Tobias Hodel

Die Maschine und die Geschichtswissenschaft

Der Einfluss von *deep learning* auf eine Disziplin

Abstract: Deep learning is a method from the field of artificial intelligence that is currently being used in many disciplines to create appraisal decisions. The form of machine learning is also being used in history, for example for text recognition or the identification of named entities. Since deep learning will become a much stronger part of the methodological apparatus in the future, it is worth taking a critical look at what is happening. The moment of training plays a crucial role within the method. There, models are created and optimized. Based on the provided data, patterns can be recognized and imitated. Significantly, the created models are only verifiable in retrospect and with test procedures and are at most partially comprehensible. Thus, hermeneutic approaches are needed to understand and classify the models. Accordingly, the use of deep learning in history will entail a new reflection on methods, which must take into account technical circumstances on the one hand and disciplinary specifications on the other.

Keywords: maschine learning, model training, corpus, text recognition, named entity recognition, text analysis

Zusammenfassung: *Deep learning* ist ein Verfahren aus dem Bereich der künstlichen Intelligenz, das aktuell in vielen Disziplinen zur Erstellung von Bewertungsentscheiden genutzt wird. Auch in der Geschichtswissenschaft wird die Form des maschinellen Lernens bereits genutzt, etwa zur Texterkennung oder der Identifikation von benannten Entitäten. Da *deep learning* zukünftig noch viel stärker Teil des Methodenapparats werden wird, lohnt sich ein kritischer Blick auf die Vorgänge. Das Moment des Trainings spielt in der Methode eine entscheidende Rolle. Dort werden Modelle erstellt und optimiert. Aufgrund der vorgesetzten Daten können Muster erkannt und imitiert werden. Bezeichnenderweise sind die erstellten Modelle nur im Nachhinein und mit Testverfahren überprüfbar und höchstens bedingt nachvollziehbar. Damit braucht es hermeneutische Herangehensweisen, um die Modelle zu verstehen und einzuordnen. Die Nutzung von *deep learning* in der Geschichtswissenschaft wird entsprechend eine neue Methodenreflexion nach sich ziehen, die einerseits technische

https://doi.org/10.1515/9783110757101-004

Gegebenheiten und andererseits fachimmanente Spezifikationen berücksichtigen muss.

Schlagwörter: Maschinelles Lernen, Modelltraining, Korpus, Texterkennung, Named Entity Recognition, Textanalyse

Die Vision einer künstlichen Intelligenz (englisch *artificial intelligence* oder kurz AI) begleitet die Computerwissenschaft seit ihrem Entstehen. Seit wenigen Jahren fokussieren die Forschungen in diesem Bereich auf *deep learning*. Das Verfahren aus dem Bereich des maschinellen Lernens wird mittlerweile für zahlreiche Bewertungsentscheide eingesetzt, die vor wenigen Jahren noch als ungeeignet für die Bearbeitung durch Algorithmen oder allgemein „den Computer" beurteilt wurden: Beispiele umfassen die Identifikation von Stimmen, selbstfahrende Autos und die Einschätzung der Rückfallgefahr bei Verurteilungen von Straftäter:innen.

Das Prinzip des *deep learning* ist an sich simpel: Neuronale Netze, dem menschlichen Gehirn nachempfundene vernetzte Speicherzellen, werden mit möglichst vielen Daten versorgt und in einem Trainingsprozess auf typischerweise eine zu lösende Aufgabe getrimmt. Von Spracherkennung über Bildanalyse zu Dokumentenauswertung – *deep learning* setzt sich als Technologie in unterschiedlichen Feldern durch. Sie wird insbesondere seit wenigen Jahren für naturwissenschaftliche Auswertungen benutzt, hält als Technologie entsprechend bereits Einzug in wissenschaftliche Disziplinen. Die Nutzung der Technologie ist aber nicht unproblematisch und in dem Kontext auftretende Probleme stehen im Zentrum dieser Seiten.[1]

In den Geisteswissenschaften wird die Technologie aktuell erst in Ansätzen genutzt. Die Texterkennung von Drucken und Handschriften ist einer der Einsatzbereiche, die sich diesen Ansatz zu Nutze macht. Dies ist jedoch erst der Anfang, denn es ist absehbar, dass in naher Zukunft weit mehr (Be-)Wertungsentscheide manuell unterstützt oder gar autonom getroffen werden, die Beschäftigung mit der Methode ist entsprechend wichtig. *Named Entity Recognition*, aber auch visuelle und textuelle Strukturanalysen zeigen gemäß ersten Tests und *proof-of-concepts* bessere Resultate, als dies rein regelgeleitete Algorithmen ver-

1 Der Aufsatz ist der Versuch einen technologischen Ansatz zu vermitteln und gleichzeitig zu problematisieren. Die angesprochenen Themen werden entsprechend nicht in aller (notwendiger) Tiefe diskutiert, sondern vielmehr angeschnitten. Der Autor dankt für die angeregte Diskussion mit vielen Anknüpfungspunkten an der virtuellen Tagung sowie Christa Schneider und David Schoch für Diskussionen und kritischen Kommentare bei der Erstellung des Textes.

mögen. Mit wenig Phantasie lassen sich gar die Einsatzmöglichkeiten noch erweitern und die Interpretation von Texten mit und dank *machine learning* modellieren.

