Renáta Németh* & Júlia Koltai**

Natural language processing: The integration of a new methodological paradigm into sociology

Intersections. EEJSP
9(1): 5–22.
https://doi.org/10.17356/ieejsp.v9i1.871
https://intersections.tk.hu

* [nemeth.renata@tatk.elte.hu] (Eötvös Loránd University)
** [koltai.julia@tk.hu] (Centre for Social Sciences; Eötvös Loránd University)

Abstract

Natural language processing (NLP) methods are designed to automatically process and analyze large amounts of textual data. The integration of this new-generation toolbox into sociology faces many challenges. NLP was institutionalized outside of sociology, while the expertise of sociology has been based on its own methods of research. Another challenge is epistemological: it is related to the validity of digital data and the different viewpoints associated with predictive and causal approaches.

In our paper, we discuss the challenges and opportunities of the use of NLP in sociology, offer some potential solutions to the concerns and provide meaningful and diverse examples of its sociological application, most of which are related to research on Eastern European societies. The focus will be on the use of NLP in quantitative text analysis. Solutions are provided concerning how sociological knowledge can be incorporated into the new methods and how the new analytical tools can be evaluated against the principles of traditional quantitative methodology.

Keywords: natural language processing; text analytics; text mining; institutialization; epistemology; causality

1 Introduction

The analysis of written texts has always been part of social research methodology, although initially, the latter was strongly associated with qualitative research. Quantitative text analysis started in the first half of the twentieth century and typically involved analyzing the frequency of occurrences of qualitatively identified categories or 'codes' (e.g., Bales, 1950). Computational power supplemented human capacity from the 1960s onwards with the invention and development of computers (Hayes, 1960). Because of the appearance of (partly) automatized content analysis, the amount of processable text increased.

In parallel with technological development, the universe of digital textual data also started to grow at an enormous speed. According to recent International Data Corporation research (Reinsel et al., 2018), the volume of this data will reach about 175 Zettabytes by 2025 (five times that of 2018). Approximately 80 percent of all data is estimated to be unstructured, text-rich data. The impact of the IT revolution is even more extensive since it has reshaped not only technologies but the whole of society. Every aspect of our lives is deeply affected by the digitalization of data. Almost every digital 'step' we take is recorded, from our mobile phone calls through text messages to posts on social media. A significant part of this data is textual – e.g., the metadata of tweets includes not only the message but the creation time, location, number of likes, and the followers of the user, etc. Thus, we are not only able to analyze the tweets but also explore users' networks, determine network roles, follow the spread of news in time and space, identify topics and opinions, and determine the factors influencing these phenomena.

Natural language processing uses computers to process and analyze large amounts of textual data. NLP is located at the intersection of computer science, artificial intelligence research and linguistics and involves employing computational methods for the purpose of analysing large amounts of text or producing human language content (Hirschberg & Manning, 2015). It also includes speech recognition and the syntactic processing of texts, but the social sciences are mainly concerned with semantic approaches to language resources. Therefore, this sub-field is also referred to using several other largely synonymous terms, such as computational linguistics, text mining, text analytics, etc. For example, Hirschberg and Manning (2015) use NLP and computational linguistics synonymously; however, in everyday use, NLP emphasizes a more empirical approach. Hereafter, we will use the term NLP throughout, as our impression is that this term is most canonized in sociological applications.

NLP primarily gained relevance during the digital revolution, which also involved a revolution in digital self-expression. While in the past, opinions and attitudes embodied in texts were almost exclusively the narratives of the cultural elite, today, the production of written (and often digital) text is no longer confined exclusively to the elite but includes everyday users. Social behaviour can be directly observed now, and not only on a self-reported basis, which supports the internal validity of sociological inquiry. In addition, the availability of digital data provides opportunities for both unbiased longitudinal and retrospective research, as behavioural data is archived, and there is a possibility for the live investigation of contemporary processes as well.

Large datasets overcome the sample-size limits of traditional analysis: small subpopulations, detailed analyses and rare phenomena can also be approached with these data and methods. Additionally, the widespread digitization of pre-existing textual archives makes it possible to investigate phenomena that existed before digital society, like historical events or cultural trends.

In parallel with the growing availability of analyzable textual data, the vast increase in computing power, the development of computer capacity, artificial intelligence, computational linguistics and more accurate machine learning algorithms are producing new tools with which these textual data can be analyzed. As these methodological tools are more and more advanced, such text analysis opportunities can provide suitable analytical depth, even fulfilling the expectation of sociologists. NLP, a whole new discipline, has emerged, allowing for the examination of corpora with billions of words, the automatization of coding (or 'annotation' – a term used by the programming community) of texts, and the discovery and processing of linguistic structures by computers. These algorithms can unfold not only the explicit content of a corpus but also its latent meaning (e.g., sarcasm and metaphors) and topical associations (e.g., values and opinions). Computers can detect semantic structures and connotations within large corpora that humans alone would not have been able to identify. The application of algorithms has also standardized the annotation processes to a high degree, substantially raising the level of reliability of text analysis. NLP methods are rapidly advancing with the development of data science and computational linguistics; new methods emerge every week. At the same time, the contribution of researchers is still very important – artificial intelligence is not that intelligent, and researchers' knowledge and theoretical conceptualization are essential in the design and implementation of analyses (Evans & Aceves, 2016). These developments have opened up entirely new opportunities for researchers interested in text analysis.

Thus, NLP fits into the trend of the endogenous evolution of social research methods. Previous qualitative analyses of textual data in sociological research required line-by-line examination of texts, hence their application to large corpora was impossible. Machine learning techniques can extend the principles of qualitative analysis, representing a promising approach to researchers. As Figure 1 shows, the popularity of NLP has been continuously growing in recent years and within each discipline that is investigated.[1] Each trend line shows persistent growth even after normalization for the total number of publications in the discipline. The proportion within sociology has increased faster than average, which suggests that NLP is becoming an increasingly recognized approach in this field.

