno, D., H. M. Wallach, E. Talley, M. Leenders and A. McCallum, 2011, Optimizing semantic coherence in topic models. Paper presented at the Proceedings of the conference on empirical methods in natural language processing.

Molina-Azorin, J. F., D. D. Bergh, K. G. Corley and D. J. Ketchen Jr., 2017, "Mixed methods in the organizational sciences: Taking stock and moving forward". *Organizational Research Methods*, **20**: 179–192.

Morgan, G., 1983. *Beyond method*. Newbury Park, CA: SAGE Publishing.

Mumby, D. K., 2011, "What's cooking in organizational discourse studies? A response to Alvesson and Kärreman". *Human Relations*, **64**: 1147–1161.

Phillips, N. and C. Oswick, 2012, "Organizational discourse: Domains, debates, and directions". *Academy of Management Annals*, **6**: 435–481.

Phillips, N., G. Sewell and S. Jaynes, 2008, "Applying critical discourse analysis in strategic management research". *Organizational Research Methods*, **11**: 770–789.

Pinto, A., H. Gonçalo Oliveira and A. Oliveira Alves, 2016, Comparing the performance of different nlp toolkits in formal and social media text. SLATE: 1–16. https://doi.org/10.4230/OASIcs.SLATE.2016.3

Potter, J. and M. Wetherell, 1987. *Discourse and social psychology: Beyond attitudes and behaviour*. Los Angeles, CA: SAGE Publications.

Puranam, D., V. Narayan and V. Kadiyali, 2017, "The effect of calorie posting regulation on consumer opinion: A flexible latent Dirichlet allocation model with informative priors". *Marketing Science*, **36**: 726–746.

Reisigl, M. and R. Wodak, 2016, "The discourse-historical approach (DHA)". In Wodak R. and A. D. Meyer (eds.), *Methods of critical discourse studies*. London: SAGE Publications, pp. 23–61.

Roberts, M. E., B. M. Stewart and E. M. Airoldi, 2016, "A model of text for experimentation in the social sciences". *Journal of the American Statistical Association*, **111**: 988–1003.

Roberts, M. E., B. M. Stewart and D. Tingley, 2019, "STM: An R package for structural topic models". *Journal of Statistical Software*, **91**: 40.

Samdanis, M. and S. H. Lee, 2019, "Access inequalities in the artistic labour market in the UK: A critical discourse analysis of precariousness, entrepreneurialism and voluntarism". *European Management Review*, **16**: 887–907.

Scherer, A. G., 1998, "Pluralism and incommensurability in strategic management and organization theory: A problem in search of a solution". *Organization*, **5**: 147–168.

Schmiedel, T., O. Müller and J. vom Brocke, 2018, "Topic modeling as a strategy of inquiry in organizational research: A tutorial with an application example on organizational culture". *Organizational Research Methods*, **22**: 941–968.

Sillince, J. A. A. and R. Suddaby, 2008, "Organizational rhetoric: Bridging management and communication scholarship". *Management Communication Quarterly*, **22**: 5–12.

Sonenshein, S., 2010, "We're changing-or are we? Untangling the role of progressive, regressive, and stability narratives during strategic change implementation". *Academy of Management Journal*, **53**: 477–512.

Suddaby, R. and R. Greenwood, 2005, "Rhetorical strategies of legitimacy". *Administrative Science Quarterly*, **50**: 35–67.

Taeuscher, K., R. B. Bouncken and R. Pesch, 2020, "Gaining legitimacy by being different: Optimal distinctiveness in crowdfunding platforms". *Academy of Management Journal*. https://doi.org/10.5465/amj.2018.0620

Tannen, D., H. E. Hamilton and D. Schiffrin, 2015. *The handbook of discourse analysis*. Malden, MA: John Wiley & Sons.

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

Toubia, O. and O. Netzer, 2017, "Idea generation, creativity, and prototypicality". *Marketing Science*, **36**: 1–20.

Vaara, E., 2010, "Taking the linguistic turn seriously: Strategy as a multifaceted interdiscursive phenomenon". *Advances in Strategic Management*, **27**: 29–50.

Vaara, E., 2014, "Struggles over legitimacy in the Eurozone crisis: Discursive legitimation strategies and their ideological underpinnings". *Discourse & Society*, **25**: 500–518.

Vaara, E., B. Kleymann and H. Seristo, 2004, "Strategies as discursive constructions: The case of airline alliances". *Journal of Management Studies*, **41**: 1–35.

Vaara, E., S. Sonenshein and D. Boje, 2016, "Narratives as sources of stability and change in organizations: Approaches and directions for future research". *Academy of Management Annals*, **10**: 495–560.

Vaara, E. and J. Tienari, 2008, "A discursive perspective on legitimation strategies in multinational corporations". *Academy of Management Review*, **33**: 985–993.

Vaara, E., J. Tienari and A. Koveshnikov, 2019, "From cultural differences to identity politics: A critical discursive approach to national identity in multinational corporations". *Journal of Management Studies*. https://doi.org/10.1111/joms.12517

Vaara, E., J. Tienari and J. Laurila, 2006, "Pulp and paper fiction: On the discursive legitimation of global industrial restructuring". *Organization Studies*, **27**: 789–813.

Vaara, E., J. Tienari, R. Piekkari and R. Säntti, 2005, "Language and the circuits of power in a merging multinational corporation". *Journal of Management Studies*, **42**: 595–623.

Van Dijk, T. A., 1998. *Ideology: A multidisciplinary approach*. London: SAGE Publications.

Van Dijk, T. A. (Ed), 2011. *Discourse studies: A multidisciplinary introduction*. London: SAGE Publications.

Van Dijk, T. A., 2016, "Critical discourse studies: A sociocognitive approach". In Wodak R. and A. D. Meyer (eds.), *Methods of critical discourse studies*. London: SAGE Publications, pp. 62–85.

Van Leeuwen, T., 2007, "Legitimation in discourse and communication". *Discourse & Communication*, **1**: 91–112.

Van Leeuwen, T., 2016, "Discourse as the recontextualization of social practice - a guide". In Wodak R. and A. D. Meyer (eds.), *Methods of critical discourse studies*. London: SAGE Publications, pp. 137–153.

Wallach, H. M., I. Murray, R. Salakhutdinov and D. Mimno, 2009, Evaluation methods for topic models. Paper presented at the Proceedings of the 26th International Conference on Machine Learning.

Wilson, A. J. and J. Joseph, 2015, "Organizational attention and technological search in the multibusiness firm: Motorola from 1974 to 1997". *Advances in Strategic Management*, **32**: 407–435.

Wodak, R., 2001, "What CDA is about: A summary of its history, important concepts and its developments". In Wodak R. and A. D. Meyer (eds.), *Methods of critical discourse analysis* Vol. **1**. London: SAGE Publications, pp. 1–13.

Wodak, R. and M. Meyer, 2016, "Critical discourse studies: History, agenda, theory and methodology". In Wodak R. and M. Meyer (eds.), *Methods of critical discourse studies*. London: SAGE Publications, pp. 1–22.