Im Rahmen dieses Papers werden drei Themenblöcke angeschnitten, die unterschiedliche Anwendungen des maschinellen Lernens im Fokus haben. Erstens, und wohl am unproblematischsten, ist die Nutzung von *deep learning* zur Handschriftenerkennung. Problematischer ist zweitens, die Entitätenerkennung (*Named Entity Recognition*), die kulturwissenschaftliche Fragen zu Praktiken der Namensgebung und zum Individuum im Generellen aufwirft. Drittens kann schließlich mit *machine learning* Ansätzen Strukturerkennung betrieben werden – dies ist eine Vorgehensweise, die in analoger Form etwa aus der Urkundenlehre bereits bekannt ist. Um die Technologie in den Fokus zu stellen, ist es jedoch nötig, dass die drei Ansätze innerhalb des Arbeitens mit neuronalen Netzen verortet und vor allem die Resultate kritisch betrachtet werden. Dieses Paper orientiert sich daher an den drei Perspektiven *Training* von neuronalen Netzen, *Interpretation* von Input und Output, sowie *Konsequenzen* des Einsatzes maschineller Lernverfahren.

1 Trainieren: Die Induktion von *bias*

Training als Basis zur komplexen, statistisch unterstützten Wertung erweist sich als größte Stärke und gleichzeitig neuralgische Stelle der Aufbereitung von Quellenmaterial als Daten, da durch das Trainingsmaterial (Vor-)Urteile übernommen und verstärkt werden. Diese Effekte wurden etwa für Suchmaschinen oder bei Bewerbungsprozessen mehrfach nachgewiesen und problematisiert.[2] Je nach Form des maschinellen Lernens werden Trainingsmaterialien zum Erlernen, etwa von Annotationen, vorgegeben und überprüft (*supervised learning*) oder die Strukturen werden selbständig erlernt (*unsupervised learning*).

Im Rahmen der Handschriftenerkennung werden Bildausschnitte einem zu erkennenden Text gegenübergestellt. Die Aufgabe des neuronalen Netzes ist es, eine Entsprechung zwischen Anhäufungen von Pixeln und zu erkennenden Zeichen zu finden. Dabei agieren die meisten Systeme unabhängig vom Vorwissen und trainieren jeweils eigenständige Modelle. Damit gibt es keine natürliche

2 Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York 2018); Aylin Caliskan, Bryson, Joanna J. und Arvind Narayanan, „Semantics derived automatically from language corpora contain human-like biases," *Science* 356 (6334), 183–186, doi: 10.1126/science.aal4230.

Verbindung zwischen Zeichen und Bild, es hängt vielmehr von den Vorgaben ab, die im Rahmen des Trainings gemacht werden. Die Distinktion von Zeichen liegt bei der trainierenden Person und auch im Umgang mit einem Zeichen in den vorhandenen Trainingsdaten. Der menschliche *bias* fließt zwar durch Transkriptionsentscheide in die Automatisierung ein, führt aber zu verhältnismäßig harmlosen Fehlern und Hyperkorrekturen.

Analog dazu verhält es sich bei der Erkennung von benannten Entitäten, einem Problem, das ebenfalls mit *supervised* Ansätzen bearbeitet wird und einen Algorithmus zur Nachahmung verleiten soll. Das System versucht *tokens* („Wörter" im Satzkontext) als einer Einheit (Person, Ort, Organisation etc.) zugehörig zu bestimmen. Im Vergleich zur Erkennung von handschriftlichem Text basiert bei der *Named Entity Recognition* ein zentraler Schritt auf der Anwendung von Sprachmodellen. Solche Modelle können große Textmengen in hochdimensionalen Vektorräumen verorten und damit Ähnlichkeiten zwischen Wörtern aufzeigen, weil diese entweder häufig im selben Kontext auftauchen, synonym verwendet werden oder aus ähnlichen Zeichenfolgen bestehen. Beim Training der Entitätenerkennung wird einem System also entsprechend vermittelt, inwiefern Wörter im Umfeld eines Vektors zu einer gemeinsamen Gruppe gehören. Zudem sehen wir, wie die Aufbereitung von Trainingsmaterial zur Einspeisung von (Vor-)Urteilen führt, indem etwas als Person oder Ort verstanden wird. Als zusätzliches Problem stehen wir vor der Herausforderung, dass Sprache nicht statisch ist, was die Erzeugung von historischen oder domänenspezifischen Sprachmodellen erforderlich macht. Diese Modelle sind natürlich aufgrund ihrer Basis auch gefärbt (dafür möchte ich den Begriff der „Korpusfärbung" beliebt machen)[3] und bilden Sprache nur entsprechend dem zugrundeliegenden Korpus ab.

Insgesamt bewegen wir uns folglich in einem Bewertungszyklus, der einerseits als Arbeitserleichterung verstanden, aber gleichzeitig zum Verstärkungsmechanismus von (Vor-)Urteilen (*bias*) wird. Der *cycle of bias* (siehe Abbildung 1) entsteht dabei aus arbeitstechnisch sinnvollen und pragmatischen Abläufen, wobei Material aufbereitet (transkribiert/annotiert) und darauf aufbauend *machine learning* basierte Modelle trainiert werden. Auf der Grundlage dieser Modelle wird weiteres Material miteinbezogen. Dadurch erhalten implizite und explizite Vorstellungen eine Verstärkung. Wertungsentscheide wie, „was wird wie transkribiert" oder „was ist ein Name" verfestigen sich als Muster. Anhand der

3 Unter Korpusfärbung verstehe ich die thematische und sprachliche Ausrichtung, die durch ein Korpus eingeführt wird und Modelle, die auf maschinellen Lernverfahren basieren, dadurch nicht nur beeinflussen, sondern geradezu definieren. Das Korpus wird dadurch mitverantwortlich für die Bewertungen, die durch ein damit trainiertes Modell erstellt werden.

häufig manuellen Aufbereitung des Materials lassen sich immerhin noch Bevorzugungen identifizieren. Noch problematischer ist indes, wenn bereits vorliegende Modelle beziehungsweise neuronale Netze direkt übernommen und durch sogenanntes *fine-tuning* angepasst werden. Insbesondere bei Sprachmodellen ist diese Vorgehensweise etabliert (siehe dazu weiter unten). Dies führt zur Übernahme von Bewertungsentscheidungen, die häufig nicht direkt und erst aus hochproblematischen Resultaten ersichtlich werden.[4] Insbesondere in den Bildwissenschaften wurde dieses Problem bereits erkannt, da aktuelle Bilderkennungsmodelle häufig auf einem kleinen Set an bereits bestehenden Modellen beruhen.

