Social Science Research 108 (2022) 102784

Contents lists available at ScienceDirect

Social Science Research

journal homepage: www.elsevier.com/locate/ssresearch

Text mining for social science – The state and the future of computational text analysis in sociology

Ana Macanovic

Utrecht University, Department of Sociology / ICS, Padualaan 14, 3584 CH, Utrecht, The Netherlands

ARTICLE INFO

Keywords:
Text mining
Text analysis
Content analysis
Machine learning
Natural language processing
Big data

ABSTRACT

The emergence of big data and computational tools has introduced new possibilities for using large-scale textual sources in sociological research. Recent work in sociology of culture, science, and economic sociology has shown how computational text analysis can be used in theory building and testing. This review starts with an introduction of the history of computer-assisted text analysis in sociology and then proceeds to discuss five families of computational methods used in contemporary research. Using exemplary studies, it shows how dictionary methods, semantic and network analysis tools, language models, unsupervised, and supervised machine learning can assist sociologists with different analytical tasks. After presenting recent methodological developments, this review summarizes several important implications of using large datasets and computational methods to infer complex meaning in texts. Finally, it calls researchers from different methodological traditions to adopt text mining tools while remaining mindful of lessons learned from working with conventional data and methods.

1. Introduction

The proliferation of big data and digitization of social life introduced social scientists to amounts of data that elude their conventional sources and methods. Keeping up with the developments in data availability and methodology, computational social science–a field that uses novel data sources and computational methods to answer questions about the social–has emerged at the intersection of several social scientific disciplines (Edelmann et al., 2020; Lazer et al., 2009). Albeit at a somewhat slower pace compared to the neighboring disciplines, the enthusiasm for the integration of new data sources and methods into sociological research has been increasing (Edelmann et al., 2020; Lazer and Radford 2017). In the last decade, as many as six papers in the *Annual Review of Sociology* have discussed new developments related to big data analysis using computational methods[1]; so did a 2016 special issue of *Social Science Research*.

Text analysis is an approach that lends itself to the use of computational methods especially well. Sociology has a long history of using computer-assisted text analysis in both quantitative and qualitative research traditions. It, thus, does not come as a surprise that notable methodological innovations were made in sociological research concerned with the analysis of large amounts of textual data. Sociologists have used novel computational methods to explore a variety of phenomena, including those clearly related to textual expression: the discourse of and surrounding the political elites (van Atteveldt et al., 2008; Bonikowski and Gidron 2016; DiMaggio

E-mail address: a.macanovic@uu.nl.

[1] These are reviews by Golder and Macy (2014), Evans and Aceves (2016), Lazer and Radford (2017), Molina and Garip (2019), and Edelmann and colleagues (2020).

https://doi.org/10.1016/j.ssresearch.2022.102784
Received 10 April 2022; Received in revised form 5 August 2022; Accepted 10 August 2022
Available online 2 September 2022
0049-089X/© 2022 The Author. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

et al. 2013), the rhetoric of online communities (Davidson et al., 2017; Törnberg and Törnberg 2016a, 2016b) and social groups and movements (Almquist and Bagozzi 2019; Nelson 2021), the varieties of event framing in media (Bastin and Bouchet-Valat 2014; Franzosi et al. 2012), and the nature of public debates surrounding different societal issues (Bail et al. 2017). Computational text analysis has supported both research concerned with the evolution of cultural meanings at a macro level (Boutyline et al. 2020; Kozlowski et al., 2019) and inference on values and cultural schemas at the micro level (Macanovic and Przepiorka 2021; Taylor and Stoltz 2020). Finally, computational methods have helped us understand the inner workings of science (McMahan and Evans 2018; Rona-Tas et al., 2019; Schwemmer and Wieczorek 2020) and supported new methodological tools (Nardulli et al. 2015; Zhang and Pan 2019).

This paper provides an overview of research leveraging the power of computational text analysis in sociological theory building and testing. To support further integration of novel text analysis methods into the sociological toolbox, this review is organized around method families and tasks they can assist with, rather than around the analyzed social phenomena (as in, for example, Edelmann et al., 2020 or Evans and Aceves, 2016). This review first introduces dictionary methods that can help researchers quantify the presence of words related to concepts of interest–such as hate speech or populism–in textual data. Further, it discusses semantic and network text analysis methods that facilitate the identification of social actors and actions in text. Next, it describes methods that allow text to be numerically represented, helping researchers analyze complex meanings in text. Finally, it introduces unsupervised text clustering methods that facilitate exploratory analysis by grouping texts based on latent topics they contain and supervised text classification approaches that can automatically expand manual text analysis onto a large number of texts.

The increasing number of applications using computational text analysis in sociology has spurred discussions concerning the reliability of large textual sources (Olteanu et al., 2019; Radford and Lazer 2019; Salganik 2017; Wagner-Pacifici et al., 2015), the practical challenges of collecting and analyzing textual data, and the consequences of introducing inductive computational methods into the sociological toolbox (Goldberg 2015). Further, these new developments have brought back to light old, but nevertheless extremely relevant discussions on the reach of sociological exploration of meaning in text (Lee and Martin 2015). As such concerns are inherent to any application using large textual data and computational methods, this review also outlines the main arguments raised in these discussions.

This paper starts with a brief overview of the history of computer-assisted text analysis in sociology and then groups the variety of computational text analysis approaches into five major method groups. It briefly outlines the capabilities of each of the method groups and presents representative sociological studies that have used them. With a few exceptions, this review is limited to research published in the last decade. Next, it discusses recent developments in text mining methodology that can further support sociological applications. Finally, this paper concludes by outlining major challenges surrounding new data sources and methods, summarizing fruitful ontological discussions in the field, and sketching the future of sociological study within the landscape of computational research. This review is not exhaustive nor technically detailed; rather, it aims to showcase the usefulness of text mining methods for different tasks at a high level with a hope of bringing them closer to a variety of researchers in the field.

2. A brief history of text analysis in sociology

Computational text analysis approaches recently used in sociology draw on existing methodological traditions. Some stem from a more quantitative text analysis tradition (e.g., dictionary methods, semantic and network analysis), some were mostly used in a more qualitative tradition (e.g., unsupervised text clustering), while others come from work at the intersection of the two. Before discussing new computational approaches, two dominant methodological traditions and their early use of computer-assisted text analysis are shortly introduced.

The roots of quantitative text analysis in social sciences can be traced back to the proliferation of printed media in the early 20th century (Krippendorff 2004). These developments led to Speed's pioneering research on newspapers in 1893 (Speed 1893) and Max Weber's proposal of a large-scale systematic analysis of the press in 1910 (Krippendorff 2004). Modern quantitative text analysis in sociology largely stems from mass communication research on press and political propaganda conducted during the 1930s and 1940s (Krippendorff 2004). In line with these developments, Berelson defined content analysis as an "objective, systematic and quantitative description of the manifest content of communication" (Berelson, 1952: 18 cited in Crano et al. 2014:303). This definition was later broadened to also encompass latent aspects of textual content (Krippendorff 2004; Roberts 1989).

Computers were used to assist with the quantitative analysis of large amounts of text (e.g., word counting) already in the late 1950s, with researchers hoping that automation could alleviate concerns surrounding the objectivity of human coders (Krippendorff 2004; Roberts 1989). Yet, the enthusiasm for computer-aided analysis has faded by the 1980s (Krippendorff 2004). In the early 2000s, Popping (2000) and Krippendorff (2004) discussed how computers can be used in quantitative text analysis for counting words and exploring semantic relationships in texts. Sections 3.1. and 3.2. discuss the contemporary successors of these computer-assisted solutions.

Qualitative text analysis emerged in the second half of the 20th century as a reply to what was seen as the superficiality of results of quantitative analyses (Krippendorff 2004). Building on the tradition of hermeneutical analysis established in the humanities in the late 19th century, qualitative approaches emphasized the importance of latent text content and the nature of texts as cultural products (Kuckartz 2014). Qualitative analyses, thus, call for deep reading by experts informed about the historical and social context the text

was generated in. Researchers should, further, avoid imposing their *a priori* theoretical expectations onto texts.[2] Due to their interpretative nature, qualitative analyses were mostly performed manually, with computers scarcely used to facilitate transcription, text handling, and manual analysis from the mid-1980s onwards (Gibbs 2014; Krippendorff 2004; Kuckartz 2014). As discussed in Section 3, computers were used more extensively in qualitative text analysis only later.

More recently, developments in the fields of Natural Language Processing and machine learning have offered new opportunities for using computers for text analysis in social sciences.[3] Despite the rich tradition of quantitative computer-assisted text analysis, the recent breakthrough in the use of text mining in sociological research mostly came from a more qualitative tradition (e.g., the 2013 issue of journal *Poetics*). Text mining has since been used to support (or replicate) diverse manual approaches to text analysis, ranging from quantitative network analysis (Goldenstein and Poschmann 2019b; Sudhahar et al., 2015a) and systematic manual coding (Macanovic and Przepiorka 2022; Rona-Tas et al., 2019) to qualitative hermeneutics (Mohr et al., 2013; Mohr et al. 2015) and grounded theory (Muller et al., 2016; Nelson 2017). Most importantly, however, computational approaches allow for both quantification and nuanced reading of texts (Wiedemann 2016), closing the gap between quantitative and qualitative approaches to text analysis (DiMaggio et al., 2013; Wiedemann 2016).

3. Text mining for sociological research

This section provides an overview of several families of text mining methods, tasks they can assist with, and studies that exemplify their usefulness for sociological exploration. Table 1 gives a summary of reviewed methods and details on their representative applications in recent sociological work in several domains. In the Appendix, readers can find a list of accessible software tools for implementing the methods discussed in this section.

3.1. Dictionary methods

In text analysis applications, dictionaries denote lists of words related to a concept of interest.[4] Dictionary methods automatically identify words from a dictionary in text and are, therefore, particularly useful in identifying and quantifying the presence of a concept of interest (e.g., offensive language) in large bodies of text. Researchers can rely on empirically validated pre-existing dictionaries, define their own dictionaries suited to a particular research question (Weber 1984), or combine these two types of resources.

Pre-existing dictionaries are particularly useful when seeking concepts that have been identified in previous research–such as, for instance, words denoting positive or negative affect. Some such pre-existing resources have been extensively empirically validated. For example, the General Inquirer contains numerous dictionary categories indicating, just to name a few, emotions, political concepts, and aspects from motivation theory (Stone et al., 1966). Similarly, the Linguistic Inquiry and Word Count (LIWC) software dictionary contains 82 categories pertaining to different psychological processes (Pennebaker et al., 2015; Tausczik and Pennebaker 2010). A study by Golder and Macy (2011) uses the LIWC dictionary to identify positive and negative words in 509 million tweets from 2.4 million individuals in 84 countries. Authors show that tweets, overall, tend to be more positive on weekends and in the mornings, finding support for the hypothesis that social media texts reflect biological rhythms. Similarly, Bail et al. (2017) use the LIWC dictionary to identify elements of affective and rational language styles in social media posts. Their analysis of the content shared on social media by 92 organizations over 1.5 years supports the hypothesis on the existence of cycles of emotional and rational language in online discussions (Bail et al., 2017).

