Chapter 6: Supervised and Unsupervised: Approaches to Machine Learning for Textual Entities

Tobias Hodel, University of Bern

Abstract

Applications that feed text into machine learning algorithms have existed for more than a decade. But it took multiple developments to make machine learning an exciting methodological approach to questions grounded in the humanities. The latest developments in handwritten text recognition (HTR) show the capabilities of supervised deep learning. However, the success of the technology comes with a price: It generates a set of methods that are complicated to grasp in theory and difficult to train algorithms in, algorithms that are not comprehensible to humans at all. By focusing on the two most frequently used approaches in machine learning (unsupervised and supervised), this paper lays out ways to critically use machine learning algorithms in the humanities. At the same time, we argue that these approaches help us to understand the epistemological assumptions of our disciplines and our methods.

Topic modeling used on large corpora of text leads to new insights into what topics occur, as well as the tendencies of a corpus. The approach uses unsupervised machine learning, through which a set of algorithms identify what words appear together frequently and so might indicate a topic. Topic modeling puts scholars at the end of the process, where they must still interpret the output of the algorithms.

In deciphering handwriting, supervised deep learning approaches have led to astonishing results, but also to new problems induced by the algorithm. The algorithm tries to adapt to the desired output, raising epistemological questions about transcribing and transliterating. The scholar is only able to alter the input, not how the algorithm manipulates it.

Based on these two examples, this paper promises a deeper understanding of a technology that is currently remodeling the way we do our research and that will increasingly intervene in our scholarship and even our daily lives in the future.

1. Machine Learning and the Humanities

It is pretty safe to assume that, in retrospect and from a computer-historical perspective, the 2010s and most probably also the 2020s will be seen as the era of the application of Artificial Intelligence (AI), or, more precisely, machine learning.[1]

From a scholarly standpoint, the arrival of AI has not, on the surface, led to a complete rethinking of research activities. Instead, established procedures have been altered slowly and sometimes impalpably. To search for literature, scholars have relied on finding aids in libraries for more than five centuries.[2] With search engines and a wide array of database infrastructures, however, the process of finding relevant primary and secondary sources is being completely changed. These changes have not been reflected in the products of our research: papers and books.[3]

The next phase of AI in the humanities is currently building up from within the humanities, as scholars have for some years now experimented with machine learning. With the advent of Tensorflow and other quite intuitively usable libraries that allow us to build deep and machine learning systems without higher degrees in engineering, we find ourselves in the first wave of applications and solutions that provide scholars with algorithms that are, at least to some degree, 'self-taught'.[4]

The growth of machine learning in the humanities disciplines has been fueled by the promises of AI, including the automatic recognition, identification and linking of text, entities, and even concepts. Algorithms that could meet these challenges not only benefit the shareholder value of companies using such tools to create adapted advertisements, but also scholars working on recently edited or digitized texts and corpora built on the web.

Although the subjects of AI and machine learning are often mentioned, we nonetheless lack much discussion of their methodological implications. We can tie

---

1 The author would like to thank the editor and the audience of the conference for its invaluable feedback. Thanks are as well due to Jake Purcell for lending his critical eye as well as his language skills. This article has benefited from work done within the READ project. READ has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement no. 674943.

2 Ann Blair, *Too Much to Know: Managing Scholarly Information before the Modern Age*,NewHaven 2010.

3 Concerning the looming overabundance of sources, see: Roy Rosenzweig, Clio Wired: The Future of the Past in the Digital Age, New York, 2011 3-27 (chapter Scarcity and Abundance? Preserving the Past). A comprehensive piece about using search engines in the humanities still needs to be written or I am just not aware of it.

4 For a broad and general introduction see Christof Schöch, Quantitative Analyse, in: Fotis Jannidis/Hubertus Kohle/Malte Rehbein (eds.), *Digital Humanities: Eine Einführung*, 279-298, Stuttgart 2017, doi:10.1007/978-3-476-05446-3_20. Tensorflow is an open-source library for machine learning: https://www.tensorflow.org/ [last accessed: April 2, 2021].

this absence to the complexity of the technology involved, as well as to the difference between machine learning and typical algorithms that follow strict orders and can be controlled much more easily. The capabilities of machine learning coupled with this black box of uncontrolled (or hard-to-control) parts leads to excitement, but also—and rightly so—to an unease about its widespread use in any discipline.

It's neither possible nor desirable to address all the problems and challenges posed by machine learning in one paper, so I will shed a light on the inner workings of this black box and put two approaches in perspective against the broader field of machine learning. In the upcoming years, we as scholars need to be able to actively engage in discussions about the potential and problems of AI at its current stage of implementation.

The paper will tackle two different approaches that fall under the umbrella of machine learning. In order to understand the differences, we will first briefly introduce how machine learning can be experienced and what trajectories are expected in the area of text analysis in the near future. From this theoretical and technical background, I will proceed to discuss two different applications of machine learning that show the gains and losses when using machine learning to analyze text. The goal of these two parts is not to showcase machine learning algorithms but to encounter methodological as well as epistemological consequences of the application of two very different forms of AI. Even before the linguistic turn, the humanities have been at the center of deliberations about what it means to "understand," and humanists have opted for hermeneutical rather than purely quantitative approaches.[5]

The paper tries to situate practices of the Digital Humanities in the quickly broadening field of critical algorithm studies. In addition to incorporating important questions brought up by recent research, I try not to focus on the analysis of implemented algorithms (like search engines)[6] or on cultural traits (like race),[7] but rather on the implementation of algorithmic solutions to problems grounded in the humanities. This approach is in line with developments within the digital humanities that emphasize the work of theorizing digital methods.[8]

---

5 On the hermeneutical method, see Hans-Georg Gadamer, *Hermeneutik I: Wahrheit Und Methode: Grundzüge Einer Philosophischen Hermeneutik*, Tübingen 2010.

6 Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York 2018.

7 Ruha Benjamin, Ruha, *Race After Technology: Abolitionist Tools for the New Jim Code*, Medford 2019.

8 Ted Underwood, Theorizing Research Practices We Forgot to Theorize Twenty Years Ago, *Representations* 127 (1/2014), 64–72, doi:10.1525/rep.2014.127.1.64.

