The Potential of Automated Text Analytics in Social Knowledge Building

Chapter · January 2021

DOI: 10.1007/978-3-030-54936-7_3

2 authors:

Renáta Németh
Eötvös Loránd University Faculty of Social Sciences
83 PUBLICATIONS **805** CITATIONS
SEE PROFILE

Julia Koltai
Centre for Social Sciences
51 PUBLICATIONS **463** CITATIONS
SEE PROFILE

The Potential of Automated Text Analytics in Social Knowledge Building

Renáta Németh and Júlia Koltai

1 Introduction

In 2007 Savage and Burrows, in one of the most highly cited sociological paper of the decade, wrote about the coming crisis of empirical sociology. They predicted that a crisis would come if sociology, known for its innovative methodological resources, could not meet the challenges put forward by big data,[1] and thus would lose its leading role. This did not come to pass. Eight years later, the first member of the book series titled *Sociological Futures*, published by the British Sociological Association (Ryan and McKie 2015), referred to the end of the crisis in its title and saw important opportunities in big data research and also in automated text analysis.

RN's work was supported by the Higher Education Excellence Program of the Ministry of Human Capacities (ELTE-FKIP).

The work of Julia Koltai was funded by the Premium Postdoctoral Grant of the Hungarian Academy of Sciences.

[1] Although the former buzzword big data is seemingly going out of fashion, it does not have a better alternative. We use it in a general sense by referring to a vast amount of digital data which, in most cases, were created for some purpose other than our analysis (see also "found data").

R. Németh (✉)
Faculty of Social Sciences, Eötvös Loránd University, Budapest, Hungary
e-mail: nemeth.renata@tatk.elte.hu

J. Koltai
Centre for Social Sciences, Hungarian Academy of Sciences Centre of Excellence, Budapest, Hungary

Faculty of Social Sciences, Eötvös Loránd University, Budapest, Hungary
e-mail: koltai.julia@tk.mta.hu

T. Rudas, G. Péli (eds.), *Pathways Between Social Science and Computational Social Science*, Computational Social Sciences,
https://doi.org/10.1007/978-3-030-54936-7_3

In this paper, we review the possibilities that automated text analytics can provide for sociology. Automated text analytics[2] refers to the automated and computer-assisted discovery of knowledge from large text corpora. It lies at the intersection of linguistics, data science, and social science and uses many tools of natural language processing (NLP).

Our aim is to encourage sociologists to enter this field. We discuss the new methods based on the classic quantitative approach, using its concepts and terminology. We also address the question of how traditionally trained sociologists can acquire new skills. We are convinced that supporting this process is of crucial importance for the future of sociology because automated text analysis is going to be a standard tool of social research within a few years.

2 Challenges

The main challenges of applying automated text analytics in sociology are methodological. Some of them are well-known for traditionally trained sociologists, like the issue of *external validity, internal validity, and reliability*. The total survey error framework provides a conceptual structure to identify and quantify these challenges. Studies of large datasets can have the same shortcomings as surveys – that is why these questions were addressed by empirical social scientists (e.g., Hargittai 2015; Kreuter and Peng 2013). Indeed, when, according to a Twitter-based text analytical tool (hedonometer.org by the University of Vermont, Complex Systems Center), Louisiana is the least happy state in the USA, while a large-sample survey for the same period finds Louisiana to be the happiest one (Oswald and Wu 2011), we must examine the mechanisms generating our data. The explanation for this contradiction may lie in the difference between the coverage of the two samples.

Factors to be considered are external validity, coverage (who does use Twitter?), biased sample composition (those who tweet more are often over-represented in the population of tweets), and sampling procedure of Twitter's API. These factors also affect many other studies on digital texts: the digital divide is a decreasing but still existing problem; replacing texts for people as the unit of analysis may cause biases since more active people are more likely to appear in digital corpora. Finally, big data are often samples themselves, resulting from an unknown or hard-to-formalize sampling procedure. See, e.g., Common Crawl (http://commoncrawl.org), an open repository of textual web page data, widely used as a source representing language usage. However, what does it precisely represent? What is the probability of a given web page to get into the dataset? Another example is Google Ngram Viewer (https://books.google.com/ngrams). The viewer was created on the top of the largest digitized collection of books published between 1500 and 2000. The viewer gives the trend analysis of any n-words-long phrase (Ngram) using a yearly count found in the corpus. It is increasingly used to measure social and cultural

[2]Similar but not synonymous terms are text mining and computer-based content analysis.

trends by everyday users and researchers as well. However, results may be affected by many different potential representation and measurement errors, e.g., changes in the book publishing industry during the centuries or the bias caused by the various number of texts published already in digital formats in the different years.

Additionally, there are other big data-specific issues related to the quality of data, which are not present in surveys or interviews. These are, for example, the *presence of noise* (irrelevant data) or *fake data*. Another problem is the lack of demographic variables, which are routinely used for post-stratification in survey research. As most Internet data do not include these variables, researchers barely know anything about the social composition of users and, thus, about the external validity of data. For the same reason, post-stratification weighting is also not possible.

Such challenges contribute to the sociological skepticism toward big data-based, social-related findings that question the potential of this knowledge production and its contribution to the scientific discourse of sociology production (e.g., Kitchin 2014). It is revealing that, using Google Scholar search terms of "big data" and "sociology," the most highly cited paper found is a skeptical one (Boyd and Crawford 2012).

Additional difficulties are neither new nor specific to automated text analysis; they can be traced back to the age-old *quantitative-qualitative dichotomy*. One of them is the *close reading – distant reading* opposition (Moretti 2013) that attracted broad attention in the literature and cultural studies recently. According to the critiques, automated text analysis cannot produce deep insights and is overly reductionist. Automated text analysis methods cannot and do not aim to substitute human reading. They are capable of extracting information from texts but do not understand them. Hedonometer uses a bag-of-words model which is not capable of understanding the text but is appropriate to measure the average sentiment of tweets. Automated text analysis can construct models of language usage (see, e.g., topic models later), which – as an inherent specification of models – do not perfectly correspond to "reality" and can hardly detect sarcasm or latent intentions of speakers. Their importance is based on the fact that large text corpora are impossible to read and summarize by humans. The traditional quanti-quali debate can be cited again: we use surveys because we cannot conduct in-depth interviews with thousands of people, even if we know, for example, that self-reported answers do not fully correspond to reality. The consequence is not to refuse surveys (or automated text analytical tools) but to articulate epistemological issues and to incorporate the answers given for them in the process of data discovery, collection, preparation, analysis, and interpretation.

3 A New Methodological Basis of Sociology

In the present chapter, we review the methodological advances sociology can apply in the new field of digital textual data. The advances include the utilization of new data sources and new statistical models based on new research logic. For those interested in more details, we recommend Aggarwal and Zhai (2012).

There is a new step in the analysis process, which is not part of traditional research: pre-processing of texts. The other steps are not new; however, some of them have novel approaches.

