2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8301

How to sort out uncategorisable documents for interpretive social science? On limits of currently employed text mining techniques

Philipps, Axel
Institute of Sociology & Leibniz Center for Science and Society (LCSS), Leibniz University Hanover, Germany.

Abstract

Current text mining applications statistically work on the basis of linguistic models and theories and certain parameter settings. This enables researchers to classify, group and rank a large textual corpus – a useful feature for scholars who study all forms of written text. However, these underlying conditions differ in respect to the way how interpretively-oriented social scientists approach textual data. They aim to understand the meaning of text by heuristically using known categorisations, concepts and other formal methods. More importantly, they are primarily interested in documents that are incomprehensible with our current knowledge because these documents offer a chance to formulate new empirically-grounded typifications, hypotheses, and theories. In this paper, therefore, I propose for a text mining technique with different aims and procedures. It includes a shift away from methods of grouping and clustering the whole text corpus to a process that sorts out uncategorisable documents. Such an approach will be demonstrated using a simple example. While more elaborate text mining techniques might become tools for more complex tasks, the given example just presents the essence of a possible working principle. As such, it supports social inquiries that search for and examine unfamiliar patterns and regularities.

Keywords: *text mining; interpretive social science; qualitative research; standardised and non-standardised methods; social science.*

1. Introduction

Before starting to answer the title of the paper, the exact nature of text mining needs to be identified. Text mining is a combination of statistical and linguistic approaches of text analysis that has lately gained attention in the field of digital humanities. An important forerunner was the Italian literary scholar Franco Moretti (2007) with his concept of "distant reading". He proposed that scholars who are used to employing in-depth interpretations (close reading) are unable to read and study the ever-increasing amount of data that is produced worldwide. Because of this, he recommends a different approach. In contrast to printed books, Moretti accesses digitally-accessible texts and identifies patterns in large corpora. This kind of distant reading includes a growing number of visualisations such as maps, graphs, and trees (Jänicke et al., 2015). Such visualisations usually show relations between such things as actors, names and places; text mining tools, in contrast, concentrate on linguistically small units: words and phrases. Text mining can be defined as a set of "computer-based methods for a semantic analysis of text that help to automatically, or semi-automatically, structure text, particular very large amounts of text" (Heyer, 2009: 2). So, such applications practically count, relate, rank, cluster, and classify single and groups of words in large text corpora and present the outcomes in frequency graphs, word clusters, and networks.

In recent years there has also been a growing interest in text mining for social science research. Various works (i.e. DiMaggio, Nag & Blei, 2013; Marres 2017; Philipps, Zerr & Herder, 2017) present mostly exploratory studies using algorithmic information extraction approaches to demonstrate the power of such tools for text analysis in the social sciences. Proponents of these computer-based methods primarily address qualitatively-oriented social scientists for two reasons (i.e. Evans & Aceves, 2016; Wiedemann, 2013). Firstly, such tools help researchers, who mainly work with textual data, to deal with the increasing number of digitally-accessible texts. Secondly, it is argued that, in a similar way to the grounded theory approach (Glaser & Strauss, 1967), text mining is employed to identify patterns. However, these propositions are slightly misleading. This is a rather unbalanced representation of qualitative and interpretive social research and might explain, to some extent, why (semi)-automatic analysis of textual data has, up to now, been widely ignored in interpretive social sciences (for more details see Philipps, 2018).

This paper therefore primarily takes a closer look at how text mining analyses textual data and in what respect that analysis differs from methods commonly employed by interpretively-oriented social scientists. In this respect, I suggest a different aim and operating procedure for text mining which is more appropriate for interpretive social science. It includes a shift from standardised procedures of classification and clustering of large text corpora to detecting documents that do not fit to applied constructed concepts. To demonstrate this approach, I am presenting an exemplary working principle of low

complexity. Later, adapted text mining techniques might become tools for more complex tasks. These seek to support interpretive social science that examine unfamiliar patterns and the regularities of socially-produced meanings.