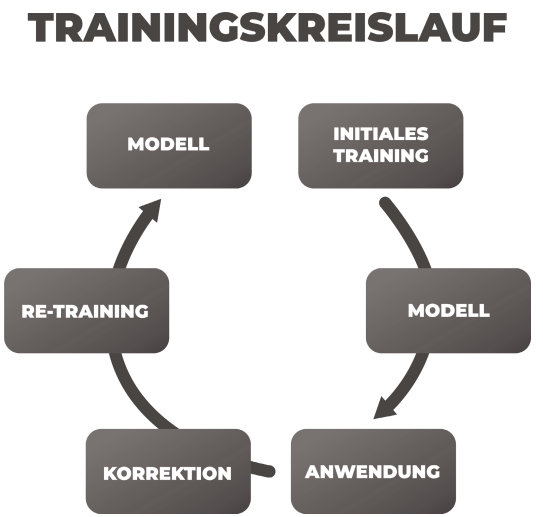

Abb. 1: „Cycle of bias." Abbildung von Viviane Blanchard und Tobias Hodel.

Wenn wir noch einen Schritt weitergehen und als Drittes versuchen Texteinheiten (Sätze, Absätze oder Sinneinheiten) einem Thema zuzuweisen oder nach semantischen Gesichtspunkten zu segmentieren, bewegen wir uns sowohl im Bereich des *supervised* als auch des *unsupervised learning*. Neben dem Trainieren

4 Siehe dazu aktuelle Eindrücke zum Sprachmodell GPT-3 Tom B. Brown et al., „Language Models are Few-Shot Learners," *arXiv:2005.14165 [cs]*, 22. Juli 2020, http://arxiv.org/abs/2005.14165, Zugriff am 16.03.2022 von OpenAI. Beispielsweise anhand des GPT-3 basierten Spiels AI Dungeon: Tom Simonite, „It Began as an AI-Fueled Dungeon Game. It Got Much Darker," *Wired*, 5. Mai 2021, https://www.wired.com/story/ai-fueled-dungeon-game-got-much-darker/, Zugriff am 28.02.2022.

von Annotationen (etwa für *sentiment analysis*), können auch Sätze (selten Satzteile) aufgrund der vorkommenden Wörter zu Themenfeldern zusammengefügt werden (etwa mit *topic modeling*).[5] Text wird in den meisten dieser Verfahren als ‚bag of words' verstanden, die Reihenfolge der Wörter also ignoriert und nur auf Frequenzen von Zeichenfolgen geachtet. Als zentrale Einheit gilt in diesen Verfahren der Satz, der durch einen Punkt (von einigen Ausnahmen abgesehen) von der nächsten Einheit abgetrennt wird. Entsprechend lässt sich an dieser Stelle ein neuralgischer Faktor identifizieren, da viele vormoderne Sprachen keine Entsprechung zum Satz kennen und die Interpunktion nicht existiert beziehungsweise nicht zwangsläufig Sinneinheiten abtrennt. Das zweite Problem ist die Aufbereitung der zu identifizierenden Teile oder Themen. Wenn wir dies anhand des Beispiels von Urkundenteilen durchdenken, werden je nach Ausgangsmaterial unterschiedliche Teile (etwa Protokoll, Kontext oder Eschatokoll) unterschiedlich stark gewichtet werden. Ein entsprechendes Training mit Übernahme der kanonisierten Wertung ist möglich, führt aber unweigerlich zur Verstärkung impliziter und expliziter Bevorzugungen. Im deutschsprachigen Raum wären solche Bevorzugungen etwa die Prägung der Diplomatik durch die Analyse ausgefertigter Königsurkunden im Gegensatz zu den zahlenmäßig massiv überwiegenden Urkunden, die unter dem Label „Privaturkunden" zusammengefasst werden.

Die drei kurz skizzierten Themenbereiche stellen unterschiedliche Phasen im Prozess der Quellenaufbereitung dar. Dabei zeigt sich sowohl in relativ simplen Erkenn- oder Identifikationsprozessen als auch in komplexen Zuordnungen das Moment des Trainings als kritischer Vorgang, da die daraus generierten Modelle je nach Korpus (Ausgangsmaterial) in einen Modus des Nachahmens übergehen. Das Verständnis der Modelle ist in der Konsequenz ein hermeneutischer Prozess, der, wie bereits von Gadamer gefordert, eine Auseinandersetzung mit (eigenen) Urteilen und insbesondere Vorurteilen miteinschließt und folglich den Prozess des Trainings nur so nachvollziehbar macht.[6]

5 Zu Topic Modeling siehe auch David M. Blei, „Introduction to Probabilistic Topic Models," Communication of the ACM, 2011.