While easy to use, pre-existing dictionaries have several limitations. First, they tend to be rather domain-specific: for instance, words usually marked as negative in general dictionary resources–such as tax or cost–do not have negative connotations in a financial context (Loughran and McDonald 2011). Therefore, dictionaries built to measure the same concept in different domains can deliver divergent results when applied to new data (Jaidka et al., 2020). Second, dictionaries have only been defined for some sociologically relevant concepts. To answer other research questions, researchers might need to develop their own dictionaries–for example, to identify populist language use by presidential candidates in the US (Bonikowski and Gidron 2016) or seek articles on economic inequality in newspapers (Nelson et al., 2018).

Finally, pre-existing resources are often available in a limited number of languages. For instance, having worked with German-language data, Spörlein and Schlueter (2021) could not rely on any of the existing English-language dictionaries of ethnic hate. Authors therefore created a custom dictionary to detect hateful posts in their dataset. Analyzing 5152 comments on YouTube videos of German political talk shows, they find evidence that previous hateful comments increase the probability of new comments also containing ethnic hate (Spörlein and Schlueter 2021). While they allow researchers to closely specify their concept of interest and fine-tune the search based on specific data, custom dictionaries can be time-consuming and challenging to create. Section 4 discusses

[2] The grounded theory approach exemplifies this tradition: rather than starting with a theoretical framework to organize texts, researchers allow categories to emerge during the process of reading (Kuckartz 2014).

[3] Natural Language Processing (NLP) denotes computational techniques for analyzing naturally occurring texts for, e.g., information retrieval, summarization, and machine translation (Liddy 2001). Machine learning (ML) is a field that uses insights from computer science and statistics to build algorithms that handle predictive tasks (Althbiti and Ma 2022). Both NLP and ML are at the intersection of the fields of Artificial Intelligence (AI) and Computer Science (CS). Text mining is a broad term denoting computer-assisted analysis of texts using techniques from NLP and ML (Hotho et al. 2005). In this review, the term "text mining" is used interchangeably with "computational" or "automatic text analysis".

[4] *Merriam-Webster*, s.v. "Dictionary", accessed March 18, 2022, https://www.merriam-webster.com/dictionary/dictionary.

Table 1
Methods and studies surveyed in this review.

Method	Text mining used to	Authors	Dataset size	Dataset source	Time span	Field
Existing dictionary	Identify concept of interest, evaluate the nature of texts (affect)	Golder and Macy (2011)	509 million short texts	Social media (twitter)	2008–2010	Online discourse, individual states
	Identify concept of interest, evaluate the nature of texts (affective and rational language)	Bail et al. (2017)	223 thousand short texts	Social media (Facebook)	2011–2012	Online discourse
Custom-made dictionary	Identify concept of interest (hate speech)	Spörlein and Schlueter (2021)	5152 short texts	Social media (YouTube)	2015–2017	Social media discourse, social influence
Semantic and network analysis	Identify actors and directionality of actions (social actors, oppressive acts)	Franzosi et al. (2012)	1332 medium-length texts	News data	1875–1930	Public discourse, social action
	Identify actors and directionality of actions (actors in political sphere, acts of speech)	Van Atteveldt et al. (2008)	5988 medium-length texts	News data	2006	Political communication
	Identify actors, directionality and polarity of actions (actors in political sphere, acts of speech)	Sudhahar et al. (2015b)	130 thousand medium-length texts	News data	2012	Political communication
Language representations	Measure similarity of language used by different actors (cultural embeddedness)	Goldberg et al. (2016)	10.2 million short texts	Communication data (internal e-mails)	2009–2014	Sociology of organizations
	Measure the likelihood of words appearing in a context (ambiguity in scientific communication)	McMahan and Evans (2018)	1.9 million short texts	Scientific data (research article abstracts)	1974–1995	Sociology of science
	Track relationships between cultural categories over time (categories of social class)	Kozlowski et al. (2019)	Millions of long texts	Books/official publications	1900–2012	Cultural sociology, social stratification
Unsupervised text clustering	Identify frames in text (public funding)	DiMaggio et al. (2013)	8000 medium-length texts	News data	1986–1997	Public discourse, cultural sociology
	Identify concept of interest in text, track its evolution (framing of Islam)	Törnberg and Törnberg (2016a, b)	50 million short texts	Social media (online forum)	2000–2013	Public discourse, cultural sociology
	Identify differences in use of concepts in texts (topics emphasized by different social actors)	Nelson (2021)	1131 medium-length/long texts	Books/official publications	1900–1975	Public discourse, cultural sociology, social action
	Identify the leaning of texts towards concepts of interest (methodological leaning)	Schwemmer and Wieczorek (2020)	8737 short texts	Scientific data (research article abstracts)	1995–2017	Sociology of science
Supervised text classification	Expand manual coding (identification of motives and moral norms)	Macanovic and Przepiorka (2021)	1.9 million short texts	Communication data (online marketplace)	2013–2017	Individual states, economic sociology
	Expand manual coding, evaluate the consistency of manual coding (identification of uncertainty in scientific communication)	Rona-Tas et al. (2019)	115 medium-length/long texts	Official documents (risk-assessment documents)	2000–2010	Public discourse, sociology of science

recent methodological developments that facilitate the creation of such dictionaries.

3.2. Semantic and network text analysis

While dictionary analysis can quantify the nature of the text (e.g., its affect) or identify the presence of concepts of interest in data (e.g., whether the text contains hate speech), other algorithms are able identify the presence of certain social actors and the nature of their social action in text. Such methods especially facilitate narrative analysis (Popping and Roberts 2015). Semantic text analysis identifies the so-called "Subject-Action-Object" (SAO) triplets in text by assigning semantic roles to words in a sentence. These triplets can then be mapped to social actors and their actions towards other actors (Franzosi 1989; Popping 2000; Roberts 1989). Network analysis further evaluates words as nodes and relationships between them as links (Carley 1999; Carley and Palmquist 1992; Lee and Martin 2015; Popping 2000), mapping the directionality and strength of relationships identified in text.[5]

A study by Franzosi et al. (2012) showcases how semantic and network analyses can assist precise identification of social actors in large textual data. Authors identify social actors and actions in 1332 newspaper articles related to 392 lynching events occurring in Georgia, US, between 1875 and 1930. By doing so, they are able to quantify the extent to which African Americans (receiving actors) were victimized (directed action) by mobs and the law enforcement (giving actors) and show how the mobs pressured law enforcement to impose the "lynch law" (Franzosi et al., 2012). Van Atteveldt, Kleinnijenhuis, and Ruigrok (2008) show how semantic text analysis can be combined with hypothesis testing. Analyzing 5988 newspaper articles, authors find that powerful political parties receive more attention in the press (van Atteveldt et al., 2008). A study by Sudhahar and colleagues (2015b) shows how network analysis can be used to evaluate the valence of relationships between actors identified in text. Analyzing 130 thousand news articles about the 2012 US presidential elections, authors show that republican politicians speak more negatively of democratic politicians than vice versa (Sudhahar et al., 2015b).[6]

3.3. Language representations

So far, this review explored how text mining can be used to capture the existence of concepts and identify social actors and their actions in textual data. Researchers can also rely on word occurrences and co-occurrences to capture differences in language used by various actors and capture complex meanings embedded in texts. The simplest approach of this kind includes converting a textual corpus into a so-called Document-Term Matrix (DTM) where every matrix row represents a text, every column a word, and every field the number of times each word appears in each text.[7] One row of such a Document-Term Matrix contains counts of all words in an individual text, effectively serving as its simple numeric vector representation. Goldberg et al. (2016) use such vectors to analyze more than 10 million emails exchanged in an organization and capture each employees' embeddedness in organizational culture. Authors capture cultural embeddedness by calculating the distance between vectors representing incoming and outcoming emails of each employee. Their analysis shows that higher levels of employee cultural embeddedness can compensate for less beneficial network positions in organizational structure and support organizational attainment (Goldberg et al., 2016).

Some research questions call for consideration of the linguistic context of each word (i.e., the few words preceding and following it). In such applications, researchers can compute the conditional probability of a word occurring in a certain position in text based on the words that precede it (Jurafsky and Martin 2009). McMahan and Evans (2018) rely on a similar model of language generation to capture the concept of ambiguity in scientific research. This model allows authors to assess the probability of a word taking a certain meaning given its linguistic context: the more synonyms of a word occur in the same context across the body of relevant texts, the more ambiguous the word's meaning. Authors combine this information with networks of scientific citations and confirm the hypothesis that scientific papers using more ambiguous language receive less fragmented citations–thus showing how ambiguity stimulates scientific discovery (McMahan and Evans 2018).

Texts can be transformed into numbers in a yet more complex manner by using the so-called distributed text representations (Mikolov et al., 2013a, b). These representations–word embedding models–represent words as vectors in a dense high-dimensional space where words that occur in similar contexts across the analyzed body of texts are positioned close to each other in space (Mikolov et al. 2013a, b). Compared to other methods, word embedding models are particularly suited for representing multifaceted associations between words and capturing cultural nuances of meaning in large, complex textual corpora (Kozlowski et al., 2019). Existing evidence indicates that word embedding models map meaning in texts in a manner comparable to human mapping (Utsumi 2020).

A study by Kozlowski and colleagues (2019) shows how word embedding models can capture complex cultural categories and

[5] For instance, take the following sentence: "A criticized B". Semantic analysis can help us identify that person B is the object of person A's criticism; network analysis will further link persons A and B with a negative tie – since criticism implies a negative relationship. Similarly, if A and B appear in a similar relationship in multiple sentences in the text, the strength of their tie can be adjusted to reflect this.

[6] See Goldenstein and Poschmann (2019b) for a text-mining implementation of map analysis as defined by Carley (1999), another approach that considers relationships between concepts in text.

[7] Such representations are called discrete text representations; they treat words independent of the context they appear in within the sentence ("bag-of-words"). While rather simple, this approach is very effective in text mining (Hopkins and King 2010; Nelson et al., 2018). DTMs can be further adjusted to account for, for example, infrequently occurring words (Aggarwal 2018). These matrices are also used as inputs for machine learning methods (see Sections 3.4. and 3.5).

relationships between them. Authors represent words found in millions of books published over a century as 300-dimensional vectors and use these vectors to construct cultural dimensions of social class–such as affluence or status–and track their positions in the language vector space over time. This approach allows for a very granular analysis of the dynamics of cultural dimensions of class: for example, while education was associated with affluence through sophistication in the 1900s, by the 1990s this association became direct (Kozlowski et al., 2019). Other studies have leveraged the power of word embedding models to trace the evolution of gender stereotypes using a large corpus of print media (Boutyline et al., 2020) and explore the relationship between the public discourse surrounding the coronavirus pandemic and the tendency of residents to socially distance in different communities (van Loon et al., 2020).

