Masterarbeit zur Erlangung des Grades
Master of Science Medieninformatik

Qualitätskriterien für die automatische Inhaltsanalyse

Zur Integration von Verfahren des maschinellen Lernens
in die Kommunikationswissenschaft

Laura Linda Laugwitz

Erster Prüfer **Prof. Dr. Joachim Trebbe**
Arbeitsstelle Medienanalyse
Freie Universität Berlin

Zweiter Prüfer **Prof. Dr. Martin Emmer**
Arbeitsstelle Mediennutzung
Freie Universität Berlin

31. August 2020

Abstract

In der automatischen Inhaltsanalyse werden die standardisierte manuelle Inhaltsanalyse der Kommunikationswissenschaft und die automatische Textklassifikation des überwachten maschinellen Lernens verbunden, um die wachsende Menge medial vermittelter Inhalte zu beschreiben, analysieren und vergleichen. Qualitätskriterien für die manuelle Inhaltsanalyse zielen auf Validität und Reliabilität ab, während sich Gütekriterien für die automatische Textklassifikation größtenteils mit Reproduzierbarkeit befassen. Zweifel daran, ob Textklassifikationsmodelle inhaltlich relevante Features lernen anstatt auf Scheinkorrelationen oder Artefakte im Datensatz trainiert zu werden, deuten auf ein Validitätsproblem für die automatische Inhaltsanalyse hin: Kommunikationswissenschaftliche Forschung, die diese Methode nutzt, muss sicherstellen, dass die automatische Textklassifikation nicht lediglich zuverlässig reproduziert, sondern gültige Ergebnisse liefert. Diese Arbeit bündelt epistemologische Differenzen in Sozialwissenschaften und Informatik und zeigt daraus resultierende Reibungspunkte im Umgang mit Theorie, Methodologie, Qualitätskriterien und dem Forschungsprozess in der Kommunikationswissenschaft und dem Machine Learning auf. Im Vergleich der Forschungsprozesse zeigt sich, dass das Kriterium der Erklärbarkeit im maschinellen Lernen als Streben nach Validität zu verstehen ist. Daraufhin wird geprüft, welche Erklärbarkeitsstrategien im maschinellen Lernen für eine Validitätsprüfung nutzbar gemacht werden können. Empfehlungen für eine Weiterentwicklung der automatischen Inhaltsanalyse umfassen die Entwicklung übergreifender Qualitätskriterien, eines interdisziplinären Forschungsprozesses und die Auseinandersetzung mit den grundlegenderen epistemologischen und methodologischen Konflikten.

Abstract (English)

In automated content analysis, communications science's standardized manual content analysis and supervised machine learning's text classification are combined to describe, analyze, and compare the growing amount of media content. Quality criteria for manual content analysis aim at validity and reliability, while quality criteria for text classification are largely concerned with reproducibility. Doubts about whether text classification models learn features relevant to the content rather than being trained on spurious correlations or artifacts in the dataset point to a validity problem for automated content analysis: communication researchers using this method must ensure that automated content analysis does not merely replicate by chance, but produces valid results. This thesis summarizes epistemological differences in social science and computer science and highlights resulting frictions in dealing with theory, methodology, quality criteria, and the research process in communication science and machine learning. In comparing research processes, it is shown that the criterion of explainability in machine learning may be understood as a pursuit of validity. Thereupon, it is examined which explainability strategies in machine learning can be leveraged for validity testing. Recommendations for further development of automated content analysis include the development of overarching quality criteria, an interdisciplinary research process, and addressing the more fundamental epistemological and methodological conflicts.

Inhaltsverzeichnis

1	Einleitung	1
2	Erkenntnistheoretische Vorannahmen	5
2.1	Sozialwissenschaften	6
2.2	Informatik	9
2.3	Konsequenzen für Kommunikationswissenschaft und Machine Learning	14
3	Qualitätssicherung automatischer Inhaltsanalyse	21
3.1	Manuelle Inhaltsanalyse: Qualitätskriterien	21
3.1.1	Validität	22
3.1.2	Reliabilität	25
3.2	Automatische Textklassifikation: Gütekriterien	28
3.2.1	Accuracy, Precision und Recall	33
3.2.2	Performanz, Sicherheit und ethische Standards	34
3.2.3	Reproduzierbarkeit	36
3.2.4	Transparenz	37
3.3	Herausforderungen für die automatische Inhaltsanalyse	39
4	Erklärbarkeit als Validitätskriterium	50
4.1	Struktur von Erklärungen	52
4.2	Erklärbare Textklassifikation	54
4.2.1	Modellunabhängige Erklärungen	57
4.2.2	Modellbasierte Erklärungen	59
4.2.3	Interpretierbare Modelle	64
4.3	Eignung für die Kommunikationswissenschaft	67
5	Empfehlungen für die Integration automatischer Textklassifikation in kommunikationswissenschaftliche Forschung	70
5.1	Reporting	72
5.2	Forschungsprozess	73
5.3	Kooperative Weiterentwicklung	76
6	Fazit und Ausblick	78
	Glossar	81

Einleitung

Als eine der Kernmethoden der Kommunikationswissenschaft bietet die Inhaltsanalyse seit Jahrzehnten das Werkzeug, mit dem medial vermittelte Inhalte beschrieben, analysiert und verglichen werden (vgl. Lacy et al., 2015; Früh, 2017). Aufgrund wachsender Datenmengen, zugänglicherer Tutorials und günstigerer Rechenkapazitäten werden seit einigen Jahren vor allem Inhaltsanalysen von Text vermehrt durch Verfahren des *maschinellen Lernens*¹ unterstützt (vgl. Boyd und Crawford, 2012; Trilling und Jonkman, 2018). Mit Hilfe der *überwachten Textklassifikation* wird beispielsweise auf Basis eines manuell codierten Korpus ein *Modell* trainiert, das die für den Korpus ausgearbeiteten Konzepte und deren Zuweisung möglichst ähnlich an neuem Material reproduziert (vgl. Boumans und Trilling, 2016). Dafür werden keine festen Regeln oder Anweisungen verwendet, vielmehr leitet ein Modell aus dem codierten Korpus *induktiv* Klassifikationsentscheidungen ab. Dieser Ansatz, in dem also manuelle Inhaltsanalyse und automatische Textklassifikation miteinander verbunden werden, findet in der Kommunikationswissenschaft unter dem Begriff der automatischen Inhaltsanalyse vermehrt Verwendung (vgl. Scharkow, 2012; Ziegele et al., 2018). Durch diese Automatisierung kann die Inhaltsanalyse besser skalieren, jedoch bringt sie auch neue Herausforderungen mit sich. Denn die Kommunikationswissenschaft hat für eine standardisierte manuelle Inhaltsanalyse bereits Qualitätssicherungsstrategien und -kriterien etabliert, im maschinellen Lernen existieren für die automatische Textklassifikation jedoch keine vergleichbaren Standards (vgl. Cohen und Howe, 1989; Stodden, 2010).

