Köln Z Soziol
https://doi.org/10.1007/s11577-026-01065-w

Kölner Zeitschrift für Soziologie und Sozialpsychologie

RESEARCH ARTICLE

Computational Text Analysis for Building and Testing Social Theory

Miriam Hurtado Bodell · Marc Keuschnigg · Ana Macanovic · Anastasia Menshikova

Received: 14 December 2025 / Accepted: 24 February 2026

Abstract Digitization and advances in natural language processing have transformed how sociologists can measure, model, and interpret social life through text. We provide an overview of computational text analysis as a methodological tool kit for building and testing social theory. The field is moving from descriptive uses toward theory-driven and causal inference approaches, though methodological standards—especially around data quality, reproducibility, and causal claims—remain inconsistent. Organizing approaches into data-first, theory-first, and theory–data integration paradigms, we highlight how different methods each balance inductive discovery with theoretical specification. We conceptualize text-analytic methods as measurement strategies that extract sociologically relevant information from unstructured language data and show how they can be incorporated into both thick descriptions and causal inference workflows. Taken together, various computational text analysis approaches offer researchers new opportunities to recover latent constructs, bridge quantitative scale with qualitative depth, and revitalize interpretive approaches in sociology.

✉ M. Hurtado Bodell · M. Keuschnigg · A. Menshikova
Institute for Analytical Sociology, Linköping University
Norrköping, Sweden
E-Mail: miriam.hurtado.bodell@liu.se

M. Keuschnigg
Institute of Sociology, Leipzig University
Leipzig, Germany

A. Macanovic
Department of Sociology/ICS, Utrecht University
Utrecht, The Netherlands

A. Menshikova
Department of Information Technology, Uppsala University
Uppsala, Sweden

Published online: 18 March 2026

Keywords Natural language processing · Large language models · Text as data · Description · Causal inference

Computergestützte Textanalyse zur Entwicklung und Prüfung sozialwissenschaftlicher Theorien

Zusammenfassung Digitalisierung und Fortschritte in der automatisierten Sprachverarbeitung veränderten die Möglichkeiten der Soziologie zur Vermessung, Modellierung und Interpretation des sozialen Lebens durch Text. Dieser Beitrag gibt einen Überblick über die computergestützte Textanalyse als methodischen Werkzeugkasten zur Entwicklung und Prüfung sozialwissenschaftlicher Theorien. Das Feld entwickelt sich von Beschreibungen hin zu theoriegeleiteten und kausalanalytischen Anwendungen; zugleich bleiben methodische Standards, insbesondere hinsichtlich Datenqualität, Reproduzierbarkeit und kausaler Inferenz, noch immer uneinheitlich. Wir ordnen bestehende Ansätze in datengetriebene, theoriegeleitete und integrierende Paradigmen ein und arbeiten heraus, wie unterschiedliche Methoden das Spannungsverhältnis zwischen induktiver Exploration und theoretischer Spezifikation ausbalancieren. Textanalytische Verfahren verstehen wir als Messstrategien zur Erschließung soziologisch relevanter Informationen aus unstrukturierten Textdaten und verdeutlichen, wie sich diese Verfahren sowohl für dichte Beschreibungen als auch für kausalanalytische Forschungsdesigns nutzen lassen. Insgesamt eröffnet die computergestützte Textanalyse neue Möglichkeiten, latente Konstrukte zu erfassen, quantitative Reichweite mit qualitativer Tiefenschärfe zu verbinden und interpretative Ansätze in der Soziologie neu zu beleben.

Schlüsselwörter Automatisierte Sprachverarbeitung · Large language models · Text als Daten · Deskription · Kausale Inferenz

1 Introduction

With the availability of large corpora of digital and digitized text and advances in computational methods, approaches to uncovering social patterns in textual data—previously confined to close reading, hand coding, and keyword counting—have become increasingly scalable. Under the banner of natural language processing (NLP; Hirschberg and Manning 2015; Jurafsky and Martin 2025) and, more recently, large language models (LLMs) in particular (Brown et al. 2020; Devlin et al. 2019; Kaplan et al. 2020; Radford et al. 2019; Vaswani et al. 2017), computational text analysis has become an integral part of sociologists' methodological tool kit (Edelmann et al. 2020; Jarvis et al. 2021; Bonikowski and Nelson 2022; Macanovic 2022; Stoltz and Taylor 2024; Davidson and Karell 2025). This chapter discusses key approaches of computational text analysis as tools for dimensionality reduction and information extraction in large and unstructured corpora of sociological relevance. We take a measurement perspective on computational text analysis, treating these methods as tools for quantifying social categories and em-

pirical phenomena that were once hard to measure (Grimmer et al. 2022). For text data, this involves compressing mostly latent linguistic information into numerical representations of specific categories. The resulting model outputs can then serve as inputs for analyses ranging from thick description to causal inference that support the development and testing of social theory.

Our selection of NLP techniques highlights their adaptability to sociological inquiry—that is, their usefulness in both describing social phenomena and uncovering the causal mechanisms underlying them. As will become clear, the downstream use of measurements derived from computational text analysis requires little departure from disciplinary conventions. Such measures can be readily integrated into standard causal inference strategies for observational data, including fixed-effects panel regression (Brüderl and Ludwig 2015), regression discontinuity designs (Lee and Lemieux 2014), and statistical matching (Stuart 2010). A prototypical sociological application might use text to measure outcomes, such as public opinion during election campaigns, individual attitudes in the wake of a natural disaster, media framing in digitized newspaper archives, or the evolution of stereotypes in books spanning centuries. Yet, like any data type, text can also serve as a treatment or operate as a confounder that must be accounted for in causal analysis.

We consider traditional text analytical methods with well-understood properties, and we highlight some key developments in (large) language models relevant to sociological research using text as data. We distinguish among data-first, theory-first, and theory–data integration approaches. Methods following the data-first approach emphasize openness to discovery, using computational text analysis to surface latent features that are then interpreted through a theoretical lens and often used for theory development. Theory-first approaches, by contrast, require researchers to specify in advance which linguistic features, categories, or relationships to measure. This leads these methods to be used in deductive ways to test theories. Theory–data integration approaches occupy the middle ground. They combine theory-driven choices in modeling with openness to unexpected patterns. In the past two decades, these distinct approaches have been implemented using specialized text analysis methods, each tailored to particular research objectives and epistemologies.

More recently, the rise of so-called foundation large language models—large, adaptable models that perform a variety of tasks with minimal modification—has begun to loosen the distinctions between these categorical boundaries in practice (Bommasani et al. 2022), as the same models can now be deployed across data-first, theory-first, and theory–data integration approaches. While discriminative language models, so-called encoder models, were primarily used to classify or extract information from text, the newer generation of generative language models has a wider range of analytical capabilities, including language generation. These generative models, also called decoder models, include GPT (OpenAI), Llama (Meta), and Claude (Anthropic). They are transforming text analysis by allowing researchers to use natural language to input text—in the form of instructions and the text to be analyzed—as well as to receive natural langage as output in the form of generated text (Chae and Davidson 2025; Radford et al. 2019). The rise of generative language models shifts computational text analysis from exclusively treating text as data to producing new textual material that can itself become an object of analysis or a com-

ponent of research design. These developments open new avenues, such as iterative adjustment of coding criteria in dialogue with generative models and simplified text analytical pipelines, making it easier than ever before to use text as data (Ibrahim and Voyer 2025; Santana and Nelson 2025). Yet these new developments also pose new challenges and require adaptation in sociologists' workflows.

Regardless of the approach or method, however, most text-as-data analyses have remained primarily descriptive, and they face common challenges when pursuing causal inference. The foremost challenge in using text as data is that the primary goal of text is communication, not recording information useful for scientific research (Benoit 2020): Most textual data are organically generated without research goals in mind. Such "found" textual data pose particular challenges for applying the standard potential outcomes framework (Rubin 1974; see Leitgöb and Keusch 2026 as well as Schwitter et al. 2026 in this issue). A central issue is that the high dimensionality of textual data forces researchers to reduce it to a lower-dimensional representation that aligns with the research question (Gentzkow et al. 2019). This often involves iterative discovery work—for example, manually coding categories for classifiers, revising codebooks, rerunning models with different parameters, or trying different dimensionality reduction techniques. Such iterative processes can complicate causal inference (Egami et al. 2022) because standard frameworks typically assume that treatments and outcomes are defined a priori and do not require discovery work. Consequently, causal inference with text data demands careful attention to workflow design, including data selection and quality, preprocessing, and curation (Hurtado Bodell et al. 2022).

An additional challenge is that computational text analysis often relies on texts from digital platforms or digitized archives. These data are frequently incomplete, nonrepresentative, and under corporate ownership (Guldi 2022; Hurtado Bodell et al. 2022). Data from digitized corpora often fail to include reliable metadata (Macanovic 2022), while online platform data typically lack detailed sociodemographic information, are subject to shifting user populations and evolving online architectures, and are often algorithmically confounded by platform features and affordances (Salganik 2018). Social media data in particular skew younger and more affluent than the general population (Murthy 2024). Such issues pose challenges for both description and causal inference using text as data, making careful attention to method selection essential.

How researchers navigate these challenges depends fundamentally on the relationship between theory and data in their analytical approach. Accordingly, in Sect. 2 we organize approaches to text analysis by whether they prioritize discovery, measurement of predefined concepts, or an integration of both logics, showing how each serves distinct epistemic aims while remaining compatible with established inferential frameworks. Moving beyond a typology of techniques, we then demonstrate how computational text analysis expands sociological inquiry along two dimensions: In Sect. 3, we show how these methods enable thick description at scale, bridging the quantitative–qualitative divide by combining pattern recognition across large corpora with interpretive close reading. In Sect. 4, we examine the growing incorporation of text data into causal inference, reviewing strategies for using text as outcome, treatment, or confounder, while highlighting persistent methodological challenges

Table 1 Categories of computational text methods

Approach	Reasoning	Objectives	Examples
Data-first	Inductive	Identify latent patterns, explore emergent concepts, generate hypotheses	Topic models, word embeddings, BERTopic
Theory-first	Deductive	Measure predefined concepts, test hypotheses	Dictionaries, supervised classifiers (e.g., naive Bayes, BERT), few- and zero-shot classification (e.g., GPT)
Theory–data integration	Mixed inductive–deductive or abductive	Combine theoretically motivated constraints with openness to unanticipated patterns, refine concepts	Structural topic models, seeded topic models, interpretative word embeddings

stemming from the high dimensionality of text and the iterative nature of text modeling. In the concluding section, we contend that computational text analysis is not merely a technical innovation but a substantive expansion of what sociology can measure, model, and interpret.