3.4. Unsupervised text clustering

In applications that seek to identify broad ways in which concepts are discussed in text, researchers can rely on unsupervised machine learning methods that automatically cluster words, groups of words, and texts into similar groups (Aggarwal and Zhai 2012a). Unlike some of the previously discussed methods that require a specification of initial theoretical expectations (i.e., defining words that correspond to a specific phenomenon of interest when using dictionary methods), unsupervised clustering allows for exploratory analyses in which concepts of interest (inductively) emerge from texts (DiMaggio et al., 2013). Following an influential issue of the journal *Poetics* in 2013,[8] sociological literature has mostly relied on algorithms that allow for both word- and text-based clustering: topic modelling algorithms.[9] The most widely-used topic modelling approach is the Latent Dirichlet Allocation (LDA) (Blei 2012). LDA sees each text as a distribution of topics, and each topic as a distribution of words. The algorithm finds the "hidden" topic structure by evaluating words that tend to occur together more frequently than they would by chance and grouping them into a topic; each text is then assigned to (some of) these topics based on the words it contains (Blei 2012; DiMaggio et al., 2013). Given their content, some topics can be mapped to meaningful analytical categories of interest (see also Section 5.3.).

In a landmark study, DiMaggio et al. (2013) use the power of unsupervised text clustering to identify topics that capture how the governmental support of the arts was framed in the US media between 1986 and 1997. By analyzing 8000 newspaper articles, authors discover that, while the governmental support for the arts started off as a noncontroversial topic, legislative shifts after the takeover of the Congress by the Republican party have led to cultural wars surrounding the arts financing. Rather than seeking theoretical concepts such as frames, Törnberg and Törnberg (2016b, 2016a) use topic modelling to identify topics (and, thus, texts) that contain discussions of Islam and Muslims in a large Swedish online forum. They complement the topic model analysis of 50 million forum posts with a manual analysis to find that Muslims tend to be portrayed as linked to conflict, terrorism, violence and sexual abuse in the analyzed forum; this portrayal is also more pronounced online than in traditional media (Törnberg and Törnberg 2016b). In another study on the same data, Törnberg and Törnberg (2016a) use topic modelling to identify texts where feminism and Islam are jointly discussed within the same topic. Analyzing these texts, authors find that forum users evoke feminism in a rhetorical strategy to attack Islam rather than to convey genuine support for gender equality.

An extension of standard topic model approaches, Structural Topic Models (STM) allow researchers to also include non-textual covariates into their models (Roberts et al. 2016). Nelson (2021) leverages the power of such a model to analyze texts published by different women's movement organizations during the first and second waves of mobilization in Chicago and New York. Using an STM allows the author to include text authorship into the model and identify topics most used by organizations across mobilization waves and cities. This analysis revealed an unexpected finding: while there were differences in discursive strategies used by organizations in two mobilization waves, the inter-city differences between movements within individual waves were even more substantial (Nelson 2021).

Sociologists can rely on another family of unsupervised clustering methods–scaling methods–to estimate the position of texts between two conceptual extremes (Slapin and Proksch, 2008) such as the political left and right. Schwemmer and Wieczorek (2020) use this method to arrange 8737 abstracts in sociology publications across the quantitative-qualitative method axis. By doing so, authors discover that methodological approaches remain closely tied to research areas and topics; yet, the preference for quantitative methodology has been marginally increasing over time across the field as a whole (Schwemmer and Wieczorek 2020).

3.5. Supervised text classification

Some sociological applications lend themselves to manual systematic text coding. This is particularly the case if researchers have already defined the theoretical concepts of interest whose presence can be best identified by trained human coders. Such concepts are often complex, co-occurring with other concepts, or difficult to infer from individual words. Researchers can leverage supervised machine learning classification methods to automatically expand (systematic) manual coding onto a large number of texts. These methods infer word patterns that characterize different coding categories (Nelson et al., 2018) and then use these patterns to code new, previously unseen texts. While such algorithms replicate trained human coding using complex schemes rather well (Hoover et al., 2020; Macanovic and Przepiorka 2022; Nelson et al., 2018), they are currently only scarcely used in the field.

Macanovic and Przepiorka (2021) manually code 2000 texts to infer motives that drive buyers in online markets to write feedback about their transactions. Using supervised machine learning, authors are able to expand this manual coding onto 2 million feedback

[8] vol 41, Issue 6 – December 2013.

[9] For an application that uses an unsupervised clustering method other than topic modelling, see, for example, Bastin and Bouchet-Valat (2014).

texts and granularly explore the role of different motives in generating feedback after various market experiences. Their analysis shows how other-regarding motives present an integral element of the successful functioning of large-scale anonymous online marketplaces (Macanovic and Przepiorka 2021). Rona-Tas et al. (2019) use supervised methods to code complex food safety risk assessments for the level of scientific uncertainty they express. Authors not only show how supervised methods can assist automatic coding of a large number of complex texts, but also use the analytical rigor of machine learning algorithms to explore whether their coding scheme contains internal inconsistencies caused by human coders (Rona-Tas et al., 2019).

Recent methodological work has also used supervised classification in combination with textual and other data to identify events of interest in social media. The Social, Political and Economic Event Database (SPEED) uses supervised machine learning to detect civil unrests in a global archive of news reports (Nardulli et al., 2015). The Collective Action from Social Media (CASM) system also uses these methods as a part of its detection of offline collective action in social media posts (Zhang and Pan 2019).

4. Promising future developments

This section discusses promising developments in text mining methodology that can be of interest to sociologists. These developments include improvements in dictionary creation and analysis, new tools that facilitate semantic and network analysis, advances in the accessibility of text representation and machine learning models, and the use of deep learning methods for text analysis.

Creating custom *dictionaries* can be difficult if researchers have to inspect a large number of texts in order to select a sufficiently wide set of words corresponding to their concepts of interest. Recent work has introduced methods that rely on the concept of word keyness from computational linguistics to facilitate discovery of words of interest in unstructured text (King et al. 2017) and simplify automatic dictionary creation from manually coded data for multiple coding categories (Macanovic and Przepiorka 2022). New software implementations also allow for more precision in dictionary analysis: they can account for words that adjust or change the valence of the word of interest.[10]

Methods that perform automatic *semantic and network analysis* can be computationally demanding. For example, the study of Franzosi et al. (2012) discussed in Section 3.2. explored as many as 7070 semantic triplets. New software implementations allow for faster and more precise extraction of relevant semantic categories from large bodies of text (Nguyen et al., 2021; Welbers et al. 2021). Researchers can also use user-friendly tools such as the ConvoKit (Cornell Conversational Analysis Toolkit) to explore deceptive behaviors, power relations, and structural differences in language used by different social actors (Chang et al., 2020).

As discussed in Section 3.3, *language representations* can assist researchers in exploring the evolution of language and cultural meanings in large textual data. Recently developed approaches facilitate such analyses: generalized word shift graphs allow researchers to identify words that contribute to differences between texts (Gallagher et al., 2021), while the Concept Mover's Distance allows the comparison of texts to an "ideal" text (Stoltz and Taylor 2019).[11] The increased popularity of word embedding models (discussed in Section 3.3.) has led to the development of models based on various textual resources. For instance, a recent project compiled language models in 157 different languages (Grave et al., 2018). Developments in transfer learning allow language models built in one domain to be used in other domains (Pan et al. 2012), providing sociologists with a range of pre-existing language representations.

In the last several years, new developments in deep learning have further improved the quality of language representations. Resembling human information processing, deep learning algorithms iteratively transform the initial (textual) information input into more and more abstract representations (Chatsiou and Mikhaylov 2020; Minaee et al., 2021). Recently developed transformer models are capable of assessing the plurality of contexts in which the same word can occur in text (Vaswani et al., 2018), thus capturing the nuances of meaning even better than simpler word embedding and deep learning models. Finally, software implementations (see the Appendix) of language representation models are becoming increasingly accessible, allowing sociologists to easily build custom models suited for answering specific research questions.

New semi-supervised *clustering models* allow researchers to specify words around which topics should be formed and include non-textual covariates into their models (Eshima et al. 2020; Watanabe 2021). Such approaches combine unsupervised topic models, STMs (see Section 3.4.), and supervised text classification (see Section 3.5.). Finally, developments in deep learning discussed in the previous paragraph have the potential to bring the performance of both supervised and unsupervised machine learning methods even closer to the quality of expert human analysts. Although there appear to be no sociological applications using these methods in text analysis at the time of writing,[12] given their growing popularity in other fields, excellent performance, and the increasing availability of accessible tools (e.g., Wolf et al., 2020), these methods will most likely replace simpler machine learning implementations.

[10] For instance, they recognize that the sentence "Person A is not successful" is not in fact positive, since the positive word "successful" is preceded by a negation (e.g., Naldi 2019).

[11] In an example study, authors define two "ideal" concepts of moral views (the conservative idea of a "strict father" and the liberal idea of a "nurturing parent") and rank the US State of Union Addresses given their similarity to these "ideal" concepts (Stoltz and Taylor 2019).

[12] The only exception being the CASM tool (Zhang and Pan 2019) that uses deep-learning recurrent neural networks for text classification as a part of its event identification procedure. Note that some widely used word embedding models, such as those discussed in Section 3.3., also use neural networks used in deep learning algorithms. Yet, since these models contain only one neural network layer, they are not considered deep learning. The Appendix lists several tools for building (deep learning) word embedding and other supervised and unsupervised machine learning models.

5. Challenges of text mining for sociological research

As discussed in Section 3, sociological studies using text mining explore large textual datasets with the help of methods adopted from computer science. This inclusion of new data sources and methods into the sociological toolbox comes with several challenges to conventional methodological practices. Using (computational) text analysis to infer knowledge about the social world also includes an (implicit) assumption that meanings can be extracted, at least to an extent, from textual data. Thus, the use of text mining in sociology has reignited discussions on the possibility of capturing implicit meaning in text (Bail 2014). Since sociological applications using text mining are bound to face such considerations pertaining to data, methods, and the limits of meaning extraction, this section provides a brief overview of related challenges and discussions in the field.

5.1. Specific considerations of working with large textual data

While social scientists conventionally had to compromise between data size and depth, developments in digitization and data generation online brought the possibility of leaving such limitations behind (Manovich 2011). This review discusses studies using large textual datasets containing as "few" as one thousand texts (Nelson 2021) or as many as 509 million tweets (Golder and Macy 2011). Especially the data generated online allow researchers to capture numerous traces of actual human behaviors in real-life situations and in real time (Radford and Lazer 2019), providing insights beyond what sociologists can collect using traditional research methods (Housley et al., 2014). Such data can be obtained from various platforms (e.g., API access provided by Twitter) or collected using user-written scripts that automatically browse ("crawl") website pages and download the data. Many online platforms, however, impose technical and legal restrictions on content collection (Olteanu et al., 2019). The Appendix lists several tools that can assist researchers in collecting textual data online.

Collecting large textual data can help overcome the obtrusiveness of other data collection methods (Golder and Macy 2014) and the problems of recall and nonresponse biases arising in survey research (Dex 1995). Yet, these data were most often not created with specific research questions in mind. Therefore, they tend to lack relevant information of interest to researchers–such as the characteristics of the authors of individual texts (e.g., their education level, socioeconomic status, individual psychological features), the context in which text was generated (Bail 2014; Boyd and Schwartz, 2021), or the relevant information about the texts themselves (e.g., the Google Books corpus that contains millions of books often lacks reliable meta-data). To address this challenge, researchers can combine different big data sources (van Loon et al., 2020), integrate big data with conventional sociological data sources (e.g., survey data) (Salganik 2017), or complement "thin" textual data with experimental work or follow-up surveys (Boyd and Schwartz, 2021). Further, researchers should collect textual data in a theory-guided manner, considering the additional information that is necessary to answer their research question.