6 Die Rolle des (Vor-)Urteils wird bei Gadamer aufgedröselt (und weniger stark als Problem aufgefasst): Hans-Georg Gadamer, *Hermeneutik I: Wahrheit und Methode: Grundzüge einer philosophischen Hermeneutik, 7. Aufl., Bd. 1, Gesammelte Werke* (Tübingen: Mohr Siebeck, 2010), 270–290. (Seitenzahlen nach der 7. Auflage).

2 Interpretieren: Quellenkritik und Hermeneutik

Über den Prozess des Trainings hinaus stellt der Umgang mit Resultaten des *machine learning* insbesondere aus Sicht der *algorithm studies* eine Herausforderung dar. Die Algorithmen lassen sich zwar an unterschiedlichen Stellen zu Ausgaben zwingen (bekannt sind die Google Image-Traum Algorithmen),[7] jedoch ist ein Nachvollzug der Entscheide innerhalb neuronaler Netze bislang nicht erfolgreich möglich. Die Kritik und die Auswertung der Resultate aus Vorgängen des maschinellen Lernens ähneln entsprechend hermeneutischen Interpretationen, die gerade durch den geschichtswissenschaftlichen Werkzeugapparat wie der Quellenkritik, aber auch andere Methoden analysiert werden müssen.[8] Erst das wechselseitige *close-* und *distant-reading* der Quellen und der Resultate macht es möglich, die Belastbarkeit der maschinell gewonnenen Wertungen zu überprüfen.

Die Überprüfung der Fehler ist bei der Texterkennung auf den ersten Blick relativ simpel, da mehr oder minder standardisierte Transkriptionskonventionen existieren. Neue Erkennalgorithmen erreichen dabei, je nach Anzahl der Trainingsseiten, unterschiedliche Resultate. Für Handschriften ist eine Erkennqualität mit Fehlerquoten im Bereich von 2,5 % technisch möglich, pro 1 000 erkannten Zeichen muss entsprechend mit 25 Fehlern gerechnet werden. In dieser Fehlerquote ist die fehlerhafte Erkennung von Satzzeichen sowie Groß-/Kleinschreibung bereits enthalten. Bei regelmäßigen Schriften wird diese Fehlerquote durch das Training eines entsprechenden Modells mit ungefähr 50 000 Wörtern erreicht.[9] Dieses Resultat lässt sich unter optimalen Bedingungen, das heißt unter Beizug von genügend Material von ähnlichen Schriften, auch für Modelle erreichen, die auf unterschiedlichen Händen basieren.[10]

---

7 Alexander Mordvintsev, Christopher Olah und Mike Tyka, „Inceptionism: Going Deeper into Neural Networks," Google AI Blog, 17. Juni 2015, https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html. Zugriff am 28.02.2022.

8 Für die Literaturwissenschaften siehe als Beispiel: Ted Underwood, „Emerging conversations between literary history and sociology," The Stone and the Shell, 02.12.2015, https://tedunderwood.com/2015/12/02/emerging-conversations-between-literary-history-and-sociology/, Zugriff am 16.03.2022.

9 Für dieses und andere Beispiele siehe Tobias Hodel, „Best-practices zur Erkennung alter Drucke und Handschriften – Die Nutzung von Transkribus large- und small-scale," in *DHd 2020. Spielräume Digital Humanities zwischen Modellierung und Interpretation*, hg. v. Christof Schöch (dhd2020, Paderborn, 2020), 84–87, doi: 10.5281/zenodo.3666689.

10 Siehe dazu auch Tobias Hodel, David Schoch, Christa Schneider, Jake Purcell, „*General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example*," Journal of Open Humanities Data, 7(13) 2021, 1–10, doi: 10.5334/johd.46.

Es bleibt die Frage offen, inwiefern durch die Quantifizierung von Fehlern Aussagen zur Leistungsfähigkeit eines Erkennmodells gemacht werden können. Zentral sind aus historischer Perspektive schließlich die Fragestellung und die (digitale) Methode, die nach dem Erkennprozess zum Einsatz kommen sollen. Je nachdem fällt auch der Fehlertyp (Satzzeichen sind für *topic modeling* Algorithmen etwa unerheblich) oder die Art eines Fehlers (die Verwechslung von Stab-s mit „f" führt im *close reading* zu keiner/wenig Verwirrung) ins Gewicht. Zukünftig wird es entsprechend wichtig sein, über quantifizierende Fehlerquoten hinaus, Angaben zur Fehleranfälligkeit eines Modells zu machen.

Der Einsatz von *Named Entity Recognition* verlangt anders gelagerte Diskussionen. Wie bereits oben angesprochen, wird dabei ebenfalls der Trainingsinput imitiert. Auch dies basiert auf Sprachmodellen, sodass die kritische Analyse eines solchen Modells Teil der Methodenkritik wird. Bei der Verwendung historischer Sprachformen entsteht jedoch zusätzlich das Problem, dass Sprachmodelle auf verhältnismäßig kleinen Datenmengen basieren.

Um die Leistungsfähigkeit bestehender Frameworks für nicht-standardisierte vormoderne Sprachen zu demonstrieren, wurde im Rahmen des Editionsprojekts Königsfelden ein Experiment zur Erkennung benannter Entitäten durchgeführt. Dabei wurde ein eigenes Sprachmodell (selbsttrainiert als FLAIR *embeddings*)[11] angelegt, das auf zeitlich nahen historischen Dokumenten aus dem 15. und 16. Jahrhundert basiert.[12] Das Training der benannten Entitäten basiert auf 645 Urkunden, für Verhältnisse des maschinellen Lernens also insgesamt eher wenig Material. Eine Besonderheit bildet das Tagging des Editionsprojekts, das die Strategie verfolgt, alle potentiell zugehörigen Informationen einem Namen zuzurechnen. Dadurch wurden auch Angaben, die heute nicht mehr als Namensteil verstanden würden, als solcher markiert und folglich auch fürs Training verwendet. Trotz des geringen Umfangs der Trainingsdaten konnten F-Scores im Bereich von 69–74 % erreicht werden.[13]

---

11 FLAIR ist ein open-source Framework für Natural Language Processing: https://github.com/flairNLP/flair. Alan Akbik et al., „FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)* (Minneapolis, Minnesota: Association for Computational Linguistics, 2019), 54–59, doi: 10.18653/v1/N19-4010.