3.1 New Data Sources

Before turning to new data sources, we have to mention that *traditional sociological textual data* like interviews, field notes, or open-ended survey questions can be analyzed by automated tools as well potentially providing new and inspiring insights into old questions. Traditional surveys with open-ended questions can also be analyzed with automated text analysis resulting in a hybrid approach.

Advances in NLP technologies allow the extension of the scope of open-ended questions. This hybrid technique increases the depth and the internal validity of surveys by moving them into the qualitative direction.

One of the most important of the new textual sources is social media. A significant and continuously increasing part of the population uses social media in their daily lives. We refer to social media as the use of technologies that turn communication into an interactive, two-way dialogue. Besides Twitter, Facebook, and Instagram, YouTube, blogs, and online forums also belong to this category. With the application of new technologies, it is possible to convert images, video, etc. to textual data and analyze the visual and textual content together (Aggarwal and Zhai (2012) give a good review).

Significant advantages of using social media for research are the low cost of data collection, an enormous amount of data, and rich metadata like geolocation, author info, exact time, and friends/followers. In addition, the network nature of the data is highly important as communication pathways can be traced through the links, which connect people. Typical analytical approaches are, e.g., social network analysis, sentiment analysis, and identification of influential/antisocial users. However, there are limitations to social media research. Social interactions within social media are mediated interactions, where individuals imagine an audience, and they build their self-representation accordingly (see, e.g., Marwick and boyd 2011). Also, the algorithms used by social media companies, which generate sampling procedures of the data, are not transparent. Therefore, researchers do not know if the data collected through APIs is a representative sample of all posts or just a biased portion of them. The question arises as to what extent the information extracted from social media can be identified with "reality."

Further textual sources, which represent more traditional, one-way communication, are *web pages and online editorial media*, which can contain multimedia contents as well.

There is a tendency for the digitalization of originally not digital textual contents. *Digitized archives* generally contain texts that originally were not produced for the public. These documents can be historical and contemporary public administration documents (birth and death certificates, healthcare patient records, property records,

military and police records), private letters, diaries, journals and newspapers, books, etc. A huge project by Google Books Library (Michel et al. 2011) is the digitalization of books from the 1500s. A good example of the analysis of such data sources is Grimmer (2010), who measured how senators explain their work in Washington to constituents based on a collection of over 24,000 press releases. Another historical analysis is the project, named "Mining the Dispatch" by Robert K. Nelson (McClurken 2012).[3] Nelson and his team analyzed the dramatic social and political changes of the Civil War through the articles of the newspaper *Richmond Daily Dispatch*.

3.2 A Brief Overview of NLP Methods

3.2.1 Pre-processing

Raw textual digital data is often unstructured, "noisy," and full of surplus information; thus, it is fundamentally different from the well-structured data sources of classical sociological research, which mainly contain the relevant information for the analysis. Therefore, it is not recommended to use the corpus itself per se, but to *pre-process the text before the analysis*. Several methods are available to prepare corpora for analysis – and as these corpora are frequently quite large, these pre-processing methods are algorithmized.

The pretreatment process usually starts with "clearing" the corpora in order to have only the relevant information in it. In practice, it means that all the punctuation marks (e.g., ., !, ?), articles (e.g., the, an), and conjunctions (e.g., of, by) are deleted. The use of stopwords (the deletion of content, where the given stop word is present) can help filter out content, which is not relevant at all for the analysis. It can be especially important if data is collected by polysemic words. Another main task is to reduce inflectional forms of words to a common base form. Stemming and lemmatization can be useful processes for this (Manning et al. 2008). Depending on the goal of the research, another step can be the detection of word classes and other linguistic or syntactic categories within the text. These tags can also help the lemmatization process. Besides these methods, the handling of multiple words/terms is also important. One such group is geographical or proper names, which usually contain several words (such as East Germany or Barack Hussein Obama), which nevertheless have to be treated together. The detection of proper names ("named entity recognition") is especially important in the case of opinion mining and sentiment analysis, where the extraction of entities has exceptional importance – considering, for example, the analysis of political texts and the politicians, who appear in them. Another such group consists of expressions, which belong together (like bus driver or carpe diem). These problems are usually solved

[3] http://dsl.richmond.edu/dispatch/pages/about

by the use of dictionaries, with which these names or expressions with multiple words can be identified and then handled together. The selection and way of application of different pre-processing steps depend on the nature of the research. Most studies fail to emphasize that these consecutively taken steps depend on each other. This dependency implies that the selection and the order of the steps have profound effects on the results (see, e.g., Denny and Spirling 2018). Accurate planning of this phase is immensely important.

3.2.2 Bag of Words and Beyond

Basic methods of NLP treat these pre-processed corpora as separated words, without their syntactical linking (*bag-of-words* model); see, e.g., topic modeling later. Basic analytic methods use the simple distribution of roots of words in the pre-processed corpora with its absolute (number of occurrences) or relative (percentage) frequency, usually treated in a vector. Other methods weight the words by their relative occurrence, for example, like the number of different documents they appear in, divided by the number of all documents (Evans and Aceves 2016). These methods are especially useful for retrieving the relevant information from a large corpus, which would be impossible to achieve solely by humans.

Nevertheless, the more refined methods of NLP do use not only the frequency of words but also the structure of the text and sentences. Syntactical analysis of the sentences can give a deep understanding of the corpora; however, it needs massive computational power and rarely provides a relevant result for questions researched by social scientists. A simpler version of the analysis of structure is the examination of the co-presence of words. This method works very well, even in large corpora. In the analysis of the co-presence of words, researchers focus on small "windows" of the text, namely, the given number of words, that are next to each other (so-called n-grams; see Google Ngram Viewer) and scan the whole document by sliding this given width window through the text. (See Fig. 1, for example, on a sentence by Max Weber.) In the end, the database of the analysis contains cases, which include all these "windows" with the given number of words. The number of words, by which this "window" is filled, can be any positive integer greater than one (e.g., bigram, trigram, fourgram, and so on). However, if this number is too high, we can lose the focus of our interest, namely, the context of the words. With the analysis of the co-presence of words and their closeness-distances, associations latently present in the corpus can be examined.