Multiple researchers have successfully applied the method in research on Eastern European societies. For example, Katona et al. (2021) and Barna and Knap (2022) conducted a thematic analysis of current online media: the former focused on the representation of corruption in Hungary, while the latter examined the discursive framing of Trianon and the Holocaust. Kmetty and Knap analyzed more than thirty million Hungarian Facebook comments to reveal the role of incivility (namely, indecent communication) as an emotional response to challenges and stress related to the COVID-19 pandemic. With the application of NLP, Bielik (2020) uncovered both the sentiment and the thematic focus of the social democratic parties of the Visegrad countries through their electoral manifestos. There is research using these methods even in the field of social history, like that of Szabó et al. (2021), who analysed temporal change in social discourses in the socialist era in Hungary using a large corpus from the journal Pártélet.

Our focus in this paper will be the use of NLP in quantitative text analysis, but it should be noted that NLP can also be successfully combined with qualitative content analysis. Such a mixed-methods analysis can respond to the depth/breadth dilemma often encountered with in-text analysis (Parks & Peters, 2022).

[1] To access publication data, we used Dimensions (https://dimensions.ai), a scholarly searchable database. We defined the search as applicable to the full text of any type of publication with the composite search terms "automated text" OR "natural language processing" OR "computer-assisted text" OR "computational linguistics" OR "text mining" OR "computational text" OR "topic model" OR "sentiment analysis" OR "text classification" OR "word embedding" OR "text clustering".

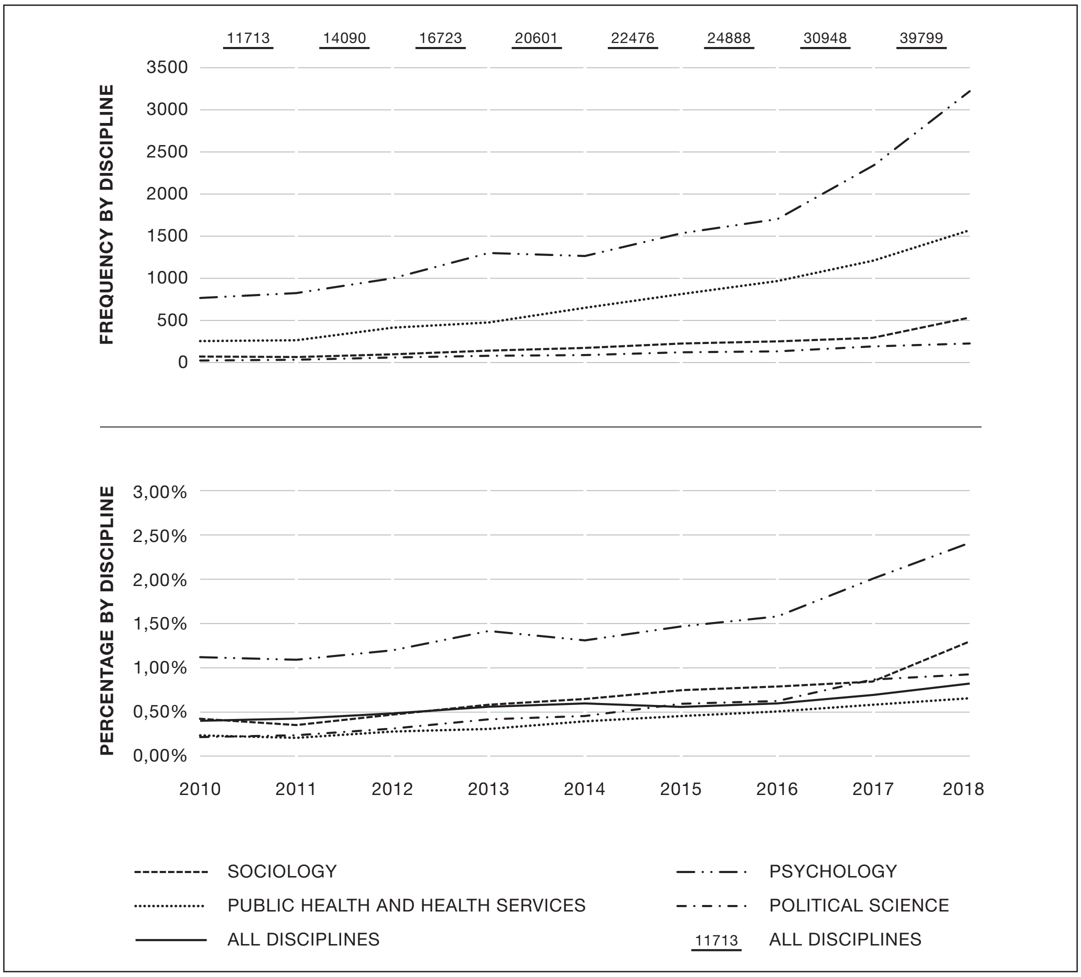

Fig. 1 Publications which use natural language processing in four disciplines (total number and percentages compared to the total number of publications in the field)

2 Analytical opportunities

NLP methods go far beyond traditional quantitative text analysis, which basically analyses the distribution of words or topics defined by the researcher. These new methods cannot and do not aim to substitute human reading and understanding but can extract specific pieces of information which help answer research questions. In simple terms, an NLP model uses a language model, which is a simplified representation of the text production process. It is designed to give an acceptable approximation of some parameters of the process that are important in the given context (e.g., the number of topics in the case of topic models; see later).

The new techniques and methods are, in some sense, just by-products of information science and business analytics; they were not developed to support social research. Therefore, it is an under-examined and still open question of which of the recently developed methods can be applied to sociological problems (some excellent examples are included in Evans & Aceves, 2016).

To illustrate the opportunities associated with these new methods, we briefly describe three of the most popular methods and describe exciting sociological research which applies them. All three methods (topic modelling, word embedding modelling, and supervised classification) have many different versions and different algorithms that can be implemented. For an easy-to-understand description of these methods and further examples of their sociological application, see Németh and Koltai (2021).