12 Verwendet wurde das Bonner Frühneuhochdeutsch Korpus (http://www.korpora.org/FnhdC/), digital vorliegende Bände der Schweizerischen Rechtsquellen (https://www.ssrq-sds-fds.ch/home/) und Urkunden und Akten des Klosters Königsfelden (https://www.hist.uzh.ch/de/fachbereiche/mittelalter/lehrstuehle/teuscher/forschung/projekte/koenigsfelden.html, Zugriff am 16.03.2022).

13 F-Scores kombinieren Recall (Ausbeute) und Precision (Präzision) und sind ein häufig genutztes Mittel, um Klassifikatoren zu beurteilen.

Hanns	B-PER	B-PER
Zender	I-PER	I-PER
ouch	O	I-PER
von	O	I-PER
Birmistorff	B-LOC	I-PER
des	O	I-PER
gerichtz	O	I-PER
daselbs	O	I-PER

unser	B-PER	O
eidgnosschaft	I-PER	O
vogt	I-PER	O
zu	I-PER	O
Baden	I-PER	B-LOC
des	I-PER	O
frommen	I-PER	B-PER
vesten	I-PER	I-PER
Hans	I-PER	I-PER
Meisen	I-PER	I-PER
des	I-PER	I-PER
räts	I-PER	I-PER
Zürich	I-PER	I-PER

Erkenvrid	O	B-PER
und	O	O
Hartman	O	B-PER
ritt[7]	O	I-PER
und	O	O
Peter	O	B-PER
gebrud[7]	O	I-PER
truchsezen	O	I-PER
*	O	I-PER
von	O	I-PER
Habspg	O	I-PER

Abb. 2: Drei Beispiele für „Fehler“ im Tagging. Da die Definition einer benannten Entität sehr weit gefasst wurde, ist auch die maschinelle rechte Spalte sinnvoll. Im rechten Beispiel zeigt sich ein Annotationsfehler. Die Auswertung erfolgte durch Ismail Prada.

Dieser Wert ist für moderne Sprachen zwar nicht besonders hoch, zeigt aber das Potential des Ansatzes. Auch für dieses Verfahren lohnt sich ein Blick auf einzelne Resultate. Dadurch lässt sich eine Vielzahl von „Fehlern“ sichtbar machen, die korrekte Resultate widerspiegeln. Die „Fehler“ stammen in diesen Fällen von Annotator:innen, die inkorrekt auszeichneten oder aber von der Maschine selbst, die gar valable alternative Annotationen (Namen können teilweise Orts- oder Personennamen bezeichnen) liefert.[14] Einschränkend muss erwähnt werden, dass die Transkription händisch erstellt und Eigennamen im Gegensatz zum restlichen Text großgeschrieben wurden. Die Algorithmen hatten entsprechend starke Indizien zur Identifikation von Entitäten.

Stärker noch als bei der Texterkennung zeigt sich für Annotationsaufgaben, wie sehr die unterschiedlichen Inputs (Sprachmodell, Transkriptionsvorgaben und Trainingsmaterial) das Resultat beeinflussen. Eine Analyse der Technologie und der Resultate muss die Komplexität und Unsicherheitsfaktoren beispielsweise von Annotationsaufgaben mitberücksichtigen, wobei aktuell der benötigte Werkzeugkasten dazu noch mehrheitlich fehlt und wiederum quantitative Angaben nur beschränkt Aussagen zur Fähigkeit eines Netzes erlauben.

14 Für diesen Versuch wurde nicht berücksichtigt, dass *pooling* von Wortvektoren (eigtl. eine Kontextualisierung innerhalb von Sätzen, indem auf vorangehende Wortvektoren zurückgegriffen wird) eine höhere Leistungsfähigkeit aufweisen. Alan Akbik, Tanja Bergmann und Roland Vollgraf, „Pooled Contextualized Embeddings for Named Entity Recognition,“ in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (NAACL-HLT 2019, Minneapolis, Minnesota: Association for Computational Linguistics, 2019), 724–28, doi: 10.18653/v1/N19-1078.

Bei der Analyse von Resultaten, die aus textbezogenen Annotationsaufgaben entstanden sind, stellt der Einbezug von Sprachmodellen eine weitere Hürde dar. Anhand von Vergleichen mit zur Verfügung gestellten Testdaten lässt sich aufzeigen, dass *language models*, die eine gewisse Größe aufweisen und mit den zu analysierenden Texten eng verwandt sind, bessere Resultate liefern. Es hängt bei dieser Aufgabe folglich nicht nur von der Menge an annotierten Daten ab, sondern auch vom Sprachmodell, welches als Ausgang gewählt wird.