Qualitätskriterien für die manuelle Inhaltsanalyse zielen auf Validität und Reliabilität ab (vgl. Merten, 1983; Krippendorff, 2004; Rössler, 2010; Früh, 2017), während sich Gütekriterien für die automatische Textklassifikation größtenteils mit Reproduzierbarkeit (vgl. Stodden, 2010; Henderson et al., 2018; Lipton und Steinhardt, 2019) befassen. Tatsächlich wird die Sicherung von Validität und Reliabilität für die automatische Inhaltsanalyse in der Kommunikationswissenschaft bereits diskutiert (vgl. Scharkow 2012, vgl. Atteveldt und Peng 2018). Diese Ansätze gehen davon aus, dass Qualität in diesem interdisziplinären Forschungsprozess größtenteils durch die Erstellung eines manuell möglichst einwandfrei codierten Trainingsdatensatzes – auch Goldstandard genannt – gesichert wird. Für die Auswertung des Schrittes der automatischen Textklassifikation wird dann lediglich auf die wenigen halbwegs

¹Diese Arbeit greift auf Wissen mehrerer Disziplinen zurück. Wichtige Begriffe werden deswegen bei erstmaliger Verwendung hervorgehoben und im Glossar festgehalten.

etablierten Gütekriterien (Accuracy, Precision und Recall) zurückgegriffen. Weniger etablierte Kriterien wie Performanz, Sicherheit oder Transparenz werden nicht berücksichtigt. Allerdings befindet sich die Informatik selbst aktuell in einer Krise bezüglich dieser Kriterien und deren Auswirkung auf die Güte ihrer Modelle. Erstens reicht die Offenlegung der Gütekriterien nicht aus, um eine Studie zuverlässig zu wiederholen (vgl. Pineau, 2020; Dodge et al., 2019; Mitchell et al., 2019). Da weitere Standards für eine Veröffentlichung fehlen, sind Ergebnisse nicht reproduzierbar, selbst wenn der genutzte *Algorithmus* theoretisch immer das selbe Ergebnis liefern sollte (vgl. Henderson et al., 2018). Zweitens entstehen vermehrt Zweifel daran, ob Modelle überhaupt inhaltlich relevante *Features* lernen und nicht lediglich auf Scheinkorrelationen oder Artefakte im Datensatz trainiert werden (vgl. Lapuschkin et al. 2019). Die Nutzung von *Machine Learning* Methoden fordert von Kommunikationswissenschaftler:innen² somit nicht nur das Erlernen neuer technischer Fähigkeiten, sondern vor allem eine neue Reflexion der Qualitätskriterien, die für eine automatische Inhaltsanalyse notwendig sind.

Als Beispiel für die verkürzte Nutzung von Verfahren des maschinellen Lernens bieten sich Sentimentanalysen, die automatisiert Stimmungen in Texten identifizieren sollen. An ihnen zeigt sich unter anderem, dass Modelle Texten eine negative Stimmung zuweisen, sobald sie Identitätsbegriffe wie „jew“ oder „black“ beinhalten und dass Hate Speech *Klassifikatoren* lernen, generell Sätze als toxisch einzuordnen, die den Begriff „islam“ beinhalten (vgl. Waseem und Hovy, 2016). Für eine Anwendung, die privatwirtschaftliche Unternehmen bei der Identifikation von potenziell problematischen Diskussionen unterstützen soll, mag diese Art von Fehler akzeptabel sein. Für eine wissenschaftliche Untersuchung jedoch ist solch eine verallgemeinernde Fehlklassifikation fatal: Die Mehrdeutigkeit von Sprache ist ein grundlegendes Merkmal von Kommunikation; ihre Komplexität muss demnach auch in der automatischen Inhaltsanalyse adäquat abzubilden sein. Und auch scheinbar sinnvolle Ergebnisse können darüber hinweg täuschen, dass Forschende nicht wissen, auf welcher Basis eine Klassifikation tatsächlich getroffen wurde. Modelle des maschinellen Lernens sind für Forschende *black boxes* und der tatsächliche Erkenntnisgewinn bleibt gering.

Es reicht nicht aus, wenn die Schnittstelle zwischen Kommunikationswissenschaft und Machine Learning für die automatische Inhaltsanalyse lediglich die Daten sind, mit denen beide arbeiten. Qualitäts- oder Gütekriterien, die in der Logik der jeweiligen Disziplin funktionieren, sind unzulänglich, wenn die epistemologische Basis in einer Untersuchung nicht die gleiche ist. Denn obwohl beispielsweise eine manuelle Inhaltsanalyse die Zuverlässigkeit ihrer Ergebnisse in Form von Reliabilitätswerten

²In dieser Arbeit wird zwecks gendergerechter Sprache die geschlechtsneutrale Pluralbildung oder der Doppelpunkt verwendet, welcher auch von Screenreadern mit einem Glottisschlag vorgelesen wird und damit barrierearmer ist als das Sternchen oder der Unterstrich. Ein Diskussion der unterschiedlichen Varianten findet sich unter anderem bei *Zwei Punkte für ein Halleluja* (2020).

festhält, geht die automatische Textanalyse darauf folgend davon aus, dass sie mit einem Goldstandard arbeitet, den es möglichst genau zu reproduzieren gilt. Was mit einem Maß für Übereinstimmung beginnt, endet in der Prüfung von Korrektheit. Während die quantitative manuelle Inhaltsanalyse somit kritisch-rational vorgeht, ist die automatische Textklassifikation maßgeblich ingenieurwissenschaftlich orientiert. So ist es keine Überraschung, dass erstere mit „Qualität“ die Produktion von zuverlässigem, gültigem und intersubjektiv nachvollziehbarem Wissen versteht (vgl. Brosius et al., 2009) und letztere mit „Güte“ die Entwicklung einer zufriedenstellenden, wiederverwendbaren Anwendung meint (vgl. Stodden, 2010). Es gilt also erstens, gemeinsame Qualitätskriterien für die automatische Inhaltsanalyse zu entwickeln.

Während Reliabilität in der Kommunikationswissenschaft und Reproduzierbarkeit im Machine Learning zwar unterschiedlich gerahmt werden, ist eine Zusammenführung dieser beiden Kriterien naheliegend. Validität hingegen scheint auf den ersten Blick im Machine Learning weniger Aufmerksamkeit zu erfahren. Allerdings ist die Frage danach, was ein Modell tatsächlich gelernt hat, sowohl in dem Kriterium der Sicherheit als auch in dem der Transparenz zu finden. Seit der Einführung der neuen Datenschutzgrundverordnung im Jahr 2018 haben User:innen ein Recht darauf, automatische Entscheidungen nachvollziehen zu können (§ 14 2 g DSGVO). Aber auch davor wurde die Transparenz von Algorithmen bereits intensiv diskutiert (vgl. Kulesza et al., 2011; Stodden und Miguez, 2013; Burrell, 2016; Biran und Cotton, 2017; Mueller et al., 2018). Bestrebungen, die Erklärbarkeit von automatisierten Ergebnissen zu erhöhen, gewinnen in den letzten Jahren unter dem Begriff der *explainability* im maschinellen Lernen an Relevanz (vgl. Samek und Müller, 2019). Teilweise ist es möglich, einzelne Klassifikationen nachvollziehbarer zu machen, indem für die Entscheidung relevante Eigenschaften des Inputs aufgezeigt werden. Aber auch die Relevanz einzelner Features für das Modell im Ganzen kann, je nach zugrunde liegendem Algorithmus, bereitgestellt werden. Ein Großteil der Forschung erarbeitet diese Transparenz, um in Endnutzer:innen Vertrauen zu generieren oder um Sicherheitsrisiken in kritischen Anwendungen zu verringern. Die Entwicklung einer zufriedenstellenden Anwendung ist demnach auch abhängig von den Anforderungen ihrer Nutzer:innen und des Anwendungsgebiets. Begreifen wir Kommunikationswissenschaftler:innen als Nutzer:innen, kann deren Anforderung an eine automatische Textklassifikation die Erklärbarkeit eines Modells sein, um dessen Validität zu prüfen. Es gilt also zweitens, für die automatische Inhaltsanalyse angemessene Erklärbarkeitsstrategien zu identifizieren.