2. About the "Intelligence" of Machines

A Google search for "artificial intelligence" yields, in the middle of 2020, more than 700 million hits. Starting with Wikipedia pages—in languages that depend on your location—and brief introductory videos from YouTube, the first pages of search results consist only of explanations of the term and its possible meanings. Although the concept of AI has been around for more than fifty years, only in the last five to ten years has it been more than a buzzword. Since the arrival of faster central processing units (CPUs) and the application of graphical processing units (GPUs, typically used to render 2D and 3D imagery) to machine learning, the term AI has become more tangible, due to its actual impact on applications and ensuing discussions about its consequences.[9]

As the Google search demonstrated by putting very broad introductions at the top of the result hierarchy, the general understanding of AI is sketchy and often far from what algorithms at the moment can actually achieve. In order to get an idea of machine learning capabilities, we need to take a small detour to briefly think about what intelligence means in the context of machines. At the same time, if we compare the impressions based on the introductions with actual results of algorithms, we will see that the current state opens up a tremendous amount of possibilities, but also that we still advance on a step-by-step basis and shouldn't overestimate the role of machines (yet). Algorithms remain in a state of being useful "helpers," and nothing more.

Although the humanities do not deal only with textual elements, a multitude of disciplines use text as one of their main foundations for furthering our knowledge within and across disciplines. The use of AI to recognize and sort or cluster text is thus a typical approach to using these technologies in the humanities. From a technical point of view, we differentiate three types of machine learning: supervised, unsupervised, and reinforcement learning.[10] The difference between the three types lies in the role of the optimization process. In supervised learning,

---

9 Discussions about "fair AI" are currently being brought forward in a wide variety of research centers, probably most notably in the AI Now institute at New York University: https://ainowinstitute.org/ [last accessed: April 2, 2021]. See also (in German): iRights lab et al., Praxisleitfaden zu den Algo. Rules - Orientierungshilfen für Entwickler:Innen und ihre Führungskräfte, Gütersloh 2020, doi:10.11586/2020029.

10 In this paper, I will not go into detail about the differences between classical machine learning and deep learning. Both approaches are addressed since topic modeling can be included in the former and handwritten text recognition to the latter. See also Ted Underwood/ Matthew L. Jockers, Text-Mining the Humanities, in: Susan Schreibman/ Ray Siemens/John Unsworth, *A New Companion to Digital Humanities*, Chichester 2016, 291–306. We do not agree with Underwood and Jockers that topic modeling belongs to text mining, as opposed to machine learning.

the algorithm aligns input with a human-determined outcome, and training processes try to get the algorithm's actual output as close as possible to the desired result. In unsupervised learning, the result is not predetermined and then imitated. Instead, algorithms search for patterns, similarities, and clusters. Finally, reinforcement learning evolves out of feedback (automatic or manual) and develops an algorithm on the fly.[11] Since reinforcement learning is currently not used in the humanities (at least not to my knowledge), I will not deal with this last approach in this paper.

For all three approaches, it is difficult to speak of machine intelligence as compared with the human capacity to learn and adapt to circumstances, social settings, languages, etc. In general, it would be easy to dismiss the notion that, at the current stage, it is advisable to talk about intelligence in regards to machines and algorithms at all.[12] At the same time, it's remarkable how trained algorithms can perform challenging tasks in a shorter time than humans.

All machine learning algorithms are based on their defined input and output; as a consequence, we can look at those two crucial parts of the process. Until about a year ago (2019), most machine learning processes started from scratch, and models were built from available data, called "Ground Truth" denominating what is correct with a certainty. In order to measure output from algorithms, computer scientists coined the term "Ground Truth" which determines a perfect or desired result. In supervised learning such data, the Ground Truth is used for the training as well as the validation process. Currently, this procedure is evolving, since certain models have been made available and can be used for refinement training or as base models.[13] The reuse of models will lead to further problems in producing algorithms, since users of pretrained models will have to deal with bias induced from training data not available to them. Currently, first experiments with pre-trained models are being conducted, and reliable statements cannot yet be made.

---

11 Best known is maybe the Super Mario algorithm MARI/O, that masters the well-known Nintendo game in a miraculous manner: URL: https://www.youtube.com/watch?v=qv6UVOQ0F44&t=115s [last accessed: April 2, 2021].

12 For an introduction, see the articles in the June 2020 issue of Wired, esp. Elizabeth Spelke, It's Called Artificial Intelligence—but What Is Intelligence?, in: *Wired* 19.05.2020, URL: https://www.wired.com/story/its-called-artificial-intelligence-but-what-is-intelligence/ [last accessed: April 2, 2021]; Kelly Clancy, Is the Brain a Useful Model for Artificial Intelligence?, in: *Wired* 19.05.2020, URL: https://www.wired.com/story/brain-model-artificial-intelligence/ [last accessed: April 2, 2021].

13 This is especially true for modern languages. Concerning language models, see Jacob Devlin et al., BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding, in: *arXiv* [preprint], 24.05.2019, https://arxiv.org/abs/1810.04805 [last accessed: April 2, 2021]; Tom B. Brown et al., Language Models Are Few-Shot Learners, in: *arXiv* [preprint], 22.07.2020, URL: https://arxiv.org/abs/2005.14165 [last accessed: April 2, 2021].

The crucial question remains: How can results be judged, apart from via statistical reports? This problem will only be addressed briefly in this paper and needs more elaboration in the future. Ways to qualify the output of machine learning will be a key issue, due to the influence of the technology not only on scholarly work with documents (as data), but also due to the embeddedness of machine learning in the algorithms of our daily life. Quantification of results using statistical techniques, such as the F1-score (a comparison of an algorithm's recall and precision) or percentages of correctly identified characters (if we think about text recognition), is one indication of the capability of an algorithm, but it doesn't show problems, uncertainties, or bias induced by the approach. F1-scores, for example, tell us about the quality and quantity of intended results, but say nothing about unintended consequences due to any imprecisions.