Besides the analysis of words, a higher-level analysis is also possible. Co-presence of words can be examined not only by n-grams but also in a paragraph; the level of analysis does not have to be on the level of words but the level of sentences. The use of these higher-level analyses can help understand the context of a word, which can shed light on several higher-level associations and meanings (e.g., in the detection of dialects or subcultural language) (Evans and Aceves 2016). A good example of higher-level analysis is a paper by Demszky et al. (2019), where the authors used sentence embedding beside the analysis of words. An example of

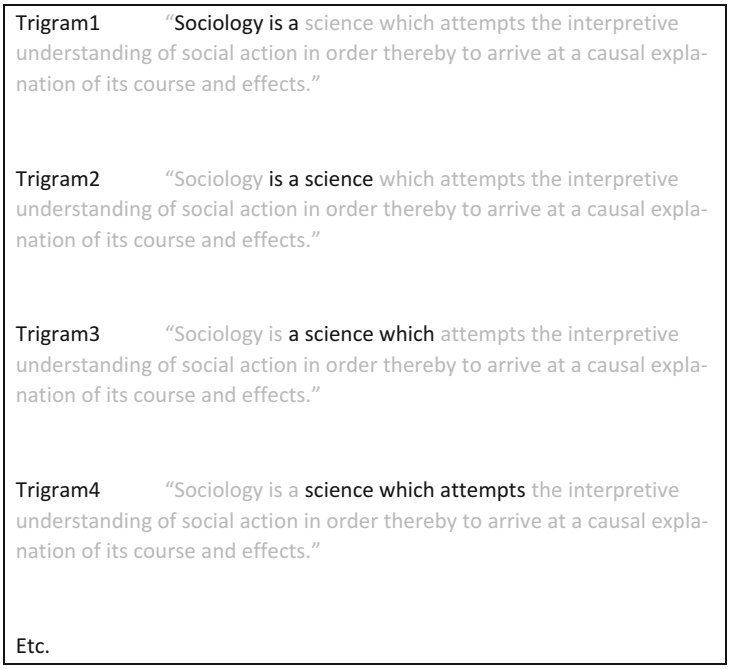

Fig. 1 An example of trigrams in a sentence

such an analytical tool is Stanford CoreNLP, which – besides other features – takes the grammatical and syntactical structure of the text into account (Hirschberg and Manning 2015).

3.3 The Goal of the Analysis and the Corresponding NLP Methods

The selection of the proper analytical NLP method is based on the same logic as in a "classical" sociological research. The choice needs to be based on the research question: (1) theory-driven approach for testing an already existing theory or (2) data-driven approach for exploring a not yet extensively studied topic and creating new theories. The group of supervised methods is the most suitable for the former, while the unsupervised methods are most suitable for the latter. These two approaches are more theoretical categories and are often used jointly in practice.

3.3.1 Supervised Methods

Supervised methods help researchers to (1) perform their analysis on larger datasets, than human capacity would be able to, or (2) expand their knowledge to external datasets they do not have background knowledge about, or (3) understand the mechanisms behind this expansion (see our discussion earlier about sociological relevance of the interpretation of "black box" parameters). A good example of the first one is the paper by Cheng et al. (2015), where the authors examined antisocial behavior in large online discussion communities. They employed humans to code the training part of the corpora, and based on these texts, they trained the computer to do the same on larger corpora. They also examined the importance of independent variables (features) of this classification, which shows the significance of the abovementioned aim (3). For the application of aim (2), Jelveh et al. (2014) provide an example. They examined the political ideology of economists based on their scientific papers. Originally, the authors were aware of the ideological attachment of a few of these economists (based on outer datasets like campaign contributions), but they expanded this knowledge by analyzing scientific papers written by the economists.

The method is supervised, as such supervision is needed to gain outside knowledge and for the training of the computer. The goal is to transmit this specific knowledge to the computer in order to be able to apply it to other datasets and texts. This outer knowledge has to be cautiously chosen by the researcher, as over-fitting of the model can result in bad classification later.

In practice, most frequently, a researcher (or a team of researchers) does the coding, the annotation of the text by hand, such as in the case of "classical" text analysis. Thus, a human-annotated text will be divided into a *train* part and a *validation* part. The "teaching" of the computer will be executed on the train part of the already coded text. The computer looks for patterns and tries to figure out the reasons why a researcher coded a sentence or word one way or another. Based on this training text, the computer creates the rules, by which it will code the text later on. To check the appropriateness of these rules and, thus, the computer coding itself, we use the other part of the already coded text, the so-called validation part. Not letting the computer know the result of the human coding, we make the computer do the annotation and compare its results with the human-coded results. This quality control usually happens by examining the area under the curve (AUC) of the receiver operator characteristic (ROC) curve or by other accuracy measures (e.g., precision or recall). If the results of the computer and the human are quite similar to each other, then the rules, which were based on the training text, are acceptable; if they are not similar, more train data is needed. If the training was acceptable, the best prediction model is used for the classification of new texts (which are not coded by human annotators) (Fig. 2).

When presented with complex annotations or codings, researchers may not have the capacity to provide enough training data. In these cases, crowdsourcing solutions, such as Amazon Mechanical Turk (Garg et al. 2018) or CrowdFlower (Kim et al. 2014), can be suitable to have enough human coded text as train or test data. Using crowdsourcing of an annotation or coding of a text can have hidden

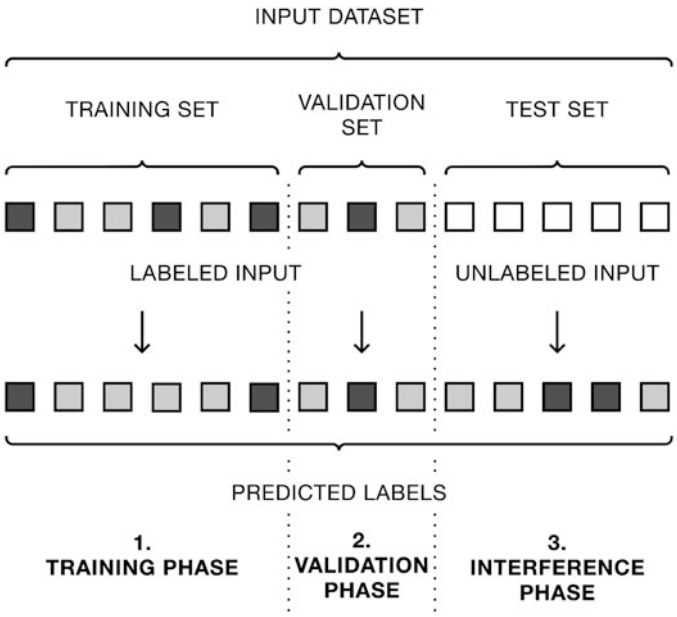

Fig. 2 The process of training, validation, and inference in supervised models

dangers, as human coders, who are not from the field of the specific research, can make different decisions from expert researchers of that field. In order to minimize this risk, it is advisable to take actions, which make the quality of the crowdsourced coding better. These actions may include the use of a very detailed codebook, using several coders for the same text and only accepting codings that are consistent across those coders, and/or continuous control of the coders with random checking. A good example of quality control can be found in the paper of Kim et al. (2014). For the coding phase, researchers can use not only their (or the crowdsource's) decision but external databases too, where the categories assigned to a text can be provided from other databases, as seen in the previously cited paper by Jelveh et al. (2014).

A special and frequently used type of supervised classification is *sentiment or emotion analysis*. The sentiment of a text is the attitude of the author toward an object (positive, negative, or neutral), while emotions are feelings from happiness to anger. A good example of its application is the Hedonometer mentioned above, where we can find the longitudinal "average happiness" for Twitter, measured with these techniques. Besides its scientific application, different classification methods can be used for the detection of sentiments and emotions of a sentence or a short text. Kharde and Sonawane (2016) introduce several techniques from the naïve Bayes method to the maximum entropy model with different data sources for the supervised part, like emoticons or external dictionaries. The goal is common: to classify the text to previously given sentiments or emotions and thus draw conclusions about the relationship of the content and the sentiment attached to it.