Gemeinsamen Qualitätskriterien für die automatische Inhaltsanalyse, die tatsächlich die unterschiedlichen epistemologischen Standpunkte reflektieren, können fehlerhafte Inferenzen aufgrund von Miss- oder Unverständnis zwischen Kommunikationswissenschaftler:innen und Machine Learning Expert:innen vermeiden. Welche gemeinsamen Qualitätskriterien braucht die automatische Inhaltsanalyse und wie

können Modelle der überwachten Textklassifikation so transparent sein, dass sie diesen Kriterien gerecht werden? Das Ziel dieser Arbeit ist erstens die Übersetzung erkenntnistheoretischer Vorannahmen zwischen Kommunikationswissenschaft und Machine Learning, um gemeinsame Qualitätskriterien zu entwickeln. Zweitens sollen Erklärbarkeitsstrategien im maschinellen Lernen auf ihre Anwendbarkeit in der automatischen Inhaltsanalyse für die Validitätsprüfung untersucht werden.

Zur Umsetzung dieser Ziele arbeitet Kapitel 2 erkenntnistheoretische Positionen in den Sozialwissenschaften und der Informatik aus und zeigt daraus resultierende Reibungspunkte im Umgang mit Theorie, Methodologie, Qualitätskriterien und dem Forschungsprozess in der Kommunikationswissenschaft und dem Machine Learning auf. Potenzielle Schnittpunkte werden ebenfalls genannt. Die Konventionen zur Sicherung der Qualität wissenschaftlicher Erkenntnisse, die sich daraus ergeben, werden in Kapitel 3 anhand des Forschungsprozesses der manuellen Inhaltsanalyse und der automatischen Textklassifikation dargestellt. Es zeigt sich hier, dass das Kriterium der Transparenz im maschinellen Lernen als Streben nach Validität zu verstehen ist, aber selten so benannt wird. Welche Erklärbarkeitsstrategien für Modelle und einzelne Klassifikationen für eine Validitätsprüfung nutzbar gemacht werden können, zeigt Kapitel 4. Kapitel 5 bündelt letztlich Empfehlungen für eine Weiterentwicklung der automatischen Inhaltsanalyse, die sich gemeinsame Qualitätskriterien auferlegt, einen interdisziplinären Forschungsprozess umsetzt und sich mit den grundlegenderen epistemologischen und methodologischen Konflikten auseinandersetzt.

Erkenntnistheoretische Vorannahmen

„ *It is one thing to identify patterns; it is another to explain them.*

— **Kitchin, 2014, S.8**

Wissenschaft ist mit den verschiedensten Aufgaben vertraut: Sie kann Phänomene beschreiben, analysieren, modellieren, kritisieren, formen oder erklären (vgl. Brühl, 2014). Manche Fachgebiete tragen ihr Ziel bereits im Namen, so beispielsweise die Mustererkennung als Unterbereich des maschinellen Lernens. Andere Disziplinen hingegen tragen den Gegenstand ihrer Betrachtung im Namen, nicht aber ihr Ziel, wie beispielsweise die Kommunikationswissenschaft. Welche Ziele oder Aufgaben können sich ergeben, wenn maschinelles Lernen und Kommunikationswissenschaft miteinander verknüpft werden? Ist es lediglich die Aufgabe des Machine Learnings, Muster zu erkennen und die der Sozialwissenschaften, Muster zu erklären? Rob Kitchin legt dies nahe, wenn er sich auf die Integration von Big-Data-Methoden in sozialwissenschaftliche Forschung bezieht. Zwar impliziert er weder eine Hierarchisierung noch eine zeitliche Abfolge, jedoch besteht er auf eine Unterscheidung in der Relevanz theoretischer Fundierung der beiden Aufgaben: Während das Erkennen von Mustern ohne Theorie möglich sei, benötigt man für die Erklärung eines Musters sozialwissenschaftliche Theorie(n) und tiefes Kontextwissen. Das Aufdecken eines Musters sei demnach nie das Ende einer Analyse, sondern vielmehr erst ihr Anfang (vgl. Kitchin, 2014, S. 8).

Ein integrativer oder auch interdisziplinärer Ansatz von Machine Learning und Kommunikationswissenschaft kann jedoch mehr fordern: dass beide Disziplinen an einem gemeinsamen Ziel arbeiten und nicht nur Teilaufgaben untereinander aufsplitten, sondern tatsächlich kooperativ forschen. Unterschiedliche Wissenschaftsverständnisse und Forschungspraxen könnten diese Form von Kooperation erschweren, aber nicht zwingend verunmöglichen. In diesem Kapitel werde ich daher Unterschiede und Gemeinsamkeiten explizieren, um potenzielle Anknüpfungs- und Reibungspunkte erkennbar zu machen.

2.1 Sozialwissenschaften

Als Sozialwissenschaft reiht sich Kommunikationswissenschaft in eine lange Tradition soziologischer Reflexionen und Auseinandersetzungen ein. Die Frage danach, was überhaupt Wissen ist und wie dieses erkannt werden kann, war und bleibt ein zentraler Punkt jener Auseinandersetzungen und lässt sich unter dem Begriff der Erkenntnistheorie, oder auch Epistemologie, fassen. Von den Dreißigerjahren bis in die Neunzigerjahre trug Karl Popper maßgeblich zur Erkenntnistheorie bei und prägte sowohl theoretische Überlegungen als auch empirische Forschungspraxen.

Popper ist Vertreter des kritischen Rationalismus, welcher ein *deduktives* Vorgehen in der Wissenschaft fordert: Thesen müssen aufgestellt und empirisch geprüft werden, bevor sie vorläufig als gültig anzunehmen oder zu verwerfen sind. Dass sie wahr seien, könne jedoch nicht bewiesen werden, man könne sie stattdessen beibehalten, solange sie nicht empirisch widerlegt wurden (vgl. Popper, 1945). Dieses Prinzip der Falsifikation geht also davon aus, dass Wissenschaft sich der Wahrheit durch empirische Prüfung nähern muss, anstatt vermeintliche Beweise zu finden. Somit habe Erkenntnis keine letzte, autoritative Quelle – die Wahrheit stehe „über jeder menschlichen Quelle“ (Popper, 1960, S. 39). Damit bot Popper eine Alternative zum induktiv vorgehenden Empirismus, der dem logischen Problem erlegen war, dass die Ansammlung einzelner Beobachtungen keinen verallgemeinernden Schluss zulässt (vgl. Popper, 1953, S. 85).

Wissenschaftliche Erkenntnis ordnet Popper dem Bereich der „dritten Welt“, oder auch der „Welt der *objektiven Gedankeninhalte*“ (Herv. im Original, Popper, 1967, S. 40) zu. Anders als die physikalische und die geistige Welt sei jene Welt die der Theorien, Probleme, Argumente und des Forschungsstandes. In ihr könne Erkenntnis ohne erkennendes Subjekt stattfinden, sie sei objektiv und selbstständig (Popper, 1967, S. 49). Er wirft der „*herkömmliche[n] Erkenntnistheorie*“ (Herv. i. O., Popper, 1967, S. 55) vor, sich mit der geistigen Welt zu beschäftigen, in der Wissen als eine Form von Glauben verstanden werde, während Wissenschaft jedoch in der dritten Welt operiere, in der sie bemüht sei, „*ihre falschen Theorien auszumerzen*“ (Herv. i. O., Popper, 1967, S. 55).