To highlight the differences among machine learning approaches, I will provide two examples that use different types of machine learning (supervised and unsupervised) and deal with questions for which machine learning yields impressive results. For the unsupervised approach, I will look at topic modeling, and for the supervised counterpart, the application of deep learning to the recognition of handwritten documents. The aim is only to introduce briefly the two approaches from a rather theoretical point of view, not to provide a how-to guide for the two methods.[14]

3. Topic Modeling: Unsupervised Clustering

One of the main advantages computers have over humans is their ability to count and compare extremely quickly. With topic modeling, scholars use these two traits and try to apply them to text. The algorithms used for topic modeling work in two directions: First, they count the appearance of strings (called "tokens") in textual entities (e.g. a letter or a document). The expected term for "token" might instead be "word," but since this term is polysemic and could mean a string of characters (a token) or the semantic meaning of the string (in a sense the lemmatized token), we will use token instead. Second, the tokens are compared to other strings appearing in the same entity. Pairs, triples… clusters of tokens appearing often across different entities are understood to belong to a distinct topic. In this perspective, a topic is nothing other than a collection of tokens appearing in context. Whether a

14 In order to get acquainted to the methods, we recommend, for topic modeling: Shawn Graham/Scott Weingart/Ian Milligan, Getting Started with Topic Modeling and MALLET, in: *Programming Historian* 02.09.2012, URL: https://programminghistorian.org/en/lessons/topic-modeling-and-mallet [last accessed: April 2, 2021], and for Handwritten Text Recognition, see URL: https://transkribus.eu/wiki/index.php/How_to_Guides [last accessed: April 2, 2021].

calculated "topic" is really congruent to the human understanding of a topic needs further discussion. The goal of the algorithm is basically to build an understanding of the rate of the occurrence of tokens in the same entity. In a third step, the process gets inverted, and for every category (or "topic"), the percentage of a textual entity that consists of this topic is calculated.

Fig 6.1: Screenshot of a section of a visual topic modeling output. Vertically on the left are the documents specified (signature, volume, document); horizontally are topics indicated. Topic 16: Warfare, topic 17: Finance, topic 18: Poverty, topic 19 Foreign policy. The color indicates the presence of a topic in a document. Screenshot by the author.

Signature_Volume_Doc	Topic 16	Topic 17	Topic 18	Topic 19	Topic 20	Topic 21	Topic 22	Topic 23
MM_2_100_RRB_1848_0528.	0.36	0.00	0.00	0.08	0.00	0.00	0.00	0.16
MM_2_100_RRB_1848_0830.	0.31	0.02	0.00	0.00	0.00	0.00	0.00	0.00
MM_2_100_RRB_1848_0926.	0.31	0.00	0.00	0.20	0.00	0.00	0.00	0.00
MM_2_101_RRB_1848_1131.	0.31	0.12	0.19	0.00	0.02	0.00	0.00	0.00
MM_2_101_RRB_1848_1594.	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MM_2_101_RRB_1848_1644.	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MM_2_102_RRB_1848_1723.	0.31	0.00	0.02	0.00	0.00	0.00	0.00	0.00
MM_2_102_RRB_1848_1826.	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MM_2_102_RRB_1848_2013.	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MM_2_102_RRB_1848_2109.	0.93	0.04	0.00	0.00	0.00	0.00	0.00	0.00
MM_2_103_RRB_1849_0018.	0.30	0.06	0.00	0.00	0.01	0.00	0.03	0.15
MM_2_103_RRB_1849_0139.	0.93	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MM_2_103_RRB_1849_0231.	0.31	0.00	0.00	0.00	0.03	0.00	0.00	0.10
MM_2_103_RRB_1849_0335.	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MM_2_103_RRB_1849_0399.	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.03
MM_2_103_RRB_1849_0468.	0.30	0.00	0.00	0.00	0.09	0.00	0.00	0.00
MM_2_103_RRB_1849_0623.	0.36	0.00	0.00	0.00	0.11	0.00	0.00	0.00
MM_2_104_RRB_1849_0649.	0.30	0.00	0.00	0.45	0.00	0.00	0.00	0.03
MM_2_105_RRB_1849_1189.	0.36	0.00	0.01	0.00	0.00	0.00	0.00	0.15
MM_2_105_RRB_1849_1192.	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.01
MM_2_105_RRB_1849_1545.	0.30	0.39	0.00	0.00	0.00	0.00	0.00	0.00
MM_2_105_RRB_1849_1660.	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MM_2_105_RRB_1849_1706.	0.36	0.08	0.00	0.00	0.02	0.00	0.00	0.06
MM_2_105_RRB_1849_1717.	0.30	0.12	0.00	0.00	0.00	0.00	0.00	0.14
MM_2_105_RRB_1849_1794.	0.31	0.02	0.00	0.00	0.00	0.00	0.00	0.25
MM_2_105_RRB_1849_2020.	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.07
MM_2_106_RRB_1849_2070.	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.24
MM_2_106_RRB_1849_2087.	0.30	0.32	0.00	0.00	0.00	0.00	0.00	0.00
MM_2_106_RRB_1849_2310.	0.32	0.00	0.00	0.00	0.00	0.00	0.00	0.22

The best way to describe the algorithm figuratively is to understand topic modeling as a scissor that cuts a text into single words (literally done in some cases)[15] and throws them into small bags (this is also where the technical term "bag of words" comes from). Now you determine the chances of two words occurring in one of the bags. If put in a table, combined with the information about the most frequent tokens, this probability is the output of a topic model.

15 Dan Hirschman, Doing Things with Bags-of-Words, in: *Scatterplot* 19.02.2020, URL: https://scatter.wordpress.com/2020/02/19/doing-things-with-bags-of-words/ [last accessed: April 2, 2021].

An entity could be a letter, a book, a chapter, whatever seems to be a useful comparator. So, if we want to build a model of topics appearing in different Wikipedia articles—a classical example where topic models are being built—the single article would be the entity.[16] For this paper, I will resort to the term "document" to indicate an entity that is of interest to us, one that can be compared and eventually shows or teaches (from the Latin "docere") something.