Yadollahi et al. (2017) give a detailed and well-defined taxonomy of sentiment analysis and emotion mining. According to them, this technique is quite widespread in business-related fields, such as customer satisfaction measurement or product recommendation. However, sentiment analysis can be easily adapted to sociological dimensions. See, e.g., Grimmer and Stewart (2013) discuss the placing of political actors on the liberal-conservative ideology scale.

Supervised methods contain well-known methods for social scientists, such as linear or logistic *regression* (Cheng et al. 2015). In these applications – contrary to the traditional theory-based approach – independent variables can be thousands of variables, which originate from the text (like a vector of word frequencies). As the number of these variables can be huge, dimension reduction techniques are frequently applied. These dimension reductions can be methods, which are familiar for social scientists, like principal component analysis or factor analysis or cluster analysis. Using sparse regression techniques, like the forward selection, is also common. Other supervised methods are more specific to NLP, such as *supervised topic modeling, support vector machine* (see, e.g., Bakshy et al. 2015), or decision tree. Regardless of the method we use, the goal is the same: to only choose those independent variables, which are important and have added value to our prediction or annotation, and to include all of them.

The application of supervised methods means that the theoretical considerations of the researcher influence the analysis, and thus, the interpretation of the predicted categories precedes the final results of the classification (Evans and Aceves 2016). As the annotation or coding is finalized, the dataset is ready to be analyzed by any statistical methods.

3.3.2 Unsupervised Methods

Unsupervised models are useful for discovering a field or topic, about which we do not have in-depth knowledge (see, e.g., Mohr et al. 2013). As the goal of these methods is exploration, there is no need for either background knowledge or prior annotation by the researcher. We allow the computer to find patterns in the data without theoretical supervision. Beyond discovering new topics, it can be useful for multigroup comparison, like comparing topics emerging from data of different cities (Nelson 2015) or countries (Marshall 2013). Unsupervised methods can also be useful as a supplementary method for supervised techniques. As we discussed above, researchers usually need data reduction techniques for the efficient application of unsupervised methods. These data reduction techniques are frequently unsupervised. Hereinafter we introduce the main unsupervised techniques, which are used in research dealing with large-scale corpora and suggest applied examples.

Some of the unsupervised methods can be familiar for social scientists, as they are frequently used in classical social research too. Widely used unsupervised classification research techniques dealing with large-scale data (Kozlowski et al. 2018; Kim et al. 2014) such as *cluster analysis, principal component analysis, and factor analysis* can be used according to the goal of the research. For example,

cluster analysis can be applied to textual documents in a vector space, where the axes stand for the words, and the values of the vector space are these documents. The position of the documents in this vector space is defined by the frequencies of different words in a given document. Cluster analysis can be applied not only to documents but also to words based on their distances from each other.

The use of *network analysis* – which was also used by social scientists before the era of big data – is also an unsupervised and useful technique when working with large textual datasets. Nodes can be entities, topics, or users; edges can be associations or co-occurrences. The network figure of a text shows the relative topological positions, the importance, and centrality of entities; network indices inform the researchers about the density and structure of a document (see, e.g., Brummette et al. 2018).

Naturally, new methods also emerged specifically for large-scale textual data. The group of *unsupervised topic models* is one of them. The goal of this method is to find latent topics in a document or across documents. Compared to a cluster analysis, where the units of analysis are words, topic models are different in their outputs. While all the words are assigned to one cluster, and cannot be assigned to several clusters, in topic models, words are assigned to each topic but with different probabilities. The theoretical concept behind topic models assumes that there are a finite number of topics, which can describe a set of documents. The probability distribution of words defines these topics, e.g., there is a higher probability that a document about sport includes the word "winner" than the word "inflation," while a document about the economy has reverse probabilities for these two words. However, documents can contain a mixture of topics. For example, a document about building a stadium could contain 80 percent economic and 20 percent sports topics. See Fig. 3 regarding the process of topic modeling.

The possible application of topic models in social sciences is very broad. We can analyze and compare the attitudes and opinions of different groups (such as political party supporters, patients having a given disease, authors of a given scientific journal, etc.) or the changes of topics (e.g., in a newspaper) over time. An exciting historical application of topic modeling is the already mentioned "Mining the Dispatch" project (McClurken 2012), where topics of "fugitive slave" were identified by topic modeling among the advertisements of a newspaper published during the Civil War. After the identification, the trend of the appearance of this topic was examined, which showed the dramatic social changes that happened at that time.

It is possible to focus only on one long document and discover the topics in it, but it is also feasible to compare several documents according to their latent meanings. As this method is unsupervised, the labeling and interpretation of the topics is the task of the researcher, and – unlike in the case of supervised topic models – this interpretational phase has to happen after the classification by the researcher. Validation of the topics is not an easy task; Chuang et al. (2012) suggest some visualization techniques that can be useful in the validation process. A good introduction for (supervised and unsupervised) topic models for social researchers is

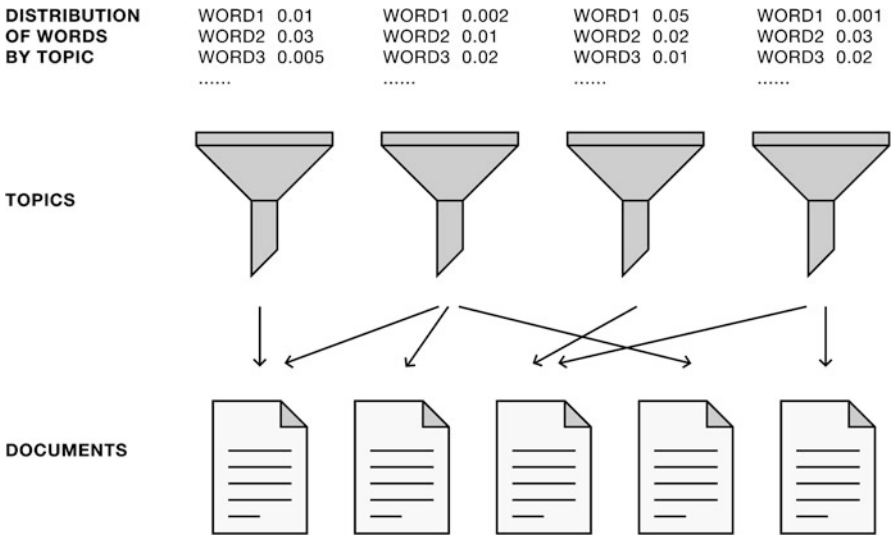

Fig. 3 The document generation process assumed by topic models

presented by Mohr and Bogdanov (2013) and Ramage et al. (2009). Wesslen (2018) presents good examples of the sociological application of topic modeling.