Topic modelling methods (Blei & Raftery, 2009) assume that a corpus is represented by latent topics that are probability distribution over words, thus treat a text as a mixture of some of these topics. Using topic models, Shah (2019) investigated the contextual framing of rape in the English language press in India before and after the 2012 gang rape case and detected important changes in the framings. Sterling et al. (2019) analyzed 3.8 million Twitter messages and revealed that, on this platform, liberals and conservatives differ significantly in their priorization of social values. Thus, these models help understand and classify the explicit content of large texts.

Another inspiring NLP model is the family of *word embedding models* (Mikolov et al., 2013), which geometrically represent the semantic structure of texts. These models position words in a high-dimensional vector space. The training of the vector space and the position of the words are based on the textual context of the words. Words with similar meanings are placed close to each other in this space. Jones et al. (2020), using millions of digitized books, quantified shifts in male gender bias over time in four domains: career, family, science, and arts. For example, they proved that gender bias still exists in science by measuring the distance between male and female word pairs (she/he, woman/men, etc.) and focused on their distance from scientific terms, such as technology and experiment.

Our last example of important NLP methods is the family of *supervised classification*. Supervised classification makes it possible to teach the result of a traditional qualitative text analysis of a small sub-corpus to an algorithm, which expands its classification to a much larger corpus. The method is a clear example of mixed methods: the result of human coding (a qualitative approach) is input into the quantitative algorithm. Florio et al. (2019) developed a supervised classification model for hate-speech detection on a corpus of Italian Twitter data. The researchers employed human coders to label (annotate) a smaller sample of their corpus and used it to teach the algorithm to automatize the labelling of a larger number of geo-tagged tweets over six years. They analysed this data together with macro-level social indicators and found correlations suggesting an association between economic and cultural factors and the presence of online hate speech. An application that applied a supervised classification to data from Eastern Europe is described in Csomor et al. (2021). In their study, they assessed the responsiveness of Hungarian local governments to requests for information by Roma and non-Roma clients, relying on a nationwide correspondence study. The authors showed that it is possible to detect discrimination in emails in an automated way without human coding.

As the above examples show, sociological knowledge and methodology can have high added value in the application of NLP tools and vice versa; these methods have strong interpretive power, and hence they have a great deal of potential in sociological knowledge building. Furthermore, recent advances in NLP methods based on artificial neural networks are developing rapidly, with outcomes such as increasingly powerful machine translation (Khan & Abubakar, 2020), chatbots capable of answering thematic questions (Tiwari et al., 2021) and humour detection (Miraj & Aono, 2021). These developments, which serve a practical purpose, will also certainly find their use in social sciences.

3 Challenges

3.1 Conceptual, disciplinary, and institutional tensions

The integration of NLP into sociology demands more than merely using new methods to address old research questions. As we show later, a clear shift can be observed in empirical expertise from academia to industry since industry generates problem-solving solutions and can also finance these developments. Previously, the role of sociology was based on its own expertise in data collection, analysis and theory creation (Savage & Burrows, 2007). However, nowadays, social and digital data are also collected and investigated by computer scientists. Social scientists, who earlier defined themselves according to specific methods of research, have had to realize that the emerging methods are not exclusive to social sciences but are being developed by experts in other fields (Savage & Burrows, 2007). The result is a loss of dominance and a need to redefine sociology's role in empirical social research.

Gaining competence in NLP as a social scientist is associated with a substantial entry cost. Sociologists have to enter a new field where language is constructed by programs, and the problems a researcher faces during data collection are not necessarily related to society or even to real people. According to Metzler et al. (2016), who conducted a worldwide survey with social scientists, approximately half of the respondents had done some kind of big data analysis, and one of the most significant barriers that was mentioned was the lack of time for learning methods in this new field. However, sociologists with strong quantitative knowledge were able to improve their skills to meet this challenge. The growing number of introductory books reflects this demand (an excellent example is Ignatow & Mihalcea, 2017). Concurrently, computer science, physics, and computational linguistics experts who have much more experience in this field have discovered that their knowledge makes them capable of researching society and have not been afraid to use it. The symbolic territorial games and the above-mentioned tensions have created further difficulties concerning the institutionalization of computational social science.

Though the tension between social sciences and the above-mentioned data science fields is decreasing, different intellectual traditions and solutions are still present in the two fields (DiMaggio, 2015). Quan-Haase and McCay-Peet (2017) discuss these challenges and difficulties, as well as the advantages of creating interdisciplinary teams. Motivations and rewards for scientists of different disciplines can vary: data- and computer scientists favour rapid publication and more application-oriented results. On the other hand, social scientists prefer publishing in conventional journals (with a slower publishing process) with no need for applied results.[2]

Nevertheless, it is worth addressing these difficulties and identifying goals that can satisfy the researchers of both disciplines. Automated research of society needs both approaches: the technical knowledge of computer- and data scientists and the social theoretical knowledge of sociologists are complementary. These collaborations can be highly successful and result in frequently cited papers (e.g., Tinati et al., 2014; Mohr et al., 2013; McFarland et al., 2013). Institutions like Harvard have realized the need to restructure their social sciences departments and promote collaboration with other fields (King 2014). Industry and academia can also be merged in such projects, such as Social Science One. This Harvard-incubated initiative enables academics to use the increasingly rich data of companies to address societal questions. In recent years, a process of consciously organized institutionalization has been observable. Several research centres and academic departments have been established (e.g. the Center for Research on Computation and Society at Harvard and the Institute for Data, Systems and Society at MIT), BSc and MSc programs on computational social science have been launched, and numerous fellowships and summer schools (e.g. the Summer Institutes in Computational Social Science in thirty renowned locations around the world) provide training in the field of computational social science and specifically in NLP. International publishers increasingly focus on the intersection of social and computational science. Book series on computational social science have emerged; newsletters and social research are being published. New journals have been established (e.g. Springer's Journal of Computational Social Science in 2018 and the Journal of Digital Social Research at Sweden's Umeå University), and conference series have been started (e.g. the International Conference on Computational Social Science, Big Data Meets Survey Science, etc.).