2 Methods Overview

Computational text analysis methods can be organized around how they handle the relationship between theory and data, a distinction that parallels broader methodological debates about deductive versus inductive reasoning in social science research (Stoltz and Taylor 2024). *Data-first methods* connect to an inductive research approach and work by discovering patterns and structures within text collections, which researchers then interpret as representing meaningful social phenomena. *Theory-first methods*, by contrast, connect to the deductive research approach and require researchers to provide clear definitions and examples of the concepts they want to measure before seeing the data. These methods are then used to identify predefined concepts in texts. *Theory–data integration methods* combine both approaches, following an abductive research logic, using theoretical guidance to focus the analysis while allowing discovered patterns to refine or extend initial concepts. This three-way distinction maps onto familiar technical categories—unsupervised, supervised, and semisupervised learning—while emphasizing the question of how researchers engage with the relationship between theory and data. Table 1 provides a summary of these method categories. Beyond these traditional method families, generative language models have introduced new methodological possibilities that can operate across the three categories and thereby soften the distinctions between approaches. Importantly, their flexibility means that they do not simply represent a theory–data integration method; instead, the same underlying model can be directed toward data-first, theory-first, or theory–data integration tasks depending on the research design. Each approach offers specific advantages for different stages of sociological inquiry, and sociological research questions can require researchers to combine methods from multiple categories to generate novel, interpretable, and theoretically meaningful insights.

2.1 Data-First Methods

Data-first methods are used by researchers open to data-driven discovery. A common use case is to find structures in big text data—aggregated from books, diaries, news media, or online posts—to trace, for example, changes in social norms or the emergence of particular understandings of concepts and events (Kozlowski et al. 2019; Nelson 2021; Voyer et al. 2022; Best and Arseniev-Koehler 2023; Hurtado Bodell et al. 2026). Here, computational text analysis can help identify latent features of the data, which researchers can then interpret in light of theory. In practice, moving from computationally discovered patterns to theoretically meaningful interpretations typically involves comparing data-driven patterns to theoretical expectations, assessing whether they align with or challenge prior understandings, and using that comparison to refine existing theories or generate new ones.

A widely adopted data-first approach is topic modeling. The traditional topic model, latent Dirichlet allocation (LDA), discovers latent thematic structures within collections of text, identifying topics as clusters of words that frequently co-occur in text documents (Blei et al., 2003, 2010). The intuition behind LDA is as follows: Imagine that each document in a corpus is written about a mixture of topics, and each topic can be characterized by the words that are likely to appear when that topic is discussed. For example, a newspaper article about sports might contain words like "game," "goal," and "player," while an article about politics might include words like "election," "candidate," and "policy." Latent Dirichlet allocation works backward from this assumption—it observes the actual words in documents and tries to infer which words go together, and thus forms topics, as well as tries to infer how much of each topic a document contains (Blei et al. 2003; Griffith and Steyvers 2007). That is, if the model observes a lot of words such as "goal" and "player" in the same documents, it will assume that these words together to form a topic and that the topic is strongly present in this document. Topics show up as collections of words that the researchers label depending on how they interpret them (Chang et al. 2009; Nelson 2020). This makes topic models a prime example of a data-first method, as theorizing occurs only after the modeling has been conducted.

Text clustering can also be done with the help of both discriminative and generative language models. Using discriminative models, the texts are first represented using word embeddings from the language model, after which the embeddings are clustered to discover semantically similar documents. Finally, topics are inferred from these clusters (Grootendorst 2022; Katz et al. 2024; Miller and Alexander 2025; e.g., BERTopic model). Researchers have also probed the capabilities of new generative LLMs in extracting relevant themes from text based on simple textual prompts (i.e., written instructions describing the task; De Paoli 2024), with rather moderate success (Mu et al. 2024). Currently, generative LLMs face several limitations in theme extraction: They tend to overemphasize general and popular topics at the expense of more specific and nuanced ones, prioritize content at the beginning or end of texts, and respond to small changes in phrasing with disproportionate changes in topic interpretations (Castellanos et al. 2025).

Beyond clustering, researchers can use different language representation approaches for the exploration of relationships between words (and concepts) in

language. Using word embeddings as multidimensional representations of words, researchers can discover semantic relationships by analyzing how words are used in context. The most common methods work on the assumption that words that appear in similar contexts tend to have similar meanings (Mikolov et al. 2013). This implies that if two words frequently appear in similar contexts, they are likely to share semantic properties. For example, "doctor" and "physician" might both appear near words like "hospital," "patient," and "treatment," suggesting they have similar meaning. Word-embedding algorithms analyze these co-occurrence patterns across large corpora to create numerical representations—embedding vectors—for each word, in which similar vector representations are attributed to semantically similar words (Mikolov et al. 2013; Rodman 2020; Arseniev-Koehler and Foster 2022). This results in a mapping where relationships between words are preserved as geometric relationships, and meaningful relationships can potentially be discovered through vector arithmetic (Kozlowski et al. 2019). For instance, the famous example "king − man + woman ≈ queen" demonstrates how these vector representations capture abstract semantic relationships. In a sociological application, researchers might measure how the semantic distance between words related to affluence and education has shifted over time (Kozlowski et al. 2019), using theoretical frameworks about social class to guide which word relationships to examine and how to interpret the patterns they find. However, recent work has raised important questions about what word embeddings actually capture, since words can align in vector space due to semantic similarity, syntactic substitutability, or mere co-occurrence patterns (Boutyline and Arseniev-Koehler 2025). Consequently, researchers must carefully assess whether the contextual patterns captured by embeddings reflect the conceptual relationships of interest or merely linguistic regularities.

Best practices for the use of word embeddings are still evolving (Boutyline and Johnston 2025; Taylor et al. 2025), and further classes of models have been developed, including contextualized word-embedding models (Jurafsky and Martin 2025). Unlike the traditional embeddings approaches (e.g., the word2vec model), which assign a single fixed vector to each word, contextualized embeddings generate different representations for the same word depending on how it is used in a particular sentence or paragraph. Models that produce contextualized embeddings often used for exploration of semantic relationships include discriminative language models such as BERT (Bidirectional Encoder Representations from Transformers; Devlin et al. 2019) and RoBERTa (Robustly Optimized BERT Pretraining Approach; Liu et al. 2019). They achieve contextualization by processing entire sequences of text rather than isolated word–context pairs and through attention mechanisms, allowing the model to dynamically weight the importance of each word in the sentence when predicting any given masked word (Vaswani et al. 2017). Rather than treating all context words equally, attention enables the model to focus more heavily on the most relevant words for each prediction task of masked words. The key insight behind contextualized embeddings is that word meaning is inherently contextual—the word "depression," for instance, means something entirely different when placed in the context of "economics" or in the context of "feelings," and these semantic differences should be captured in the word's numerical representation (Arseniev-Koehler 2024). This shift from static to contextual embeddings expands the methodological

tool kit available to sociologists, offering richer ways of capturing how meaning varies with context, while still requiring careful theoretical framing to ensure that these representations map onto substantively interpretable constructs. Beyond static and contextualized word embeddings, researchers can also extract vector embeddings of whole sentences or entire texts, reducing the high dimensionality of textual data to a limited number of dimensions. These vector embeddings can be used as input for different analyses, such as text similarity (Lin 2025), classification (Bestvater and Monroe 2023; Widmann and Wich 2023), and treatment effect estimation (Veitch 2020).

Data-first approaches require careful validation to determine whether the patterns they uncover correspond to meaningful sociological constructs rather than modeling artifacts. The form of validation depends on the method. In topic modeling, for instance, researchers can use statistical criteria—such as goodness-of-fit measures for different numbers of topics—to evaluate model performance (Arun et al. 2010), and they can assess interpretability by examining whether the inferred topics correspond to theoretically relevant concepts (Grimmer and Stewart 2013). For word embeddings, relationships among concepts can be compared to human judgments or theoretically expected associations (Boutyline and Johnston 2025). Across these approaches, validation serves to ensure that computationally derived patterns are substantively coherent and theoretically informative, rather than merely reflections of linguistic regularities or modeling assumptions.

2.2 Theory-First Methods

In contrast to data-first methods, theory-first approaches require researchers to specify in advance which linguistic features, categories, or relationships they want to measure, rather than discovering them through statistical patterns in data. Word frequencies and dictionary-based methods have the longest traditions in analyses of text as data in sociology (Lasswell 1934; Franzosi 2004). The core logic of such approaches is that researchers specify words that indicate a theoretical concept or can tell something about the context of which the text was written (van Loon 2022). The frequency of these words then serves as a measure of the theoretical concept's presence in a text. For example, if theory suggests that anxiety manifests through certain linguistic markers, researchers can create a dictionary containing words like "worried," "nervous," "anxious," and "stressed," and then count how often these words appear in texts to generate an "anxiety score." Researchers can either create their own dictionaries or rely on dictionaries curated by others. In sociology, such measures are often used to track emotions, political orientations, or cultural frames across time and contexts (see Macanovic 2022 for an overview).

Supervised classification methods—whether traditional or language model based—represent another class of theory-first approaches because they require researchers to make explicit theoretical decisions about the categories they want to identify before any modeling begins (Bonikowski and Nelson 2022; Hurtado Bodell 2024). That is, researchers need to have a good idea of what their theoretical concepts look like in their textual data and be able to answer questions such as "Is this text about migration? Does this speech contain populist rhetoric? Does this review

express positive or negative sentiment?" The theoretical frameworks that define these categories fundamentally shape the entire analytical process. Conventionally, researchers are then required to produce a sample of texts that have been manually evaluated (coded) for the presence of the theoretical concept of interest. Some of the methods we list below rely on these manually evaluated texts to learn patterns of interest in texts and then apply them to the remainder of the corpus. These manually coded texts also provide the basis for validating model performance.

Traditional classification models—for example, logistic regression, naive Bayes, support vector machines, and random forest (see Murphy 2012 for an overview of these models)—typically require the researcher to select which feature representation to use for analyzing the text. In practice, this means that theory informs not only the categories to predict but also how texts are represented. For example, when predicting whether a text is about migration, words are represented as features, but the model learns which ones (such as "migrants" or "refugee") are predictive as features. Beyond using words as features, researchers can also incorporate metadata such as online engagement metrics, author characteristics, temporal information, or contextual variables into their analyses (Grimmer et al. 2022). Feature selection can also be automated, for example using least absolute shrinkage and selection operator (LASSO) regression (Tibshirani 1996), which applies a penalty to estimated feature effects, shrinking coefficients toward zero based on their magnitude such that features with weak predictive power are excluded while stronger predictors are retained. Yet, even when automated, performance still depends heavily on researchers' theoretical understanding of both categories and textual features.