Finally, while researchers have control over sampling in survey research, large textual data obtained online suffer from the illusion of completeness. While appearing comprehensive, they contain biases stemming from platform design, (self-)selection into platforms and databases, and restrictions in data accessibility imposed by the platforms (Manovich 2011; McFarland et al. 2016; Pechenick et al. 2015). Olteanu et al. (2019) provide an extensive survey of issues arising when using such data.

5.2. Text differs from conventionally used quantitative data

To make use of large amounts of text in sociological analyses–whether descriptive or hypothesis-based–textual data needs to be transformed into manageable input. Yet, text substantively differs from conventional quantitative data used by sociologists. Textual data are high dimensional (Gentzkow et al., 2019) and sparse (Aggarwal and Zhai 2012b). Concerning the former, even a small number of texts can contain thousands of unique words: capturing the position of each word in its context within each text requires a high-dimensional data representation. Concerning the latter, each text might contain only a small number of words, while the number of words across all considered texts can be rather large. While modern text mining algorithms can handle sparse high-dimensional data, including the abundance of textual information into the analysis does not necessarily improve the performance of text mining methods (Grimmer and Stewart 2013). On the contrary, working with sparse data and too many variables can easily lead to model overfitting (Aggarwal 2018; Aggarwal and Zhai 2012b; Gentzkow et al., 2019).

Researchers must, therefore, first reduce the dimensionality and sparsity of the data. To decrease the number of features (e.g., words) entering the analysis, one can consider only some textual elements (e.g., removing infrequently occurring words) and use automatic methods to retain only those features that are useful in text mining analyses (Grimmer et al. 2021). Researchers can further simplify text representations by disregarding the context in which each word appears (e.g., the "bag-of-words" approach, see also note 7) (Gentzkow et al., 2019) or using word vector representations such as those described in Section 3.3. Especially when working with machine learning methods, researchers need to be aware of the dangers of overfitting to the data they use to train their models. To check for overfitting, researchers can split their dataset into several subsets–one on which the model is "trained" and one on which it is "validated". If the model performs rather well on the former, but poorly on the latter subset, there is likely overfitting (Domingos 2012). On top of the proper data preparation discussed above, researchers can also choose algorithms that are less prone to overfitting (Hartmann et al., 2019) or look for more complex solutions to combat this issue (see, for instance, Aggarwal and Zhai 2012b).

5.3. Validity of constructs extracted from text can be challenging to establish

As this survey shows, automatic text analysis can be used for a variety of tasks in social science research–from quantifying concepts

of interest to performing an exploratory analysis. The prospect of using text mining to identify and quantify relevant concepts in text has reignited the deeply-rooted sociological discussions on whether the meaning expressed in text can be summarized into quantifiable categories that reflect ontologically real social entities (Ignatow 2016; Lee and Martin 2015) or if the embeddedness of meaning in the ever-changing cultural and historical context can be understood only holistically, through nuanced interpretation and deep reading (Biernacki 2012; Shklar 1986). This question is not unique to text mining applications–it poses a broader question as to whether the outcome of any text analysis accurately captures the meaning contained in textual content. The majority of sociological work using text mining appears to take a middle-ground stance along the following lines: while texts indeed reflect subtle socially situated considerations, they also do encompass some "objective" truths of social life, allowing inference on social phenomena (Ignatow 2016; Ignatow and Mihalcea 2017; Taylor and Stoltz 2020).[13] This approach has the potential to dissolve many barriers between these somewhat contrasting "positivist" and "hermeneutic" views (Edelmann and Mohr 2018; Mohr et al., 2015; Wagner-Pacifici et al., 2015).

Some researchers emphasize how automatic analysis can support transparent and objective identification and quantification of (socially) meaningful phenomena in text (DiMaggio et al., 2013; Mohr 1998; Mohr and Bogdanov, 2013). In this view, for example, topics obtained with topic models discussed in Section 3.4. correspond to actual frames, narratives, or discourses inherently present in text (Pääkkönen and Ylikoski 2021). Other researchers suggest that, rather than helping identify "ontologically real" social phenomena in texts, text mining can facilitate the interpretative analysis with the researcher "in charge" (Breiger et al. 2018). Using the topic modelling parallel again, topics are not mapping "real" social phenomena, but rather assisting researchers with "debiased" deeper reading and qualitative analyses (Mohr et al., 2013; Pääkkönen and Ylikoski 2021; Törnberg and Törnberg 2016a).

Grimmer et al. (2022) suggest that researchers are unlikely to recover a single "true" underlying concept of interest contained in texts. There is no single best way to seek meaning: different text mining methods can help capture relevant aspects of text as per one's theoretical framework. Automatic methods are merely tools to enhance and augment human analysis; their performance, therefore, needs to be validated against a (human) reference (Grimmer et al., 2022). The validation step depends on the approach and the research question: it can range from comparing the performance of automatic methods to a golden standard of expert manual coding (e.g., when it comes supervised text classification) to deep reading of a sample of texts by researchers (e.g., when it comes to unsupervised text clustering). Therefore, text mining methods are as good at extracting and quantifying meaning from texts as the researcher who implements them. Further, recent work has taken up more extensive validation of concepts measured in text against comparable measurements obtained through survey methods (Kross et al., 2019; Pellert et al., 2022; Tay et al., 2020).

5.4. *Methodological and ethical challenges surrounding text mining*

Especially the introduction of text mining methods that inductively seek patterns in text (to capture relationships between words using vector representations described in Section 3.3, cluster texts into topics as discussed in Section 3.4, or find word patterns that best match different coding categories as shown in Section 3.5.) has raised questions about their compatibility with conventional research practices in sociology (Mohr et al., 2013). Yet, the inductive nature of methods themselves does not imply inductive analysis. Nelson (2017) shows how text mining can be used at different stages of the research process, from exploratory analyses to the validation of researchers' interpretations.[14] Once text mining helps researchers identify (and quantify) concepts of interest in text, the identified information can be used in either more exploratory (and deductive) analyses, or as input for strict deductive hypothesis testing (see, in this review, van Atteveldt et al., 2008; Bail et al., 2017; van Loon et al., 2020; also see van de Rijt et al., 2013).[15] More generally, researchers have also discussed how using inductive methods in general could initiate a rethinking of conventional variable-based approaches in sociology (Wagner-Pacifici et al., 2015), inspire new ways of conceptualizing theoretical discovery (Goldberg 2015), and ignite sociological imagination (DiMaggio 2015).

When it comes to practical applications, text mining analyses are often complex and call for numerous decisions regarding data preparation, model choice, result interpretation, and validation (Macanovic and Przepiorka 2022; Nelson 2019; Nelson et al., 2018). For instance, it has been shown that different data preparation pipelines can lead to vastly different outcomes when working with text mining methods (Denny and Spirling, 2018; Uysal and Gunal 2014). Similarly, different choices of the number of topics to be identified by topic models can lead to completely different topic structures and interpretations. Newer sociological work explicitly emphasizes the importance of transparency in text mining implementations (Goldenstein and Poschmann 2019a; Nelson 2019). In this vain, recent contributions have evaluated how data preparation, model choice and specification, and software implementation decisions all affect the performance of text mining methods (Macanovic and Przepiorka 2022; Nelson et al., 2018). This is especially relevant when working with complex models–such as deep learning algorithms discussed in Section 4. By calling for clear reporting of such researcher-made decisions, these contributions set the stage for a systematic development of best practices for sociological research using text mining.

Due to their inductive nature, language models and machine learning methods (see Sections 3.3, 3.4, and 3.5) are prone to

[13] Researchers holding different views all approach this question in a very nuanced manner. For instance, while they take a somewhat more "positivist" approach when using topic modelling, Mohr and Bogdanov clearly emphasize the importance of well-informed heuristic work at the stage of model interpretation (Mohr and Bogdanov, 2013).

[14] Baden et al. (2020) discuss a similar approach in communication science.

[15] Work in other social scientific fields showcases the potential of this approach (Gentzkow et al. 2019; Ghose et al. 2009; Kacewicz et al., 2014; Tausczik and Pennebaker 2010).

algorithmic bias stemming from the fact that these algorithms are designed by humans and learn patterns from human-generated data (Chatsiou and Mikhaylov 2020; Waseem 2016). For example, research has shown that hate speech detection systems that use supervised classification learn and reproduce biases from manually coded data they received as input, effectively discriminating against the groups they were initially designed to protect (see, for example, Davidson and Bhattacharya 2020). In this light, calls for "algorithmic fairness" emphasize the need to remedy discriminatory and harmful representations of certain groups in big data (Whittaker et al., 2018). Researchers should, therefore, be mindful of potential biases present in their data prior to using machine learning methods.

Finally, new ethical questions arise around the collection, use, and storage of large data sources collected online (Radford and Lazer 2019; Salganik 2017). Unlike participants in survey, interview, or experimental studies, individuals who engage in everyday online activities do not provide their consent to researchers as explicitly (Salganik 2017).[16] Further, data generated online can be easily linked across online platforms and sources if not adequately anonymized (Salganik 2017), possibly allowing for the identification of individuals without their consent. Finally, the ethics boards of research institutions, at times, lag behind the dynamic needs of new digitalized environments and computational applications (Neuhaus and Webmoor 2012).

6. Conclusions

Advances in digitization and computing have inspired a range of applications using computational methods to analyze large-scale behavioural data in the emerging field of computational social science (Lazer et al., 2020). The proliferation of digitalized textual data about the present and the past provides exciting opportunities for advancing the exploration of the social through text. While the inclusion of text mining into the sociological toolbox has not always been straightforward (Ignatow 2016), sociologists have been increasingly relying on large textual data and powerful computational tools to test a range of sociological theories. This review summarizes how different text mining tools have been used by sociologists in various analytical tasks across research questions and methodological traditions.

As discussed in Section 3, sociologists have used text mining in a range of theory testing applications–from exploring individual values and cultural embeddedness, to analyzing shifts in public discourse and cultural categories over time. Text mining has proven useful in enhancing both more quantitative and more qualitative text analysis approaches. As this review shows, computational text analysis has supported sociological inquiry in multiple ways: by assisting the identification of texts of interest in large datasets (e.g., with the help of dictionary methods in Spörlein and Schlueter 2021 or topic models in Törnberg and Törnberg 2016a), facilitating the understanding of agency through the identification of social agents and their actions in text (e.g., using semantic analysis in Franzosi et al., 2012), tracking complex concepts by evaluating word occurrences in specific contexts (e.g., evaluating ambiguity of meaning as in McMahan and Evans (2018) or capturing cultural categories as in Kozlowski et al., 2019), discovering the underlying topic structure in large textual datasets (e.g., frame identification in DiMaggio et al., 2013), or expanding complex patterns of manual coding onto a large number of texts (e.g., automatic text coding in Macanovic and Przepiorka 2022). These tasks were performed on a variety of large textual sources, including vast collections of short social media texts from social media platforms and large collections of digitized news, political texts, and books. Increased accessibility and availability of advanced computational methods discussed in Section 4 and the Appendix promises an even wider adoption of text mining methods in sociological applications.