Als *modus operandi* in dieser Welt der Gedankeninhalte entwickelt Popper die wissenschaftliche Methode. Er versteht Erkenntnistheorie als Methodenlehre und stellt deswegen methodologische Regeln auf, die Forschende befolgen, wenn sie Wissenschaft – unabhängig der Disziplin – betreiben (vgl. Popper, 1934). Diese Einheitslogik für Wissenschaft besagt, dass Theorien mit der Realität konfrontiert werden müssen, indem aus ihnen logisch widerspruchsfreie, empirisch überprüfbare und falsifizierbare Hypothesen abgeleitet werden, die dann einer intersubjektiv nachvollziehbaren

empirischen Überprüfung unterzogen werden (vgl. Scheufele und Engelmann, 2009). Sozial- und naturwissenschaftliche Forschungen können somit den gleichen Regeln folgen, was besonders den Sozialwissenschaften zu einem objektiveren, solideren Fundament verhelfen soll. Die Konventionen, die sich aus dieser Methodenlehre ableiten, werden im späteren Verlauf dieses Kapitels aufgegriffen.

Nachdem Popper Erkenntnisgewinn wie oben gezeigt konkretisierte, entwickelten sich in den 1960er Jahren alternative Vorstellungen davon, wie Wissenschaft Erkenntnisse erlangen kann. Thomas Kuhn zeigte historisch auf, dass die Entwicklung von Theorien bestimmten Paradigmen unterliegt, die sich im Laufe der Zeit verändern. Ursprünglich unterschied er zwischen Normalwissenschaft, die sich durch das Anhäufen von weiteren Erkenntnissen entwickelt und wissenschaftlichen Revolutionen, die es erfordern, Teile des als wahr geglaubten Wissens zu verwerfen. Dies korrigierte er jedoch später in die Unterscheidung zwischen Entwicklungen, die zu einer Veränderung in wissenschaftlicher Taxonomie führen und jenen, bei denen dies nicht geschieht (vgl. Kuhn, 1990).

Diesem Wissenschaftsverständnis zufolge wäre auch Poppers kritischer Rationalismus lediglich ein Paradigma, welches mit der Zeit überholt und revolutioniert werden kann. Kuhn beschreibt außerdem, wie sich nach einer wissenschaftlichen Revolution neue Disziplinen abspalten und neue Communities von Forschenden bilden, die sich spezialisieren (vgl. Kuhn, 1990). Im Rahmen ihrer neuen Taxonomie vollziehe sich jedoch auch eine Isolation. Daraus folge unweigerlich ein Bruch in der Kommunikation zwischen diesen wissenschaftlichen Communities – die unterschiedlichen Taxonomien würden dies erzwingen. Anders als Popper versteht Kuhn Wahrheit als relativ für jene Communities: „Conceptually, the world is *our* representation of *our* niche, the residence of the particular human community with whose members we are currently interacting“ (Herv. i. O., Kuhn, 1990, S. 11). Somit legt Kuhn auch einen Blick auf erkennende Subjekte, vor allem als Gruppe mit einem geteilten Lexikon. Er verwehrt sich dabei der Bewertung unterschiedlicher Disziplinen und deren wissenschaftlichen Praktiken, postuliert sogar, dass sich diese Heuristiken einer Prüfung auf Korrektheit oder Falschheit entziehen (vgl. Kuhn, 1990, S. 12). Trotzdem fordert er für die bewusste Auswahl von Theorien Kriterien wie Genauigkeit, Konsistenz, Reichweite, Einfachheit und Ergiebigkeit (vgl. Kuhn, 1977). Dem Vorwurf des Relativismus konnte Kuhn dennoch nicht entgehen (vgl. Shapere, 1984).

Parallel zu Kuhns Positionen entwickelten sich seit den Sechzigerjahren auch feministische Kritiken am wissenschaftlichen System, wissenschaftlicher Praxis und Epistemologie. Der Ausschluss sowie die geringe Wertschätzung von Frauen und marginalisierten Gruppen in der Wissenschaft führte dazu, dass ganze Lebensrealitäten unterschlagen wurden – als Gegenstand von Forschung, aber auch als Perspektiven

in der Forschung. Feministische Kritik zeigt auf, dass das Verhältnis von Gesellschaft und Wissenschaft entgegen Poppers Vorstellungen eben nicht vorwissenschaftlich ist, sondern Einfluss an diversen Stellen im Forschungsprozess nimmt (vgl. Harding, 1986). Daraus folgen zwei Ziele: Einerseits soll Wissenschaft in der Form, in der sie aktuell betrieben wird, anhand ihrer eigenen Qualitätskriterien herausgefordert und verbessert werden, indem Leerstellen aufgedeckt werden. So kann (feministische) Kritik zu einer präziseren wissenschaftlichen Praxis beitragen. Andererseits soll eine neue Form von Wissenschaft hervorgebracht werden, die von jenen Strukturen losgelöst ist, die überhaupt erst Ausschlüsse und Leerstellen in der Wissenschaft bewahren und reproduzieren.

Der Versuch des kritischen Rationalismus, Sozialwissenschaften in das Korsett naturwissenschaftlicher Forschung zu zwängen und gesellschaftliche Zusammenhänge in den Bereich des Vorwissenschaftlichen zu drängen, ist problematisch. Während beispielsweise die Physik „entweder simple Systeme oder simple Aspekte komplexer Systeme“ (Harding, 1986, S. 44) betrachtet und diese als deterministisch und abgeschlossen versteht, können die Sozialwissenschaften den komplexen Faktor Mensch nicht ausblenden. Anstatt also die Physik als das Paradebeispiel für empirische Wissenschaft zu nehmen und alle anderen Disziplinen als davon abweichend und unzulänglich zu betrachten, weil ihre Untersuchungsgegenstände ungünstigerweise durch soziale Zusammenhänge beeinflusst werden, schlägt Harding eine kritische und selbstreflexive Sozialwissenschaft als Modell für Wissenschaft vor (Harding, 1986, S. 48).

In diesen unterschiedlichen Konzeptionen von Wissenschaft finden sich verschiedene Werte. Popper fordert für die Wissenschaft eine Einheitslogik mit hypothetisch-deduktiver Methode und Objektivität in Form von Unabhängigkeit vom forschenden Subjekt. Kuhn verabschiedet sich von einem einheitlichen Wissenschaftsverständnis und der Idee einer Objektivität außerhalb der Subjekte. Objektivität ist wiederum für Harding ein wichtiges Kriterium von Wissenschaft, soll jedoch nicht als Wertneutralität missverstanden werden. Das Vertreten bestimmter Werte könne überhaupt erst zu wissenschaftlicher Objektivität führen:

„A maximally objective science, natural or social, will be one that includes self-conscious and critical examination of the relationship between the social experience of its creators and the kinds of cognitive structures favored in its inquiry“ (Harding, 1986, S. 250).

Es ist diese Positionierung, die sich von Poppers methodologisch einheitlichem oder Kuhns unüberbrückbar vereinzelter Wissenschaftsverständnis löst und Alternativen entwickelt, in denen unterschiedliche Standpunkte zusammenkommen können.