The more in-depth research of the semantic structure of texts became possible with *word embedding models*, which gained great popularity in the field of computational linguistics in recent years. The success of these models is based on their ability to help the understanding of large corpora. Earlier mentioned techniques do not model the meaning of words, while word embedding models do. The models identify the meanings with the contexts of the words. This distributional semantic approach has language philosophical antecedents: as its earliest representations, see the work of Wittgenstein from the 1930s and the paper by John R. Firth from 1957.

In word embedding models, words are embedded in a multidimensional semantic vector space, where words are positioned by their meanings, which is defined by their narrow textual environment. Two words will be positioned close or far from each other according to the similarity of their environment in the corpus. The multidimensional (usually a couple of hundred) vector space is created by the training of the words of a corpus by neural networks. These neural networks take into account the different semantical specifications of the corpus to reduce the space of words, where the semantical distances of words can be analyzed. Figure 4 presents a graphical illustration of a vector space, and as it shows, it is possible to grab semantic relationships (variations of greetings); partition the clusters of words with different meanings (sociologist and computational social scientist are separated from the other words); analogies can be created by vector differences (what Trump is for the USA, what can be for Russia?). However, the most inspiring possibility for sociologists is the identification of latent dimensions in the vector

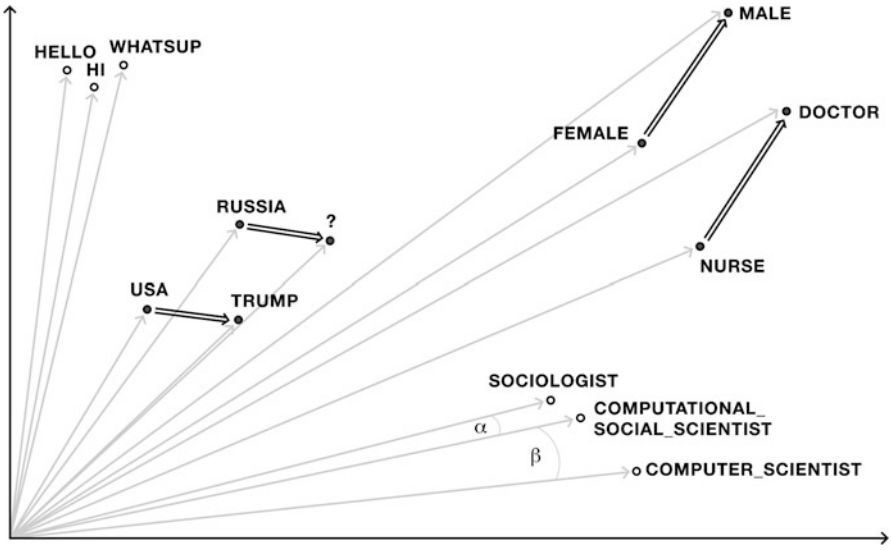

Fig. 4 An example for the vector space of a word embedding model

space, where dimensions represent some social difference. For example, if the vector from the nurse to the doctor is parallel with the vector from female to male, it refers to the male dominance of the profession of physicians. Using similar pairs to female-male (like girl-boy or women-men), one can identify the latent dimension of gender in this vector space. By projecting different expressions to this dimension, it is possible to grab the gender inequalities in different fields (e.g., differences in cultural consumption).

Word embedding methods can be used for several research-related questions, like sentiment analysis, classification of documents, or the detection of proper nouns. However, its most interesting feature is its ability to provide information about the associations between words, which can show the cultural analogies through texts. Kozlowski et al. (2018) provide one of the most interesting applications of this method. The authors identify latent dimensions (such as the gender dimension in Fig. 4) and show how cultural phenomena can be detected in various texts and how these phenomena change over time. Kulkarni et al. (2015) also show that longitudinal analysis can be efficiently processed with this method, which can help researchers detect and understand the change in the meaning of different expressions (and the cultural change behind it). The latest publications imply that word embedding models can detect such smooth distinctions as the identification of euphemistic coded words used in hate speeches (Magu and Luo 2018), or the detection of sarcasm (Joshi et al. 2016). Different computational methods, such as GloVe, fastText, and Word2vec, are available.

Unsupervised methods are different from supervised ones in the phase of interpretation of categories. While the interpretation of categories and theoretical

considerations happen before the automatized classification in the case of supervised methods, these tasks have to be completed after the automatized analysis in the case of unsupervised methods. Nevertheless, it is true for both methods that the classical statistical analyses are conducted on the dataset, which is the result of these methods. The analysis goes the same way on this structured data as in classical social research; and the topics or groups of cases usually serve as independent variables in the analysis.

3.3.3 Which Method to Choose

One group of the methods contains tools for *classification for previously given categories*. Logistic regression or supervised topic models are part of this group of methods. It can help detect the frequency of the defined topics in different documents. These tools are especially useful for longitudinal analysis, where researchers would like to discover the topic changes over time. As we mentioned above, sentiment and emotion analysis is also part of this group of methods. With these latter techniques, the relationship between the content and the sentiment of the text can also be analyzed.

Another group of NLP methods is the *classification of text without previously given categories*. Unsupervised topic model and cluster analysis are typical examples of these methods. The goal is the exploration, where researchers do not have a priori knowledge about the examined field and would like to have a comprehensive picture of the topics that arise in the documents. In the case of text data from different sources, countries, or timepoints, unsupervised classification can help compare the different groups.

The third type of methods includes those techniques, which *show the associations, the latent relationships* that are in the corpus. Vector space-based word embedding models belong here. These methods are based on the phenomenon that the language mirrors the user's cultural frame of mind. Thus, the associations and relationships of words in a corpus show the associations and relations that are present in a culture, or in a society. The method allows researchers to explore the distances between different concepts or the detection of the interconnection of connotations. Word embedding methods are substantive methods, but can also be used for arranging other methods, such as cluster analysis or network analysis, which can be applied to the resulted vector space.

4 New Possibilities for Sociological Research

4.1 How to Approach Automated Text Analysis as a Social Scientist

As evidenced in the preceding chapters, the use of NLP methods requires more skills than conventional social science qualifications provide. Reasonable familiarity with programming, pre-processing, and data science analytical tools are important prerequisites. Thus, the *cost of entry is* relatively high for those who try to acquire the necessary skills. Therefore it is very common that social scientists lead *interdisciplinary research*, where computer scientists and/or computational linguists are co-authors of a paper, and they are the ones who conduct the data generation part of the research. The papers of Kim et al. (2014), Kozlowski et al. (2018), McFarland et al. (2013), Mohr et al. (2013), Niculae et al. (2015), Srivastava et al. (2018), and Tinati et al. (2014) are all good examples of this kind of collaborations.

Nevertheless, we would not like to dissuade anyone from getting deeper into this field. Possessing basic programming knowledge and with an openness for new methods, it is worth delving deeper into automated text analysis. For those who decide to start, we recommend the book by Ignatow and Mihalcea (2017), which summarizes the actual development of automated text analysis with actual data sources, program languages, software, and analytical tools, especially for social scientists. Aggarwal and Zhai (2012) provide a more technical but very detailed synthesis of NLP methods. These books present a broad overview and helpful start for the application of NLP methods.