However, most of the initiatives described above come from data- and computer science and are less likely to be facilitated by social scientists, especially sociologists. Indeed, until recent years, sociology has made only minor contributions to such research. At the same time, researchers who analyze social digital data have made few attempts to engage with the social science literature (Mützel, 2015; Watts, 2013). In 2016, in a paper entitled 'Theoretical Foundations for Digital Text Analysis', Gabe Ignatow discussed the lack of stable institutionalization of digital text analysis in the social sciences. The question is how to expand this interdisciplinary approach to the traditional institutions of sociology. For example, although some departments (mainly at graduate schools) teach methods of computational social science, these new techniques are still not part of general sociological curricula. According to Jager et al. (2020), programs in computational social science are still quite rare in Europe, both at the bachelor's and master's levels. The latter detects two main reasons for this lack of training: students' mathematics-related anxiety and the

[2] Results from the database of Microsoft Academic Research tend to support these conclusions. Most of the results for the search term 'natural language processing' are from conference papers; while the main type of publication for 'social network' is journal publication; going forward to a more classical social topic, 'feminist theory', the ratio of conference papers compared to all publications in this topic is even fewer.

limited knowledge of academic instructors. The latter claim is consistent with the finding of Metzler et al. (2016) that this knowledge is not evidently available among social scientists, and significant time will be needed to build it. The main question is how can education provide an appropriate knowledge of social theory as well as specific computational skills that students can apply (Boyd & Crawford, 2012).

3.2 Do 'numbers speak for themselves'? Measurement and data quality

3.2.1 Objectivity

The claim that 'numbers speak for themselves' comes from a widely cited and widely criticized pamphlet written by Chris Anderson, former editor-in-chief of Wired magazine (2008). Anderson introduced a new paradigm of empiricism that assumes that the vast volume of data offers objectivity and precision, suggesting that scientific hypotheses and modelling are not needed anymore. Many different disciplines were represented in the responses provoked by the article, including biology (Pigliucci, 2009), biochemistry (White, 2009), and, with some delay, social science (a well-known response was provided by boyd & Crawford, 2012, and a more recent one by Resnyansky, 2019).

The question of objectivity also arises in the field of NLP. The challenge is epistemological in nature: digital data – compared to survey data – are not the product of traditional operationalization processes but are like 'digital footprints'. Researchers cannot plan the concrete measurement of their concepts but 'have to cook using what they find'. For example, let us take the Hedonometer, a Twitter-based text analysis tool that measures the happiness of large populations in near real-time (hedonometer.org) by the University of Vermont, Complex Systems Center). A traditional survey with a similar aim would be based on responses from a representative sample (e.g. to the question 'How happy do you feel on a scale of 1 to 4?'). The Hedonometer measures the happiness of regions/time intervals based on the *average happiness* of the respective tweets (subjective decision 1). The happiness of tweets is based on *average happiness scores* assigned to the words in the tweet (subjective decision 2). In this language model, tweets are not analyzed syntactically, but their words *as a set* are considered, which is called the bag-of-words model (Németh & Koltai, 2021) (subjective decision 3). To quantify the happiness of words, Amazon's Mechanical Turk service (that is, combined *human decisions,* were used). For example, words like 'hope', 'hero', and 'to win' score highly (subjective decision 4). The procedure is an example of lexicon-based sentiment analysis. It is clear from the example that digital footprints do not have meaning in themselves; it is researchers who construct meaning.

3.2.2 Context

The role of context is another concern that is often raised (Shaw, 2015; Lewis, 2015; Törnberg & Törnberg, 2018). Context is a specific phenomenon that is hard to capture on a large scale. In the case of the Hedonometer, the question of context trivially emerges: the happiness of words is measured without taking their context (= the tweet they are mentioned in) into account, and the happiness of tweets is also measured without assessment of their context. However, messages are culturally embedded and are affected by social norms, which can vary in different societies. Some of the Hedonometer's results (e.g., older bloggers are less happy, and the happiness of music lyrics decreases over time) might be traced back to these factors and do not indicate real shifts in well-being.

3.2.3 Internal and external validity

Basic methods for evaluating the quality of surveys – e.g., assessments of external and internal validity – can also be easily implemented in this field (for a complete adaption of total survey error to Big Data, see Amaya et al., 2020). Turning back to the Hedonometer: do American Twitter users represent the whole American population? Is the random sample of tweets provided by the Twitter API representative of the whole Twitter stream? Such questions are all related to the classical aspects of the quality of measurement. Regarding external validity, intuition and scientific research suggest that widely used social media data cannot be treated as a representative sample of the general population (Blank & Lutz, 2017). As for internal validity, traditional quantitative social research works with numerical data supplied by survey respondents and is subject to potential measurement errors like recall bias or social desirability bias. Computational methods, however, are often applied to 'found texts' that were usually created for some other purpose than scientific analysis. Thus, these data often refer not to self-reported but observed behaviour, hence are free from recall bias and social desirability bias. These data have higher internal validity in this sense than classic survey data. However, during the analysis, texts are translated into numbers (see earlier for how the Hedonometer defines and measures happiness), and the translation process includes many decision points, which – as with any quantitative research – have to be handled by the researcher.

We conclude that in the case of NLP, there is some distance between results and 'reality' that decreases internal validity, just like in the case of traditional survey research. The consequence is also similar: this distance does not invalidate 'translated texts' but highlights the importance of understanding our data that are 'social' in origin. Including conceptual frameworks and social theories and integrating social scientific knowledge into computational analytics could support this understanding. With detailed documentation of proper methodology and conscious interpretation that deals with representation error, measurement error, and context, reports like Hedonometer's can also give compelling insights into society.

4 'Change the instruments, and you will change the entire social theory that goes with them' – A new methodological paradigm?