Wie bereits oben angesprochen, wird aus pragmatischen Gründen und teilweise auch aus Gründen der Performanz auf vortrainierte Netze zurückgegriffen. Für moderne und von großen Firmen als zentral erachtete Sprachen (Englisch, Französisch, Deutsch etc.) ist dies etwa BERT[15] beziehungsweise darauf aufbauende Varianten wie CamemBERT.[16] Auch die bereits in den Fußnoten erwähnten Modelle *GPT-2* und *GPT-3* fallen in die Kategorie der vortrainierten Netze. Allen Sprachmodellen ist gemein, dass sie jeweils zwar eine erhöhte Leistungsfähigkeit in der Problemlösung (typischerweise ausfüllen von Lückentexten) und der Annotation aufweisen, jedoch auf (Trainings-)Materialien zurückgreifen, die entweder überhaupt nicht publiziert oder insofern nicht nachvollziehbar sind, als dass die Daten zwar zugänglich aber nicht mit Metadaten angereichert sind. Von Datenpublikation unter FAIR-Kriterien kann in keinem der Fälle gesprochen werden.[17] Die Konsequenz ist, dass wir weder die Grundlagen beurteilen noch das Funktionieren der Modelle nachvollziehen können.[18]

In einem weit experimentelleren Stadium als die Identifikation von Entitäten befindet sich die Zuordnung von Annotationen, die Sinneinheiten klassifizieren. Bereits etwas etabliert, vor allem da kommerziell interessant, ist dabei die *sentiment analysis*, die Sätzen meistens positiven oder negativen Gefühlsausdrücken zuordnet. In der deutschsprachigen Digital Humanities Community

15 Jacob Devlin et al., „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805 [cs], 24. Mai 2019, http://arxiv.org/abs/1810.04805, Zugriff am 16.03.2022.

16 Louis Martin et al., „CamemBERT: A Tasty French Language Model," 9. Oktober 2019, HAL Id: hal-02445946.

17 Mark D. Wilkinson et al., „The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data* 3, Nr. 1 (Dezember 2016): 160018, https://doi.org/10.1038/sdata.2016.18.

18 Ansätze in diese Richtung zum Verständnis von Sprachmodellen über Visualisierung werden aktuell getestet, siehe bspw. hier Challenge 4, https://www.cnd.philnat.unibe.ch/ueber_uns/aktivitaeten/nlp_hackathon/ und das damit verbundene Poster: https://www.dh.unibe.ch/unibe/portal/fak_historisch/fsuf/d_dh/content/e330319/e336052/e1074527/PosterBDSD_ger.pdf, Zugriff am 16.03.2022.

werden entsprechende Ansätze bereits intensiv bearbeitet und auch für ältere Sprachstufen vorbereitet.[19]

Analog dazu können auch andere, etwa thematische Labels vergeben und trainiert werden. Die bereits oben beschriebenen Probleme werden dabei übernommen und die Komplexität nochmals um eine Stufe gesteigert, da die Sprachmodelle auf der Ebene „Satz" angewandt werden. Das Verfahren wird dabei vom Wort (eigentlich *token*) auf eine Zeichenkette erweitert, die durch vordefinierte Stoppzeichen (Komma und [Doppel-]Punkte) abgetrennt werden. Das bedeutet, dass aus den hochdimensionalen Vektoren von Wörtern ein Vektor pro Satz errechnet wird, der gar auf ganze Texte erweitert werden kann. Dadurch lassen sich Textähnlichkeiten mathematisch über Cluster, also die Zusammenführung von nahen Vektoren, aufzeigen. Obwohl solche Verfahren bereits seit einigen Jahren verfügbar und auch die informatischen Anforderungen an die Infrastruktur nicht unermesslich sind, sind es tendenziell literarische Texte, die diesbezüglich mit solchen Verfahren behandelt wurden.[20] Mit spezifischeren Datensets, die etwa historische Textgattungen abbilden und entsprechend das Erstellen von Sprachmodellen erleichtern, ist indes die Erweiterung der Einsatzgebiete in Sicht.

Da typischerweise auf den Satz als einfach zu segmentierende Einheit Bezug genommen wird, entstehen etwa für vormoderne Texte oder wenig gepflegte textuelle Formen (Stichwort: Kurznachrichten oder Social Media Posts) Her-

19 Siehe Thomas Schmidt, Manuel Burghardt und Katrin Dennerlein, „‚Kann man denn auch nicht lachend sehr ernsthaft sein?' – Zum Einsatz von Sentiment Analyse-Verfahren für die quantitative Untersuchung von Lessings Dramen" (Vortrag auf der DHd 2018 Kritik der digitalen Vernunft. 5. Tagung des Verbands „Digital Humanities im deutschsprachigen Raum" (DHd 2018), Köln, Zenodo, 26. Februar 2018), doi: 10.5281/zenodo.4622557; David Wodausch et al., „Hinterlistig – schelmisch – treulos – Sentiment Analyse in Texten des 19. Jahrhunderts: Eine exemplarische Analyse für Länder und Ethnien" (Vortrag auf der DHd 2018 Kritik der digitalen Vernunft. 5. Tagung des Verbands „Digital Humanities im deutschsprachigen Raum" (DHd 2018), Köln, Zenodo, 26. Februar 2018), doi: 10.5281/zenodo.4622483. Siehe dazu auch die Ressource S. Clematide und M. Klenner, „Evaluation and Extension of a Polarity Lexicon for German," in *Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*, hg. v. A. Montoyo et al. (Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA); Held in conjunction to ECAI 2010 Portugal, Lisbon, Portugal: University of Zurich, 2010), 7–13, doi: 10.5167/uzh-45506.