Beyond engaging in interdisciplinary cooperation or learning new skills, *using software, which needs entry-level knowledge,* is another strategy of getting involved with NLP usage. One such software is the Stanford Topic Modeling Toolbox (Ramage et al. 2009). There are also examples for easy-to-use platforms, which deliver a free run of the results of deep and rich NLP analysis. These platforms allow traditional sociologists to use these NLP results for the analysis of their own research questions. One of these platforms is a word embedding demo of Turku NLP group,[4] where queries can be performed without the knowledge of programming. These queries produce information about the meaning of words in vector spaces trained by other researchers. Another platform enables interactive examination of gender bias through the expressions of different topics – also with the use of word embedding methods.[5] The earlier mentioned Google Ngram Viewer is also part of this software group. It offers researchers the opportunity to follow historical social and cultural trends embedded in the wording of books. The Viewer's advanced wildcard search features make expressively meaningful analysis possible. A good example is provided by Michel et al. (2011) or Ophir (2016), who discovered hidden

[4] http://bionlp-www.utu.fi/wv_demo/

[5] http://wordbias.umiacs.umd.edu

patterns of centuries-long conceptual trends, using the Viewer's wildcard search feature. The terms "truth," "love," and "justice" were analyzed, and the results were paralleled with Sorokin's (1937) findings of the connection between "systems of truth" and "ethics of love." These examples prove that the use of these entry-level platforms can also be an avenue for sociologists to utilize the advantages of NLP in their research.

4.2 Combining the New with the Traditional: Mixed Approaches

The new techniques and methods we presented above are, in some sense, just by-products of information science and business analytics; they were not meant for supporting social research. Therefore it is an under-examined and open question which of the recently developed methods can be applied to sociological problems outside the scope of business applications (excellent examples are Evans and Aceves (2016) and Ignatow and Mihalcea (2017)). In the present chapter, we discuss how new methods and new textual data sources can be used jointly with the traditional ones, which are the sociological questions not yet studied in this area, and how the new approach can offer new insight into old research questions.

The most natural and widely used example for mixed methods in automated text analysis is the case of supervised learning when *models are trained on human-coded data*. Human coding requires close reading, and hence practically, it is a kind of qualitative text analysis. Such approaches directly extend qualitative text analysis to investigating large corpora. The problem of inter-rater reliability is of great importance here, as it profoundly affects not only label quality but predictive power as well. Two examples using crowdsourcing platforms for human coding are Iyyer et al. (2014), who aimed to identify political ideology from congressional debates, books, and magazine articles, and Cheng et al. (2015), who tried to detect trolls in online discussion communities.

Another field to mention is *qualitative content analysis*, a traditional and well-elaborated approach in social sciences that has been utilizing quantitative tools for a long time. Hence automated text analytics naturally emerges in this field, and its use could produce a more generalizable but interpretative approach. Ignatow and Mihalcea (2017) present a good summary of integrating NLP methods into classic sociological research. This approach could support the knowledge-driven applications of automated analytical tools, but it has some methodological challenges as well. Chen et al. (2016) provide a review of the challenges of applying automated analysis in qualitative coding. A theoretically elaborated example is Murthy (2016), who studied Twitter to rely on established grounded theory to give input to a computational Twitter analysis. According to Murthy's conclusion, mixed methods can offer new ontologies and epistemologies, an entirely new knowledge.

Grounded theory is the most elaborated qualitative content analysis method which has great potential in automated text analytics. Nelson (2017) proposes integration as a three-step methodological framework called *computational grounded theory*. The first step is a computational, unsupervised pattern detection step. The second step is a pattern refinement step that involves close reading. The third, pattern confirmation step, uses NLP techniques again to access the inductively identified patterns. Nelson (2015) provides a great example of her theory, a historical study of women's organizations in Chicago and New York City through the documents they left behind (e.g., internal memos and newsletters).

Other qualitative approaches present in the analysis of large textual data are ethnography of online communities, e.g., Facebook groups. An example is given by Baym (1999) with an ethnographic study of an online soap opera fan group. "*Netnography*"[6] adapts the traditional participant observation technique to the study of digital communications. It could be effectively integrated into automated text analysis (or vice versa) to get an exciting and balanced means of analysis. Di Giammaria and Faggiano (2017) gave an excellent example of the mixed analysis of this kind. They analyzed the Roman Five Star Movement Blog, by combining Facebook quantitative text analysis, ethnographic analysis of online conversations, social network analysis, and semiotic analysis of visual materials.

Additionally, integration of different approaches may be realized by *combining different data sources* like digital text corpora with survey data and/or census data. An example of combining data sources is Jelveh et al. (2014), who extracted political ideology from economists' papers, where ideologies in the training set were determined from datasets of political campaign contributions and petition signing activities. Another example is Garg et al. (2018), where word embedding models were trained on textual data (news and books) spanning a 100-year period in order to detect changes in stereotypes and attitudes in the USA. Human annotators were recruited on Amazon Mechanical Turk to assign gender labels to occupations. External validation was conducted by comparing the trends with shifts in US Census data, as well as with changes in historical surveys of stereotypes. They used digitized textual data, census data, survey data, and qualitative, human-coded data simultaneously.

4.3 What the Approach Can Offer to Classic Sociological Questions

The most important epistemological advantage of digital data is that it provides *observed* instead of self-reported *behavior*. This type of data offers real-time observation with continuous follow-up. The new approach offers access to *new data sources* (social media, digitized archives) and *new text analytic methods* (e.g., topic

[6]The term comes from Kozinets (1998).

modeling, word embedding modeling) not known before even by quantitative text analysis in sociology. NLP *technologies* developed for industry can be *transformed to answer sociological questions*.

Large digital corpora are spread over time, space, and topic. Therefore, (1) they provide the opportunity to conduct studies, which are otherwise impossible or at least hard-to-conduct within the traditional approach, like *longitudinal studies* by utilizing the dynamic flow of data. (2) They make possible *cross-country or cross-region comparisons* without conducting costly data collection far from the place of research. Finally, (3) given their size, they make possible the investigation of *small subpopulations* of societies (like soap opera fans) or subpopulations, which are *hard-to-contact otherwise* (e.g., drug users, members of illegal movements).

By definition, social sciences are less about individuals than interactions. When having textual data, *social phenomena* like identity, norm, conformity, deviance, status, conflict, or cooperation *emerge from communicative interactions*, not from separate individual statements. Large digital corpora present an outstanding opportunity for this analysis. Good examples are Lindgren and Lundström (2011) on rules of a global movement denoted by the #WikiLeaks hashtag on Twitter, Danescu et al. on power relations among Wikipedia editors (2012), Cheng et al. (2015) on antisocial behavior in online communities, and Srivastava et al. (2018) on adaptation to organizational culture analyzing internal employee emails.