4.1 Prediction vs. causation

A search for Latour's quote (2010) in the title, together with the expression 'big data', generates hundreds of Google hits. Indeed, digital data reframe the measurement process (as discussed in the previous chapter) and the logic of scientific research. Because of its industrial origin, analytical methods associated with the new paradigm are optimized for prediction, while explanation and causation – the focus of sociology – tend to be outside its scope. The discourse that frames this change as a paradigmatic one (creating a tension between 'new' and 'old') may make it more difficult for automated data analytics to be integrated into sociology.

Consider the paper of Tsakalidis et al. (2015), who predicted election results using Twitter data and polls. The authors defined different machine learning models and, based on some measures, selected the best one. They implemented this selection process without interpreting the model and identifying the factors most likely to influence the election result – as would routinely have been done in traditional sociology.

In general, a specific feature of computational text analysis inherited from data science is that it aims to optimize predictions, as opposed to traditional social research, which tries to define causal models. This difference is crucial. The former concentrates on the outcome (e.g., classification), looking for the optimal function of the predictors while trying to avoid over-fitting. The latter is interested in the effect of a given predictor (a social determinant) adjusted to potential confounders. Predictive models are not interested in interpretable effects (e.g., they often transform or re-categorize predictors if this increases model fit, even if this results in noninterpretable effects). On the contrary, for traditional social scientists, proper operationalization and interpretation of the effects of predictors are very important, and predictive power is less so. A good example of such non-interpretable effects is the Netflix Grand Prize winner model. In this contest, the goal was to predict the evaluation of movies. In the winning model, one of the main explanatory variables was 'if there is a number in the title of the movie' (Töscher et al., 2009). NLP applications selected the best-performing predictive models based on their predictive accuracy: the focus was on the model's performance, and how the model worked did not matter. From the traditional causal point of view, these models are like black boxes.

4.2 Two cultures of statistical modelling

The prediction vs. causation contrast yields a significant dichotomy from the perspective of the philosophy of science and has statistical relevance as well. However, it fits more generally with the contrast that is identified between data-driven vs theory-driven approaches. Indeed, the essence of the dispute sparked by Anderson (2008) can be captured as a conflict between new data-driven and explorative approaches and traditional, theory-driven, model-based ones. There have been two decades of reflection on the statistical elements involved in this contrast, provoked by a paper by Leo Breiman (2001). Breiman writes about 'two cultures' of statistical modelling. The traditional approach assumes the presence of a stochastic model for the data generation process: i.e., how are the data distributed, and how does the outcome relate to the predictors? The model is validated by residuals and goodness-of-fit statistics, while the model parameters are interpreted to answer the research questions. Algorithmic modelling, on the contrary, assumes nothing about the data and is not engineered to create an interpretable model. It just creates a 'black box' model and evaluates its performance according to its accuracy – namely, its fit to the data. This dichotomy anecdotally provoked the British statistician Brian D. Ripley to state that 'machine learning is statistics minus any checking of models and assumptions'.[3]

Additionally, data science uses optimization procedures to choose from many machine learning models to select the best one according to its predictive power. Since a vast number of predictors are available in most cases, variable selection methods are used to overcome the effect of noise and obtain relatively few variables relevant to the predictive task. In contrast, traditional social researchers only define one model, but they describe it in a theory-based way. The selection of predictors is also based on theory. In doing this, both approaches aim to create a robust model that works not just in the given context or with specific data but also in other cases. Therefore, integrating sociological knowledge into each step of the automated text NLP analysis process may increase the robustness of the results.

4.3 Reconciling the two seemingly conflicting approaches

The above-described features of machine learning algorithms are understandable, considering the large amount of data they use, the many parameters they estimate, and the complexity of the nonlinear functions and interactions, which are all part of the outcome prediction process. Breiman's point is that data scientists' modelling approach should be added to the standard statistical toolbox as an option. Indeed, the contrast between the two approaches is neither antagonistic nor paradigmatic. Predictive and causal approaches are not substitutes: they can help answer different research questions and even supplement each other. In biomedical sciences, predictive (etiologic) and causal (diagnostic) approaches have long co-existed. Their relationship is well-articulated and is even part of the standard training curriculum (e.g., the widely used university literature by Kleinbaum et al., 2008). The interpretation of predictive models helps us open up the black box. For example, finding the most important predictors (in NLP terms, the 'features') of logistical regression and interpreting their coefficients brings us closer to understanding the measured social phenomenon. Cheng et al.'s (2015) paper provides an excellent application of NLP where prediction is applied in addition to an explanation in a study of antisocial behaviour in online discussion communities. Lately, data scientists have also tried to open the 'black box', arguing that even the developers of these algorithms should not 'trust the model' and ignore why it made certain decisions (Molnar, 2019).

4.4 Causal analysis in NLP

Extracting causal relations from data, specifically textual data, has also received attention recently (Feder et al., 2022). The dichotomy between theory-driven and data-driven approaches mentioned above can be seen here in the fact that a causal analysis is based on

[3] One of the fortune cookies in R (package 'fortunes', fortune 50) includes this quote: 'To paraphrase provocatively, "machine learning is statistics minus any checking of models and assumptions" – Brian D. Ripley (about the difference between machine learning and statistics) useR! 2004, Vienna (May 2004).'

domain knowledge, while an analysis that seeks to identify correlations requires only data. As one big-data bestseller (Mayer-Schönberger & Cukier, 2013) summarizes: correlation analysis is quicker and cheaper than causal methods. Moreover, the causal relationship does not deepen our understanding of the world. These sentences may remind traditionally trained sociologists of the threat of identifying spurious (only seemingly existing) correlations. Indeed, spurious correlation is one of those methodological pitfalls that is listed when speaking about the analytical features of digital data (see Gandomi's and Haider's prolifically cited paper from 2015). Spurious correlation may also be present in an NLP analysis. For example, we can detect it in the results of the Hedonometer that older bloggers are less happy, or the happiness of music lyrics decreases over time. A potential confounder is language usage, which changes by age group and period. Sociologists interested in causal explanations routinely try to operationalize and adjust for potential confounders. However, analysts often do not have access to such background variables, including demographic variables, when they use digital data. A perspective solution is to extract this background information from the text by identifying document covariates (e.g., masculine or feminine style).