Language model–based classification models, by contrast, typically operate on the raw textual input, learning internal representations that capture relevant features automatically. This shifts the burden of feature selection and engineering away from the researcher. While researchers still define the theoretical categories of interest, the model itself identifies the textual patterns most relevant for classification (Devlin et al. 2019; Grimmer et al. 2022). This represents a significant methodological shift—whereas traditional classifiers require theoretical decisions about both categories and features, language model–based approaches primarily require theoretical decisions about which categories the model should predict. Both discriminative language models and generative LLMs such as generative pretrained transformers (GPTs; Radford et al. 2019; Brown et al. 2020; OpenAi et al. 2024) can be used in text classification. In these models, it is the attention mechanisms that automatically identify which parts of texts are most relevant for the predefined classification task. For instance, when classifying whether a social media post is about migration, the attention mechanism might automatically learn to pay more attention to specific terms like "illegal aliens" or "border security" than to more generic language that can be found in posts about many different topics.

Traditional classifiers and language model–based classification (especially generative LLMs) call for different training procedures. When working with traditional classifiers and simpler language models, researchers would train a classifier model from scratch, which would often require a substantive amount of coded data. Newer generations of language models instead allow us to use the capabilities of models that have already been trained on large amounts of different data and adapt them to

the task at hand. There are several approaches to doing this: fine-tuning, few-shot learning, and zero-shot learning. Discriminative language models such as BERT and RoBERTa have often been used in combination with the fine-tuning approach that involves starting with a pretrained model and adjusting it (that is, fine-tuning) on labeled data for a specific classification task (Radford et al. 2018; Devlin et al. 2019; Howard and Ruder 2018). This process typically requires a substantial number of labeled examples, but it can achieve high performance because the model starts with rich representations of language learned from massive corpora (Do et al. 2022). In contrast, few-shot learning requires only a handful of labeled examples, relying more heavily on the model's preexisting capabilities (Brown et al. 2020). Zero-shot learning goes even further, requiring no labeled examples at all; instead, researchers provide carefully crafted prompts that describe the classification task in natural language, asking the model to classify texts based solely on its pretraining knowledge and the task description (Xian et al. 2017; Ollion et al. 2024). Few-shot and zero-shot learning are usually performed with generative LLMs. Yet another procedure that can help achieve higher classification performance with generative LLMs is so-called instruction tuning, in which the model is provided with a set of instructions and example outputs in order to learn how to follow particular coding instructions (Chae and Davidson 2025). For instance, a researcher will provide the model with a task instruction (e.g., "Denote if this text is positive or negative: [text]") alongside examples of the preferred format of the output (e.g., "Positive"). This can help streamline the model output so that it aligns well with researchers' goals. The choice of how to use language models ultimately reflects a trade-off among available labeled data, computational resources, time constraints, and desired accuracy. Fine-tuning offers the highest performance when sufficient labeled data exists, while few-shot and zero-shot methods provide strong alternatives when data are limited or unavailable. Given the pace of methodological development, few-shot and zero-shot approaches are achieving increasingly competitive accuracy (Brown et al. 2020; Kojima et al. 2022).

Further theory-first approaches include information extraction and dependency parsing. The former retrieves relevant pieces of information about certain concepts or actors from unstructured text (Grishman 2019). The latter identifies grammatical relationships between these targets by representing sentences as networks of relationships (Jurafsky and Martin 2025). Take the sentence "The researcher analyzed the data carefully." Dependency parsing would identify "analyzed" as the main verb (the head or root of the sentence), with "researcher" depending on it as the subject, "data" depending on it as the object, and "carefully" depending on it as an adverb. These relationships form a head-dependent, treelike structure that captures not just which words appear together but also how they function grammatically in relation to each other. Although generative LLMs appear to perform rather well in such general information extraction tasks (Stuhler et al. 2025), their performance lags behind specialized models when it comes to dependency parsing (Lin et al. 2023; Ezquerro et al. 2025). Dependency parsing holds particular promise for sociological inquiry because it can extract semantically rich relations from text data to capture "who did what to whom" in systematic ways (Stuhler 2022). For example, dependency parsing could be used to understand gender differences in attributed agency in the

ways that men and women are discussed online by identifying whether women and men are typically positioned as agents performing actions or as being subjected to them.

To ensure that computational methods successfully capture the concepts of interest in text, researchers should always validate their outputs. The standard procedure involves dividing the manually coded set of texts into three subsets: training, validation, and test sets. The training set is used to train or fine-tune the model, after which the validation set guides adjustments to model parameters for improved performance. The final model is then applied to the test set, and its outputs are compared to human coding (Grimmer and Stewart 2013). Some approaches, such as zero-shot classification with generative LLMs, do not require a training set but still benefit from validation to select appropriate model variants or parameters. Agreement between human and machine coding is typically measured using accuracy (the proportion of texts correctly classified) or F-score (the harmonic mean of precision and recall, balancing the model's ability to identify true positives while avoiding false negatives; Macanovic and Przepiorka 2024).

2.3 Theory–Data Integration Methods

A third family of methods available to analyze texts lies somewhere between the theory-first and data-first approaches. We call them theory–data integration methods, since they—to varying degrees—allow researchers to make both theory-driven choices in the modeling process and allow data to disclose potentially unexpected patterns.

Topic models have been extended in two important ways that make them relevant in this intermediate modeling category. First, the structural topic model allows researchers to include document-level features as "explanatory variables," for which the model estimates how these features affect both topic prevalence and topic content (Roberts et al. 2014, 2019; Grimmer et al. 2022). For instance, researchers might include variables such as the author's political affiliation or the publication date to examine how these characteristics shape the topics discussed and the language used to discuss them. This makes structural topic models particularly valuable for testing theoretical hypotheses about how social, political, or temporal contexts influence discourses. Unlike traditional LDA, which treats all documents as interchangeable, structural topic models explicitly model the relationship between document metadata and topical content, enabling researchers to ask questions such as "Do conservative and liberal politicians discuss immigration using different topics?" or "Has the prevalence of migration-related topics changed over time in news coverage?"

Second, seeded topic models (Jagarlamudi et al. 2012; Watanabe and Zhou 2022) allow researchers to guide the topic model toward identifying specific topics that relate to their research interests by placing informative priors on particular words (Hurtado Bodell et al. 2026). This means that researchers can specify a priori which words they expect to relate to their theoretical constructs of interest, and then examine whether these expectations align with patterns the model discovers in the data. Researchers begin with theoretical assumptions about which words constitute meaningful topics (similar to dictionary-based methods), but unlike fully theory-first

approaches, seeded topic models are not deterministic. Even when researchers guide the model by specifying seed words they believe will cluster together to form a theoretical construct, the model still inductively identifies additional words that co-occur with these seeds in the data. For instance, if a researcher studying migration discourse seeds a topic with words like "asylum-seeker," "refugee," and "migrant," the model might discover that words like "deportation" or "illegal aliens" also belong to this topic based on their co-occurrence patterns in the corpus—words the researcher may not have anticipated. This semisupervised approach ensures that topics align with theoretical interests while remaining open to data-driven discovery; it can improve topic coherence and interpretability compared to fully unsupervised LDA, and it allows researchers to test whether their theoretical expectations about topic structure are supported by actual language use patterns in their corpus. To perform such analyses, similar methodologies now also exist for language model–based approaches such as BERTopic (Grootendorst 2022). However, regardless of the method used, the choice of seed words remains crucial and reflects theoretical commitments that shape the model's output (Hurtado Bodell et al. 2026).

Theory–data integration approaches also build on the word-embedding methodology. Similar to the seeded topic models, prior-informed embeddings allow researchers to guide the estimation of vectors to isolate an interpretable dimension into a specific part of the embeddings space (Hurtado Bodell et al. 2019). This approach exemplifies theory–data integration because researchers use theoretical expectations to shape how the model learns from data, while simultaneously allowing the data to reveal patterns beyond these theoretical priors. In effect, the theory guides which dimensions are interpretable, but the data determine where other words fall along these dimensions. In practice, this allows researchers to specify theoretically motivated relationships among selected words and encode these expectations as priors during the embedding training process. For example, if researchers are interested in studying how words have been gendered over time, they can place priors on word pairs related to men and women (such as "he/she," "man/woman," "father/mother") in such a way that at least one dimension of their embedding represents gender, with male-associated and female-associated words positioned at opposite ends of that dimension. In this way, researchers know before running the word-embedding model which dimension will represent their theoretical construct of interest (gender) and can then study how other words—such as occupation terms or personality adjectives—position themselves along that theoretically anchored dimension. This reveals whether terms like "doctor," "nurse," "ambitious," or "caring" are more associated with men or women in the corpus, while allowing the remaining embedding dimensions to capture other semantic relationships discovered inductively from the data.

Pushing this logic further, recent work has sought to integrate theoretical expectations with a data-driven analysis by using chatbots relying on generative LLMs in qualitative research as tools for iteratively refining theoretical ideas in light of empirical insights (Ibrahim and Voyer 2025; Than et al. 2025). Some propose deploying generative LLMs throughout the analytical pipeline, beginning with LLM-driven code discovery in text. Researchers then refine and update these codes before giving the model more targeted instructions and proceeding with thematic analysis

and data classification (Drápal et al. 2023). Other approaches begin with a clearly defined concept of interest and ask the generative model to extract themes related to that concept, iteratively adjusting the instructions to ensure alignment of model output with theoretical expectations without imposing too many theoretical constraints on the dataset (see Ibrahim and Voyer 2025 for an overview).

Like theory-first and data-first approaches, theory–data integration methods require validation. In practice, researchers typically compare the models' theory-guided components to human annotations or theoretically grounded expectations, while the inductively learned components are evaluated using procedures analogous to those in data-first methods—for example, assessing whether discovered patterns are coherent, interpretable, and consistent with domain knowledge (Hurtado Bodell et al. 2026).

Together, these various methods illustrate the promise of theory–data integration approaches. By embedding theoretical expectations into models that still learn from data, they offer sociologists a way to preserve interpretability and test substantive hypotheses while remaining open to the discovery of unanticipated patterns. In this sense, integration approaches avoid the rigidity of purely theory-first designs and the opacity of purely data-first ones, letting theory and data mutually inform one another within a single modeling process rather than being applied in separate, sequential stages (Nelson 2020; Hurtado Bodell 2024).