The level of analysis (network/group or individual) can be freely selected when analyzing digital data. Sociological theories operating either at the *macro* (large-scale social processes) *or the micro* (face-to-face interactions or individual-level values, attitudes, and acts) *level* can be approached through digital data, and the syntheses of the two levels can also be realized. Sociology had an important role here since, as Resnyansky (2019) highlights, standard social media research concentrates on processes that are manifested at the micro level while forgets about structure. Researchers should not treat the text as standing alone in a vacuum, but they should pay attention to its social construction by taking the users' context and network as a starting point.

5 Summary

The analysis of a vast amount of digital textual data offers sociologists a broad perspective. Classic questions may be tested on new empirical bases; new insights and theories may also be generated. This new approach should be regarded as a complement and not a replacement for the traditional methods of sociology. Therefore, all of these data sources together may present sociology a flexible and broad enough empirical base.

Beyond encouraging interdisciplinary collaborations and revision of university training, further methodological developments are needed to decrease the currently high entry cost. Well-elaborated and easy-to-follow guidelines and rule of thumbs would make sociologists' task easier in data collection, preparation, and model

training. Analytical platforms that do not require advanced programming skills also support the transformation.

Fast development of NLP is expected in the next decade. Sociology will exploit the potential of this development if it is able to update its research culture while preserving its critical reflections. A renewed discipline like this will be hopefully able to understand better the profound changes in our contemporary society.

References

C. C. Aggarwal, C. Zhai (eds.), *Mining Text Data* (Springer, New York, 2012)

E. Bakshy, S. Messing, L.A. Adamic, Exposure to ideologically diverse news and opinion on Facebook. Science **348**(6239), 1130–1132 (2015). https://doi.org/10.1126/science.aaa1160

N.K. Baym, *Tune in, log on: soaps, fandom, and online community*, 1st edn. (SAGE Publications, Inc., Thousand Oaks, 1999)

D. Boyd, K. Crawford, Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. Inf. Commun. Soc. **15**(5), 662–679 (2012). https://doi.org/10.1080/1369118X.2012.678878

J. Brummette, M. DiStaso, M. Vafeiadis, M. Messner, Read all about it: the politicization of "fake news" on twitter. J. Mass Commun. Q. **95**(2), 497–517 (2018). https://doi.org/10.1177/1077699018769906

N.-C. Chen, R. Kocielnik, M. Drouhard, V. Peña-Araya, J. Suh, K. Cen, et al. *Challenges of Applying Machine Learning to Qualitative Coding*. Presented at the ACM SIGCHI workshop on human-centered machine learning, 2016

J. Cheng, C. Danescu-Niculescu-Mizil, J. Leskovec, Antisocial behavior in online discussion communities (2015). arXiv:1504.00680 [cs, stat]. http://arxiv.org/abs/1504.00680. Accessed 30 Oct 2018

J. Chuang, D. Ramage, C. Manning, J. Heer, Interpretation and trust: designing model-driven visualizations for text analysis, in *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems – CHI'12*. Presented at the 2012 ACM Annual Conference (ACM Press, Austin, 2012), p. 443. https://doi.org/10.1145/2207676.2207738

C. Danescu-Niculescu-Mizil, L. Lee, B. Pang, J. Kleinberg, Echoes of power: language effects and power differences in social interaction, in *Proceedings of the 21st International Conference on World Wide Web – WWW'12*. Presented at the 21st international conference (ACM Press, Lyon, 2012), p. 699. https://doi.org/10.1145/2187836.2187931

D. Demszky, N. Garg, R. Voigt, J. Zou, M. Gentzkow, J. Shapiro, D. Jurafsky, Analyzing polarization in social media: method and application to tweets on 21 mass shootings (2019). arXiv:1904.01596 [cs]. http://arxiv.org/abs/1904.01596. Accessed 4 Apr 2019

M.J. Denny, A. Spirling, Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it. Polit. Anal. **26**(2) (2018). https://doi.org/10.1017/pan.2017.44

L. Di Giammaria, M.P. Faggiano, Big text corpora & mixed methods – the roman five star movement blog. Bull. Sociol. Methodol./Bulletin de Méthodologie Sociologique **133**(1), 46–64 (2017). https://doi.org/10.1177/0759106316681088

J.A. Evans, P. Aceves, Machine translation: mining text for social theory. Annu. Rev. Sociol. **42**(1), 21–50 (2016). https://doi.org/10.1146/annurev-soc-081715-074206

J.R. Firth, *A Synopsis of Linguistic Theory. Studies in Linguistic Analysis* (Blackwell, Oxford, 1957)

N. Garg, L. Schiebinger, D. Jurafsky, J. Zou, Word embeddings quantify 100 years of gender and ethnic stereotypes. Proc. Natl. Acad. Sci. **115**(16), E3635–E3644 (2018). https://doi.org/10.1073/pnas.1720347115

J. Grimmer, A Bayesian hierarchical topic model for political texts: measuring expressed agendas in senate press releases. Polit. Anal. **18**(1), 1–35 (2010). https://doi.org/10.1093/pan/mpp034

J. Grimmer, B.M. Stewart, Text as data: the promise and pitfalls of automatic content analysis methods for political texts. Polit. Anal. **21**(3), 267–297 (2013)

E. Hargittai, Is bigger always better? Potential biases of big data derived from social network sites. Ann. Am. Acad. Pol. Soc. Sci. **659**(1), 63–76 (2015). https://doi.org/10.1177/0002716215570866

J. Hirschberg, C.D. Manning, Advances in natural language processing. Science **349**(6245), 261–266 (2015). https://doi.org/10.1126/science.aaa8685

G. Ignatow, R.F. Mihalcea, *An Introduction to Text Mining: Research Design, Data Collection, and Analysis*, 1st edn. (SAGE Publications, Inc., Los Angeles, 2017)

M. Iyyer, P. Enns, J. Boyd-Graber, P. Resnik, Political ideology detection using recursive neural networks, in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Association for Computational Linguistics, Baltimore, 2014), pp. 1113–1122. http://www.aclweb.org/anthology/P14-1105. Accessed 30 Oct 2018

Z. Jelveh, B. Kogut, S. Naidu, Detecting latent ideology in expert text: evidence from academic papers in economics, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, Doha, 2014), pp. 1804–1809. http://www.aclweb.org/anthology/D14-1191. Accessed 30 Oct 2018

A. Joshi, V. Tripathi, K. Patel, P. Bhattacharyya, M. Carman, Are Word Embedding-Based Features Useful for Sarcasm Detection? Presented at the conference on empirical methods in natural language processing, 2016

A.V. Kharde, S.S. Sonawane, Sentiment analysis of twitter data: a survey of techniques. Int. J. Comput. Appl. **139**(11), 5–15 (2016). https://doi.org/10.5120/ijca2016908625