The problem with the generation of topic lists becomes clear once people are confronted with a specific topic model. As the main input, we need to define how many topics are to be found. However, according to the explanation of topic modeling above, clusters are only found near the end of the process, so the system obviously cannot know how many of these clusters to expect (though there are some ways to check if the number of clusters is statistically "ideal"). For a scholar, this input is counterintuitive, since we don't know at the beginning of the research process how many different topics to expect from a corpus. Working with students using topic modeling for the first time, I in general had the experience that they set the number of expected topics too low, since people were interested in very general topics. A smaller number of topics doesn't mean a more accurate depiction of the most common topics, but rather a clustering of very general terms (maybe on level of adverbs or adjectives).

For this experiment, I worked with a dataset of about 150,000 nineteenth-century documents, called *Regierungsratsprotokolle*, that record decisions by the highest executive of the canton of Zürich (a district in modern Switzerland) and its administrative predecessor. At first glance, the tokens in the generated list did not form coherent "topics," but rather were just lists of terms that seemed to have some relation. If we rethink the process of their generation, this result can be explained: The clearest expression of a topic might not occur that often in a corpus. Even in mass produced literature in genres like romance, we might not encounter the token "love" as often as "hugging" or "kissing," or "felt" or "feeling."[17] Probably the most striking topic in the *Regierungsratsprotokolle* deals with warfare, but the token "Krieg" is missing completely; instead we find "Militärs" (military), "Regierungsrath" (the executive, appearing in almost all generated topics), "Kriegsrathe" (war council), "Infanterie" (infantry), "Truppen" (troops) and so forth. This example is quite intuitively interpreted, but several other "topics" can only be identified from token

16 For an example of a combined approach of categorization and Latent Dirichlet Allocation, see Xu Kang et al., Incorporating Wikipedia Concepts and Categories as Prior Knowledge into Topic Models, in: *Intelligent Data Analysis* 21 (2/2017), 443-461, doi:10.3233/IDA-160021.

17 Regarding the use of topic modeling for literary genres, see Christof Schöch, Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama, in: *Digital Humanities Quarterly* 11/2 (2017), URL: http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html [last accessed: April 2, 2021].

lists with difficulty, and for certain lists it is nearly impossible. With some background knowledge about the corpus, it should eventually be possible to make sense of the lists, but until we hit something like topic 22, the list for this topic reads as follows: "zeit, diesem, weise, während, dann, sollte, darauf, zwar, möglich, seit, namentlich, alle…" (time, this, ways, while, then, should, thereupon, possible, since, namely, all…). Even with in-depth knowledge about the corpus, it's not possible to interpret some of the topics since they consist of auxiliary vocabulary with limited testimony.

Approaches to streamlining topic modeling do exist: 1) Text can be pre-processed according to part of speech, especially using normalization by lemmatization, 2) Elaborate stop-word lists can be implemented, containing words that do not have "meaning." The lemmatization of tokens as a preprocessing step will lead to text that is less variable and to the federation of tokens, especially when plural and singular forms of nouns or verbs are treated as identical. For highly standardized languages, such as most western languages including modern English, automatic lemmatization is already possible and leads to excellent results (some of the approaches are rule-based; others, more efficient, tend to use machine learning as well). As soon as we switch our interest to most languages pre-1900 (most historical documents) or to languages that so far have not attracted commercial enterprises such as search engines, vendors, or social media giants, the situation is completely different. The lack of annotated corpora, the quantity of variations in spelling, and, frankly, the lack of interest from researchers yield subpar results and consequently we are denied this method. With the second approach, the usage of stop-word lists, we could also resort to using ready-made files, which might result in unconsciously erasing tokens. Maybe such approaches even take us into philosophical or theoretical-linguistic territories, since we need to determine what words do reveal, or at least nudge us towards, the meaning of a document.

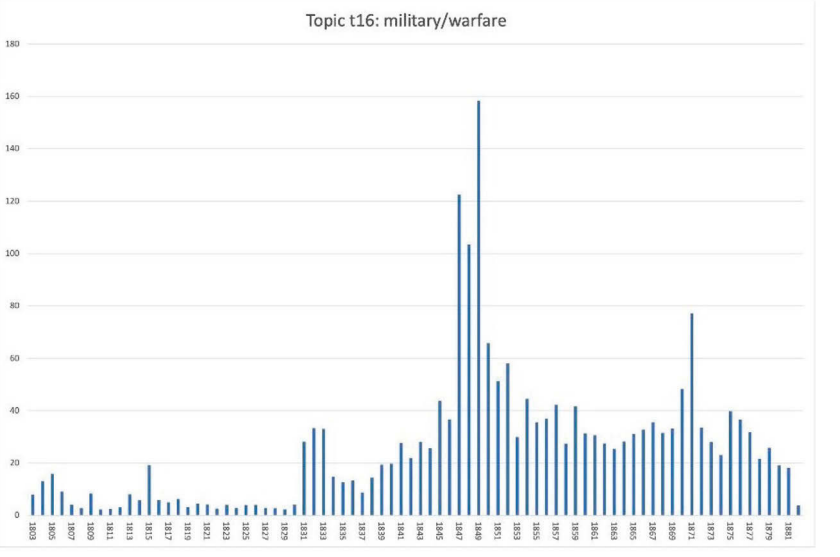
This leaves us for the moment in a difficult situation, since the algorithm indicates a variety of occurring topics only through a list in need of interpretation. It is through visualizations and repeated (re-)interpretation of topic lists that we will get an understanding of both the corpus we are analyzing and the method we are applying. In a sense, the cycle of interpretation can be understood as a hermeneutical cycle (here understood in a positive way) bringing us step-by-step closer to the corpus and thus the subject of our research.[18]

If we make temporal visualizations for the example corpus, we find striking traces of how some clusters indicate meaningful topics, helping us to understand what was decided by Zürich's executive in the 19th century. The topic indicating

18 For a broader introduction to topic modeling see also: Scott Weingart, Topic Modeling for Humanists: A Guided Tour, in: *The scottbot irregular* 25.07.2012, URL: http://www.scottbot.net/HIAL/?p=19113 [last accessed April 2, 2021].