Dieser Ansatz ist zuvorderst politisch, da Hardings Konzept von Standpunkten an soziale Positionen gebunden ist, die aus Arbeitsteilung und darin resultierenden Machtverhältnissen entstehen. Auch für diese Arbeit steht an erster Stelle die Einsicht, dass Wissen ein soziales Produkt ist. In diesem Sinne sind kritische Arbeiten, die Fehler und Gefahren in gängiger wissenschaftlicher Praxis und genutzten Technologien aufdecken¹, ein Versuch, andere Standpunkte sichtbar zu machen. Anstatt jedoch vereinzelte Gegenpositionen aufzuzeigen und lediglich im Nachhinein Kritik zu bündeln, bietet Pohlhaus das Konzept der *knowing communities*. Diese Gemeinschaften definieren sich nicht über gemeinsame Eigenschaften, sondern über das strategische, bewusste Streben nach gemeinsamem Wissen. Darin liegt, wenn auch vermutlich nicht so beabsichtigt, eine Chance für interdisziplinäre Forschung:

„[U]nderstanding knowledge as always formed within the relations of a particular community leaves open the possibility of new members entering that community by forming relations within it. Our questions concerning how to understand the world more objectively will require us to ask, not how I might see from a particular standpoint, but rather how I may move with (or forge) new knowing communities“ (Pohlhaus, 2002, S. 292).

Mit dem Konzept der *knowing communities* werden der Austausch zwischen unterschiedlichen Wissenschaftsverständnissen und das Bilden eines gemeinsamen Vokabulars möglich. Interdisziplinäre Forschung kann so fundierter agieren. Im Rahmen dieser Arbeit sollen also zwei Ziele berücksichtigt werden: Die Verbesserung wissenschaftlicher Arbeiten durch das Aufdecken von Lücken in der Umsetzung eigener Qualitätsansprüche und das Anregen einer neuen *knowing community*, in der Sozialwissenschaftler:innen und Informatiker:innen aktiv daran arbeiten, die eigenen Positionierungen sowie die der anderen zu verstehen, kritisch zu betrachten und miteinander weiter zu entwickeln.

2.2 Informatik

Die Zuordnung der Informatik in eine wissenschaftliche Tradition ist eine Herausforderung. Einzelne Aspekte der Disziplin gehören in die Mathematik (Entwicklung und Prüfung neuer Algorithmen), andere in die Ingenieurwissenschaften (Software Engineering), wieder andere nähern sich den Sozialwissenschaften (Human Computer Interaction) an (vgl. Marcos und Marcos, 1999; Freitas, 2009; Warfield, 2010). Wegner (1976) sieht auch einen naturwissenschaftlichen Strang in der Informatik und

¹Mit Harding (1986) ist davon auszugehen, dass sowohl die Definition von Objektivität als wertneutral als auch von Technik als objektiv im gleichen Sinne, durch androzentrische Strukturen in Wissenschaft und Technik gewachsen sind.

Lipton und Steinhardt beschreiben die Ziele computerwissenschaftlicher Arbeiten wie folgt:

„In a given paper, researchers might aspire to any subset of the following goals, among others: to theoretically characterize what is learnable, to obtain understanding through empirically rigorous experiments, or to build a working system that has high predictive accuracy“ (Lipton und Steinhardt, 2019).

Im Diskurs um die Informatik als wissenschaftliche Disziplin stellt sich jedoch nicht nur die Frage, *welcher* Wissenschaft sie unterzuordnen ist, sondern auch, *ob* sie überhaupt als Wissenschaft zu verstehen ist. Tedre (2011) zeichnet unterschiedliche Perspektiven auf die Disziplin nach. Sie könne sowohl als Naturwissenschaft (vgl. Denning, 2005), *unnatürliche* Wissenschaft (vgl. Knuth, 2003), *artifizielle* Wissenschaft (vgl. Simon, 1990), als Nicht-Wissenschaft (vgl. Hartmanis, 1993; McKee, 1995) sowie auch als schlechte Wissenschaft (vgl. Holloway, 1995) gerahmt werden. Tedre identifiziert drei Herausforderungen in der Definition der Informatik als Wissenschaft: Generell bestünde keine Einigung unter den Positionen darüber, was überhaupt Wissenschaft ausmacht, welche Eigenschaften, Ziele und Methoden die Informatik gemeinsam hat und wie grundlegende wissenschaftliche Begriffe für das Fach zu definieren sind (vgl. Tedre, 2011, S. 371). Unter anderem begründet er diese Undeutlichkeit darin, dass das Fach noch vergleichsweise jung sei (vgl. Tedre, 2011, S. 362). Die epistemologische Verortung des Faches ist nicht unmöglich, aber auch nicht frei von Widersprüchen.

Die Ursprünge der Informatik finden sich in der Mathematik und dem Bestreben, Berechnungen zu automatisieren (vgl. Tedre, 2011). Eine Möglichkeit der epistemologischen Verortung der Informatik besteht also darin, jene Diskussionen in der Mathematik zu betrachten. Die Frage danach, was die Mathematik beschreiben und wissen kann, zeigt sich besonders in dem metamathematischen Konflikt des frühen zwanzigsten Jahrhundert, in dem sich Logizismus, Formalismus und Intuitionismus gegenüber standen. Der Logizismus nach Frege geht von einer „*a priori*-Natur der Mathematik [...] [als] eine Folge der uneingeschränkten Allgemeinheit des logischen Wissens“ (George und Velleman, 2018, S. 195) aus. Die Idee, dass die Mathematik letztendlich auf Logik zurückgeführt werden könne, wurde jedoch 1902 durch Russell (vgl. Horsten, 2019) und dann 1931 durch Gödel (vgl. George und Velleman, 2018, S. 196) widerlegt. Im scharfen Kontrast zum Logizismus entwickelte sich der Intuitionismus, der die Mathematik vielmehr als ein Ergebnis menschlicher Konstruktion versteht (vgl. Horsten, 2019). Dieser Standpunkt hat vor allem Auswirkung auf die Mathematik im Bereich der Unendlichkeit, indem der Intuitionismus „das aktual Unendliche verdammt und nur das potentiell Unendliche billigt“ (George

und Velleman, 2018, S. 197). Diese Perspektive war jedoch unvereinbar mit klassischer Mathematik und Logik und wurde deswegen nicht breit angenommen. Der Formalismus nach Hilbert nimmt einen weniger konstruktivistischen Standpunkt ein. Natürliche Zahlen versteht er als Symbole oder abstrakte Objekte, die jedoch in konkreten Objekten verkörpert werden können (vgl. Parsons, 2007). Höhere Mathematik sei lediglich ein Spiel nach festgelegten Regeln, Axiome genannt, die sich aber letztendlich auch in simpler Arithmetik beweisen ließen (vgl. Horsten, 2019). Hilberts *Entscheidungsproblem* sucht nach einem Algorithmus, der eine Aussage als Input nimmt und am Ende eine universell valide Entscheidung darüber trifft, ob diese wahr oder falsch ist. Hilberts Formalismus ging davon aus, dass dies möglich sei. Doch auch dieser Standpunkt wurde durch Gödels Unvollständigkeitssätze – dem Beweis, dass manche Aussagen weder als wahr noch falsch bewiesen werden können – in Frage gestellt.