3 Thick Description

Many canonical applications in sociology use computational text analysis to create thick descriptions of social reality, whether they adopt data-first, theory-first, or theory–data integration approaches to drawing insights from text corpora. The concept of thick description originated in Clifford Geertz's (1973) anthropological work and has been adapted by Hedström and Udehn (2011) in the context of middle-range theorizing. They distinguish "thin descriptions"—factual accounts of a chain of events—from "thick descriptions," which situate those events within their broader social, cultural, and economic contexts. Thick descriptions become especially valuable when they enable deeper, more generalizable explorations of the mechanisms underlying observed phenomena, thereby paving the way for theory development. Although thick description has traditionally been associated with qualitative approaches, computational methods can extend its reach: By analyzing vast amounts of unstructured textual data systematically and reproducibly, at a scale far beyond what individual researchers can achieve through manual reading alone, computational methods can capture complex relationships within social systems. Yet without scholars who interpret and contextualize identified patterns in light of sociological theory and existing empirical evidence, these regularities cannot become scientific knowledge.

When the pattern-detecting capacities of algorithms combine with sociological interpretation (Santana and Nelson 2025) and theoretical knowledge informs each step of working with "found" textual data (Yung et al. 2025), computational text analysis becomes a tool for refining and extending sociological concepts and the-

ories. Social theorists and text analytical tools thus complement and augment one another, as researchers use computational representations to formulate and refine hypotheses about underlying social systems. Thick description based on computational methods can thereby connect large-scale, transparent, and reproducible analyses with the interpretive depth traditionally associated with qualitative research. Exploratory methods of pattern recognition in unstructured texts are therefore not opposed to interpretive modes of inquiry but can scale and complement approaches such as grounded theory (Glaser and Strauss 1967), abductive analysis (Tavory and Timmermans 2014), and forensic social science (McFarland et al. 2016).

Several workflows have been proposed for how computational text analysis can be used to refine and build theory through iterative engagement with data and interpretation. Computational grounded theory (Baumer et al. 2017; Nelson 2020) exemplifies one such workflow, using computational methods to identify patterns in large corpora before developing theoretical insights with close reading. The process involves three steps: (1) automated pattern detection using data-first methods as described above, (2) manual inspection and interpretation of a subset of documents to understand what automated patterns mean in context, and (3) refining of the computationally identified patterns based on these interpretations, for example using theory-first methods or theory–data integration approaches. This workflow enhances rigor, reproducibility, and scalability while emulating the interpretive depth of qualitative close reading (Nelson 2020).

Abductive approaches (Brandt and Timmermans 2021; Tavory and Timmermans 2014) offer an alternative workflow that combines inductive and deductive reasoning. Abduction begins with existing theoretical and empirical knowledge but remains open to previously unknown or unexpected patterns through data exploration. When a new insight emerges from the data that challenges existing theory, theory can be revised and extended. For example, research analyzing parliamentary speeches on migration might begin with established theories of how restrictive migration policies are justified through economic, security, or cultural threat narratives, but through a data-first or theory–data integration approach, it could identify that far-right parties frequently employ humanitarian framing—emphasizing the need to help people in their home countries rather than through migration—prompting theoretical refinements about how restrictive immigration policies gain legitimacy through humanitarian appeals. Brandt and Timmermans (2021) argue that the abductive approach augments computational text analysis because the large scale of text corpora allows patterns that might remain hidden in smaller datasets to appear systematically and because text-as-data analyses are closely linked to powerful machine-learning techniques for pattern detection.

What unites both inductive and abductive pathways in computational text analysis is their fundamentally iterative nature. Researchers move back and forth between theory and data to gradually refine their measurement strategies and calibrate them to the intended theoretical conceptualization (Bonikowski et al. 2022). Equally important is the iterative movement between quantitative and qualitative modes of analysis: Researchers zoom out to capture overarching patterns in a corpus using computational techniques and a standardized analytical pipeline and zoom in to interpret those patterns through interpretative close reading of a small subset of

documents (Voyer et al. 2022; Pournaki and Willaert 2025). The importance of the iterative workflow in computational text analysis has been emphasized across the full range of computational text analysis methods—from more conventional topic models (Karell and Freedman 2020) and word embeddings (Boutyline and Arseniev-Koehler 2025) to LLM-based tools (Stuhler et al. 2025).

Such iterative workflows have enabled advances in producing thick descriptions and generating new theoretical insights into the structure of social relations. First, text-based research has documented systematic relations between concepts that constitute structures of shared cultural meaning (Mohr 1998). For example, Nelson (2020) used computational text analysis to identify the political logics of historical women's movements across U.S. cities. By going back and forth between computational pattern detection using topic models and close reading of literature produced by women's movements in two cities, the author discovered how Chicago-based groups focused more on organization and the concrete needs of the community, whereas New York groups emphasized individuals and their diverse lived experiences to make claims about how social structures affected women's lives. This result refined existing theories of how local cultures exist within and shape social movements. Second, computational text analysis has expanded researchers' ability to examine the interdependencies among different groups of actors within a social system. The large corpora produced across societal domains—newspaper articles, literature, political speeches, digital trace data from online communities—provide a foundation for systematically analyzing relations between groups (Menshikova 2025). For example, Barberá et al. (2019) demonstrated how the political agenda of high-ranking U.S. legislators responded to public discourse by drawing connections between two corpora of Twitter posts—one from members of Congress and one from the public. By analyzing the saliency of topics in tweets posted by both groups, they uncovered that politicians were more likely to follow the agenda set by the public rather than lead it. This result contradicted the long-standing theory of agenda-setting power held by political parties. Such analyses illustrate how computational text analysis can be used to refine theories about how groups interact and influence each other.

Lastly, exploring patterns emerging from texts has provided novel ways to relate different units of analysis, similar to the logic of bipartite networks. We can draw connections between actors who produce texts based on the similarity of the opinions and interpretations they express, while also linking opinions or interpretations according to the number of actors who share them. Illustrating this approach, Karell and Freedman (2020) used topic models to measure the cultural elements that different Afghan militant groups use. By measuring overlap in cultural elements used, they demonstrated that groups that share common cultural elements were less likely to engage in conflict with one another later, thereby narrowing theoretical predictions about how the mechanisms of cultural change operate in the context of social conflict. Their computational text analysis approach also provided a stronger theoretical integration between cultural sociology and conflict studies. Together, these examples demonstrate how computational text analysis advances sociological thick description by enabling systematic explorations of social mechanisms at scale while retaining interpretative depth, helping bridge the qualitative–quantitative divide.

4 Causal Analysis

Computational text analysis methods have expanded the capacity for theory testing by allowing social scientists to design more precise measurements of complex theoretical concepts and test hypotheses at larger scales or within previously hard-to-access populations (e.g., Edelmann et al. 2020; Mohr et al. 2020; Stoltz and Taylor 2024). Still, computational text analysis thrives on large-scale observational data that are well-suited for mapping and measuring social phenomena but that pose challenges for causal inference.

Moving from thick description to causal inference creates tension, as statements of cause and effect typically require controlled interventions and random assignment, which are difficult to implement when analyzing naturally produced text at scale. Yet sociologists have increasingly sought to use measures extracted from text data for purposes of causal inference—by using text as either outcome, treatment, or confounder (Grimmer et al. 2022; Schwitter et al. 2026, this issue)—employing different research strategies that make trade-offs between causal identification and the practical constraints of working with process-produced textual data. Among these approaches, the text-as-outcomes paradigm most directly leverages computational text analysis's core strength of extracting measurements that capture variations in textual patterns. We review some highlights of this research, with a focus on studies of political discourse and public opinion.

Within the *text-as-outcome* paradigm, researchers have pursued causal inference through two primary approaches relying on exogenous variation to study changes in texts: (1) leveraging naturally occurring observational data and (2) conducting experiments. Among researchers using observational data, event-based designs are especially common, treating external events as natural experiments to approximate causal inference without randomized experimental manipulation. These studies use changes in macro-level textual patterns as proxies for shifting public opinions, shared understandings, or public discourses. One early example is Bail (2012), who studied the consequences of the 9/11 terrorist attacks on mass media reporting about Muslims and the role of civil society organizations in such shifts. Using plagiarism detection software, he identified that messages from anti-Muslim fringe organizations dominated mass media reporting following the attacks. Garcia and Rimé (2019) identified increasing negative emotions in online discussions following the 2015 terrorist attacks in Paris. Hurtado Bodell et al. (2026) examined how terrorist attacks and other emotional events during the European "refugee crisis" impacted which topics were salient in mass media reporting on migration. The study used measures of pretheorized concepts, extracted with seeded topic models, in a regression–discontinuity kind of design, comparing the salience of different interpretations of immigration before and shortly after the event. These studies analyzed aggregate textual corpora—collections of media articles, social media posts, or public documents—to trace how collective discourse shifted in response to exogenous shocks.

Drawing strict causal inferences from aggregate-level event studies, however, remains challenging. The estimation strategy typically rests on two key assumptions: that events prior to the focal event do not directly affect current textual patterns, and that past textual patterns do not affect the occurrence of the focal event (Imai

and Kim 2019). In the context of macro-level discourse, both assumptions are difficult to satisfy. Past events may have longer-lasting consequences on media framing through institutional adaptation and the gradual consolidation of interpretive frameworks. Conversely, past topic saliencies may impact actors' decisions to create or amplify subsequent news events. Moreover, aggregate-level data obscure important compositional dynamics, as observed shifts may reflect changes in who speaks rather than genuine transformations (e.g., opinion change) among the panel of people who produced the texts before and after the event.