"warfare" peaks in the middle/end of the 1840s, right at the time when Switzerland struggled towards a civil war (the so-called *Sonderbundskrieg*, in 1847). [19] And, even after the Swiss were pacified, the topic remained virulent: as we know, there were uprisings in the rest of Europe, especially in the bordering German lands in 1848. A second, smaller peak indicates the time of the French-German war (with battles in nearby Belfort) and the surrender of the Bourbaki army on Swiss soil in 1871. At the time, the Swiss army (including troops from Zürich) was mobilized and present at the borders.

Fig 6.2: Visualization of topic 16, concerning military/warfare in the corpus "Zürcher Regierungsratsprotokolle". The graph demonstrates how often a topic occurs (y-axis) in the documents produced in a certain year (x-axis).



Of course, historians knew about all those encounters, and searching for consequences of mobilization would have been possible in the minutes even without topic modeling. But the one topic only covers a small part of what can be found in the minutes. Some topics focus on the building of infrastructures to facilitate the use of railways. Quite frequently, financial decisions had to be made, another topic reflected in this group of documents.

19 A classical reference for the war in 1847 is Edgar *Der Sonderbundskrieg*, Zürich, 1947. For later research, see Pierre *La Guerre Du Sonderbund: La Suisse de 1847*, Neuchâtel 2018; Hans Rudolf Fuhrer/Jean Paul Loosli/Christian Moser, *Sonderbundskrieg 1847*, Wettingen 1997.

The main challenge—and, fittingly, this is a general problem in the humanities—is discerning relevant from irrelevant incidences, not on the level of sources but rather on a meta or mediated level. The algorithm forces the user to deal with processed documents—or data—and urges them to shift to a multitude of perspectives: From in-depth analysis and close reading of single pages in a source to a birds-eye view, scanning hundreds and thousands of pages, trying to make sense of a corpus.

The approach is of course not free of flaws. From a methodological point of view, we can identify a multitude of problems in addition to the exceptional opportunities. For one, the input, the corpus, very much influences the topics harvested. Furthermore, the number of topics needs to be defined beforehand, since the optimization process needs to optimize toward a value. The number of topics has therefore to be played around with, since some of the topics generated are mere lists of "non-topical vocabularies" that probably will not be taken into consideration when making statements about the content of the corpus.

In a sense, the approach can thus not be generalized. There are some (ongoing) attempts in that direction, but they raise a multitude of underlying questions (what's a general topic? are there topics independent of genres?) and consequently are mostly used within narrow typological frames.

From a technical standpoint and compared with deep learning approaches (we will tackle those later on), topic modeling is very much explainable and based on a set of clustering algorithms.[20]

Still, we recognize bias—not only in negative meaning—on different levels. First, the corpus biases the output in terms of topics. Documents from a narrow field will lead to a number of similar-seeming topics from that field (in the Regierungsrathsprotokolle "finances" as an ever important topic is present in at least three calculated topics). Second, the bias is induced by the user trying to make sense of the list of tokens. Depending on the knowledge of the corpora, expected topics will be read into the list.

The question arising from this is how to build data sets that are not biased. Is it possible to broaden the input, in order not to have the material focus on specific aspects? At the end of my consideration of topic modeling, we find ourselves right at the heart of issues of bias in machine learning. Topic modeling has often been advertised as a new way to approach digitized documents. The capability to visualize a multitude of documents quite easily would seem to make this method a

20 For latent dirichlet allocation, see David M. Blei, Introduction to Probabilistic Topic Models, in: *Communication of the ACM* (2011); David M. Blei/Andrew Y. Ng/Michael I. Jordan, Latent Dirichlet Allocation, in: *Journal of Machine Learning Research*, (3/2003), 993–1022.

preferred one for carving meaning out of a corpus. When considering the problems of the approach, however, the ease vanishes and critical stances open up.[21]

Despite the embedded problems, topic modeling offers a new way to treat sources heuristically, and it supports scholarly endeavors that want to deal with the abundance of sources that arises from mass digitization and born-digital data. Topic modeling allows this first steps toward big data in the humanities,[22] without necessarily performing quantitative analyses in later study stages.

In topic modeling, the clustering algorithm uses statistical methods that belong to the realms of machine learning. Although the algorithm starts at a more-or-less random point of word distribution, it is still very much understandable how the results are generated. It would even be possible, in theory at least, to replicate the result using analogue methods.

If we switch to deep learning, it would not be possible to do so. As we will see, there's a significant difference in the handling of input and output in the context of neural networks. Deep learning gained a lot of traction in the past decade and is currently implemented in a wide variety of algorithms (from image identification to self-driving cars, digital medicine and hiring applications). In the humanities, we currently stand at the starting point of this movement, with the first algorithms based on the networks being used on a regular basis, for example (and especially successfully) for handwritten text recognition.[23]

4. Handwritten Text Recognition: Supervised Training

With the advent of text recognition for print in the 1990s based on optical character recognition (OCR), similar results for handwriting seemed only a few years away. But the technology, based on the idea of isolating single characters, was never capable of delivering meaningful results, spare some success in recognizing neatly painted letters from the Middle Ages. Only in 2010 did the introduction of neural networks and the field of deep learning lead to a stark and astonishing improvement in handwritten text recognition (HTR), at that time as a proof of concept.

---

21 Ben Schmidt, Words Alone: Dismantling Topic Models in the Humanities, in: *Journal of Digital Humanities* 2 (1/2013), URL: http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/ [last accessed: April 2, 2021].

22 Shawn Graham/Ian Milligan/Scott Weingart, *Exploring Big Historical Data: The Historian's Macroscope*, London 2015, doi:10.1142/p981. A project using topic modeling reasonable on a large scale is: Impresso – Media Monitoring of the Past, URL: https://impresso-project.ch/) [last accessed: April 2, 2021].

23 See Guenter Muehlberger et al., Transforming Scholarship in the Archives through Handwritten Text Recognition: Transkribus as a Case Study, in: *Journal of Documentation* 75 (5/2019), 954-976, doi:10.1108/JD-07-2018-0114.

Eight years later, HTR entered the scene such that scholars could train and use specific models.