This review reveals several regularities in sociological work using text mining in theory testing. First, despite the existence of systematic resources on text mining for social science applications (e.g., Grimmer et al., 2022; Ignatow and Mihalcea 2017), most studies still retain a strong methodological focus. This reflects a lack of broad consensus on the place of text mining in the sociological toolbox. Yet, rich methodological and ethical discussions outlined in Section 5 have helped formulate the limits of and best practices in using computational text mining methods in sociological exploration. Second, despite the challenges arising from the character of (large) textual data and the inductive nature of some text mining methods, work in the field shows how information can be meaningfully extracted from text and combined with other variables of interest for use in both inductive and deductive theory-based research. Finally, while Section 5 indicates that the consensus on all these questions seems to be forming, extensive discussions remain essential for the broader integration of new text analysis methods into the field at large.

On the one hand, new developments in data and methods are bound to keep expanding the horizon of sociological research. Training students to use programming languages and tools that can tackle the abundance of textual data will help advance sociological exploration of text and customize methods to suit the needs of our field. Once acquired, these skills can help researchers extend their analyses to other types of complex data that can be processed using similar methods–such as image (Torres and Cantú 2022), video, location, and audio data (Lazer et al., 2020). Better integration of big data with conventional data collection infrastructures (such as survey or interview data) can further help adapt new sources to researchers' needs. Furthermore, the nature of computational social science has the potential to stimulate (even) more interdisciplinary collaborations. More broadly, powerful computational tools do not only assist us in data preparation and analysis, but can also support our quest for relevant scientific puzzles (Evans 2020).

On the other hand, the methodological and ethical challenges will likely become even more salient as data generated online become more and more suited to particular (private) platform goals (Lazer et al., 2020). With the advancement of new web technologies (e.g.,

[16] Some social media platforms have been developing clear guidelines on the possible uses of their data. For example, Twitter's Privacy Policy outlines that, by posting tweets to the platform, users direct Twitter to "disclose the information as broadly as possible, including through our (Twitter's) APIs, and directing those accessing the information through our APIs to do the same" (Twitter n.d.). This includes use by researchers, who can easily access such Twitter data (Twitter Developer Platform n.d.).

Web 3.0) the data are bound to become even more complex, the sampling procedures more difficult to track, and the integration with external data sources more challenging. The increasing complexity of cutting-edge computational text analysis methods can deem them more difficult to understand and implement for social scientists–and the rapid development of new tools will require researchers to constantly acquire new skills at a much faster pace than before. These challenges become even more salient in the light of somewhat slow initiatives within universities to strengthen interdisciplinary collaborations, develop new data and research infrastructures, and design ethical and regulatory frameworks that follow developments in computational social science (Lazer et al., 2020).

The goal of this review was to inspire sociologists from different fields and research traditions to consider large textual data and various text mining tools and find them a place in their toolboxes. How can our field, with the existing methodological practices so established, best embrace these new developments? It is useful to consider suggestions that inductive approaches can spark sociological imagination (Evans and Aceves, 2016; Goldberg 2015) and that the wealth of new data sources can allow us to overcome the limitations of variable-based sociology and tedious manual text analysis (DiMaggio et al., 2013; Wagner-Pacifici et al., 2015). Text mining methods can augment our analytical capabilities and help overcome many problems arising in manual text analysis (Grimmer et al., 2022). Yet, more data and more method complexity do not guarantee success on their own. This is why it is important to keep in mind the importance of theory testing, precise measurement, and causal inference in scientific discovery (Grimmer et al., 2021). In conclusion, as Grimmer et al. (2021) note, as researchers, we need to be careful not to confuse data abundance with progress and remain aware of the lessons learned from working with scarce data in social sciences.

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

None.

Acknowledgements

I thank Wojtek Przepiorka and Vincent Buskens for their helpful comments on several versions of this paper. I thank two anonymous reviewers for their very insightful comments and suggestions.

Appendix: Text Mining for Social Science – The State and the Future of Computational Text Analysis in Sociology

This appendix presents a number of accessible tools that can assist researchers in using computational methods for text analysis of large datasets. Tables A.1. to A.5. mirror the structure of Section 3 in the main paper. That is, they list tools that can assist analyses using: dictionary approaches, semantic and network analysis methods, language representations, unsupervised text clustering, and supervised text classification machine learning methods. Additionally, Table A.6. includes tools for obtaining (textual) data from online platforms and web pages. These tools are written in widely-used programming languages (mostly R and Python), but also include some stand-alone programs. For each of the listed tools, readers can find the programming language it is implemented in, tasks it can be used for, and a link to an online reference containing further details. Given the dynamic development of new tools, this list of tools is in no way extensive; it should, however, include some of the most widely utilized and most user-friendly tools for large-scale text mining applications.

Table A.1
Tools for dictionary method implementations

Tool	Language	Purpose	Link
CoreNLP	Java	Evaluate texts using a sentiment dictionary	https://stanfordnlp.github.io/CoreNLP/
Umigon	Java/Web application	Evaluate texts using a sentiment dictionary	https://github.com/seinecle/Umigon; https://nocodefunctions.com/
SentiStrength	Java/Web application	Evaluate texts using a sentiment dictionary	http://sentistrength.wlv.ac.uk/
nltk	Python	Evaluate texts using a range of pre-existing dictionaries and create custom dictionaries of any kind	https://www.nltk.org/book/ch02.html
textblob	Python	Evaluate texts using a range of pre-existing dictionaries	https://textblob.readthedocs.io/en/dev/
VADER	Python	Evaluate texts using the vader sentiment dictionary	https://github.com/cjhutto/vaderSentiment
tidytext	R	Evaluate texts using a range of pre-existing dictionaries and create custom dictionaries of any kind	https://www.tidytextmining.com/sentiment.html
quanteda	R	Evaluate texts using a range of pre-existing dictionaries and create custom dictionaries of any kind	https://quanteda.io/
quanteda.dictionaries	R	Evaluate texts using 10 existing dictionaries; replicate the analysis approach used by the LIWC software	https://github.com/kbenoit/quanteda.dictionaries
tm	R	Select words that match a vector of user-defined words	

(*continued on next page*)

Table A.1 (*continued*)

Tool	Language	Purpose	Link
			https://cran.r-project.org/web/packages/tm/
corpustools	R	Search words in text that match words in a dictionary; match tokens (e.g., words) with corresponding dictionary categories	https://cran.r-project.org/web/packages/corpustools/
sentimentr	R	Evaluate texts using a range of pre-existing dictionaries	https://cran.r-project.org/web/packages/sentimentr/
sentometrics	R	Evaluate texts using a range of pre-existing dictionaries and create custom dictionaries	https://doi.org/10.18637/jss.v099.i02
SentimentAnalysis	R	Evaluate texts using a range of pre-existing dictionaries and create custom dictionaries of any kind	https://cran.r-project.org/web/packages/SentimentAnalysis/
syuzhet	R	Evaluate texts using a range of pre-existing dictionaries and create custom dictionaries of any kind	https://cran.r-project.org/web/packages/syuzhet/
The General Inquirer	Standalone Java software	Evaluate texts using a range of pre-existing dictionaries	http://www.mariapinto.es/ciberabstracts/Articulos/Inquirer.htm
WordStat	Standalone software	Evaluate texts using a range of pre-existing dictionaries and create custom dictionaries of any kind	https://provalisresearch.com/products/content-analysis-software/
LIWC	Standalone software	Evaluate texts using the LIWC dictionary	https://www.liwc.app/
WordNet	Standalone software/ Web application	A lexical database of the English language that can be used in combination with other software solutions (e.g., nltk in Python)	https://wordnet.princeton.edu/

Table A.2
Tools for semantic and network analysis

Tool	Language	Purpose	Link
CoreNLP	Java	POS tagging; semantic tagging	https://stanfordnlp.github.io/CoreNLP/
OpenNLP	Java	POS tagging; semantic tagging	https://opennlp.apache.org/
HanLP	Python	POS tagging; semantic tagging (104 languages)	https://github.com/hankcs/HanLP
coreNLP	R	POS tagging; semantic tagging	https://cran.r-project.org/web/packages/coreNLP/
openNLP	R	POS tagging; semantic tagging	https://cran.r-project.org/web/packages/openNLP/
VARD	Standalone Java software	POS tagging; semantic tagging	https://ucrel.lancs.ac.uk/vard/about/
PC-ACE	Standalone Software	Semantic analysis; network analysis	https://pc-ace.com/
WordNet	Standalone software/Web application	A lexical database of the English language that can be used in combination with other software solutions (e.g., nltk in Python)	https://wordnet.princeton.edu/
CLAWS	Web application	POS tagging	https://ucrel.lancs.ac.uk/claws/
Lydia/ TextMap	Web Application	POS tagging; semantic tagging	https://www3.cs.stonybrook.edu/~skiena/lydia/

Table A.3
Tools for language representations

Tool	Language	Purpose	Link
scikit-learn	Python	Build a document-term matrix	https://scikit-learn.org/stable/
textmining	Python	Build a document-term matrix	https://www.christianpeccei.com/textmining/
NumPy	Python	Build a document-term matrix	https://numpy.org/
tm	R	Build a document-term matrix	https://cran.r-project.org/web/packages/tm/
quanteda	R	Build a document-term matrix	https://quanteda.io/
tidytext	R	Build a document-term matrix	https://www.tidytextmining.com
corpustools	R	Build a document-term matrix	https://cran.r-project.org/web/packages/corpustools/
textmineR	R	Build a document-term matrix	https://www.rtextminer.com/
textTinyR	R	Build a document-term matrix	https://cran.rstudio.com/web/packages/textTinyR/
udpipe	R	Build a document-term matrix	https://cran.r-project.org/web/packages/udpipe/
textrecipes	R	Build a document-term matrix	https://cran.r-project.org/web/packages/textrecipes/
superml	R	Build a document-term matrix	https://cran.r-project.org/web/packages/superml/
NLTK	Python	Build and use n-gram language models (e.g., bigram models)	https://www.nltk.org/
kgrams	R	Build and use n-gram language models (e.g., bigram models)	https://cran.r-project.org/web/packages/kgrams/
doc2vec	R	Build and use word embedding models (doc2vec)	https://cran.r-project.org/web/packages/doc2vec/
fastText	C++/Python	Build and use word embedding models (fastText)	https://fasttext.cc/
fasttext	R	Build and use word embedding models (fastText)	https://cran.r-project.org/web/packages/fastText/

(*continued on next page*)

Table A.3 (*continued*)

Tool	Language	Purpose	Link
text2vec	R	Build and use word embedding models (GloVe)	https://cran.r-project.org/web/packages/text2vec/
h2o	Java/R/Python	Build and use word embedding models (word2vec)	https://docs.h2o.ai/
word2vec	R	Build and use word embedding models (word2vec)	https://cran.r-project.org/web/packages/word2vec/
TensorFlow	Python	Build and use word embedding models (word2vec, BERT)	https://www.tensorflow.org/
GENSIM	Python	Build and use word embedding models (word2vec, doc2vec)	https://radimrehurek.com/gensim/
Hugging Face Transformers	Python	Build and use transformer models	https://huggingface.co/
Keras	Python	Build and use deep learning models	https://keras.io/
Keras	R	Build and use deep learning models	https://keras.rstudio.com/