2. Analysing textual data with text mining and in interpretive social science

Text mining techniques comprise of a wide range of methods from frequency and co-occurrence analysis to sentiment analysis and then to more complex approaches such as topic models and machine learning (Marres 2017; Wiedemann, 2016, 2013). While frequency and co-occurrence counts and identifies the use of words and the relationship between groups of words in large text corpora topic models, machine learning transforms words into numbers and computes statistical interferences in textual data. By no means can these methods be successfully employed to detect thematic shifts or networks of knowledge structures on a trans-textual level in social research studies (i.e. Adam & Roscigno, 2005; Blei & Lafferty, 2006). However, applying text mining requires the setting of some parameters before research is started. For frequency and co-occurrence analyses, for example, researchers need to determine relevant words or groups of words in advance. For a sentiment analysis they have to define classes, ranging from extremely negative to extremely positive. In addition, most machine-learning algorithms demand supervised training (intermediate results are controlled and evaluated by analysts during processing) and even for unsupervised topic models (without interference of external data or human control) researchers have to determine the exact number of clusters to be computed. Hence, current text mining methods have certain characteristics in common; before analysis, researchers define, even to the smallest degree, what is relevant and can potentially be found in textual data. Based on these (standardised) parameter settings, whole text corpora are classified, ranked, or grouped.

However, standardised approaches are, for a great deal of interpretively-oriented social scientists, the opposite to how they were trained. For the most part, they learned and share the basic premise of interpretive social science working with non-standardised methods. This means that a researcher should approach their object of investigation with an open mind and be prepared for surprises. Hence, these researchers seek to situationally understand meanings produced in interactional settings – being ready to overcome previous classifications and schemes. They aim to generate assumptions based on identified content-related, functional and formal aspects of the examined empirical material (for more details see Soeffner, 1999). Nonetheless, while these interpretively-oriented social researchers avoid standardised settings, they employ heuristic models to interpret textual data. They work with commonly-known (scientific) classifications and typifications in order to see how useful this knowledge is for understanding the meaning of given textual data and, at

the same time, they search for unfamiliar regularities and patterns. Thus, these researchers translate and describe the world of the observed "into one that we find comprehensible" (Abbott, 2004: 31) and only if they discover so far incomprehensible phenomena do they seek to grasp the underlying working principle and meaning in the form of new but empirically-grounded typifications, hypotheses, and theories.

Against this background, I presume that currently-operating text mining applications for classification and information extraction are often insufficient to be "complementing techniques" (Wiedemann, 2013: no page) for most social scientists with special training in interpretive methods. Under certain circumstances text mining might enable qualitatively-oriented researchers to learn about the variety and development of relevant categories. It is also reasonable to assume that machine learning algorithms which demonstrate knowledge about statistical characteristics of language and text-external knowledge manually coded by analysts (e.g. categories or example sets) will help to retrieve or annotate information in unknown material. However, in all these cases text mining is used to classify and group the entire textual data based on determined parameter settings. We therefore need to think of additional text mining strategies more adjusted to interpretative social science and its basic premise.

3. Adjusting text mining for interpretive social science

Text mining applications might become more relevant for interpretive social science, I suppose, if they enable researchers to divide a large corpus of documents into those with and without comprehensible patterns and components. Such information will stimulate the power of interpretive social inquiry, interpretively explore hidden patterns and unveil unfamiliar meaning. The working principle of such a search strategy might be best described with Max Weber's (1949) limiting concept of ideal types: "It is a conceptual construct (*Gedankenbild*) which is neither historical reality nor even the 'true' reality. It is even less fitted to serve as a schema under which a real situation or action is to be subsumed as one *instance*. It has the significance of a purely ideal *limiting* concept with which the real situation or action is *compared* and surveyed for the explication of certain of its significant components" (Weber, 1949: 93, italics in the original work). Thus, ideal types are not the final outcome of empirical investigations but are used as an heuristic limiting concept to identify the significant aspect of real situations or actions. Practically, if an ideal type has not fully-grasped all aspects of the social phenomena, the researcher will pay full attention to this and mark it for further interpretation. In Weber's book *Economy and Society* (2013) he, for example, applied ideal types in a "procedure of the 'imaginary experiment'" (10) comparing a purely rational constructed course of actions with the concrete course of events: "By comparison with this it is possible to understand the ways in

which actual action is influenced by irrational factors of all sorts, such as affects and errors, in that they account for the deviation from the line of conduct which would be expected on the hypothesis that the action were purely rational" (Weber 2013: 6). Thus, he intellectually constructs an ideal type of pure rationality to grasp favouring or hindering circumstances which are devoid of subjective meaning "if they cannot be related to action in the role of means or ends" (Weber 2013: 7). Generally speaking, with ideal types as limiting concepts he describes a common strategy among interpretive social scientists to approach their object of investigation in that one employs conceptual constructs to understand social phenomena and by paying attention to unfamiliar regularities and patterns (in Weber's terms: deviations). The latter phenomena are of special interest because their interpretation offers a chance to broaden or even to rewrite established scientific knowledge. However, one has to note that Weber was interested in understanding and explaining social action motivationally. The construction of ideal types thus is not restricted to a rational course of actions.