Beyond the 'control variable' approach, other causal approaches may be applied in the field of NLP. For example, follow-up studies yield stronger causal evidence than repeated cross-sectional ones. However, their sociological analogues – panel studies – are expensive and burdened by the dropout rate (panel attrition). Digital data include the needed temporal dimension, making them more available for causal analyses. For example, Barberá et al. (2014) employed Granger causality (a temporal causal framework for determining whether one time series is useful in forecasting another) to analyze whether members of Congress are more likely to lead or follow their constituents on political issues. The authors analyzed Twitter messages of legislators and the public during the 113th US Congress. Hedonometer data could be used in a similar way to investigate the relationship between spatial aggregates (e.g. unemployment) and the happiness levels of people living in the same area. However, note that only aggregate data can be analyzed based on the Hedonometer, as we cannot link Twitter data with unemployment data at the individual level. In contrast, the following paragraph provides an example of extracting causal relationships at the individual level using the Hedonometer.

In a randomized controlled experiment, the gold standard of causality can also be applied in an online context. One example is the famous large-scale Facebook experiment (Kramer et al., 2014), where authors found causal evidence for emotional contagion. Another example is described in Walther et al. (2008), who examined observers' impressions of profile owners by manipulating what other users post on their profiles. Using the Hedonometer also represents an opportunity to experiment, and the easiest way to do this is by natural experiment when experimental and control conditions are determined by nature or by other factors outside the control of the investigators. For example, a fascinating question is whether environment or culture has a greater influence on happiness. Following Twitter users for a longer period could help identify whether the country of birth or country of later move is a more important determinant of happiness, whether moving later in life plays a lesser role, and how long one has to live in a new country for its effects on happiness to be felt.

5 Incorporating traditional quantitative methodology into the field of automated text analysis

Traditional quantitative sociological methodology could also enrich NLP research. Earlier, we wrote about including explanations (and causality) in research goals. In contrast to DiMaggio (2015, p. 4), who states that '[e]ngagement with computational text analysis […] requires social scientists to relax some of our own disciplinary biases, such as our preoccupation with causality', we believe that descriptions and predictive models alone are not sufficient to satisfy the interest of social scientists, thus the latter should make efforts to create explanatory models with causal assumptions. (For methodological solutions that achieve this aim, see our earlier recommendations.)

A paradigmatic methodological feature of quantitative social science is its deep-rooted connection with statistics. Applied statistics and social sciences evolved closely and in parallel, whereby the latter promoted the development of new statistical models for addressing substantive sociological questions (see, e.g., the successive generations of intergenerational social mobility research [Ganzeboom et al., 1991]). Sociologists and data scientists know about different facets of statistics, thus, their cooperation could be inspiring. Additionally, quantitative sociologists are trained in a broader context of social data analysis to discover domain-related knowledge. For example (as mentioned earlier), the century-long sociological reflection on representation and measurement errors can be directly implemented into computational social science. A further example is the group of classic multivariate methods that are frequently used in sociology. Their straightforward geometric interpretation (such as through the different variants of cluster analysis, principal component analysis, factor analysis, and multidimensional scaling) may provide new insight into the study of NLP methods such as word embedding vector spaces.

6 Summary

The digitalization of society and new methods of social research are together creating fundamental changes in science. In this paper, we gave an overview of the social scientific context of NLP, review the main opportunities associated with this, and discuss the challenges of integrating the new generational toolbox into sociology.

One of the challenges is epistemological in nature: digital data are not products of a traditional operationalization process but are like 'digital footprints', which raises questions related to their objectivity and validity. More generally, the challenge of the new data and methods can be formulated as a new research paradigm: new analytical methods are optimized for prediction, while explanation and causation – which are the focus of sociology – are somewhat outside their scope. The new, data-driven approach often approximates causality with simple correlation.

Our point is that the contrast between the old and the new approaches is neither antagonistic nor paradigmatic. On the contrary, predictive and causal approaches can give answers to different research questions, and the approaches may be complementary.

Based on our earlier argument, integrating sociological knowledge into NLP enhances our understanding and makes it possible to broaden sociological knowledge by discovering new insights. Numbers do not speak for themselves, whatever complex analytical methods we use. It is the researcher who makes the calls 'behind' the numbers when making decisions at each step of the analysis and when interpreting the results. The depth and reliability of new insights depend on the domain-specific knowledge applied throughout the whole analytical process. Sociological expertise is required to formulate a research question, select the proper corpora, understand its formation and context, select and specify data preparation and pre-processing procedures, evaluate model validity, interpret the results, and position the interpretation in the scientific discourse. Without a theoretical framework, our model may not contribute to understanding the problem being investigated. As the above-mentioned case of the Netflix Prize also shows, solely searching for patterns is no more than 'data fishing', which is of only slight relevance in sociology and data science. However, NLP analyses which attempt to interpret models and adapt methods that provide stronger causal evidence can lead to correct results and more effective ways to build sociological knowledge.

Turning to the institutional context, several research projects that used digital social data have been implemented by data scientists rather than sociologists. However, the discovery of new knowledge regarding society also needs sociological knowledge, which makes dialogue between the two disciplines necessary. The hope is that these methods and perspectives will be organically integrated into sociology, which requires much more than 'picking up' some new methods. New institutional conditions are needed that are built on a conscious process of institutional renewal (e.g., aimed at making the different interests of disciplines compatible). Quantitative sociology is undergoing a transformation from the work of single researchers who write the questionnaires and analyze the data alone to interdisciplinary collaborations. The new skills that are needed to analyze digital textual data should be added to the quantitative sociological curriculum in parallel with the expansion of pre-existing courses that cover issues related to digital data. If computational methods are more widely known and taught, sociology in general can take a huge step forward.

Acknowledgements

This work was supported by the Lendület 'Momentum' Program of the Hungarian Academy of Sciences to JK, and by NKFIH (National Research, Development and Innovation Office) grant K-134428 to NR. The funding sources had no involvement in conducting the research or the preparation of the article.