Table A.4
Tools for unsupervised text clustering

Tool	Language	Purpose	Link
stats	R	Unsupervised text clustering (various)	https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00
scikit-learn	Python	Unsupervised text clustering (various)	https://scikit-learn.org/stable/modules/clustering.html
kernlab	R	Unsupervised text clustering (various)	https://cran.r-project.org/web/packages/kernlab/
SuperLearner	R	Unsupervised text clustering (various)	https://cran.r-project.org/web/packages/SuperLearner/
superml	R	Unsupervised text clustering (e.g., k-means)	https://cran.r-project.org/web/packages/superml/
mlr	R	Implement unsupervised machine learning algorithms (10 models)	https://mlr.mlr-org.com/; https://mlr3.mlr-org.com
Keras	R	Implement deep learning approaches	https://keras.rstudio.com/
Keras	Python	Implement deep learning approaches	https://keras.io/
h2o	Java/R/Python	Implement deep learning approaches	https://docs.h2o.ai/
deepnet	R	Implement deep learning approaches	https://cran.r-project.org/web/packages/deepnet/
PyTorch	Python	Implement deep learning approaches	https://pytorch.org/
Hugging Face Transformers	Python	Implement deep learning approaches	https://huggingface.co/
wordfish	R	Fit the wordfish model	http://www.wordfish.org/
stm	R	Fit topic models (STM)	https://cran.r-project.org/web/packages/stm/
lda	R	Fit topic models (LDA, sLDA, corrLDA)	https://cran.r-project.org/web/packages/lda/
textmineR	R	Fit topic models (LDA, CTM, LSA)	https://www.rtextminer.com/
topicmodels	R	Fit topic models (LDA, CTM)	https://cran.r-project.org/web/packages/topicmodels/
GENSIM	Python	Fit topic models (LDA, author-topic model, hierarchical dirichlet model)	https://radimrehurek.com/gensim/apiref.html
CorEx	Python	Fit topic models (hierarchical topic modelling)	https://github.com/gregversteeg/corex_topic
BTM	R	Fit topic models (BTM)	https://cran.r-project.org/web/packages/BTM/
seededlda	R	Fit semi-supervised topic models (seeded LDA)	https://cran.r-project.org/web/packages/seededlda/
keyATM	R	Fit semi-supervised topic models (keyword-assisted topic models)	https://keyatm.github.io/keyATM/

Table A.5
Tools for supervised text classification

Tool	Language	Purpose	Link
randomForest	R	Implement supervised machine learning algorithms (Random Forest); including multilabel classification	https://cran.r-project.org/web/packages/randomForest/
randomForestSRC	R	Implement supervised machine learning algorithms (Random Forest); including multilabel classification	https://cran.r-project.org/web/packages/randomForestSRC/
LiblineaR	R	Implement supervised machine learning algorithms (Logistic regression, linear SVM); including multi-class classification	https://cran.r-project.org/web/packages/LiblineaR/
LIBLINEAR	Python/java/MATLAB	Implement supervised machine learning algorithms (Logistic regression, linear SVM); including multi-class classification	https://www.csie.ntu.edu.tw/~cjlin/liblinear/
kernlab	R	Implement supervised machine learning classification algorithms (e.g., SVM)	https://cran.r-project.org/web/packages/kernlab/
SuperLearner	R	Implement supervised machine learning classification algorithms (e.g., Random Forest, SVM)	https://cran.r-project.org/web/packages/SuperLearner/
party	R	Implement supervised machine learning classification algorithms (e.g., Random Forest); including multilabel classification	https://cran.r-project.org/web/packages/party/

(*continued on next page*)

Table A.5 (*continued*)

Tool	Language	Purpose	Link
scikit-learn	Python	Implement supervised machine learning classification algorithms (e.g., Naïve Bayes, SVM, Random Forest)	https://scikit-learn.org/stable/supervised_learning.html#supervised-learning
e1071	R	Implement supervised machine learning classification algorithms (e.g., Naïve Bayes, SVM)	https://cran.r-project.org/web/packages/e1071/
NLTK	Python	Implement supervised machine learning classification algorithms (e.g., Naïve Bayes, SVM)	https://www.nltk.org/
superml	R	Implement supervised machine learning classification algorithms (e.g., Naïve Bayes)	https://cran.r-project.org/web/packages/superml/
mlr	R	Implement supervised machine learning classification algorithms (84 models); including multi-class and multi-label classification	https://mlr.mlr-org.com/; https://mlr3.mlr-org.com
caret	R	Implement supervised machine learning classification algorithms (238 models)	https://cran.r-project.org/web/packages/caret/
Stanford Classifier	Java	Implement supervised machine learning algorithm (maximum entropy)	https://nlp.stanford.edu/wiki/Software/Classifier
Keras	R	Implement deep learning approaches	https://keras.rstudio.com/
Keras	Python	Implement deep learning approaches	https://keras.io/
h2o	Java/R/Python	Implement deep learning approaches	https://docs.h2o.ai/
deepnet	R	Implement deep learning approaches	https://cran.r-project.org/web/packages/deepnet/
PyTorch	Python	Implement deep learning approaches	https://pytorch.org/
Hugging Face Transformers	Python	Implement deep learning approaches	https://huggingface.co/

Table A.6
Tools for large textual data collection from online sources

Tool	Language	Purpose	Link
twitteR	R	Obtain twitter data from twitter API	https://cran.r-project.org/web/packages/twitteR/twitteR.pdf
tweepy	Python	Obtain twitter data from twitter API	https://www.tweepy.org/
URS	Python	Obtain reddit data from reddit API	https://github.com/JosephLai241/URS
PRAW	Python	Obtain reddit data from reddit API	https://praw.readthedocs.io/en/stable/
socalreaper	Python	Scrape social media sites (e.g., Facebook, Twitter, Reddit, Youtube, Pinterest, Tumblr)	https://github.com/ScriptSmith/socialreaper
httr	R	Send requests to websites to download data	https://cran.r-project.org/web/packages/httr/httr.pdf
Requests	Python	Send requests to websites to download data	https://requests.readthedocs.io/en/latest/
Rcrawler	R	Crawl websites and extract structured data	https://github.com/salimk/Rcrawler
rvest	R	Crawl websites and extract structured data	https://rvest.tidyverse.org/
pyspider	Python	Crawl websites and extract structured data	http://docs.pyspider.org/en/latest/
Scrapy	Python	Crawl websites and extract structured data	https://scrapy.org/
Heritrix	Java	Crawl websites and extract structured data	https://github.com/internetarchive/heritrix3/wiki
Jaunt	Java	Crawl websites and extract structured data	https://jaunt-api.com/
Apify ADK v2	JavaScript	Crawl websites and extract structured data	https://sdk.apify.com/
MechanicalSoup	Python	Interact with websites	https://mechanicalsoup.readthedocs.io/en/stable/
XML	R	Parse content of scraped websites	https://cran.r-project.org/web/packages/XML/index.html
Beautiful Soup	Python	Parse content of scraped websites	https://www.crummy.com/software/BeautifulSoup/
twitteR	R	Obtain twitter data from twitter API	https://cran.r-project.org/web/packages/twitteR/twitteR.pdf
tweepy	Python	Obtain twitter data from twitter API	https://www.tweepy.org/

References

Aggarwal, Charu C., 2018. Machine Learning for Text. Springer International Publishing, Cham, Switzerland.
Aggarwal, Charu C., Zhai, ChengXiang, 2012a. A survey of text clustering algorithms. In: Aggarwal, C.C., Zhai, C. (Eds.), Mining Text Data. Springer US, Boston, MA, pp. 77–128.
Aggarwal, Charu C., Zhai, ChengXiang, 2012b. Mining Text Data. Springer US, Boston, MA.
Almquist, Zack W., Bagozzi, Benjamin E., 2019. Using radical environmentalist texts to uncover network structure and network features. Socio. Methods Res. 48 (4), 905–960.
Althbiti, Ashrf, Ma, Xiaogang, 2022. Machine learning. In: Encyclopedia of Big Data. Springer International Publishing, Cham, Switzerland, pp. 633–637.
Baden, Christian, Kligler-Vilenchik, Neta, Yarchi, Moran, 2020. Hybrid content analysis: toward a strategy for the theory-driven, computer-assisted classification of large text corpora. Commun. Methods Meas. 14 (3), 165–183.
Bail, Christopher A., 2014. The cultural environment: measuring culture with big data. Theor. Soc. 43 (3–4), 465–482.