20 Ted Underwood, „The Literary Uses of High-Dimensional Space," Big Data & Society 2, Nr. 2 (1. Dezember 2015): 2053951715602494, doi: 10.1177/2053951715602494. Siehe auch einführend Ted Underwood und Matthew L. Jockers, „Text-Mining the Humanities," in *A New Companion to Digital Humanities*, hg. v. Susan Schreibman, Ray Siemens und John Unsworth (John Wiley & Sons, 2016), 291–306.

ausforderungen. Gerade für die angesprochenen Urkunden ist der Satz keine sinnvolle Einheit, um Zuordnungen zu erstellen.

Mit Blick auf diesen dritten Themenbereich (Textteile einem Thema zuordnen), stehen wir heute in einer initialen Findungsphase. Erste Modelle führen zu vielversprechenden Eindrücken, liefern jedoch noch zu wenig belastbare Resultate. Auch die Anwendung von *topic modeling* auf einzelne Sätze ist möglich, führt aber zu einem Clustering von ähnlichen Wortkonstruktionen und mahnt an die Auswertung von Kookkurrenzen. Aufschlüsse zu semantischen oder gar thematischen Feldern werden damit nur mittelbar gegeben.[21]

3 Konsequenzen: Von einer neuen Heuristik zu einer neuen Epistemologie?

Das Oszillieren zwischen praktischen Umsetzungen und theoretischen Überlegungen führt zu neuen Problemstellungen, die Epistemologie und heuristische Methoden der Geschichtswissenschaften betreffen. Maschinelles Lernen zeigt sich dabei bereits heute als nützliche Erweiterung der Disziplin an der Schwelle des Einsatzes von *big data*, die es kritisch zu betrachten und zu verfolgen gilt. Die Einsichten dienen dabei nicht nur der intradisziplinären Methodendiskussion, sondern führen darüber hinaus zu kritischen Positionen für den Einsatz von *deep learning* im alltäglichen Leben.

Die Nutzung von *deep learning* in einer hochgradig reflexiven Wissenschaft wie der Geschichtswissenschaft, bedeutet die Explizierung erkenntnistheoretischer Grundannahmen. Was etwa als „Text" verstanden wird, muss offengelegt sein, wenn ein Algorithmus zur Erkennung von „Text" gebracht wird.[22] Dabei regen auch einzelne Vorstufen von Texten die Diskussion an, etwa wenn identifiziert werden muss, wo sich auf einem Artefakt Text befindet. Die Identifikation von Personen oder Orten in textuellen Strukturen greift ebenso auf Vorannah-

21 Siehe dazu auch Tobias Hodel, „Supervised and Unsupervised: Approaches to Machine Learning for Textual Entities," in *Archives, Access and AI: Working with Born-Digital and Digitised Archival Collections*, hg. v. Lise Jaillant (Transcript, 2021), 162–168.

22 Tobias Hodel, „Konsequenzen automatischer Texterkennung – Ein Aufriss zur Texterkennung mit Machine Learning," in *DHd 2018. Kritik der digitalen Vernunft* Konferenzabstracts. Universität zu Köln 26. Februar bis 2. März 2018, hg. v. Georg Vogeler (Köln, 2018), 249–251, http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf, Zugriff am 16.03.2022; Patrick Sahle, Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 3: Textbegriffe und Recodierung., Bd. 3, 3 Bde., (Norderstedt: BoD, 2013), http://kups.ub.uni-koeln.de/5352/. Zugriff am 28.02.2022.

men zurück, indem Fragen nach bedeutungstragenden Namen gegenüber von Zuschreibungen abgewogen werden müssen. Zentral wird dabei die Dokumentation der Aufbereitung von Grundlagen, die Bewertungsentscheide nachvollziehbar macht. Indirekt lassen sich darauf aufbauend die Entscheide eines Modells nachvollziehen.

Gleichermaßen ist das Teilen der zugrunde liegenden Daten zentral, die eine Analyse der Annotations- und Bewertungsentscheide erlauben und somit das Problem des *bias* aktiv angehen. Die bereits existierenden und von diversen Forschungsförderanstalten eingeforderten FAIR-Richtlinien geben nicht nur diese Richtung vor, sondern lassen im akademischen Kontext fast keinen anderen Weg mehr zu. Anders sieht es bei der Nutzung kommerzieller Produkte oder Modelle aus, auf die weit weniger Zugriff besteht, die aber aufgrund ihrer Leistungsfähigkeit nicht komplett missachtet werden können. Der Umgang mit solchen Modellen muss im Kontext der *ethical AI* diskutiert und mit entsprechenden Ansätzen angegangen werden, etwa durch die Zertifizierung (*auditing*) von Algorithmen.[23]

Die Anwendung von maschinellen Lernverfahren erfordert somit nicht eine komplett neue historische Methode, sondern eine Erweiterung des technischen Horizonts, indem zumindest im Grundsatz die Verfahren verstanden werden müssen. Überdies ist eine konsequente Erweiterung der Hermeneutik auf eingesetzte Methoden notwendig, da nicht mehr nur das erforschte Material, sondern auch die technischen Herangehensweisen nie vollständig überblickt und auch nur in (langsamer) Annäherung verstanden werden können.

Der *machine learning turn* führt nicht zu einer Abkehr von der historischen Methode, sondern vielmehr zu einer neuen Art der Beschäftigung mit Quellen, die nicht nur den Aussagewert beurteilt, sondern gleichzeitig auch die (automatisierte) Beschäftigung damit berücksichtigt.

Bibliographie

Akbik, Alan, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter und Roland Vollgraf. „FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP.“ In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 54–59. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-4010.

23 Ruha Benjamin, *Race After Technology: Abolitionist Tools for the New Jim Code*, 1. (Medford, MA: Polity, 2019).

Akbik, Alan, Tanja Bergmann und Roland Vollgraf. „Pooled Contextualized Embeddings for Named Entity Recognition." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 724–28. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1078.