Applying this search strategy to text mining, a modified variant might become central for interpetative social research working with large digitally-accessible text corpora. In contrast to currently operating mining techniques which classify and group an entire text corpus, an adaptation would use constructed concepts to identify documents which show characteristics assumed in the formulated concept and those that do not fit. Therefore, in contrast to present computer-based applications working with linguistic models and theories, an adjusted text mining technique would operate with preliminary ideas and assumptions, formulated by interpretively-oriented researchers. In particular, for a large corpus of documents the latter will come up with a constructed concept after analysing some selected documents and heuristically employ this to sort out documents that display conceptually anticipated features and relations. In the next step, researchers examine and interpret the specificity of the remaining documents. In this process they might adjust existing concepts or formulate others.

In addition, from the perspective of the humanities one could also say such a modified text mining technique mimics the hermeneutic circle (see Gadamer, 2004). Suggestions formulated in a first round of interpreting textual data are used to identify what is comprehensible and what is not. Incomprehensible textual data will be analysed in further interpretive rounds producing altered or additional suggestions which become the basis for more interpretive sequences. The process will come to an end with working interpretations (constructed concepts) to understand the textual data of interest. Nonetheless, like the hermeneutic circle the process will be impossible to finish as other researchers might find more appropriate readings for understanding certain textual data in the future.

4. An example for sorting out uncategorisable documents

Often interpretively-oriented social scientists work with and interpret a small number of documents. However, sometimes they are confronted with a large corpus of textual data such as an archive of interview transcripts, protocols, letters and other forms of written documents. There are various ways of dealing with such conditions. With Merkens (2004), one might select some documents according to specific characteristics (i.e. relevant for the research goal) and concentrate on these cases or apply the theoretical sampling strategy starting with a few documents and selecting further documents for interpretation based on minimal and maximal contrasts. Theoretical sampling comes to an end if additional analyses of documents reveal no further information. However, there always remain documents that are not interpreted and may contain unexplored patterns and meanings. Under such circumstances an adapted version of text mining technique would offer an opportunity to search these documents for deeper analyses.

In the following paragraphs, I present an instance of low complexity to give an idea of how such an variant of text mining can support interpretively-oriented research projects. It does not involve a reprogrammed text mining application but rather it demonstrates a possible working principle. The case in point is an investigation of applied approaches to promote unconventional ideas in 93 grant proposals sent to a major research-funding organisation in Germany in 2013 (for more detail on method and findings see Philipps, forthcoming). The study started by skimming through the textual data and selecting proposals for deeper analysis. Without any predefined assumptions about specific approaches to unconventional ideas, I began to read a number of grant proposals to get an idea of these. Based on a preliminary impression of the material, I then employed closer readings in a contrastive manner. Using maximal and minimal contrast cases, I searched for specific structural and rhetorical patterns in the rationales of the grant proposals. My interpretation of research proposals continued until typical approaches to unconventional ideas could be identified and separated. After scrutinising 20 proposals and skimming through further applications I came up with a typology of distinct approaches. In an additional and laborious step the typology of identified argumentative patterns was separated into segments and described in a codebook. After a group of interpreters applied segment descriptions to a randomly selected sample of proposals and discussed disagreements and questions, amended codes were used by the author to annotate all 93 research grant proposals. Finally, the manual coding process enabled us to categorise all documents and search for cases with different argumentative patterns or other aspects.

Especially for studies with a greater corpus, automatic text mining would be another option searching for empirically-identified patterns before establishing a codebook and manually annotating the remaining documents. However, such a search strategy requires a limiting concept to sort out documents that show conceptually-suggested patterns and those that do

not. In my study, such a concept might, for example, be typical wordings that appear with the identified approaches. Applicants who promoted ideas of solving practical problems typically discussed "drawbacks" or "disadvantages" of earlier solutions and what "benefit" or "advantage" their solution offers in contrast. Concentrating on these wordings, of course, is one-sided and does not fully capture all possible variants and other typical aspects. However, by producing two groups of documents (with and without these certain wordings), one can reduce the number of proposals demanding deeper analysis. In the case of this research project, a simple retrieval of these terms shows that 48 grant proposals used at least one of the terms if not all of them. Combining this result with the already examined proposals (n=20) 33 uncategoriseable documents remain. Hence, this procedure already condenses the number of non-examined documents from 73 down to 33. Apart from applying additional limiting concepts to further reduce the amount of these documents it should be clear that such a search strategy assists interpretively-oriented social scientists to single out documents for further examination.