References

Amaya, A., Biemer, P. P. & Kinyon D. (2020). Total Error in a Big Data World: Adapting the TSE Framework to Big Data. *Journal of Survey Statistics and Methodology, 8*(1), 89–119. https://doi.org/10.1093/jssam/smz056

Anderson, C. (2008, June 23). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired.* https://www.wired.com/2008/06/pb-theory

Bales, R. F. (1950). A set of categories for the analysis of small group interaction. *American Sociological Review, 15*(2), 257–263. https://doi.org/10.2307/2086790

Barberá, P., Bonneau, R., Egan, P., Jost, J. T., Nagler, J. & Tucker, J. (2014). *Leaders or Followers? Measuring Political Responsiveness in the U.S. Congress Using Social Media Data.* Presented at the Annual Meeting of the American Political Science Association.

Barna, I. & Knap, Á. (2022). Analysis of the Thematic Structure and Discursive Framing in Articles about Trianon and the Holocaust in the Online Hungarian Press Using LDA Topic Modelling. *Nationalities Papers,* online first. https://doi.org/10.1017/nps.2021.67

Bielik, I. (2020). Application of natural language processing to the electoral manifestos of social democratic parties in Central Eastern European countries. *Politics in Central Europe, 16*(1), 259–282. https://doi.org/10.2478/pce-2020-0012

Blei, D. M. & Lafferty, J. D. (2009). Topic Models. In A. Srivastava & M. Sahami (Eds.), *Text Mining: Classification, Clustering, and Applications* (pp. 71–93). Chapman & Hall; CRC Press.

boyd, danah & Crawford, K. (2012). Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Csomor, G., Simonovits, B. & Németh, R. (2021). Hivatali diszkrimináció? Egy online terepkísérlet eredményei [Discrimination at local governments? Results of an online field experiment]. *Szociológiai Szemle, 31*(1), 4–28. https://doi.org/10.51624/SzocSzemle.2021.1.1

Grant, B. & Lutz, C. (2017). Representativeness of Social Media in Great Britain: Investigating Facebook, LinkedIn, Twitter, Pinterest, Google+, and Instagram. *American Behavioral Scientist, 61*(7), 741–756. https://doi.org/10.1177/0002764217717559

Cheng, J., Danescu-Niculescu-Mizil, C. & Leskovec, J. (2015). Antisocial Behavior in Online Discussion Communities. *Proceedings of the International AAAI Conference on Web and Social Media, 9*(1). https://doi.org/10.1609/icwsm.v9i1.14583

DiMaggio, P. (2015). Adapting computational text analysis to social science (and vice versa). *Big Data & Society, 2*(2), 2053951715602908. https://doi.org/10.1177/2053951715602908

Evans, J. A. & Aceves, P. (2016). Machine Translation: Mining Text for Social Theory. *Annual Review of Sociology, 42*(1), 21–50. https://doi.org/10.1146/annurev-soc-081715-074206

Feder, A., Keith, K. A., Manzoor, E., Pryzant, R., Sridhar, D., Wood-Doughty, Z., Eisenstein, J., Grimmer, J., Reichart, R., Roberts, M. E., Stewart, B. M., Veitch, V. & Yang, D. (2022). Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics, 10*, 1138–1158. https://doi.org/10.1162/tacl_a_00511

Florio, K., Basile, V., Lai, M. & Patti V. (2019). Leveraging Hate Speech Detection to Investigate Immigration-related Phenomena in Italy, *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, Cambridge, United Kingdom. https://doi.org/10.1109/ACIIW.2019.8925079

Gandomi, A. & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Ganzeboom, Harry B. G., Treiman, D. J. & Ultee, W. C. (1991). Comparative Intergenerational Stratification Research: Three Generations and Beyond. *Annual Review of Sociology, 17*, 277–302. https://doi.org/10.1146/annurev.so.17.080191.001425

Hays, D. C. (1960). *Automatic Content Analysis.* Santa Monica, CA, Rand Corp.

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

Ignatow, G. (2016). Theoretical Foundations for Digital Text Analysis. *Journal for the Theory of Social Behaviour, 46*(1), 104–120. https://doi.org/10.1111/jtsb.12086

Ignatow, G. & Mihalcea, R. F. (2017). *An Introduction to Text Mining: Research Design, Data Collection, and Analysis.* Sage.

Jager, W., Abramczuk, K., Komendant-Brodowska, A., Baczko-Dombi, A., Fecher, B., Sokolovska, N. & Spits, T. (2020). Looking into the educational mirror: why computation is hardly being taught in the social sciences, and what to do about it. In H. Verhagen, M. Borit, G. Bravo & N. Wijermans (Eds.), *Advances in Social Simulation* (pp. 239–245). Springer. https://doi.org/10.1007/978-3-030-34127-5_22

Jones, J. J., Amin, M. R., Kim, J. & Skiena, S. (2020). Stereotypical Gender Associations in Language Have Decreased Over Time. *Sociological Science, 7,* 1–35. https://doi.org/10.15195/v7.a1

Katona, E., Kmetty, Z. & Németh, R. (2021). A korrupció hazai online médiareprezentációjának vizsgálata természetes nyelvfeldolgozással. *Médiakutató, 22*(2), 69–88.

Khan, F. A. & Abubakar, A. (2020). Machine translation in natural language processing by implementing artificial neural network modelling techniques: An analysis. *International Journal on Perceptive and Cognitive Computing, 6*(1), 9–18.

King, G. (2014). Restructuring the Social Sciences: Reflections from Harvard's Institute for Quantitative Social Science. *PS: Political Science and Politics*, *47*(1), 165–172. https://doi.org/10.1017/S1049096513001534

Kleinbaum, D. G., Kupper, L. L., Nizam, A., Muller, K. E., Curns, A. T. & Nizam, Z. G. (2008). *Applied regression analysis and other multivariable methods: Student solutions manual.* Duxbury.