Bail, Christopher A., Brown, Taylor W., Mann, Marcus, 2017. Channeling hearts and minds: advocacy organizations, cognitive-emotional currents, and public conversation. Am. Socio. Rev. 82 (6), 1188–1213.
Bastin, Gilles, Bouchet-Valat, Milan, 2014. Media corpora, text mining, and the sociological imagination - a free software text mining approach to the framing of Julian assange by three news agencies using R.TeMiS. Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique 122 (1), 5–25.
Biernacki, Richard, 2012. Reinventing Evidence in Social Inquiry: Decoding Facts and Variables. Palgrave Macmillan, New York, NY.
Blei, David M., 2012. Probabilistic topic models. Commun. ACM 55 (4), 77–84.
Bonikowski, Bart, Gidron, Noam, 2016. The populist style in American politics: presidential campaign discourse, 1952–1996. Soc. Forces 94 (4), 1593–1621.
Boutyline, Andrei, Arseniev-Koehler, Alina, Cornell, Devin J., 2020. School, Studying, and Smarts: Gender Stereotypes and Education across 80 Years of American Print Media, 1930-2009. University of Michigan, MI. Unpublished manuscript. https://osf.io/preprints/socarxiv/bukdg.
Boyd, Ryan L., Schwartz, Andrew H., 2021. Natural Language analysis and the psychology of verbal behavior: the past, present, and future states of the field. J. Lang. Soc. Psychol. 40 (1), 21–41.
Breiger, Ronald L., Wagner-Pacifici, Robin, Mohr, John W., 2018. Capturing distinctions while mining text data: toward low-tech formalization for text analysis. Poetics 68, 104–119.
Carley, Kathleen M., 1999. Extracting team mental models through textual analysis. J. Organ. Behav. 18 (S1), 533–558.
Carley, Kathleen, Palmquist, Michael, 1992. Extracting, representing, and analyzing mental models. Soc. Forces 70 (3), 601–636.
Chang, Jonathan P., Chiam, Caleb, Fu, Liye, Wang, Andrew, Zhang, Justine, Danescu-Niculescu-Mizil, Cristian, 2020. ConvoKit: a Toolkit for the analysis of conversations. In: Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Association for Computational Linguistics, pp. 57–60.
Chatsiou, Kakia, Mikhaylov, Slava Jankin, 2020. Deep learning for political science. In: Cuirini, L., Franzese, R. (Eds.), The SAGE Handbook of Research Methods in Political Science and International Relations. Sage, Thousand Oaks, CA, 1053–78.
Crano, William D., Brewer, Marilynn B., Lac, Andrew, 2014. Principles and Methods of Social Research, third ed. Routledge, New York, NY.
Davidson, Thomas, Bhattacharya, Debasmita, 2020. Examining racial bias in an online abuse corpus with structural topic modeling.". In: Proceedings of 2020 ICWSM Data Challenge.
Davidson, Thomas, Warmsley, Dana, Macy, Michael, Weber, Ingmar, 2017. Automated hate speech detection and the problem of offensive language. In: Proceedings of the International AAAI Conference on Web and Social Media, 11, pp. 512–515.
Denny, Matthew J., Spirling, Arthur, 2018. Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it. Polit. Anal. 26 (2), 168–189.
Dex, Shirely, 1995. The reliability of recall data: a literature review. Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique 49 (1), 58–89.
DiMaggio, Paul, 2015. Adapting computational text analysis to social science (and vice versa). Big Data Soc. 2, 1–5.
DiMaggio, Paul, Nag, Manish, Blei, David, 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: application to newspaper coverage of U.S. Government arts funding. Poetics 41 (6), 570–606.
Domingos, Pedro, 2012. A few useful things to know about machine learning. Commun. ACM 55 (10), 78–87.
Edelmann, Achim, Mohr, John W., 2018. Formal studies of culture: issues, challenges, and current trends. Poetics 68, 1–9.
Edelmann, Achim, Wolff, Tom, Montagne, Danielle, Bail, Christopher A., 2020. Computational social science and sociology. Annu. Rev. Sociol. 46 (1), 61–81.
Eshima, Shusei, Imai, Kosuke, Sasaki, Tomoya, 2020. Keyword Assisted Topic Models. Department of Government, Harvard University, United States. Unpublished manuscript.
Evans, James, 2020. Social computing unhinged. J. Soc. Comput. 1 (1), 1–13.
Evans, James A., Aceves, Pedro, 2016. Machine translation: mining text for social theory. Annu. Rev. Sociol. 42 (1), 21–50.
Franzosi, Roberto, 1989. From words to numbers: a generalized and linguistics-based coding procedure for collecting textual data. Socio. Methodol. 19, 263–298.
Franzosi, Roberto, De Fazio, Gianluca, Vicari, Stefania, 2012. Ways of measuring agency: an application of quantitative narrative analysis to lynchings in Georgia (1875–1930). Socio. Methodol. 42, 1–42.
Gallagher, Ryan J., Frank, Morgan R., Mitchell, Lewis, Schwartz, Aaron J., Reagan, Andrew J., Danforth, Christopher M., Dodds, Peter Sheridan, 2021. Generalized word shift graphs: a method for visualizing and explaining pairwise comparisons between texts. EPJ Data Sci. 10 (1).
Gentzkow, Matthew, Kelly, Bryan, Taddy, Matt, 2019. Text as data. J. Econ. Lit. 57 (3), 535–574.
Ghose, Anindya, Ipeirotis, Panagiotis G., Sundararajan, Arun, 2009. The dimensions of reputation in electronic markets. In: NYU Center for Digital Economy Research Working Paper No. CeDER-06-02.
Gibbs, Graham R., 2014. Using software in qualitative analysis. In: Flick, U. (Ed.), The SAGE Handbook of Qualitative Data Analysis. Sage Publications Ltd, London, UK, pp. 277–294.
Goldberg, Amir, 2015. In defense of forensic social science. Big Data Soc. 2 (2), 1–3.
Goldberg, Amir, Srivastava, Sameer B., Govind Manian, V., Monroe, William, Potts, Christopher, 2016. Fitting in or standing out? The tradeoffs of structural and cultural embeddedness. Am. Socio. Rev. 81 (6), 1190–1222.
Goldenstein, Jan, Poschmann, Philipp, 2019a. A quest for transparent and reproducible text-mining methodologies in computational social science. Socio. Methodol. 49 (1), 144–151.
Goldenstein, Jan, Poschmann, Philipp, 2019b. Analyzing meaning in big data: performing a map analysis using grammatical parsing and topic modeling. Socio. Methodol. 49 (1), 83–131.
Golder, Scott A., Macy, Michael W., 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. Science 333 (6051), 1878–1881.
Golder, Scott A., Macy, Michael W., 2014. Digital footprints: opportunities and challenges for online social research. Annu. Rev. Sociol. 40 (1), 129–152.
Grave, Edouard, Bojanowski, Piotr, Gupta, Prakhar, Armand, Joulin, Mikolov, Tomas, 2018. Learning word vectors for 157 languages. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation. LREC 2018, pp. 3483–3487.
Grimmer, Justin, Stewart, Brandon M., 2013. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. Polit. Anal. 21 (3), 267–297.
Grimmer, Justin, Roberts, Margaret E., Stewart, Brandon M., 2021. Machine learning for social science: an agnostic approach. Annu. Rev. Polit. Sci. 24 (1), 395–419.
Grimmer, Justin, Roberts, Margaret E., Stewart, Brandon M., 2022. Text as Data: A New Framework for Machine Learning and the Social Sciences. Princeton University Press, Princeton, NJ and Oxford, UK.
Hartmann, Jochen, Huppertz, Juliana, Schamp, Christina, Heitmann, Mark, 2019. Comparing automated text classification methods. Int. J. Res. Market. 36 (1), 20–38.
Hoover, Joe, Portillo-Wightman, Gwenyth, Yeh, Leigh, Havaldar, Shreya, , Aida Mostafazadeh Davani, Lin, Ying, Kennedy, Brendan, et al., 2020. Moral foundations twitter corpus: a collection of 35k tweets annotated for moral sentiment. Soc. Psychol. Personal. Sci. 11 (8), 1057–1071.
Hopkins, Daniel J., King, Gary, 2010. A method of automated nonparametric content analysis for social science. Am. J. Polit. Sci. 54 (1), 229–247.
Hotho, Andreas, Nürnberger, Andreas, Paaß, Gerhard, 2005. A brief survey of text mining. Ldv Forum 20 (1), 19–62.
Housley, William, Procter, Rob, Edwards, Adam, Burnap, Peter, Williams, Matthew, Sloan, Luke, Rana, Omer, Morgan, Jeffrey, Voss, Alex, Greenhill, Anita, 2014. Big and broad social data and the sociological imagination: a collaborative response. Big Data Soc. 1 (2), 205395171454513.
Ignatow, Gabe, 2016. Theoretical foundations for digital text analysis. J. Theor. Soc. Behav. 46 (1), 104–120.
Ignatow, Gabe, Mihalcea, Rada, 2017. Text Mining: A Guidebook for the Social Sciences. SAGE Publications, Thousand Oaks, CA.
Jaidka, Kokil, Giorgi, Salvatore, Schwartz, H. Andrew, Kern, Margaret L., Ungar, Lyle H., Eichstaedt, Johannes C., 2020. Estimating geographic subjective well-being from twitter: a comparison of dictionary and data-driven language methods. Proc. Natl. Acad. Sci. USA 117 (19), 10165–10171.
Jurafsky, Daniel, Martin, James H., 2009. Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Pearson/Prentice Hall, Upper Saddle River, NJ.