Benjamin, Ruha. *Race After Technology: Abolitionist Tools for the New Jim Code*. 1. Medford, MA: Polity, 2019.

Blei, David M. „Introduction to Probabilistic Topic Models." *Communication of the ACM*, 2011.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, u. a. „Language Models are Few-Shot Learners." *arXiv:2005.14165 [cs]*, 22. Juli 2020. http://arxiv.org/abs/2005.14165, Zugriff am 16.03.2022.

Caliskan, Aylin, Joanna J. Bryson und Arvind Narayanan. „Semantics Derived Automatically from Language Corpora Contain Human-like Biases." *Science* 356, Nr. 6334 (14. April 2017): 183–86. doi: 10.1126/science.aal4230.

Clematide, S. und M. Klenner. „Evaluation and Extension of a Polarity Lexicon for German." In *Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*, hg. v. A. Montoyo, P. Martínez-Barco, A. Balahur und E. Boldrini, 7–13. Lisbon, Portugal: University of Zurich, 2010. doi: 10.5167/uzh-45506.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee und Kristina Toutanova. „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *arXiv:1810.04805 [cs]*, 24. Mai 2019. http://arxiv.org/abs/1810.04805, Zugriff am 16.03.2022.

Gadamer, Hans-Georg. *Hermeneutik I: Wahrheit und Methode: Grundzüge einer philosophischen Hermeneutik*. 7. Aufl. Bd. 1. Gesammelte Werke. Tübingen: Mohr Siebeck, 2010.

Hodel, Tobias. „Konsequenzen automatischer Texterkennung – Ein Aufriss zur Texterkennung mit Machine Learning." In *DHd 2018. Kritik der digitalen Vernunft Konferenzabstracts. Universität zu Köln 26. Februar bis 2. März 2018*, hg. v. Georg Vogeler, 249–51. Köln, 2018. http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf. Zugriff am 16.03.2022.

Hodel, Tobias. „Best-practices zur Erkennung alter Drucke und Handschriften – Die Nutzung von Transkribus large- und small-scale." In *DHd 2020. Spielräume Digital Humanities zwischen Modellierung und Interpretation*, hg. v. Christof Schöch, 84–87. Paderborn, 2020. doi: 10.5281/zenodo.3666689.

Hodel, Tobias, David Schoch, Christa Schneider und Jake Purcell, „General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example," in *Journal of Open Humanities Data* 7 (2021). doi: 10.5334/johd.46.

Hodel, Tobias, „Supervised and Unsupervised: Approaches to Machine Learning for Textual Entities," in *Archives, Access and AI: Working with Born-Digital and Digitised Archival Collections*, hg. v. Lise Jaillant. Bielefeld: Transcript, 2021, 157–177.

Martin, Louis, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah und Benoît Sagot. „CamemBERT: A Tasty French Language Model," 9. Oktober 2019. Hal Id: 02445946.

Mordvintsev, Alexander, Christopher Olah und Mike Tyka. „Inceptionism: Going Deeper into Neural Networks." *Google AI Blog*, 17. Juni 2015. https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html, Zugriff am 16.03.2022.

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

Sahle, Patrick. *Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 3: Textbegriffe und Recodierung.* Bd. 3. 3 Bde. Norderstedt: BoD, 2013. http://kups.ub.uni-koeln.de/5352/, Zugriff am 28.02.2022.

Schmidt, Thomas, Manuel Burghardt und Katrin Dennerlein. „‚Kann man denn auch nicht lachend sehr ernsthaft sein?' – Zum Einsatz von Sentiment Analyse-Verfahren für die quantitative Untersuchung von Lessings Dramen." Vortrag auf der DHd 2018 Kritik der digitalen Vernunft. 5. Tagung des „Verbands Digital Humanities im deutschsprachigen Raum" (DHd 2018), Köln, 26. Februar 2018. doi: 10.5281/zenodo.4622557.

Simonite, Tom. „It Began as an AI-Fueled Dungeon Game. It Got Much Darker." *Wired*, 5. Mai 2021. https://www.wired.com/story/ai-fueled-dungeon-game-got-much-darker/, Zugriff am 28.02.2022.

Underwood, Ted. „Emerging Conversations between Literary History and Sociology." *The Stone and the Shell* (blog), 2. Dezember 2015. https://tedunderwood.com/2015/12/02/emerging-conversations-between-literary-history-and-sociology/, Zugriff am 16.03.2022.

Underwood, Ted. „The Literary Uses of High-Dimensional Space." *Big Data & Society* 2, Nr. 2 (1. Dezember 2015): 2053951715602494. doi: 10.1177/2053951715602494.

Underwood, Ted und Matthew L. Jockers. „Text-Mining the Humanities." In *A New Companion to Digital Humanities*, hg. v. Susan Schreibman, Ray Siemens und John Unsworth, 291–306. John Wiley & Sons, 2016.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, u. a. „The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3, Nr. 1 (Dezember 2016): 160018. doi: 10.1038/sdata.2016.18.

Wodausch, David, Maik Fiedler, Ben Heuwing und Thomas Mandl. „Hinterlistig – schelmisch – treulos – Sentiment Analyse in Texten des 19. Jahrhunderts: Eine exemplarische Analyse für Länder und Ethnien." Vortrag auf der DHd 2018 Kritik der digitalen Vernunft. 5. Tagung des Verbands „Digital Humanities im deutschsprachigen Raum" (DHd 2018), Köln, 26. Februar 2018. doi: 10.5281/zenodo.4622483.