5. Conclusion

In this paper, I discussed how standardising procedures of current text mining techniques differs in respect of methodological premises commonly employed by interpretively-oriented social scientists. Without question, text mining features such as ranking, grouping or classifying textual data are useful for many research questions in social sciences. However, I presume an adjusted mining techique will greatly support interpretive social science if it shifts from standardised procedures of classification and the clustering of large text corpora to detect documents that do not fit into applied constructed concepts. It is also important to note that such a mining technique would not be based on linguistic theories and information management concepts but on suggestions offered by interpretively-oriented social scientists. As demonstrated at a low level, such an approach can help interpretive social inquiries to single out documents and examine them for unfamiliar patterns and regularities of socially-produced meanings. Nonetheless, as the complex topic of this paper shows it is still a long way from translating the methodological premises of interpretive social sciences into working additional text mining techniques.

References

Abbott, A. (2004). *Methods for Discovery. Heuristics for the Social Sciences*. New York: W. W. Norton & Company, Inc.

Adams, J., & Roscigno, V. J. (2005). White Supremacists, Oppositional Culture and the World Wide Web. *Social Forces*, 84(2), 759–778.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic Topic Models. *Proceedings of the 23rd International Conference on Machine Learning, ACM*, 113–120.

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting Affinities Between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Art Fundings. *Poetics*, 41(6), 570–606.

Evans, J. A., & Aceves, P. (2016). Machine Translation: Mining Text for Social Theory. *Annual Review of Sociology*, 42, 21–50.

Gadamer, H.-G. (2004). *Truth and Method*. 2nd rev. ed. Trans. J. Weinsheimer & D. G. Marshall. New York: Crossroad.

Glaser, B. G., & Strauss, A. L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. New Brunswick & London: AldineTransaction.

Heyer, G. (2009). Introduction to TMS 2009. In G. Heyer (Ed.), *Text Mining Services. Building and Applying Text Mining Based Service Infrastructures in Research and Industry. Proceedings of the Conference on Text Mining Services 2009 at Leipzig University* (pp.1–14). Leipzig: LIV.

Jänicke, S., Franzini, G., Cheema, M. F., & Scheuermann, G. (2015). On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. In R. Borgo, F. Ganovelli & I. Viola (Eds.), *Eurographics Conference on Visualization (EuroVis)-STARs.* The Eurographics Association.

Marres, N. (2017). *Digital Sociology: The Reinvention of Social Research*. Hoboken: John Wiley & Sons.

Merkens, H. (2004). Selection Procedures, Sampling, Case Construction. In U. Flick, E. von Kardoff, & I. Steinke (Eds.), *A Companion to Qualitative Research* (pp. 165–171). London: Sage.

Moretti, F. (2007). *Graphs, Maps, Trees. Abstract Models for Literary History*. London: Verso.

Philipps, A. (forthcoming). Wissenschaftliche Orientierungen. Empirische Rekonstruktionen an einer Ressortforschungseinrichtung. München & Weinheim: Juventa.

Philipps, A. (2018). Text Mining-Verfahren als Herausforderung für die rekonstruktive Sozialforschung. *Sozialer Sinn*. *Zeitschrift für hermeneutische Forschung*, 19(1): 191–210.

Philipps, A., Zerr, S., & Herder, E. (2017). The Representation of Street Art on Flickr. Studying Reception with Visual Content Analysis. *Visual Studies*, 32(4), 382–393.

Soeffner, H.-G. (1999). Verstehende Soziologie und sozialwissenschaftliche Hermeneutik. In R. Hitzler, J. Reichertz, & N. Schröer (Eds.), *Hermeneutische Wissenssoziologie* (pp. 39-49). Konstanz: UVK.

Weber, M. (2013). *Economy and Society: An Outline of Interpretive Sociology*. (Translation of *Wirtschaft und Gesellschaft*, 4th ed., 1956). Berkeley, Los Angeles, & London: University of California Press.

Weber, M. (1949). Objectivity in Social Science and Social Policy. (Translation of *Die 'Objektivität' sozialwissenschaftlicher Erkenntnis*, 1904) In E. Shils, & H. Finch (Eds.), *The Methodology of the Social Sciences* (pp. 49–112). Glencoe: The Free Press.

Wiedemann, G. (2016). *Text Mining for Qualitative Data Analysis in the Social Sciences: A Study on Democratic Discourse in Germany*. Wiesbaden: Springer.

Wiedemann, G. (2013). Opening up to Big Data: Computer-Assisted Analysis of Textual Data in Social Sciences [54 paragraphs], in: *Forum Qualitative Sozialforschung/Forum Qualitative Research*, 14, Art. 13, http://nbn-resolving.de/urn:nbn:de:0114-fqs1302231.