Es waren diese Unvollständigkeitsgesetze und das Entscheidungsproblem, die Alan Turing in den Dreißigerjahren dazu motivierten, eine universelle Maschine zu modellieren, die das Entscheidungsproblem lösen sollte. 1936, viele Jahre bevor Computer erstmals in Militär und Forschung eingesetzt wurden, veröffentlichte Alan Turing „On Computable Numbers, with an Application to the Entscheidungsproblem“. Anhand der eigens dafür entwickelten Turingmaschine zeigt seine Arbeit, dass es keinen Algorithmus gibt, der den Wahrheitsgehalt für jede ausreichend formalisierte Aussage der Mathematik entscheiden kann (vgl. Petzold, S. 260, 2008). Für die Informatik bedeutete Turings Arbeit auch, dass es kein Programm geben kann, welches im Vorhinein entscheiden kann, was ein anderes Programm tun oder ob es abbrechen wird. Mit der theoretischen Modellierung eines Computers, der nicht nur spezifische Aufgaben, sondern generelle lösen kann, legte Turing den Grundstein der theoretischen Informatik. Damit einhergehend entstand der Forschungsbereich der Komplexitätstheorie von Algorithmen sowie das Paradigma, dass Programme elektronisch auf Computern gespeichert werden (vgl. Tedre, 2011, S. 378). Eden ordnet diese stark mathematisch geprägte Perspektive einem rationalistischen Paradigma zu, das Programme als mathematische Objekte versteht und durch deduktives Vorgehen nach *a priori* Wissen über deren Korrektheit strebt (vgl. Eden, 2007, S. 3).

Mit der wachsenden Komplexität von Computern und ihrer Expansion in betriebliche Anwendung entwickelten sich in den 1970er Jahren neben der Beweisführung zur Korrektheit eines Algorithmus auch Methoden, um die Zuverlässigkeit von Systemen zu prüfen (vgl. Eden, 2007). Hier produziert die Informatik die Objekte und Modelle selbst, die gleichzeitig Gegenstand ihrer Betrachtung sind. Als Ingenieurswissenschaft interessiert sie sich nicht für die Entdeckung von Wahrheit(en), sondern für die Entdeckung von (Zahl-)werten, mit denen sie ihre Ergebnisse optimieren kann (vgl. Colburn und Shute, 2010). Damit rückt eine ingenieurswissenschaftliche Perspektive in den Fokus, die Programme als Daten versteht und durch empirische

Testverfahren wahrscheinliches, *a posteriori* Wissen über deren Reliabilität anstrebt (vgl. Eden, 2007, S. 3). Reliabilität ist hier aber nicht als das sozialwissenschaftliche Qualitätskriterium der Reproduzierbarkeit wissenschaftlicher Ergebnisse zu verstehen, sondern vielmehr als ein zuverlässiges Erreichen von zuvor definierten Anforderungen der Anwender:innen (vgl. Eden, 2007).

Neben seinem Einfluss auf die theoretische Informatik trieb Turing auch die Frage nach der Möglichkeit künstlicher Intelligenz an. In „Computing Machinery and Intelligence“ fragte er nicht, ob Maschinen denken können, sondern, ob Maschinen ausreichend überzeugend Intelligenz imitieren können (vgl. Turing, 1950). Der Turing Test, in dem die Intelligenz einer Maschine angenommen wird, wenn ein Mensch im Austausch mit ihr nicht feststellen kann, ob sie Mensch oder Maschine ist, beschäftigt auch heute noch Forschungsbereiche der Informatik. Das Feld der künstlichen Intelligenz ordnet Eden der empirischen Naturwissenschaft zu, die Programme als mentale Prozesse versteht und durch eine Kombination aus Deduktion und Experimenten sowohl *a priori* als auch *a posteriori* Wissen über diese Prozesse anstrebt (vgl. Eden, 2007, S. 3). In diesem Kontext wird häufig das Prinzip der *Abduktion* nach Peirce (1839–1914) aufgegriffen. Diese Form logischer Inferenz und Begründung geht davon aus, dass Epistemologie nicht nur mit der Begründung für das Akzeptieren – beziehungsweise das Verwerfen – einer Hypothese befasst ist, sondern auch mit der Begründung der Auswahl der Hypothese (vgl. Chauviré, 2005). Abduktion verläuft so, dass nach einer Beobachtung verschiedene Thesen für deren Existenz aufgestellt und auf ihre Plausibilität geprüft werden. Daraufhin wird die beste oder auch wahrscheinlichste, ausgewählt und empirisch geprüft (vgl. Kitchin, 2014). In einem Vergleich von Poppers kritischem Rationalismus und Peirces Abduktion zeigt Chauviré auf, dass die Differenzen zwischen beiden Vorgehen hauptsächlich sprachlich sind, jedoch nicht konzeptionell. Begründet sei dies in Poppers kategorischer Ablehnung des Begriffs der Induktion, den Peirce nutzt, jedoch anders definiert:

„[W]hat Peirce calls induction corresponds in Popper to the last phase of the deductive method of control, and that what he describes by the term abduction covers a set of procedures accepted and described by Popper, who gives them no special name. Beyond these terminological differences, they therefore reconstruct more or less the same procedures“ (Chauviré, 2005, S. 216).

Mit der in den letzten Jahren wieder wachsenden Forschung im Bereich der künstlichen Intelligenz erfährt das Konzept der Abduktion erneut Aufmerksamkeit (vgl. Marcos und Marcos, 1999; Kitchin, 2014; Margolin, 2019). Dieses Vorgehen, das sowohl eine *a priori* Akzeptanz via Begründung einer Hypothese als auch die *a*

posteriori Akzeptanz via empirischer Prüfung dieser Hypothese in den Fokus stellt (vgl. Chauviré, 2005, S. 214), bietet eine Brücke zum kritischen Rationalismus in der Kommunikationswissenschaft. Eden stellt diesen Ansatz als den für die Informatik am produktivsten heraus und begründet es ontologisch: Anders als mathematische Objekte seien Computerprozesse, welche den Hauptgegenstand der Disziplin ausmachen, temporal begrenzt, an eine physikalische Manifestation gebunden (den Computer), kausal im Sinne der Auswirkung auf physikalische Gerätschaften und metabolisch im Sinne des Verbrauchs von Energie (vgl. Eden, 2007, S. 21). Eine mathematisch-rationalistische Informatik kann er demnach nicht annehmen. Auch das ingenieurswissenschaftliche Selbstverständnis lehnt er ab, da Computerprozesse keine greifbaren, physikalischen Objekte sind, aber sehr wohl eine greifbare Auswirkung auf die physikalische Realität haben, was der empiristisch-ingenieurswissenschaftliche Blick jedoch auch nicht berücksichtigt (vgl. Eden, 2007, S. 22).

Eine einheitliche epistemologische Verortung des Faches ist somit nicht möglich und mit Blick auf diese Arbeit auch nicht zielführend. Vielmehr ist es sinnvoll, den Aspekt der Informatik aufzugreifen, der im Rahmen dieser Arbeit in kommunikationswissenschaftliche Forschung integriert werden soll: Die Anwendung von Algorithmen zur Erkennung von Mustern, also der praktische Part des maschinellen Lernens, der sowohl ingenieurs- als auch naturwissenschaftlich gerahmt sein kann. Eden identifiziert jedoch einen Widerspruch mit weitreichenden Folgen:

„[W]hile most computer scientists pledge allegiance to the scientific position, at least in principle, mainstream computer science is yet to concede the ontological commitments of the scientific paradigm“ (Eden, 2007, S. 24).