Causal inference may prove more tractable when individual-level textual data are available, as with social media platforms where posts can be linked to specific users. This enables researchers to leverage within-individual variation and employ techniques such as difference-in-differences or fixed effects. Flores (2017), for instance, used Twitter data to study how changes in Arizona's anti-immigrant law shaped public opinion, comparing users in the affected state with those in similar unaffected states to establish causality. This difference-in-differences approach, combined with dictionary-based sentiment measurement, exploits the individual-level structure of social media data to strengthen the causal identification. Czymara et al. (2023) showed how hostility on YouTube increased following terrorist attacks around Europe as a result of shifting user composition. Similarly, Czymara and Gorodzeisky (2024) employed interrupted time series analysis to examine how jihadist terror attacks affected ethnoreligious hostility on Twitter. Their individual-level approach enabled them to decompose the overall effect into intra-user changes (individuals becoming more hostile) versus compositional changes (different users participating). Hurtado Bodell and Menshikova (2024) extended this analytical strategy from studying sentiments to examining mechanisms of meaning-making. Analyzing posts from Sweden's largest discussion forum across 37 jihadist terrorist attacks, they distinguished between within-individual shifts in how online users talk about immigration and compositional shifts of the speakers themselves. Although these studies vary in their claims about causality, they share a commitment to leveraging temporal patterns to move beyond pure descriptions. These studies collectively demonstrate how found individual-level data with appropriate temporal variation can approximate causal inference through ex post facto designs. This offers a valuable complement to experimental approaches using text as data (see Schwitter et al. 2026, this issue), interesting in particular for the much larger scale and external validity this approach supports (Breen and Pan 2026, this issue).

In contrast to the text-as-outcome tradition, research treating *text as treatment* relies predominantly on experimental designs that use controlled exposure to exogenous variation in textual content for clear causal identification. While this work shares the broader interest in understanding how discourses shape social outcomes or how statements of others influence people's beliefs, attitudes, and behaviors, this research typically does not involve computational text analysis. Instead, textual content serves as the experimental stimulus, and outcomes are typically measured through survey responses. Prominent examples include Bail et al.'s (2018) experiment on how exposure to ideologically misaligned online messages can radicalize the political opinions of recipients. Guess and Coppock (2020) tested how polarization changed when individuals encountered confirming or opposing information

to previously held beliefs. They randomly assigned subjects to receive positive or negative information about gun control, minimum wage, and capital punishment, finding that both proponents and opponents updated their views in the direction of the textual information presented. Relatedly, texts have been used in web experiments that studied the spreading of different types of messages, such as true and false news, in social networks (Stein et al. 2023).

The *text-as-confounder* approach, finally, allows researchers to extract information from text in order to identify relevant features, reduce dimensionality, and incorporate textual measures into causal models. In other words, computational text analysis is used to extract granular measures of previously hard-to-capture concepts to improve the internal validity of traditional methods of causal inference. This is crucial when researchers expect that latent information from textual data confounds the main effect of interest. In an early demonstration, Roberts et al. (2020) analyzed, in one study, Chinese social media censorship, estimating whether experiencing censorship would increase the likelihood of users getting censored in the future (it did), adjusting for both the content of the post as well as nontextual confounders such as previous posting rates. In a second study, focusing on the science of science, they examined the gender citation gap in international relations scholarship, testing how accounting for differences in female and male scholars' writing could explain citation disparities without invoking gender bias. Studying peer-to-peer influence in music diffusion, Arvidsson et al. (2025) analyzed digital traces from Spotify users to estimate the causal effect of social exposure on music adoption. The key confounder was music taste such that separating homophily from influence required measuring users' listening prior to exposure. To capture latent musical preferences, they applied topic models to user playlists, treating playlists as documents and artists as words. The resulting topics captured groups of co-occurring artists, corresponding to genres and cultural eras. Users' topic proportions, i.e., their listening to specific genres, then represented music taste. The topic proportions enabled high-dimensional statistical matching (Roberts et al. 2020; Eckles and Bakshy 2021) to control for homophily. After adjusting for confounding in this way, the researchers still found substantial social-influence effects on music adoption, although smaller ones than naive estimates would imply (Arvidsson and Keuschnigg 2025).

Another way to handle potential confounders contained in textual data is by using a so-called deconfounder method (Wang and Blei 2019; see also Leitgöb and Keusch 2026, this issue). This approach infers a latent variable that acts as a substitute for unobserved confounders, allowing for the estimation of causal effects net of these confounders. In a similar vein, Veitch et al. (2020) used an embedding-based approach to deal with textual confounding. They examined whether adding a theorem to an academic paper affected its acceptance rate. However, papers with theorems may systematically differ in subject matter, writing quality, or complexity from papers without theorems. They developed "causally sufficient embeddings" that created low-dimensional text representations preserving information needed for causal adjustment. Their embeddings from a discriminative language model captured textual features that predicted both theorem presence and paper acceptance, enabling the researchers to control for the thematic focus and writing quality of papers and isolate the causal effect of including a theorem. Veitch et al. (2020) found

that including a theorem had a positive effect on acceptance, though the effect was smaller after adjusting for textual confounding than naive estimates suggested. Imai and Nakamura (2025) have proposed a similar framework in which generative LLMs are used instead of discriminative models to retrieve low-dimensional text features. The generative LLM is used to generate the exact same text as the original, after which researchers can retrieve the model's internal low-dimensional representation of the text (i.e., the embeddings). These embeddings can then be used to estimate the deconfounder before moving on with the causal treatment effect estimation. The authors used this approach to reestimate the effects from Roberts et al. (2020) and found that using this approach led to lower bias in effect estimates. The embedding-plus-deconfounder method can be useful in both text-as-treatment and text-as-confounder designs.

Generative LLMs can further be useful in causal inference because they can generate counterfactual texts that differ from the original text only in the dimension of interest (Nguyen et al. 2024; Wang et al. 2024). While such textual counterfactuals are often generated with the goal of understanding the models themselves, they can be useful in text-as-treatment designs where researchers want to study the effect of a single feature on an outcome, holding all the other text features constant (Leitgöb and Keusch 2026, this issue).

Beyond the challenges of identifying causal effects in specific empirical contexts, recent methodological research has identified further problems that arise when using high-dimensional textual data for causal inference—problems that apply regardless of whether data are aggregate- or individual-level or whether researchers use found data or experiments. A fundamental challenge stems from the high dimensionality of textual data. As in the Spotify example, researchers typically require creation of lower-dimensional representations of textual content before incorporating them into analyses. As seen in previous sections of this article, the mapping between raw text and a lower-dimensional representation can be done in different ways: by manually reading and annotating texts, using data-first approaches to identify previously unknown patterns, or by applying theory-first methods to classify texts into predefined categories. Regardless of which mapping approach is used, it risks violating the stable unit treatment value assumption (SUTVA), which requires that the outcome of any unit must not vary with the treatment assignment of any other unit (Imbens and Rubin 2015; Egami et al. 2022; see also Breen and Pan 2026 as well as Leitgöb and Keusch 2026 in this issue).

When dealing with lower-dimensional representations of texts, this is problematic because the specific units used to create the mapping influence how treatment is assigned to all other documents, thereby influencing their potential outcomes. Egami et al. (2022) call this the fundamental problem of causal inference with text data. Consider an example in which a researcher randomly selects 500 social media posts to create a codebook defining what hateful content looks like in the studied context. This codebook is then used to map all other posts to a binary variable with value 1 if it contains hateful content and 0 otherwise. The decision of how to classify the remaining posts depends on how hateful content appeared in the randomly sampled 500 posts. If the researcher had instead selected 500 different posts, the codebook might have looked different, leading to potentially different classification decisions

for the remaining posts. Hence, in violation of SUTVA, the assignment for any given post depends on the set of posts present in the initial sample. Egami et al. (2022) further suggest that modeling high-dimensional text data is sensitive to overfitting. Texts contain thousands or even millions of unique words, phrases, and linguistic patterns, creating a vast feature space where spurious correlations easily arise. When these features are used to estimate treatment effects, models can capture patterns specific to the training data that fail to generalize to new texts.

As a principled solution, Egami et al. (2022) propose data splitting. Data are divided into one subset exclusively for creating the lower-dimensional mapping of texts and into another reserved for estimating causal effects. By separating the two stages with sufficient data, researchers ensure that units used to estimate causal relationships are not the same units that determined how the text-to-treatment mapping was constructed. This approach helps address both SUTVA violations and overfitting concerns, providing clearer guidance for future sociological work that seeks to draw causal inferences from textual data.

5 Discussion

Large corpora of text now serve as social sensors, and the text analytical methods discussed here are one way of capturing what is happening in society. They offer new and refined ways to measure phenomena that quantitative sociology has traditionally studied through survey research. Because textual expressions contain richer, more multidimensional information than survey items—and large corpora enable their collection at scale—text as data can support more precise theory building and testing, especially when relevant constructs involve emotions, sentiment, or interpretations. Moreover, text data such as digital traces from social media are often well suited for developing and testing theories of social dynamics and collective phenomena, as they capture individual behaviors, expressed opinions and sentiments, and social outcomes within the interactive social contexts within which they emerge. In this context, we have also emphasized how computational text analysis bridges quantitative scale with qualitative depth. By enabling researchers to combine distant and close reading, work through larger data, and enhance transparency and reproducibility, these methods and their further development will also provide valuable tools for qualitative research designs and the scaling-up of rich interview studies.

Analyses of text as data are surely not without their own limitations. Like many forms of digital trace data, text corpora are rarely created for research purposes and often provide "thin" information on the individual-level characteristics of the populations that generated them (Salganik 2018). This limitation may be less pronounced for corpora generated by well-documented social groups such as media or political elites (e.g., U.S. Congress representatives [Gentzkow et al. 2018]) or literary authors (e.g., English-language fiction writers [Underwood et al. 2022]). However, when studying the general public's opinions through social media data, sociodemographic information is typically sparse (Flores 2017; Hastings and Pesando 2024). Moreover, text corpora put to research uses disproportionately represent Western populations while underrepresenting women and ethnic minorities (Brown et al.

2016). These limitations complicate efforts to contextualize and generalize findings, underscoring the need for careful consideration of whether the theories tested and developed through computational text analysis are equally applicable across diverse social settings. Beyond these demographic biases, text data inherently reflect the social contexts in which they are produced, including prevailing norms and stereotypes (Stoltz and Taylor 2024). While computer scientists develop debiasing algorithms to reduce inherent biases in text data and language models, social scientists leverage these biases to analyze inequalities and stereotypes at scale (e.g., Boutyline et al. 2023 on gender stereotypes; Luo et al. 2024 on ethnic prejudice). Against this backdrop of likely biases and data quality issues, we point to recent contributions by Hurtado Bodell et al. (2022) and Mützel and Ollion (2025) on evaluating text corpus quality at different stages of the sociological analysis of textual data.

Clearly, research that uses text as data has matured beyond its initial descriptive impetus, and sociological research is transitioning from treating text analysis as an "end," i.e., reporting novel empirical regularities, to treating it as a "means" of advancing social theory (Bonikowski and Nelson 2022). The path toward rigorous causal inference, however, remains uneven. Few current studies, including our own, implement methodological safeguards such as the data-splitting framework of Egami et al. (2022), and claims about causality vary widely across published empirical work. Although the use of generative models for causal inference is exploding, its use in the social sciences is still in its infancy (see also Jeon and Brand 2026 as well as Leitgöb and Keusch 2026 in this issue). As the field develops further, bridging the gap between current practices and emerging methodological standards represents a key challenge.