Kacewicz, Ewa, Pennebaker, James W., Davis, Matthew, Jeon, Moongee, Graesser, Arthur C., 2014. Pronoun use reflects standings in social hierarchies. J. Lang. Soc. Psychol. 33 (2), 125–143.
King, Gary, Lam, Patrick, Roberts, Margaret E., 2017. Computer-assisted keyword and document set discovery from unstructured text. Am. J. Polit. Sci. 61 (4), 971–988.
Kozlowski, Austin C., Taddy, Matt, Evans, James A., 2019. The geometry of culture: analyzing the meanings of class through word embeddings. Am. Socio. Rev. 84 (5), 905–949.
Krippendorff, Klaus, 2004. Content Analysis: an Introduction to its Methodology, 2nd ed. SAGE Publications, Thousand Oaks, CA.
Kross, Ethan, Verduyn, Philippe, Boyer, Margaret, Drake, Brittany, Gainsburg, Izzy, Vickers, Brian, Ybarra, Oscar, Jonides, John, 2019. Does counting emotion words on online social networks provide a window into people's subjective experience of emotion? A case study on Facebook. Emotion 19 (1), 97–107.
Kuckartz, Udo, 2014. Qualitative Text Analysis: A Guide to Methods, Practice & Using Software. SAGE Publications, London, UK.
Lazer, David, Radford, Jason, 2017. Data ex machina: introduction to big data. Annu. Rev. Sociol. 43 (1), 19–39.
Lazer, David, Pentland, Alex, Adamic, Lada, Aral, Sinan, Barabási, Albert-László, Brewer, Devon, Christakis, Nicholas, et al., 2009. Computational social science. Science 323 (5915), 721–723.
Lazer, David, Pentland, Alex, Watts, Duncan J., Aral, Sinan, Athey, Susan, Contractor, Noshir, Freelon, Deen, et al., 2020. Computational social science: obstacles and opportunities. Science 369 (6507), 1060–1062.
Lee, Monica, Martin, John Levi, 2015. Coding, counting and cultural cartography. Am. J. Cult. Sociol. 3 (1), 1–33.
Liddy, Elizabeth D., 2001. Natural Language Processing. Encyclopedia of Library and Information Science.
Loughran, Tim, McDonald, Bill, 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. J. Finance 66 (1), 35–65.
Macanovic, Ana, Przepiorka, Wojtek, 2021. The Moral Embeddedness of Cryptomarkets: Text Mining Feedbacks on Economic Exchanges in the Darknet. Department of Sociology/ICS, Utrecht University, The Netherlands. Unpublished manuscript.
Macanovic, Ana, Przepiorka, Wojtek, 2022. A Systematic Evaluation of Text Mining Methods for Short Texts: Mapping Individuals' Internal States from Feedback Texts and Tweets. Department of Sociology/ICS, Utrecht University, The Netherlands. Unpublished manuscript: https://osf.io/preprints/socarxiv/cerzs/.
Manovich, Lev, 2011. Trending: the promises and the challenges of big social data. In: Gold, M.K. (Ed.), Debates in the Digital Humanities. University of Minnesota Press, Minneapolis, MN, 460–75.
McFarland, Daniel A., Lewis, Kevin, Goldberg, Amir, 2016. Sociology in the era of big data: the ascent of forensic social science. Am. Sociol. 47 (1), 12–35.
McMahan, Peter, Evans, James, 2018. Ambiguity and engagement. Am. J. Sociol. 124 (3), 860–912.
Mikolov, Tomas, Wen-tau, Yih, Zweig, Geoffrey, 2013a. Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 746–751.
Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg S., Dean, Jeffrey, 2013b. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, pp. 3111–3119.
Minaee, Shervin, Kalchbrenner, Nal, Cambria, Erik, Nikzad, Narjes, Chenaghlu, Meysam, Gao, Jianfeng, 2021. Deep learning-based text classification: a comprehensive review. ACM Comput. Surv. 54 (3), 1–40.
Mohr, John W., 1998. Measuring meaning structures. Annu. Rev. Sociol. 24 (1), 345–370.
Mohr, John W., Bogdanov, Petko, 2013. Introduction—topic models: what they are and why they matter. Poetics 41 (6), 545–569.
Mohr, John W., Wagner-Pacifici, Robin, Breiger, Ronald L., 2015. Toward a computational hermeneutics. Big Data Soc. 2 (2).
Mohr, John W., Wagner-Pacifici, Robin, Breiger, Ronald L., Bogdanov, Petko, 2013. Graphing the grammar of motives in national security strategies: cultural interpretation, automated text analysis and the drama of global politics. Poetics 41 (6), 670–700.
Molina, Mario, Garip, Filiz, 2019. Machine learning for sociology. Annu. Rev. Sociol. 45 (1), 27–45.
Muller, Michael, Guha, Shion, Baumer, Eric P.S., Mimno, David, Shami, Sadat N., 2016. Machine learning and grounded theory method. In: *Proceedings of the 19th International Conference on Supporting Group Work*. ACM, New York, NY, USA, pp. 3–8.
Naldi, Maurizio, 2019. A Review of Sentiment Computation Methods with R Packages. University of Rome Tor Vergata, Rome, Italy. Unpublished manuscript.
Nardulli, Peter F., Althaus, Scott L., Hayes, Matthew, 2015. A progressive supervised-learning approach to generating rich civil strife data. Socio. Methodol. 45 (1), 148–183.
Nelson, Laura K., 2017. Computational grounded theory: a methodological framework. Socio. Methods Res. 49 (1), 3–42.
Nelson, Laura K., 2019. To measure meaning in big data, don't give me a map, give me transparency and reproducibility. Socio. Methodol. 49 (1), 139–143.
Nelson, Laura K., 2021. Cycles of conflict, a century of continuity: the impact of persistent place-based political logics on social movement strategy. Am. J. Sociol. 127 (1), 1–59.
Nelson, Laura K., Burk, Derek, Knudsen, Marcel, McCall, Leslie, 2018. The future of coding. Socio. Methods Res. 50 (1), 202–237.
Neuhaus, Fabian, Webmoor, Timothy, 2012. Agile ethics for massified research and visualization. Inf. Commun. Soc. 15 (1), 43–65.
Nguyen, Minh Van, Lai, Viet Dac, Veyseh, Amir Pouran Ben, Nguyen, Thien Huu, 2021. Trankit: A Light-Weight Transformer-Based Toolkit for Multilingual Natural Language Processing. Department of Computer and Information Science University of Oregon. Unpublished manuscript: https://arxiv.org/pdf/2101.03289.pdf.
Olteanu, Alexandra, Castillo, Carlos, Diaz, Fernando, Kıcıman, Emre, 2019. Social data: biases, methodological pitfalls, and ethical boundaries. Frontiers in Big Data 2, 13.
Pääkkönen, Juho, Ylikoski, Petri, 2021. Humanistic interpretation and machine learning. Synthese 199 (1–2), 1461–1497.
Pan, Weike, Zhong, Erheng, Yang, Qiang, 2012. Transfer learning for text mining. In: Aggarwal, C.C., Zhai, C. (Eds.), Mining Text Data. Springer, Boston, MA, 223–57.
Pechenick, Eitan Adam, Danforth, Christopher M., Dodds, Peter Sheridan, 2015. Characterizing the Google books corpus: strong limits to inferences of socio-cultural and linguistic evolution. PLoS One 10 (10), e0137041.
Pellert, Max, Metzler, Hannah, Matzenberger, Michael, Garcia, David, 2022. Validating daily social media macroscopes of emotions. Sci. Rep. 12 (1), 11236.
Pennebaker, J.W., Boyd, R.L., Jordan, K., Blackburn, K., 2015. *The Development and Psychometric Properties of LIWC2015*. Software Manual. University of Texas at Austin, Austin, TX. https://repositories.lib.utexas.edu/handle/2152/31333.
Popping, Roel, 2000. Computer-Assisted Text Analysis. SAGE Publications, London, UK.
Popping, Roel, Roberts, Carl W., 2015. Semantic text analysis and the measurement of ideological developments within fledgling democracies. Soc. Sci. Inf. 54 (1), 23–37.
Radford, Jason, Lazer, David, 2019. Big data for sociological research. In: Ritzer, G., Murphy, W.W. (Eds.), The Wiley Blackwell Companion to Sociology, 2nd ed. John Wiley & Sons, Hoboken, NJ and Chichester, UK, pp. 417–443.
Roberts, Carl W., 1989. Other than counting words: a linguistic approach to content analysis. Soc. Forces 68 (1), 147–177.
Roberts, Margaret E., Stewart, Brandon M., Airoldi, Edoardo M., 2016. A model of text for experimentation in the social sciences. J. Am. Stat. Assoc. 111 (515), 988–1003.
Rona-Tas, Akos, Cornuéjols, Antoine, Blanchemanche, Sandrine, Duroy, Antonin, Martin, Christine, 2019. Enlisting supervised machine learning in mapping scientific uncertainty expressed in food risk analysis. Socio. Methods Res. 48 (3), 608–641.
Salganik, Matthew J., 2017. Bit by Bit: Social Research in the Digital Age. Princeton University Press, Princeton, NJ.
Schwemmer, Carsten, Wieczorek, Oliver, 2020. The methodological divide of sociology: evidence from two decades of journal publications. Sociology 54 (1), 3–21.
Shklar, Judith N., 1986. Squaring the hermeneutic circle. Soc. Res. 53 (3), 449–473.
Slapin, Jonathan B., Proksch, Sven-Oliver, 2008. A scaling model for estimating time-series party positions from texts. Am. J. Polit. Sci. 52 (3), 705–722.
Speed, John Gilmer, 1893. Do newspapers now give the news? Forum 15, 705–711.
Spörlein, Christoph, Schlueter, Elmar, 2021. Ethnic insults in YouTube comments: social contagion and selection effects during the German "refugee crisis". Eur. Socio Rev. 37 (3), 411–428.
Stoltz, Dustin S., Taylor, Marshall A., 2019. Concept Mover's distance: measuring concept engagement via word embeddings in texts. J. Comput. Soc. Sci. 2 (2), 293–313.

Stone, Philip J., Dunphy, Dexter C., Smith, Marshall S., Ogilvie, Daniel M., 1966. The General Inquirer: A Computer Approach to Content Analysis. MIT Press, Cambridge, Massachusetts and London, England.
Sudhahar, Saatviga, De Fazio, Gianluca, Franzosi, Roberto, Cristianini, Nello, 2015a. Network analysis of narrative content in large corpora. Nat. Lang. Eng. 21 (1), 81–112.
Sudhahar, Saatviga, Veltri, Giuseppe A., Cristianini, Nello, 2015b. Automated analysis of the US presidential elections using big data and network analysis. Big Data Soc. 2 (1), 1–28.
Tausczik, Yla R., Pennebaker, James W., 2010. The psychological meaning of words: LIWC and computerized text analysis methods. J. Lang. Soc. Psychol. 29 (1), 24–54.
Tay, Louis, Woo, Sang Eun, Hickman, Louis, Rachel, M., Saef, 2020. Psychometric and validity issues in machine learning approaches to personality assessment: a focus on social media text mining. Eur. J. Pers. 34 (5), 826–844.
Taylor, Marshall, Stoltz, Dustin, 2020. Concept class Analysis: a method for identifying cultural schemas in texts. Sociol. Sci. 7, 544–569.
Törnberg, Anton, Törnberg, Petter, 2016a. Combining CDA and topic modeling: analyzing discursive connections between islamophobia and anti-feminism on an online forum. Discourse Soc. 27 (4), 401–422.
Törnberg, Anton, Törnberg, Petter, 2016b. Muslims in social media discourse: combining topic modeling and critical discourse analysis. Discourse, Context & Media 13, 132–142.
Torres, Michelle, Cantú, Francisco, 2022. Learning to see: convolutional neural networks for the analysis of social science data. Polit. Anal. 30 (1), 113–131.
Twitter. n.d.. Twitter privacy policy. Retrieved. https://twitter.com/en/privacy. (Accessed 23 March 2022).
Twitter Developer. Platform. n.d. "academic research access. Retrieved. https://developer.twitter.com/en/products/twitter-api/academic-research. (Accessed 23 March 2022).
Utsumi, Akira, 2020. Exploring what is encoded in distributional word vectors: a neurobiologically motivated analysis. Cognit. Sci. 44 (6).
Uysal, Alper Kursat, Gunal, Serkan, 2014. The impact of preprocessing on text classification. Inf. Process. Manag. 50 (1), 104–112.
van Atteveldt, Wouter, Kleinnijenhuis, Jan, Ruigrok, Nel, 2008. Parsing, semantic networks, and political authority using syntactic analysis to extract semantic relations from Dutch newspaper articles. Polit. Anal. 16 (4), 428–446.
van de Rijt, Arnout, Shor, Eran, Ward, Charles, Skiena, Steven, 2013. Only 15 minutes? The social stratification of fame in printed media. Am. Socio. Rev. 78 (2), 266–289.
van Loon, Austin, Stewart, Sheridan, Waldon, Brandon, Lakshmikanth, Shrinidhi K., Shah, Ishan, Guntuku, Sharath Chandra, Sherman, Garrick, Zou, James, Eichstaedt, Johannes, 2020. Explaining the Trump gap in social distancing using COVID discourse. In: Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020. Association for Computational Linguistics.
Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Łukasz, Polosukhin, Illia, 2018. Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30.
Wagner-Pacifici, Robin, Mohr, John W., Breiger, Ronald L., 2015. Ontologies, methodologies, and new uses of big data in the social and cultural sciences. Big Data Soc. 2 (2), 1–11.
Waseem, Zeerak, 2016. Are you a racist or Am I seeing things? Annotator influence on hate speech detection on twitter. In: Proceedings of the First Workshop on NLP and Computational Social Science, pp. 138–142.
Watanabe, Kohei, 2021. Latent semantic scaling: a semisupervised text analysis technique for new domains and languages. Commun. Methods Meas. 15 (2), 81–102.
Weber, Robert P., 1984. Computer-aided content analysis: a short primer. Qual. Sociol. 7 (1–2), 126–147.
Welbers, Kasper, van Atteveldt, Wouter, Jan, Kleinnijenhuis, 2021. Extracting semantic relations using syntax: an R package for querying and reshaping dependency trees. Comput. Commun. Res. 3 (2), 180–194.
Whittaker, Meredith, Crawford, Kate, Dobbe, Roel, Fried, Genevieve, Kaziunas, Elizabeth, Mathu, Varoon, , Sarah Mysers West, Richardson, Rashida, Schultz, Jason, Schwartz, Oscar, 2018. AI Now Report 2018.
Wiedemann, Gregor, 2016. Text Mining for Qualitative Data Analysis in the Social Sciences: A Study on Democratic Discourse in Germany. Springer VS, Leipzig, Germany.
Wolf, Thomas, Debut, Lysandre, Sanh, Victor, Chaumond, Julien, Clement, Delangue, Anthony, Moi, Cistac, Pierric, et al., 2020. HuggingFace's transformers: state-of-the-art Natural Language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp. 38–45.
Zhang, Han, Pan, Jennifer, 2019. CASM: a deep-learning approach for identifying collective action events with text and image data from social media. Socio. Methodol. 49 (1), 1–57.