Ein standardisiertes wissenschaftliches Vorgehen ist weder in der Informatik allgemein noch im maschinellen Lernen verankert. Dieses wissenschaftliche Selbstverständnis der Disziplin ohne die Umsetzung wissenschaftlicher Standards ist nicht haltbar. Im Folgenden werde ich möglichst jene Ansätze hervorheben, die sich um einen deduktiv-empirischen Standpunkt bemühen, da diese sowohl nach Selbsteinschätzung innerhalb des Faches als auch im Hinblick auf die Integration in den kommunikationswissenschaftlichen Forschungsprozess am produktivsten erscheinen. Dennoch werden auch ingenieurswissenschaftlich-technokratische Arbeiten aufgegriffen, da diese – eine realistische Konsequenz aus Edens Analyse – wesentlich prominenter sind.

2.3 Konsequenzen für Kommunikationswissenschaft und Machine Learning

Nachdem die unterschiedlichen epistemologischen Standpunkte innerhalb der Disziplinen dargestellt wurden, wird nun genauer betrachtet, wie sich diese auf die Integration des Machine Learnings in die Kommunikationswissenschaft auswirken. Ziel der Kommunikationswissenschaft als empirische Sozialwissenschaft ist die Deskription, Analyse und Vorhersage öffentlicher Kommunikation (vgl. Karmasin, 2008). Ziel des maschinellen Lernens in diesem Kontext ist die Unterstützung jener Deskription, Analyse und Vorhersage durch Algorithmen (vgl. Scharnow, 2012).

Mehler und Lücking identifizieren insgesamt fünf Scheren zwischen den Sozial- und Technikwissenschaften, die miteinander verknüpft sind. Das *epistemological gap* bezieht sich auf grundsätzliche Herangehensweisen an die Produktion von Wissen (vgl. Mehler und Lücking, 2014, S. 93). Darüber hinaus bestünden Unterschiede im Interesse an tiefen oder breiten Daten (*data-related gap*). Das *semiotic gap* beschreibt konkret das Verständnis von Text, welcher in der Informatik lediglich eine Zeichenkette sei, während Text und Zeichen in den Sozialwissenschaften jedoch tiefere Bedeutung haben. Letztlich erwähnen die Autor:innen eine Kluft im Verständnis der Funktionalität von Methoden (*methodological gap*) und in der Bezugnahme statistischer Ergebnisse auf theoretische Erkenntnisse (*interpretation gap*). Im Folgenden werden vor allem die aus dem *epistemological gap* resultierenden konkreten Unterschiede im Umgang mit Theorie, Methodologie, Methoden, Qualitätskriterien und interdisziplinären Bestrebungen ausgearbeitet.

Theorie

Für die Kommunikationswissenschaft hat die Theorie einen forschungsleitenden oder auch rahmenden Charakter. Da sie über „kein einigendes Paradigma“ (Karmasin, 2008, S. 230) verfügt, sind Theorien ein zentraler Ankerpunkt des Faches. Für deduktives Vorgehen bilden sie den Ausgangspunkt, für induktives Vorgehen sind sie das Ergebnis (vgl. Krotz et al., 2008). Bestenfalls ergänzen sich beide Vorgehen, um sowohl die Sicherung als auch die Weiterentwicklung von Erkenntnissen in der Kommunikationswissenschaft zu ermöglichen (vgl. Baumann und Scherer, 2012). Für den theoretischen Anteil im maschinellen Lernen sind mathematische und logische Axiome die Grundlage. Ullman (1995) zählt hier beispielsweise die Automatentheorie, die Komplexität von Algorithmen und die Theorien zur Semantik von Programmiersprachen. Für die Anwendung des maschinellen Lernens hingegen spielt Theorie eine untergeordnete Rolle. Ullmann beschreibt eine kuriose Entwicklung innerhalb theoretischen Arbeiten in der Informatik:

„[A]s the theory field grew there was a tendency to attack problems obliquely, assuming that we were contributing to an important goal. Papers would be written to address the issues raised by other papers. Then another round of work would address these papers, and so on. In the process, the real problems that spawned the original paper were forgotten, and the community began to imagine that the research embodied in the papers was the ultimate goal“ (Ullman, 1995).

Auch Stol und Fitzgerald (2015) schließen sich der Kritik an, dass der Stellenwert von Theorie unterschätzt wird. Sie attestieren dem Software Engineering (SE) einen Mangel im Verständnis davon, was Theorie ist und welche Funktion sie in der Forschung einnimmt. Sie beschreiben Software Engineering in seiner aktuellen Forschung als evidenz- statt theoriebasiert und warnen davor, dass das Berichten von Ergebnissen nicht automatisch zu Erkenntnissen führt und deswegen nicht ausreicht, um sich Forschung zu nennen. Die Synthese von Ergebnissen eigener und anderer Studien sehen sie oft als mangelhaft in der Forschung:

„[T]heory is not a luxury, to be left to 'philosophers', but an essential element of SE research and thus should be considered when presenting research results in papers. Theory *should* inform the design of new studies, which will help to converge the research literature on a particular topic (or research question)“ (Stol und Fitzgerald, 2015, S. 95).

Sie zeigen außerdem Studien auf, in denen vor allem sozialwissenschaftliche Theorien aufgegriffen werden, was auf zwei Ebenen sinnvoll erscheint: Auf der Ebene der Softwareentwicklung und -anwendung sind Menschen beteiligt, sodass deren Verhalten in diesem Prozess sozialwissenschaftlich betrachtet werden kann. Auf der Ebene der Softwareinhalte kann ein Programm beliebige Datensätze verarbeiten, die menschliches Verhalten erfassen. Hier sind die Sozialwissenschaften bereits besonders gut mit Theorien ausgestattet, um die Aussagekraft von Ergebnissen zu stärken. Die Kommunikationswissenschaft bietet hier der Informatik die Chance, ihre Ergebnisse fundierter und ganzheitlicher auszuwerten und somit auch die Synthese von Wissen zu verstärken. Als Konsequenz für Kooperationen zwischen Kommunikationswissenschaft und Machine Learning folgt, dass kommunikationswissenschaftliche Theorien der Ankerpunkt für interdisziplinäre Forschung sein können. Vor allem sollte aber auch Wert darauf gelegt werden, dass die Erkenntnisse zur theoretischen Weiterentwicklung beider Bereiche beitragen können.

Methodologie

Methodologisch ist die Kommunikationswissenschaft eine empirisch-analytische Wissenschaft, die größtenteils den Prinzipien des kritischen Rationalismus folgt: Sie nutzt deduktiv-nomologische Erklärungsschemata und arbeitet größtenteils quantitativ (vgl. Scheufele und Engelmann, 2009). Aber auch qualitative Vorgehensweisen finden sich in kommunikationswissenschaftlicher Forschung (vgl. Meyen et al., 2011). Für den Rahmen dieser Arbeit, die sich letztlich auf die quantitative Inhaltsanalyse fokussiert, werde ich jedoch qualitative Forschung größtenteils ausklammern.