Kmetty Z. & Knap Á. (2022). Trágárság mint érzelmi válasz a COVID-19-járvány idején [Obscenity as emotional response during the COVID-19 pandemic]. In G. Szabó (Ed.), *Érzelmek és járványpolitizálás. Politikai érzelemmenedzserek és érzelemszabályozási ajánlataik Magyarországon a COVID-19 pandémia idején* (pp. 173–190). ELTE Eötvös Kiadó.

Kramer, A. D. I., Guillory, J. E. & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790. https://doi.org/10.1073/pnas.1320040111

Latour, B. (2010). Tarde's Idea of Quantification. In M. Candea (Ed.), *The Social after Gabriel Tarde: Debates and Assessments* (pp. 145–162). Routledge.

Lewis, K. (2015). Three fallacies of digital footprints. *Big Data & Society*, *2*(2), 2053951715602496. https://doi.org/10.1177/2053951715602496

Mayer-Schönberger, V. & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

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

Metzler, K., Kim, D. A., Allum, N. & Denman, A. (2016). *Who is doing computational social science? Trends in big data research (White paper)*. SAGE Publishing. https://doi.org/10.4135/wp160926

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Proceedings of the Conference on Advances in Neural Information Processing Systems, 2,* 3111–3119. https://doi.org/10.48550/arXiv.1310.4546

Miraj, R. & Aono, M. (2021). Combining BERT and Multiple Embedding Methods with the Deep Neural Network for Humor Detection. In H. Qiu, C. Zhang, Z. Fei, M. Qiu & S.-Y. Kung (Eds.), *Knowledge Science, Engineering and Management: 14th International Conference, KSEM 2021 Tokyo, Japan, August 14–16, 2021 Proceedings,* Part III (pp. 53–61). Springer. https://doi.org/10.1007/978-3-030-82153-1_5

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

Molnar, C. (2019). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (eBook). Morrisville, Lulu.

Mützel, S. (2015). Facing Big Data: Making sociology relevant. *Big Data & Society*, *2*(2), 205395171559917. https://doi.org/10.1177/2053951715599179

Németh, R. & Koltai, J. (2021). The potential of automated text analytics in social knowledge building. In T. Rudas & G. Péli (Eds.), *Pathways Between Social Science and Computational Social Science: Theories, Methods, and Interpretations* (pp. 49–70). Springer. https://doi.org/10.1007/978-3-030-54936-7_3

Parks, L. & Peters, W. (2022). Natural Language Processing in Mixed-methods Text Analysis: A Workflow Approach. *International Journal of Social Research Methodology,* online first. https://doi.org/10.1080/13645579.2021.2018905

Pigliucci, M. (2009). The end of theory in science? *EMBO reports*, *10*(6), 534–534. https://doi.org/10.1038/embor.2009.111

Reinsel, D., Gantz, J. & Rydning J. (2018). *The Digitization of the World From Edge to Core.* An International Data Corporation Whitepaper. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

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

Savage, M. & Burrows, R. (2007). The Coming Crisis of Empirical Sociology. *Sociology*, *41*, 885–899. https://doi.org/10.1177/0038038507080443

Shah, A. H. (2019). How episodic frames gave way to thematic frames over time: A topic modeling study of the Indian media's reporting of rape post the 2012 Delhi gang-rape. *Poetics, 72*, 54–69. https://doi.org/10.1016/j.poetic.2018.12.001

Shaw, R. (2015). Big Data and reality. *Big Data & Society*, *2*(2), 2053951715608877. https://doi.org/10.1177/2053951715608877

Sterling, J., Jost, J. T. & Hardin, C. D. (2019). Liberal and Conservative Representations of the Good Society: A (Social) Structural Topic Modeling Approach. *SAGE Open, 9*(2), https://doi.org/10.1177/2158244019846211

Szabó, M. K., Ring, O., Nagy, B., Kiss, L., Koltai, J., Berend, G., Vidács, L., Gulyás, A. & Kmetty, Z. (2021). Exploring the dynamic changes of key concepts of the Hungarian socialist era with natural language processing methods. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 54*(1), 1–13. https://doi.org/10.1080/01615440.2020.1823289

Tinati, R., Halford, S., Carr, L. & Pope, C. (2014). Big Data: Methodological Challenges and Approaches for Sociological Analysis. *Sociology*, *48*(4), 663–681. https://doi.org/10.1177/0038038513511561

Tiwari, V., Verma, L. K., Sharma, P., Jain, R. & Nagrath, P. (2021). Neural network and NLP based chatbot for answering COVID-19 queries. *International Journal of Intelligent Engineering Informatics, 9*(2), 161–175. https://doi.org/10.1504/IJIEI.2021.10040085

Törnberg, P. & Törnberg, A. (2018). The limits of computation: A philosophical critique of contemporary Big Data research. *Big Data & Society*, *5*(2), 2053951718811843. https://doi.org/10.1177/2053951718811843

Töscher, A., Jahrer, M. & Bell, R. M. (2009). *The BigChaos Solution to the Netflix Grand Prize*. Technical Report. commendo research & consulting. https://www.asc.ohio-state.edu/statistics/statgen/joul_aut2009/BigChaos.pdf

Tsakalidis, A., Papadopoulos, S., Cristea, A. I. & Kompatsiaris, Y. (2015). Predicting Elections for Multiple Countries Using Twitter and Polls. *IEEE Intelligent Systems*, *30*(2), 10–17. https://doi.org/10.1109/MIS.2015.17

Walther, J. B., Heide, B. V. D., Kim, S.-Y., Westerman, D. & Tong, S. T. (2008). The Role of Friends' Appearance and Behavior on Evaluations of Individuals on Facebook: Are We Known by the Company We Keep? *Human Communication Research*, *34*(1), 28–49. https://doi.org/10.1111/j.1468-2958.2007.00312.x

Watts., D. J. (2013). Computational Social Science: Exciting Progress and Future Directions. *The Bridge*, *43*(4), 5–10.

White, M. (2009, August 22). Networks Are Killing Science. *Science 2.0.* https://www.science20.com/adaptive_complexity/networks_are_killing_science