In the READ (short for "Recognition and Enrichment of Archival Documents") project, running from 2016 to 2019, scientists and scholars constructed a platform allowing them to train algorithms to recognize specific scripts with an error rate of around 3 percent.[24] This meant that, on the character level, out of 100 characters, 97 would be recognized correctly, including punctuation and space. Although the character error rate cannot be compared to a scholarly edition (typically around 0.1 to 0.2 percent character error rate), the result is still very much a legible and recognizable text that can be read, searched, and mined.[25]

Currently, we stand at the brink of training what are called general models that can recognize entire "styles" of handwriting, such as English in a Latin script of the 18th and 19th centuries or German current scripts of the 16th and 17th centuries. We thus can already conclude that deep learning can broaden access to historical material and helps us dive deeper into the subjects we're interested in. However, this comes at a price: By training and by selecting material for training, we strengthen machine learning algorithms, but we also bias them.

A very brief look at the architecture of the system of a deep learning algorithm, focused here on text recognition, explains the problem, since its success is based on training. In deep learning, we use systems modeled in ways similar to our understanding of the brain, consisting of a layered network of neurons. These cells are capable of either amplifying or reducing a signal coming from the preceding layer. The cells get "weighted" through training, meaning that the behavior of the cell (amplifying/reducing certain signals) will be optimized in a training process.

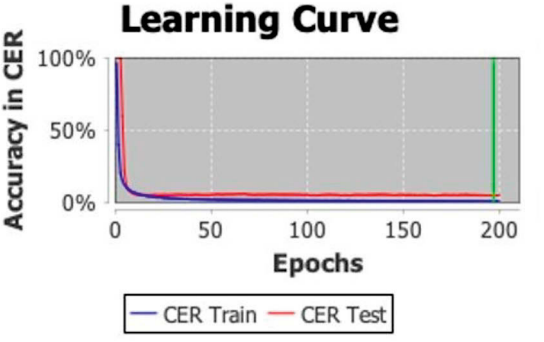
In the training process, the algorithm tries to align an input (in the case of text recognition, the image of a line of text) with a desired output (a string of characters). The training set is processed by the algorithm a certain number of times (called "epochs"), while optimizing the results towards the desired output. The main task of deep learning experts is thus not to know about the context of a certain problem like paleographical discussions for text recognition, but to decide on the number of layers, the (mathematical) optimizing processes, and the forms of decoding. This makes the technology very adaptable but also problematic.

---

24 Guenter Muehlberger et al., Transforming Scholarship in the Archives through Handwritten Text Recognition: Transkribus as a Case Study, in: *Journal of Documentation* 75 (5/2019), 954-976, doi:10.1108/JD-07-2018-0114.

25 Since we discussed topic modeling earlier in the paper, it must be stated that the step from HTR to topic modeling is still quite difficult: Stephen Mutuvi et al., Evaluating the Impact of OCR Errors on Topic Modeling, in: *Lecture Notes in Computer Science*, 11279, Cham 2018, 3-14, doi:10.1007/978-3-030-04257-8_1.

Fig 6.3: Image from a learning curve, demonstrating how the neural networks reaches better results after each iteration. Own screenshot, made in Transkribus (text recognition platform and software): transkribus.eu.



If we train the algorithm on a set of pages from a specific hand, we will receive our first usable result after about 1,000 lines.[26] But, whatever decisions are made in the transcription process (for example, expanding abbreviations or the use of certain characters, such as long ſ vs. round s), the algorithm will "learn" the decisions implicitly. Even on the level of text recognition, this can become problematic, for example, in Latin scripts that tended to introduce a multitude of signs indicating abbreviation, signs that need to be expanded in context. One example would be "ꝝ" [Unicode A75D, Latin Small Letter Rum Rotunda], indicating the genitive plural ending of a word and resulting in different expansions depending on the gender (male/female/impersonal).[27] Since deep learning tends to rely on quantity, the instance most occurring will most probably be chosen. The inclusion or lack of a character will also lead to a different character set, making the recognition

26 For a more in-depth view on best practices (in German), see: Tobias Hodel, Best-Practices zur Erkennung alter Drucke und Handschriften – Die Nutzung von Transkribus large- und small-scale, in: Christof Schöch (ed.), *DHd 2020 Spielräume. Digital Humanities zwischen Modellierung und Interpretation*, Paderborn 2020, 84–87, doi:10.5281/zenodo.3666689 The recordings of recent Transkribus User Conferences might also give some insight: https://readcoop.eu/transkribus-user-conference-2018/ and https://readcoop.eu/transkribus-user-conference-2020/ [last accessed: April 2, 2021].

27 See also the talk by Estelle Gueville/David Wrisley, Rethinking the Abbreviation: Questions and Challenges of Machine Reading Medieval *Scripta*, in: *Dark archives 2020 conference*, URL: https://www.youtube.com/watch?v=p38lvPRRNmA [last accessed: April 2, 2021].

process more or less likely to succeed, since a minimal number of occurrences of a certain character is necessary to provide a reliable identification.[28]

With regard to text recognition, the consequences of such flawed output won't create too many problems. However, the issue is exacerbated if we use the same technology for things like named entity recognition that lead to interpretations involving questions of identity (what/who is a person) or typology (what is a text)—basically, everything that falls in the context of natural language processing.[29]

One of the main problems in dealing with these kinds of deep learning algorithms is the black-box that exists in the process of training. Since none of them is based on any (let's call it contextual) knowledge of the field or corpus in question, the algorithms are trained on a particular input-output or specific corpus. Accordingly, the resulting models are highly biased by the material they are trained on. In each training session, the algorithm "learns" to deal with a world consisting only of the training material. In Foucauldian terms, we could speak of an episteme that lays in the model and is different from the episteme of all other models.[30]

Let's look, for example, at a recognition model that has been trained on letters from the 15th century. The corpus presented consists of letters sent from the city council of Bern to a bailiff residing in nearby Thun, a small city under the dominion of Bern. As is often the case with medieval administrative writing, the content is quite formulaic and repetitive. The letters were written by different scribes, and we collected about 100 such letters, which we transcribed by hand and used as the material to train an HTR model.[31] Even when we apply the model on some random part of an image with no text written on it whatsoever (as far as I can tell), the HTR model gives us as a result parts of a typical letter, with the salutation and signature even appearing in the correct order. We could therefore conclude that the model is trained and keen to recognize what it was trained to "read." Of course, the example here is forced, and the preceding algorithm—the layout analysis—would not actually identify the part of the image with no text as a text region. The result is also striking due to its coupling with a generated vocabulary. Nonetheless, it

28 In addition, current HTR systems have been produced with alphabet languages in mind, leading to problems in decoding in the cases where not 100 characters but some thousand signs are part of the character set.