As noted earlier, the rise of the new generation of generative LLMs and chatbots that build on them has significantly democratized the use of text as data. The ease of their use has brought upon a wave of studies relying on generative language models. Results show that while generative LLMs enable researchers to achieve excellent performance on certain analytical tasks (Gilardi et al. 2023; Törnberg 2023), they also introduce challenges such as inconsistent results (Ollion et al. 2024; Pangakis et al. 2023), model hallucinations (Huang et al. 2025), and sensitivity to subtle variations in prompting (Reiss 2023; Savelka et al. 2023), where prompting refers to the input text and instructions provided to the language model. Additional concerns include limited transparency and reproducibility (Liesenfeld et al. 2023), biased representations of minority social groups and languages (Guilbeault et al. 2025; Wang et al. 2025), and risks related to sharing sensitive data with companies behind proprietary models (Spirling 2023; Ollion et al. 2024; Rytting et al. 2023). While the current wave of excitement around generative artificial intelligence may soon give way to a more measured assessment of its strengths and weaknesses, researchers must remain committed to reproducible, traceable, and precise measurement of concepts grounded in social–scientific theory and practice (Grimmer et al. 2021; Macanovic 2022).

The methods we have touched upon primarily extract information from text, such that they use text as a sensor for capturing social patterns. With the rise of generative LLMs, however, researchers are increasingly exploring *text generation*—and with it the possibility to create counterfactual actors and alternative worlds. The text-

generation application that so far has garnered the most enthusiasm is using generative models to simulate human behavior across contexts (see Kozlowski and Evans 2025 for a recent discussion). Early studies in survey research suggest that prompting generative LLMs with characteristics of real respondents can yield answers that resemble those of actual respondents (Argyle et al. 2023), though subsequent work has shown that these results do not generalize across contexts (von der Heyde et al. 2025). Another emerging direction involves using generative LLMs to emulate agents in agent-based computer simulations (Gao et al. 2024). The premise is that, compared to the simple agents typically employed, generative LLMs can introduce richer behavioral complexity: They can take actions without explicit step-by-step instructions, respond and adapt to stimuli, and interact with other agents or humans (Park et al. 2023; Kozlowski and Evans 2025). Related work has also evaluated the extent to which decisions made by generative LLMs in social scientific experiments resemble those made by human participants (Chen et al. 2023; Mei et al. 2024). In other words, the door is once again open to run social–scientific research on entirely "silicon" populations. Although agents equipped with some sort of artificial intelligence have been used productively in a range of applications since as early as the 1950s (Holme and Tsvetkova 2025), the deployment of generative LLM-based agents requires particular caution (Dillion et al. 2023). Currently, such generative models tend to underestimate the diversity of responses given by actual humans (Bisbee et al. 2024; Kozlowski and Evans 2025), produce responses biased toward majority opinions (Crockett and Messeri 2025), and reflect certain cultural and social backgrounds more than others (Alvero et al. 2024; Tao et al. 2024). Moreover, it remains unclear to what extent models trained on broad, generalized knowledge can accurately represent perspectives stemming from specific, embodied experiences of individuals (Kozlowski and Evans 2025).

Challenges and open questions notwithstanding, computational text analysis in sociology has opened new avenues for quantifying hard-to-measure concepts, uncovering elusive patterns, and interpreting a wealth of novel insights through theoretical frameworks (Mohr et al. 2020; Borch and Pablo Pardo-Guerra 2025; Santana and Nelson 2025). However, validating the outputs of chosen methods by comparing them to human judgments or theoretically guided assessments and ensuring that extracted patterns accurately reflect their intended concepts remains essential for realizing the potential of computational text analysis (Grimmer et al. 2022; Schwitter et al. 2026, this issue). With this in mind, computational text analysis has the potential to revitalize theory building and spark a renaissance of cultural sociology and the interpretative paradigm within disciplinary communities that were previously exclusively quantitatively oriented—and thus often confined to studying a narrow range of phenomena supported by systematic but often mundane data. As computational methods continue to mature and sociologists refine their approaches to working with textual data, these limitations are rapidly becoming challenges of the past.

Acknowledgements We thank Julius Hehenkamp and Felix Lennert for helpful comments. We are grateful for funding by the Swedish Research Council (2018-05170) and Riksbankens Jubileumsfond (M21-0021). This research was partly carried out at the Swedish Excellence Center for Computational Social Science, also funded by the Swedish Research Council (2022-06611).

Funding Open access funding provided by Linköping University.

References

Alvero, A. J., Jinsook Lee, Alejandra Regla-Vargas, René F. Kizilcec, Thorsten Joachims, and Anthony Lising Antonio. 2024. Large language models, social demography, and hegemony: comparing authorship in human and synthetic text. *Journal of Big Data* 11:138.

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis* 31:337–351.

Arseniev-Koehler, Alina. 2024. Theoretical foundations and limits of word embeddings: What types of meaning can they capture? *Sociological Methods & Research* 53:1753–1793.

Arseniev-Koehler, Alina, and Jacob G Foster. 2022. Machine learning as a model for cultural learning: Teaching an algorithm what it means to be fat. *Sociological Methods & Research* 51:1484–1539.

Castellanos Arturo, Haoqiang Jiang, Paulo Gomes, Debra Vander Meer, and Alfred Castillo. 2025. Large language models for thematic summarization in qualitative health care research: Comparative analysis of model and human performance. *Journal of Medical Internet Research AI* 4:e64447.

Arun, R., V. Suresh, C. Veni Madhavan, M. Narasimha Murthy, 2010. On finding the natural number of topics with latent Dirichlet allocation: Some observations. In *Advances in knowledge discovery and data mining*, eds. Mohammed J. Zaki, Jeffrey Xu Yu, B. Ravindran, and Vikram Pudi, 391–402. Berlin: Springer.

Arvidsson, Martin, and Marc Keuschnigg. 2025. Estimating social influence using machine learning and digital trace data. In *The Oxford handbook of the sociology of machine learning*, eds. Christian Borch and Juan Pablo Pardo-Guerra. Oxford: Oxford University Press.

Arvidsson, Martin, Peter Hedström, and Marc Keuschnigg. 2025. Wide social influence and the emergence of the unexpected: An empirical test using spotify data. *Sociological Science* 12:715–742.

Bail, Christopher A. 2012. The fringe effect: Civil society organizations and the evolution of media discourse about Islam since the September 11th attacks. *American Sociological Review* 77:855–879.

Bail, Christopher A., Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* 115:9216–9221.

Barberá, Pablo, Andreu Casas, Jonathan Nagler, Patrick J. Egan, Richard Bonneau, John T. Jost, and Joshua A. Tucker. 2019. Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data. *American Political Science Review* 113:883–901.

Baumer, Eric, David Mimno, Shion Guha, Emily Quan, and Geri 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.

Benoit, Kenneth. 2020. Text as data: An overview. In *SAGE handbook of research methods in political science and international relations*, eds. Luigi Curini and Robert Franzese, 1–55. London: Sage.

Best, Rachel Kahn, and Alina Arseniev-Koehler. 2023. The stigma of diseases: unequal burden, uneven decline. *American Sociological Review* 88:938–969.

Bestvater, Samuel E., and Burt L. Monroe. 2023. Sentiment is not stance: Target-aware opinion classification for political text analysis. *Political Analysis* 31:235–256.

Bisbee, James, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. 2024. Synthetic replacements for human survey data? The perils of large language models. *Political Analysis* 32:401–416.

Blei, David, Lawrence Carin, and David Dunson. 2010. Probabilistic topic models. *IEEE Signal Processing Magazine* 27:55–65.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Bommasani, Rishi et al. 2022. On the opportunities and risks of foundation models. http://arxiv.org/abs/2108.07258.

Bonikowski, Bart, and Laura K Nelson. 2022. From ends to means: The promise of computational text analysis for theoretically driven sociological research. *Sociological Methods & Research* 51:1469–1483.

Bonikowski, Bart, Yuchen Luo, and Oscar Stuhler. 2022. Politics as usual? Measuring populism, nationalism, and authoritarianism in US presidential campaigns (1952–2020) with neural language models. *Sociological Methods & Research* 51:1721–1787.

Borch, Christian, and Juan Pablo Pardo-Guerra, eds. 2025. *The Oxford handbook of the sociology of machine learning*. Oxford: Oxford University Press.

Boutyline, Andrei, and Alina Arseniev-Koehler. 2025. Meaning in hyperspace: Word embeddings as tools for cultural measurement. *Annual Review of Sociology* 51:89–107.

Boutyline, Andrei, and Ethan Johnston. 2025. Forging better axes: Evaluating and improving the reliability of semantic dimensions in word embeddings. https://doi.org/10.31235/osf.io/576h3_v2.

Boutyline, Andrei, Alina Arseniev-Koehler, and Devin J Cornell. 2023. School, studying, and smarts: Gender stereotypes and education across 80 years of american print media, 1930–2009. *Social Forces* 102:263–286.

Brandt, Philipp, and Stefan Timmermans. 2021. Abductive logic of inquiry for quantitative research in the digital age. *Sociological Science* 8:191–210.

Breen, Richard, and Guanghui Pan. 2026. Bringing external validity into sociological research. *This issue.*

Brown, Nicole M., Ruby Mendenhall, Michael L. Black, Mark Van Moer, Assata Zerai, and Karen Flynn. 2016. Mechanized margin to digitized center: black feminism's contributions to combatting erasure within the digital humanities. *International Journal of Humanities and Arts Computing* 10:110–125.

Brown, Tom B. et al. 2020. Language models are few-shot learners. http://arxiv.org/abs/2005.14165.

Brüderl, Josef, and Volker Ludwig. 2015. Fixed-effects panel regression. In *The SAGE Handbook of Regression Analysis and Causal Inference*, 327–357. London: Sage Publications.

Chae, Youngjin, and Thomas Davidson. 2025. Large language models for text classification: From zero-shot learning to instruction-tuning. *Sociological Methods & Research* https://doi.org/10.1177/00491241251325243.

Chang, Jonathan, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, *NIPS'09*, 288–296. Red Hook: Curran Associates.