A. Kim, J. Murphy, J. Richards, A. Hansen, J. Murphy, R. Haney, Can tweets replace polls? A U.S. health-care reform case study, in *Social Media, Sociality, and Survey Research,* ed. by C.A. Hill, E. Dean, J. Murphy (Wiley, Hoboken, 2014), pp. 61–86. https://www.rti.org/publication/can-tweets-replace-polls-us-health-care-reform-case-study. Accessed 1 Nov 2018

R. Kitchin, Big data, new epistemologies and paradigm shifts. Big Data Soc. **1**(1), 2053951714528481 (2014). https://doi.org/10.1177/2053951714528481

R.V. Kozinets, On Netnography: initial reflections on consumer research investigations of cyberculture. ACR North Am. Adv. **NA-25** (1998) http://acrwebsite.org/volumes/8180/volumes/v25/NA-25. Accessed 30 Mar 2019

A.C. Kozlowski, M. Taddy, J.A. Evans, The Geometry of Culture: Analyzing Meaning Through Word Embeddings (2018). arXiv:1803.09288 [cs]. http://arxiv.org/abs/1803.09288. Accessed 30 Oct 2018

F. Kreuter, R. Peng, Extracting information from big data: issues of measurement, inference and linkage, in *Privacy, Big Data, and the Public Good: Frameworks for Engagement* (2013), pp. 257–275. https://doi.org/10.1017/CBO9781107590205.016

V. Kulkarni, R. Al-Rfou, B. Perozzi, S. Skiena, Statistically significant detection of linguistic change, in *Proceedings of the 24th International Conference on World Wide Web – WWW'15*. Presented at the 24th International Conference (ACM Press, Florence, 2015), pp. 625–635. https://doi.org/10.1145/2736277.2741627

S. Lindgren, R. Lundström, Pirate culture and hacktivist mobilization: the cultural and social protocols of #WikiLeaks on twitter. New Media Soc. **13**(6), 999–1018 (2011). https://doi.org/10.1177/1461444811414833

R. Magu, J. Luo, Determining code words in euphemistic hate speech using word embedding networks, in *Proceedings of the Second Workshop on Abusive Language Online*, Brussels, 2018, pp. 93–100

C.D. Manning, P. Raghavan, H. Schütze, *Introduction to Information Retrieval*, 1st edn. (Cambridge University Press, New York, 2008)

E.A. Marshall, Defining population problems: using topic models for cross-national comparison of disciplinary development. Poetics **41**(6), 701–724 (2013). https://doi.org/10.1016/j.poetic.2013.08.001

A.E. Marwick, D. Boyd, I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. New Media Soc. **13**(1), 114–133 (2011). https://doi.org/10.1177/1461444810365313

J.W. McClurken, Richmond daily dispatch, 1860–1865 and mining the dispatch. J. Am. Hist. **99**(1), 386–388 (2012). https://doi.org/10.1093/jahist/jas157

D.A. McFarland, D. Ramage, J. Chuang, J. Heer, C.D. Manning, D. Jurafsky, Differentiating language usage through topic models. Poetics **41**(6), 607–625 (2013). https://doi.org/10.1016/j.poetic.2013.06.004

J.-B. Michel, Y.K. Shen, A.P. Aiden, A. Veres, M.K. Gray, Google Books Team, et al., Quantitative analysis of culture using millions of digitized books. Science (New York, N.Y.) **331**(6014), 176–182 (2011). https://doi.org/10.1126/science.1199644

J.W. Mohr, P. Bogdanov, Introduction—topic models: what they are and why they matter. Poetics **41**(6), 545–569 (2013). https://doi.org/10.1016/j.poetic.2013.10.001

J.W. Mohr, R. Wagner-Pacifici, R.L. Breiger, P. Bogdanov, 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 (2013). https://doi.org/10.1016/j.poetic.2013.08.003

F. Moretti, *Distant Reading* (Verso, London, 2013)

D. Murthy, The ontology of tweets: mixed methods approaches to the study of twitter, in *The SAGE Handbook of Social Media Research Methods*, ed. by L. Sloan, A. Quan-Haase (SAGE, London, 2016), pp. 559–572

L. Nelson, *Political Logics as Cultural Memory: Cognitive Structures, Local Continuities, and Women's Organizations in Chicago and New York City* (2015). https://www.academia.edu/10250788/Political_Logics_as_Cultural_Memory_Cognitive_Structures_Local_Continuities_and_Womens_Organizations_in_Chicago_and_New_York_City. Accessed 31 Oct 2018

L. Nelson, *Computational Grounded Theory: A Methodological Framework* (2017). https://doi.org/10.1177/0049124117729703. Accessed 30 Mar 2019

V. Niculae, S. Kumar, J. Boyd-Graber, C. Danescu-Niculescu-Mizil, Linguistic harbingers of betrayal: a case study on an online strategy game (2015). arXiv:1506.04744 [physics, stat]. http://arxiv.org/abs/1506.04744. Accessed 31 Oct 2018

S. Ophir, Big data for the humanities using Google Ngrams: discovering hidden patterns of conceptual trends. First Monday **21**(7) (2016). https://doi.org/10.5210/fm.v21i7.5567

A.J. Oswald, S. Wu, Well-Being Across America (2011). https://doi.org/10.1162/REST_a_00133

D. Ramage, E. Rosen, J. Chuang, C.D. Manning, D.A. McFarland, *Topic modeling for the social sciences*. Presented at the workshop on applications for topic models, neural information processing system, Stanford Computer Science (2009)

L. Resnyansky, Conceptual frameworks for social and cultural big data analytics: answering the epistemological challenge. Big Data Soc. **6**(1), 2053951718823815 (2019). https://doi.org/10.1177/2053951718823815

L. Ryan, L. McKie (eds.), *An End to the Crisis of Empirical Sociology? Trends and Challenges in Social Research* (Routledge, London, 2015)

M. Savage, R. Burrows, The coming crisis of empirical sociology. Sociol. J. British Sociol. Assoc. **41**, 885–899 (2007)

P.A. Sorokin, *Fluctuation of Systems of Truth, Ethics, and Law*, vol 2 (American Book Co., New York, 1937)

S.B. Srivastava, A. Goldberg, V.G. Manian, C. Potts, Enculturation trajectories: language, cultural adaptation, and individual outcomes in organizations. Manag. Sci. **64**(3), 1348–1364 (2018). https://doi.org/10.1287/mnsc.2016.2671

R. Tinati, S. Halford, L. Carr, C. Pope, Big data: methodological challenges and approaches for sociological analysis. Sociology **48**(4), 663–681 (2014). https://doi.org/10.1177/0038038513511561

R. Wesslen, Computer-assisted text analysis for social science: topic models and beyond. arXiv:1803.11045 [cs] (2018). http://arxiv.org/abs/1803.11045. Accessed 17 Feb 2019

A. Yadollahi, A.G. Shahraki, O.R. Zaïane, Current state of text sentiment analysis from opinion to emotion mining. ACM Comput. Surv. **50**, 25–25 (2017). https://doi.org/10.1145/3057270