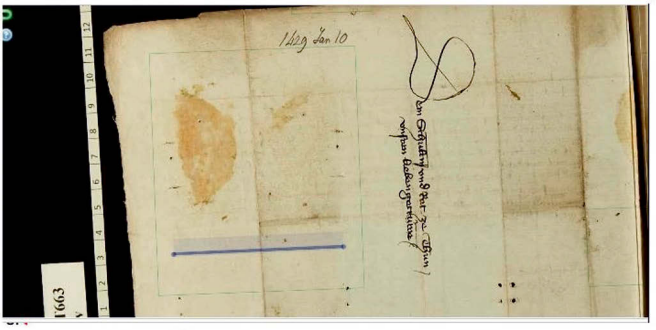
29 Tobias Hodel, Konsequenzen automatischer Texterkennung – Ein Aufriss zur Texterkennung mit machine learning, in: Vogeler, Georg (ed.), *DHd 2018. Kritik der digitalen Vernunft. Konferenzabstracts. Universität zu Köln, 26. Februar Bis 2. März 2018*, 249–51, Köln 2018, URL: http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf [last accessed: April 2, 2021].

30 Michel Foucault, *The Archaeology of Knowledge: And the Discourse on Language*, New York 1982.

31 The model was trained on 21,682 words and leads to a sub-par model with a character error rate on the validation set of 22.54 percent.

demonstrates how closely related the result of the recognition process is to the training set.

Fig 6.4: Screenshot from Transkribus, with the model recognizing (non-existent) text in the green square. The document transcribed is from the City Archives of Thun (BAT), BAT 663, 296v.



Since deep learning relies on large amounts of data to train specific models, we end up in a cycle of bias. In order to strengthen the model's ability to perform, in this case, text recognition, as much training data as possible is needed, usually gained from recognition processes that have undergone correction. The decisions of the annotators and the recognition process will thus fire back into the model and reinforce any problems. The whole issue becomes obvious if we look at the field's nomenclature. The term "Ground Truth" refers both to the material used to train a system and to the material used to evaluate the algorithm. The two sets are technically completely separated and not overlapping, with one subset used for training and another subset for evaluation. But both need to be prepared and checked by a human, thus the so-called Ground Truth is a quite far-reaching term for something influenced by someone (or rarely but ideally a group of people) who

needs to determine, in the case of text recognition, the correct transcription of a particular document.

What we see in the realms of deep learning is a reinforcement of bias introduced by the selection of training material. In a way, this issue is quite similar to what we observed in topic modeling, where the corpus at the beginning strongly influenced the calculated topics. Furthermore, deep learning is basically impossible to grasp as an algorithm. What happens in the neural network can only be observed; it is not possible to influence the process (save for creating a completely new alignment of the layers of the neural network). This barrier means that deep learning is in the difficult position that every model has to be checked for its bias and consequently for implicit and explicit problems through both input and output. Even epistemological questions must be brought up as part of the methodological discussion: What do we deem recognizable or sortable for an algorithm?

This brings us back to the initial question of what consequences machine learning approaches have for the humanities, looking at two very different approaches under the umbrella of Artificial Intelligence.

5. Addressing and (Not Yet) Solving Bias in Machine Learning: An Initial Conclusion

The most notable conclusion from the two approaches, text recognition and topic modeling, is they include—inevitably—bias. At the same time, the use of machine learning algorithms opens up a wide array of research possibilities that would have been unthinkable only ten years ago. For example, results from masses of handwritten documents were, before automatic recognition became a scalable process, accessible on only a very limited basis. And we only stand at the brink to grasp how we can use this new research tool.[32] Accessing masses of documents presorted by topic, or at least by clusters, opens up new ways to sift through material.

Topic modeling offers the opportunity to cluster together similar documents and extract characteristic tokens out of a cluster, allowing for an in-depth engagement with the documents as well as the method. In comparison, the controllability of topic modeling remains quite high, although the approach of unsupervised machine learning algorithms would seem to hint into another direction.

The case is quite the opposite if we look at the supervised deep learning approach, used here for handwritten text recognition. Although the model tries to adapt to and approximate a trained, desired output, the determining factors that

32 The consequences of the development of Google books are only starting to be recognized, see Lara Putnam, The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast, in: *The American Historical Review* 121 (2/2016), 377–402, doi:10.1093/ahr/121.2.377.

lead to a specific output are unknown to the user and part of a black box consisting of artificial neurons defying interpretation.

Both approaches, supervised as well as unsupervised, insert forms of bias at different stages, whether through the corpus employed to build topics or through training material.

It is thus the connection between the capabilities of and the deeply embedded bias in machine learning that makes it a divided and torn approach in dire need of contextualization. Understanding the method—similar to the methods and theories of any discipline—is consequently a first step to interacting with its technological aspects critically and cracking open the parts of the method that are stored away in a black box.

The use of AI in scholarly endeavors is not only a way to yield results (be they recognized text or clustered topics), but moreover an intriguing means of engaging with a set of technologies that govern our everyday lives, including a multitude of processes we might be subjected to.

In conclusion, scholars need to treat machine learning methods as they would source material or research methods, by adding a layer of methodological critique to the research process. Alongside the publication of training and validation data (if possible), this will lead to a deeper level of interaction with a method that has been deemed inexplicable.

Algorithms and data should nonetheless be approached playfully, so that results of machine learning approaches are not taken too seriously, and data from the humanities can form a playground with a rich body of research for identifying and critiquing biased and thus problematic results.

Bibliography

Benjamin, Ruha, *Race After Technology: Abolitionist Tools for the New Jim Code*, Medford 2019.