Chen, Yiting, Tracy Xiao Liu, You Shan, and Songfa Zhong. 2023. The emergence of economic rationality of GPT. *Proceedings of the National Academy of Sciences* 120: e2316205120.

Crockett, M.J., and Lisa Messeri. 2025. AI surrogates and illusions of generalizability in cognitive science. *Trends in Cognitive Sciences* https://doi.org/10.1016/j.tics.2025.09.012.

Czymara, Christian S., and Anastasia Gorodzeisky. 2024. Hostility on Twitter in the aftermath of terror attacks. *Journal of Computational Social Science* 7:1305–1325.

Czymara, Christian S., Stephan Dochow-Sondershaus, Lucas G. Drouhot, Müge Simsek, and Christoph Spörlein. 2023. Catalyst of hate? Ethnic insulting on YouTube in the aftermath of terror attacks in France, Germany and the United Kingdom 2014–2017. *Journal of Ethnic and Migration Studies* 49:535–553.

Davidson, Thomas, and Daniel Karell. 2025. Integrating generative artificial intelligence into social science research: Measurement, prompting, and simulation. *Sociological Methods & Research* 54:775–793.

De Paoli, Stefano. 2024. Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach. *Social Science Computer Review* 42:997–1019.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for Language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1*, eds. Jill Burstein, Christy Doran and Thamar Solorio, 4171–4186. Minneapolis: Association for Computational Linguistics.

Dillion, Danica, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences* 27:597–600.
Do, Salomé, Étienne Ollion, and Rubing Shen. 2022. The augmented social scientist: Using sequential transfer learning to annotate millions of texts with human-level accuracy. *Sociological Methods & Research* 53:1167–1200.
Drápal, Jakub, Hannes Westermann, and Jaromir Savelka. 2023. Using large language models to support thematic analysis in empirical legal studies. https://doi.org/10.2139/ssrn.4617116.
Eckles, Dean, and Eytan Bakshy. 2021. Bias and high-dimensional adjustment in observational studies of peer effects. *Journal of the American Statistical Association* 116:507–517.
Edelmann, Achim, Tom Wolff, Danielle Montagne, and Christopher A Bail. 2020. Computational social science and sociology. *Annual Review of Sociology* 46:61–81.
Egami, Naoki, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2022. How to make causal inferences using texts. *Science Advances* 8:1–13.
Ezquerro, Ana, Carlos Gómez-Rodríguez, and David Vilares. 2025. Better benchmarking LLMs for zero-shot dependency parsing. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies*, 121–135, Tallinn: University of Tartu Library.
Flores, René D. 2017. Do anti-immigrant laws shape public sentiment? A study of Arizona's SB 1070 using twitter data. *American Journal of Sociology* 123:333–384.
Franzosi, R. 2004. *From words to numbers: Narrative, data, and social science*. Cambridge: Cambridge University Press.
Gao, Chen, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: a survey and perspectives. *Humanities and Social Sciences Communications* 11: 1259.
Garcia, David, and Bernard Rimé. 2019. Collective emotions and social resilience in the digital traces after a terrorist attack. *Psychological Science* 30:617–628.
Geertz, Clifford. 1973. *The interpretation of cultures.* New York: Basic.
Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. Text as data. *Journal of Economic Literature* 57:535–574.
Gentzkow, Matthew., Shapiro, Jesse M., & Taddy, Matt. Congressional Record for the 43rd-114th Congresses: Parsed Speeches and Phrase Counts., *Stanford Libraries*, congress_text (2018).
Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* 120:1–3.
Glaser, Barney G., and Anselm L. Strauss. 1967. *The discovery of grounded theory: strategies for qualitative research*. Chicago: Aldine Publishing.
Grimmer, Justin, and Brandon M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis* 21:267-297.
Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2021. Machine learning for social science: An agnostic approach. *Annual Review of Political Science* 24:395–419.
Grimmer, Justin, Margaret E Roberts, and Brandon M Stewart. 2022. *Text as data: A new framework for machine learning and the social sciences*. Princeton: Princeton University Press.
Grishman, Ralph. 2019. Twenty-five years of information extraction. *Natural Language Engineering* 25:677–692.
Grootendorst, Maarten. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. https://arxiv.org/abs/2203.05794.
Guess, Andrew, and Alexander Coppock. 2020. Does counter-attitudinal information cause backlash? Results from three large survey experiments. *British Journal of Political Science* 50:1497–1515.
Guilbeault, Douglas, Solène Delecourt, and Bhargav Srinivasa Desikan. 2025. Age and gender distortion in online media and large language models. *Nature* 646:1129–1137.
Guldi, Jo. 2022. *The dangerous art of text mining: A methodology for digital history.* Cambridge: Cambridge University Press.
Hastings, Orestes P., and Luca Maria Pesando. 2024. What's a parent to do? Measuring cultural logics of parenting with computational text analysis. *Social Science Research* 124: 103074.
Hedström, Peter, and Lars Udehn. 2011. Analytical sociology and theories of the middle range. In *The Oxford Handbook of Analytical Sociology*, eds. Peter Hedström and Peter Bearman, 25–48. Oxford: Oxford University Press.
von der Heyde, Leah, Anna-Carolina Haensch, and Alexander Wenz. 2025. Vox Populi, Vox AI? Using large language models to estimate German vote choice. *Social Science Computer Review* https://doi.org/10.1177/08944393251337014.

Hirschberg, Julia, and Christopher D. Manning. 2015. Advances in natural language processing. *Science* 349:261–266.

Holme, Petter, and Milena Tsvetkova. 2025. Artificially intelligent agents in the social and behavioral sciences: A history and outlook. https://arxiv.org/abs/2510.05743.

Howard, Jeremy, and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics Vol. 1*, eds. Iryna Gurevych and Yusuke Miyao, 328–339. Melbourne: Association for Computational Linguistics.

Huang, Lei, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* 43:1–55.

Hurtado Bodell, Miriam. 2024. Mining for meaning: Using computational text analysis for social inquiry. Doctoral dissertation. Linköping: Linköping University Electronic Press.

Hurtado Bodell, Miriam, and Anastasia Menshikova. 2024. Mechanisms of change: what explains shifts in online immigration discourse after terror attacks? https://osf.io/23cmx_v1/.

Hurtado Bodell, Miriam, Martin Arvidsson, and Måns Magnusson. 2019. Interpretable word embeddings via informative priors. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6323–6329. Hong Kong: Association for Computational Linguistics.

Hurtado Bodell, Miriam, Måns Magnusson, and Sophie Mützel. 2022. From documents to data: A framework for total corpus quality. *Socius: Sociological Research for a Dynamic World* 8. https://doi.org/10.1177/23780231221135523.

Hurtado Bodell, Miriam, Måns Magnusson, and Marc Keuschnigg. 2026. Seeded topic models in digital archives: Analyzing interpretations of immigration in Swedish newspapers, 1945–2019. *Sociological Methods & Research* 55:120–156.

Ibrahim, Elida Izani, and Andrea Voyer. 2025. Qualitative research with LLM chatbots: Technological reflexivity for interpretative technology. *Qualitative Research.* https://doi.org/10.1177/14687941251390794.

Imai, Kosuke, and In Song Kim. 2019. When should we use unit fixed effects regression models for causal inference with longitudinal data? *American Journal of Political Science* 63:467–490.

Imai, Kosuke, and Kentaro Nakamura. 2025. GenAI-powered inference. https://arxiv.org/abs/2507.03897.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge: Cambridge University Press.

Jagarlamudi, Jagadeesh, Hal Daumé III, and Raghavendra Udupa. 2012. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 204–213. Avignon: Association for Computational Linguistics.

Jarvis, Benjamin F., Marc Keuschnigg, and Peter Hedström. 2021. Analytical sociology amidst a computational social science revolution. In *Handbook of computational social science*, eds. Uwe Engel, Anabel Quan-Haase, Sunny Xun Liu and Lars Lyberg, 33–52. London: Routledge.

Jeon, Nanum, and Jennie E. Brand. 2026. Causal machine learning: A deductive-inductive framework for sociological research. *This issue.*

Jurafsky, Daniel, and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd edition online draft, https://web.stanford.edu/~jurafsky/slp3/.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. http://arxiv.org/abs/2001.08361.

Karell, Daniel, and Michael Freedman. 2020. Sociocultural mechanisms of conflict: Combining topic and stochastic actor-oriented models in an analysis of Afghanistan, 1979–2001. *Poetics* 78:101403.

Katz, Andrew, Gabriella Coloyan Fleming, and Joyce Main. 2024. Thematic analysis with open-source generative AI and machine learning: A new method for inductive qualitative codebook development. http://arxiv.org/abs/2410.03721.

Kojima, Takeshi, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, *NIPS '22*, 22199–22213. Red Hook: Curran Associates.

Kozlowski, Austin C., and James Evans. 2025. Simulating subjects: The promise and peril of artificial intelligence stand-ins for social agents and interactions. *Sociological Methods & Research* 54:1017–1073.

Kozlowski, Austin C., Matt Taddy, and James A. Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review* 84:905–949.
Lasswell, Harold D. 1934. *World Politics and Personal Insecurity*. Glencoe: Free Press.
Lee, David S., and Thomas Lemieux. 2014. Regression discontinuity designs in social sciences. In *The SAGE Handbook of Regression Analysis and Causal Inference*, 301–326. London: Sage Publications.
Leitgöb, Heinz, and Florian Keusch. 2026. Causal inferences from digital behavioral data. Methodological implications. *This issue.*
Liesenfeld, Andreas, Alianda Lopez, and Mark Dingemanse. 2023. Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, *CUI '23*, 1–6. New York: Association for Computing Machinery.
Lin, Boda, Xinyi Zhou, Binghao Tang, Xiaocheng Gong, and Si Li. 2023. ChatGPT is a potential zero-shot dependency parser. http://arxiv.org/abs/2310.16654.
Lin, Gechun. 2025. Using cross-encoders to measure the similarity of short texts in political science. *American Journal of Political Science* 69:1600–1616.
Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. http://arxiv.org/abs/1907.11692.
van Loon, Austin. 2022. Three families of automated text analysis. *Social Science Research* 108:102798.
Luo, Yiwei, Kristina Gligorić, and Dan Jurafsky. 2024. Othering and low status framing of immigrant cuisines in US restaurant reviews and large language models. In *Proceedings of the International AAAI Conference on Web and Social Media* , Vol. 18, eds. Yu-Ru Lin, Yelena Mejova, Meeyoung Cha, 985–998, Washington, DC: AAAI Press.
Macanovic, Ana. 2022. Text mining for social science: The state and the future of computational text analysis in sociology. *Social Science Research* 108: 102784.
Macanovic, Ana, and Wojtek Przepiorka. 2024. A systematic evaluation of text mining methods for short texts: Mapping individuals' internal states from online posts. *Behavior Research Methods* 56:2782–2803.
McFarland, Daniel A., Kevin Lewis, and Amir Goldberg. 2016. Sociology in the era of big data: The ascent of forensic social science. *The American Sociologist* 47:12–35.
Mei, Qiaozhu, Yutong Xie, Walter Yuan, and Matthew O. Jackson. 2024. A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences* 121: e2313925121.
Mendelsohn, Julia, Ceren Budak, and David Jurgens. 2021. Modeling framing in immigration discourse on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2219–2263. Association for Computational Linguistics.
Menshikova, Anastasia. 2025. *Cultural change : Studying social interdependencies in public discourse with computational text analysis*. Linköping: Linköping University Electronic Press.
Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. http://arxiv.org/abs/1310.4546.
Miller, Justin K. and Tristram J. Alexander. 2025. Human-interpretable clustering of short text using large language models. *Royal Society Open Science* 12:241692.
Mohr, John W. 1998. Measuring meaning structures. *Annual Review of Sociology* 24:345–370.
Mohr, John W., Christopher A. Bail, Margaret Frye, Jennifer C. Lena, Omar Lizardo, Terence E. McDonnell, Ann Mische, Iddo Tavory, and Frederick F. Wherry. 2020. *Measuring culture*. New York: Columbia University Press.
Mu, Yida, Chun Dong, Kalina Bontcheva, and Xingyi Song. 2024. Large language models offer an alternative to the traditional approach of topic modelling. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, eds. Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, Nianwen Xue, 10160–10171. Torino: ELRA and ICCL.
Murphy, Kevin P. 2012. *Machine Learning: A probabilistic perspective*. Cambridge: MIT Press.
Murthy, Dhiraj. 2024. Sociology of Twitter/X: Trends, challenges, and future research directions. Annual Review of Sociology 50:169–190.
Mützel, Sophie, and Étienne Ollion. 2025. Machine learning and the analysis of culture. In *The Oxford Handbook of the Sociology of Machine Learning*, eds. Christian Borch and Juan Pablo Pardo-Guerra. Oxford: Oxford University Press.

Nelson, Laura K. 2020. Computational grounded theory: A methodological framework. *Sociological Methods & Research* 49:3–42.
Nelson, Laura K. 2021. Leveraging the alignment between machine learning and intersectionality: Using word embeddings to measure intersectional experiences of the nineteenth century U.S. South. *Poetics* 88:101539.
Nguyen, Van Bach, Paul Youssef, Christin Seifert, and Jörg Schlötterer. 2024. LLMs for generating and evaluating counterfactuals: A comprehensive study. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 14809–14824. Miami: Association for Computational Linguistics.
Ollion, Étienne, Rubing Shen, Ana Macanovic, and Arnault Chatelain. 2024. The dangers of using proprietary LLMs for research. *Nature Machine Intelligence* 6:4–5.
OpenAI et al. 2024. GPT-4 Technical report. http://arxiv.org/abs/2303.08774.
Pangakis, Nicholas, Samuel Wolken, and Neil Fasching. 2023. Automated annotation with generative AI requires validation. https://arxiv.org/abs/2306.00176.
Park, Joon Sung, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–22. New York: Association for Computing Machinery.
Peters, Matthew E., Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, eds. Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, 1499–1509. Brussels: Association for Computational Linguistics.
Pournaki, Armin, and Tom Willaert. 2025. Extracting narrative signals from public discourse: a network-based approach. *Humanities and Social Sciences Communications* 12:1774.
Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.
Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
Reiss, Michael V. 2023. Testing the reliability of ChatGPT for text annotation and classification: A cautionary remark. http://arxiv.org/abs/2304.11085.
Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. Structural topic models for open-ended survey responses. *American Journal of Political Science* 58:1064–1082.
Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. 2019. stm: An R package for structural topic models. *Journal of Statistical Software* 91:1–40.
Roberts, Margaret E., Brandon M. Stewart, and Richard A. Nielsen. 2020. Adjusting for confounding with text matching. *American Journal of Political Science* 64:887–903.
Rodman, Emma. 2020. A timely Intervention: Tracking the changing meanings of political concepts with word vectors. *Political Analysis* 28:87–111.
Rubin, Donald B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66:688–701.
Rytting, Christopher Michael, Taylor Sorensen, Lisa Argyle, Ethan Busby, Nancy Fulda, Joshua Gubler, and David Wingate. 2023. Towards coding social science datasets with language models. https://arxiv.org/abs/2306.02177v1.
Salganik, Matthew J. 2018. *Bit by bit: Social research in the digital age*. Princeton: Princeton University Press.
Santana, Jessica J., and Laura K. Nelson. 2025. How machine learning is reviving sociological theorization. In *The Oxford Handbook of the Sociology of Machine Learning*, eds. Christian Borch and Juan Pablo Pardo-Guerra. Oxford: Oxford University Press.
Savelka, Jaromir, Kevin D. Ashley, Morgan A. Gray, Hannes Westermann, and Huihui Xu. 2023. Can GPT-4 support analysis of textual data in tasks requiring highly specialized domain expertise? In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education Vol. 1*, 117–123. New York: Association for Computing Machinery.
Schwitter, Nicole, Ruben L. Bach, and Christopher Klamm. 2026. Text-as-data and causal inference in sociology. *This issue.*
Spirling, Arthur. 2023. Why open-source generative AI models are an ethical way forward for science. *Nature* 616:413–413.

Stein, Jonas, Marc Keuschnigg, and Arnout van de Rijt. 2023. Network segregation and the propagation of misinformation. *Scientific Reports* 13:917.

Steyvers, Mark, and Tom Griffiths. 2007. Probabilistic topic models. In *Handbook of latent semantic analysis*, 439-460. New York: Psychology Press.

Stoltz, Dustin S, and Marshall A Taylor. 2024. *Mapping texts: Computational text analysis for the social sciences*. Oxford University Press.

Stuart, Elizabeth A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical Science* 25: 1–21.

Stuhler, Oscar. 2022. Who does what to whom? Making text parsers work for sociological inquiry. *Sociological Methods & Research* 51:1580–1633.

Stuhler, Oscar, Cat Dang Ton, and Etienne Ollion. 2025. From codebooks to promptbooks: Extracting information from text with generative large language models. *Sociological Methods & Research* 54:794–848.

Tao, Yan, Olga Viberg, Ryan S. Baker, and René F. Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus* 3: pgae346.

Tavory, Iddo, and Stefan Timmermans. 2014. *Abductive analysis: Theorizing qualitative research*. Chicago: University of Chicago Press.

Taylor, Marshall A., Dustin S. Stoltz, Heather Harper, Sanuj Kumar, Sumanth Reddy Nandhikonda, and Luke Burks. 2025. A simulation-based slope metric for anchor list reliability in word embedding spaces. https://osf.io/preprints/socarxiv/sc2ub_v2/.

Than, Nga, Leanne Fan, Tina Law, Laura K. Nelson, and Leslie McCall. 2025. Updating "the future of coding": Qualitative coding with generative large language models. *Sociological Methods & Research* 54:849–888.

Tibshirani, Robert. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B* 58:267–288.

Törnberg, Petter. 2023. ChatGPT-4 outperforms experts and crowd Workers in annotating political twitter messages with zero-shot learning. https://arxiv.org/abs/2304.06588.

Underwood, Ted, Kevin Kiley, Wenyi Shang, and Stephen Vaisey. 2022. Cohort succession explains most change in literary culture. *Sociological Science* 9:184–205.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, *NIPS'17*, 6000–6010. Red Hook: Curran Associates.

Veitch, Victor, Dhanya Sridhar, and David Blei. 2020. Adapting text embeddings for causal inference. In *Conference on Uncertainty in Artificial Intelligence,* 919–928. Proceedings of Machine Learning Research. Oregon: AUAI Press.

Voyer, Andrea, Zachary D Kline, and Madison Danton. 2022. Symbols of class: A computational analysis of class distinction-making through etiquette, 1922–2017. *Poetics* 94: 101734.

Wang, Yixin, and David M. Blei. 2019. The blessings of multiple causes. *Journal of the American Statistical Association* 114:1574–1596.

Wang, Angelina, Jamie Morgenstern, and John P. Dickerson. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence* 7:400–411.

Wang, Yongjie, Xiaoqi Qiu, Yu Yue, Xu Guo, Zhiwei Zeng, Yuhong Feng, and Zhiqi Shen. 2024. A survey on natural language counterfactual generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, eds. Yaser Al-Onaizan, Mohit Bansal and Yun-Nung Chen, 4798–4818. Miami: Association for Computational Linguistics.

Watanabe, Kohei, and Yuan Zhou. 2022. Theory-driven analysis of large corpora: Semisupervised topic classification of the UN speeches. *Social Science Computer Review* 40:346–366.

Widmann, Tobias, and Maximilian Wich. 2023. Creating and comparing dictionary, word embedding, and transformer-based models to measure discrete emotions in German political text. *Political Analysis* 31:626–641.

Xian, Yongqin, Bernt Schiele, and Zeynep Akata. 2017. Zero-shot learning—The good, the bad and the ugly. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 3077–3086. Los Alamitos: IEEE Computer Society.

Yung, Vincent, Jeanette. A. Colyvas, and Hokyu Hwang. 2025. Quality control for quality computational concepts: Wrangling with theory and data wrangling as theorizing. In *The Oxford Handbook of the Sociology of Machine Learning*, eds. Christian Borch and Juan Pablo Pardo-Guerra. Oxford: Oxford University Press.

Miriam Hurtado Bodell is a postdoctoral researcher in analytical sociology at Linköping University, Sweden. Her research focuses on meaning-making dynamics, interpretative heterogeneity, cultural change, and the development of computational text-analytical methods for sociological inquiry.

Marc Keuschnigg is professor of sociology at Leipzig University, Germany, and at Linköping University, Sweden. His research interests include cultural dynamics, normative change, and spatial inequality.

Ana Macanovic is an assistant professor of sociology at Utrecht University, the Netherlands. She studies processes of cumulative advantage, inequality, social influence, and innovation with the help of computational analyses of large textual and register data, formal models and experimental methods.

Anastasia Menshikova is a postdoctoral researcher in computational social science at Uppsala University, Sweden. She studies individual and collective cultural change, applying computational approaches to measure culture.