Blair, Ann, *Too Much to Know: Managing Scholarly Information before the Modern Age*, New Haven 2010.

Blei, David M., Introduction to Probabilistic Topic Models, *Communication of the ACM*

Blei, David M./Ng, Andrew Y./Jordan, Michael I., Latent Dirichlet Allocation, in: *Journal of Machine Learning Research*

Bonjour, Edgar, *Der Sonderbundskrieg*, Zürich 1947.

Brown, Tom B., et al., Language Models Are Few-Shot Learners, in: *arXiv* [preprint], 22.07.2020, URL: arxiv:2005.14165 [last accessed: April 2, 2021].

Clancy, Kelly, Is the Brain a Useful Model for Artificial Intelligence?, in: *Wired* 19.05.2020, URL: https://www.wired.com/story/brain-model-artificial-intelligence/ [last accessed: April 2, 2021].

Devlin, Jacob, et al., BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding, in: *arXiv* [preprint], 24.05.2019, URL: arxiv:1810.04805 [last accessed: April 2, 2021].

Du Bois, Pierre, *La Guerre Du Sonderbund: La Suisse de 1847*, Neuchâtel 2018.

Foucault, Michel, *The Archaeology of Knowledge: And the Discourse on Language*

Fuhrer, Hans Rudolf/Loosli, Jean Paul/Moser, Christian, *Sonderbundskrieg 1847*, Wettingen 1997.

Gadamer, Hans-Georg, *Hermeneutik I: Wahrheit Und Methode: Grundzüge Einer Philosophischen Hermeneutik*, Tübingen 2010.

Graham, Shawn/Milligan, Ian/Weingart, Scott, *Exploring Big Historical Data: The Historian's Macroscope*, London 2015, doi:10.1142/p981.

Graham, Shawn/Weingart, Scott/Milligan, Ian, Getting Started with Topic Modeling and MALLET, in: *Programming Historian* 02.09.2012, URL: https://programminghistorian.org/en/lessons/topic-modeling-and-mallet [last accessed: April 2, 2021].

Gueville, Estelle/Wrisley, David, Rethinking the Abbreviation: Questions and Challenges of Machine Reading Medieval *Scripta*, in: *Dark archives 2020 conference*, URL: https://www.youtube.com/watch?v=p38lvPRRNmA [last accessed: April 2, 2021].

Hirschman, Dan, Doing Things with Bags-of-Words, in: *Scatterplot* 19.02.2020, URL: https://scatter.wordpress.com/2020/02/19/doing-things-with-bags-of-words/ [last accessed: April 2, 2021].

Hodel, Tobias, Best-Practices zur Erkennung alter Drucke und Handschriften – Die Nutzung von Transkribus large- und small-scale, in: Christof Schöch (ed.),

DHd 2020 Spielräume. Digital Humanities zwischen Modellierung und Interpretation, Paderborn 2020, 84–87, doi:10.5281/zenodo.3666689.

Hodel, Tobias, Konsequenzen automatischer Texterkennung – Ein Aufriss zur Texterkennung mit machine learning, in: Vogeler, Georg (ed.), *DHd 2018. Kritik der digitalen Vernunft. Konferenzabstracts. Universität zu Köln, 26. Februar Bis 2. März 2018*, 249–51, Köln 2018, URL: http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf [last accessed: April 2, 2021].

iRights lab, et al., Praxisleitfaden zu den Algo. Rules - Orientierungshilfen für Entwickler:Innen und ihre Führungskräfte, Gütersloh 2020, doi:10.11586/2020029.

Muehlberger, Guenter, et al., Transforming Scholarship in the Archives through Handwritten Text Recognition: Transkribus as a Case Study, in: *Journal of Documentation* 75 (5/2019), 954-976, doi:10.1108/JD-07-2018-0114.

Mutuvi, Stephen, et al., Evaluating the Impact of OCR Errors on Topic Modeling, in: *Lecture Notes in Computer Science*, 11279, Cham 2018, 3-14, doi:10.1007/978-3-030-04257-8_1.

Noble, Safiya Umoja, *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York 2018.

Putnam, Lara, The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast, in: *The American Historical Review* 121 (2/2016), 377–402, doi:10.1093/ahr/121.2.377.

Rosenzweig, Roy, *Clio Wired: The Future of the Past in the Digital Age*, New York 2011.

Schmidt, Ben, Words Alone: Dismantling Topic Models in the Humanities, in: *Journal of Digital Humanities* 2 (1/2013), URL: http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/ [last accessed: April 2, 2021].

Schöch, Christof, Quantitative Analyse, in: Fotis Jannidis/Hubertus Kohle/Malte Rehbein (eds.), *Digital Humanities: Eine Einführung*, 279-298, Stuttgart 2017, doi:10.1007/978-3-476-05446-3_20.

Schöch, Christof, Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama, in: *Digital Humanities Quarterly* 11/2 (2017), URL: http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html [last accessed: April 2, 2021].

Spelke, Elizabeth, It's Called Artificial Intelligence—but What Is Intelligence?, in: *Wired* 19.05.2020, URL: https://www.wired.com/story/its-called-artificial-intelligence-but-what-is-intelligence/ [last accessed: April 2, 2021].

Underwood, Ted, Theorizing Research Practices We Forgot to Theorize Twenty Years Ago, *Representations* 127 (1/2014), 64–72, doi:10.1525/rep.2014.127.1.64.

Underwood, Ted/Jockers, Matthew L., Text-Mining the Humanities, in: Susan Schreibman/ Ray Siemens/John Unsworth, *A New Companion to Digital Humanities*, Chichester 2016, 291–306.

Weingart, Scott, Topic Modeling for Humanists: A Guided Tour, in: *The scottbot irregular* 25.07.2012, URL: http://www.scottbot.net/HIAL/?p=19113 [last accessed April 2, 2021].

Xu, Kang, et al., Incorporating Wikipedia Concepts and Categories as Prior Knowledge into Topic Models, in: *Intelligent Data Analysis* 21 (2/2017), 443-461, doi:10.3233/IDA-160021.