Die methodologische Einordnung des maschinellen Lernens wird erschwert durch die doppelte Bedeutung des Begriffs „Methode“ in der Informatik. Einerseits ist die Methode bekannt als Werkzeug, mit dem eine Forschungsfrage beantwortet werden soll. In der Informatik jedoch wird „Methode“ vor allem als Begriff für einen Baustein in einer Software-Implementierung genutzt, also die technische Umsetzung einer Problemlösung. Auf der Meta-Ebene sind beide „Methoden“ das Handwerkszeug, mit dem eine Aufgabe gelöst wird, aber für die Kommunikation zwischen verschiedenen Disziplinen kann dies dennoch zu Missverständnissen führen. Aufgrund dieser Doppeldeutigkeit nutzen Informatiker:innen den Begriff „Methodologie“ anstelle von „Methode“ (z.B. Colburn, 2008; Amaral et al., 2007). Methodologie beschreibt in computerwissenschaftlichen Arbeiten also viel mehr konkrete Methoden zur Beantwortung wissenschaftlicher Fragestellungen, während Auseinandersetzungen mit den Regeln, „welche Methoden wissenschaftlich anerkannt sind und warum welche Methoden eingesetzt werden“ (Scheufele und Engelmann, 2009, S. 26) wenig Aufmerksamkeit erfahren. Stattdessen existieren viele Anleitungen zum praktischen, induktiven Vorgehen (vgl. Amershi et al., 2019; Chaoji et al., 2016). Dazu kommt die (Selbst-)kritik zum Mangel an Theoriebezügen (vgl. Ullman, 1995; Lipton und Steinhart, 2019) sowie Edens Ausführungen zum ingenieurwissenschaftlichen Zweig der Informatik. Gemeinsam legen diese Punkte den Schluss nahe, dass maschinelles Lernen in der Anwendung als positivistisch eingeordnet werden kann. Die Schere zwischen theoretischer Informatik und der Anwendung maschinellen Lernens scheint für die Disziplin ein bis dato anerkanntes, aber ungelöstes Problem zu sein.

Für die Kooperation von Kommunikationswissenschaft und Machine Learning können die Parallelen zwischen abduktiven Verfahren und kritischem Rationalismus produktiv genutzt werden. Ein bewusster Vergleich der beiden Vorgehensweisen schafft die Grundlage für eine methodologische Auseinandersetzung in interdisziplinärer Forschung und ermöglicht es, über einzelne Schritte im Vorgehen zu sprechen, trotz unterschiedlichem Vokabular.

Methoden und Forschungsprozess

Der kommunikationswissenschaftliche Forschungsprozess geht deduktiv und hypothesengeleitet vor. Methoden zur Datenerhebung zwecks Prüfung von Hypothesen sind klassischerweise die Inhaltsanalyse, Beobachtung, Befragung (vgl. Scheufele und Engelmann, 2009) oder das Experiment (vgl. Koch et al., 2019). Das maschinelle Lernen kennt formale Beweise für Algorithmen, Experimente zur Evaluation neuer Problemlösungen, die konkrete Entwicklung von Lösungen, die Modellierung von Prozessen sowie die Modellierung von Systemen (vgl. Amaral et al., 2007). Am häufigsten operiert das Feld jedoch mit Experimenten (vgl. Freitas, 2009), wenngleich die Kritik an deren Durchführung scharf ausfällt. Experimente werden nicht ausreichend dokumentiert, sodass die Forschenden selbst ihre eigenen Arbeiten ein paar Wochen später nicht reproduzieren können (vgl. Amaral et al., 2007). Dieses Problem wird bereits seit den späten 1980er Jahren diskutiert:

„Where other sciences have standard experimental methods and analytic techniques, we have faith — often groundless and misleading — that building programs is somehow informative. Where other sciences expect specific aspects of research to be presented [...], empirical AI [Artificial Intelligence] has no comparable standards“ (Cohen und Howe, 1988, S. 35).

Diese elementare Kritik zielt auf zwei Aspekte wissenschaftlicher Forschung: Standardisierung von Methoden und Prozessen sowie deren Dokumentation in Veröffentlichungen. Sowohl der Forschungsprozess als auch die Qualitätssicherung im maschinellen Lernen weisen demnach bedeutende Lücken auf. In Kooperationen kann die Kommunikationswissenschaft, die den Schritt der Standardisierung bereits vielfach gegangen ist, die Richtung weisen. Da die Informatik jedoch auch neue Methoden mitbringt, ist davon auszugehen, dass es mehrerer Iterationen bedarf, bis Kommunikationswissenschaft und Machine Learning aufeinander abgestimmt sind und einen gemeinsamen Forschungsprozess entwickelt haben, der Standards beider Bereiche zufriedenstellend vereint.

Qualitätskriterien

Zusätzlich zu der oben erwähnten Kritik finden sich fachintern auch Lösungsvorschläge für die Standardisierung und Qualitätssicherung eines Machine Learning Prozesses. So überspitzt Cohen und Howes Aussage auch klingen mag, Ende der 1980er Jahre beschrieb sie den Mangel wissenschaftlicher Standardisierung im Feld der Künstlichen Intelligenz (KI) laut Lipton und Steinhardt (2019) scheinbar ange-

messen. Cohen und Howes Antwort war ein ausführlicher Fragenkatalog, mithilfe dessen die Evaluation von KI-Systemen ermöglicht werden sollte, um eine Grundlage für das Kumulieren von Wissen zu schaffen (vgl. Cohen und Howe, 1989). Er beinhaltet 31 Fragen entlang des Forschungs- und Entwicklungsprozesses, die wiederum in diverse Unterfragen unterteilt sind. Sie streifen Konzepte von Reliabilität, Validität, intersubjektiver Nachvollziehbarkeit, Relevanz und Repräsentativität, dennoch distanzieren sich die Autor:innen von den Qualitätskriterien etablierter empirischer Wissenschaft:

„We advocate evaluation not out of envy for “real science” but because we believe AI needs evaluation to move forward. Thus as a field we are not obliged to adopt “scientific” evaluation criteria [...] to move forward, but we should design our own“ (Cohen und Howe, 1989, S. 634).

Cohen und Howe wehren sich also gegen eine generelle Verpflichtung zu wissenschaftlichen Qualitätskriterien und argumentieren stattdessen für die Entwicklung eigener Evaluationskriterien, um den eigenen Fortschritt zu unterstützen. Diese Abgrenzung scheint grotesk, ist doch ein Grund für die Ausbildung wissenschaftlicher Qualitätskriterien genau jene Expansion wissenschaftlicher Erkenntnisse. Vorschläge für Evaluationskriterien sowie ein standardisiertes Vorgehen im Machine Learning sind mittlerweile vielfach vorhanden und zielen größtenteils auf Reproduktion, Performanz, Sicherheit und ethische Standards ab. Sie werden in Kapitel 3.2 vertieft.

Aus der Kommunikationswissenschaft kennen wir Qualitätskriterien wie Reliabilität, Validität und Objektivität, die sich aus dem epistemologischen Standpunkt ergeben und für deren Einhaltung bestimmte Konventionen entwickelt wurden. In Kapitel 3.1 werden diese Qualitätskriterien aufgearbeitet und mit den Kriterien des maschinellen Lernens verglichen. Es wird sich in 3.3 zeigen, dass es für die Kooperation von Kommunikationswissenschaft und Machine Learning nicht ausreicht, lediglich die eigenen Kriterien zu erheben.

Interdisziplinarität

Die Informatik kann aufgrund der Breite ihrer Themen und Methoden bereits als interdisziplinäres Feld verstanden werden (vgl. Tedre, 2011), auch innerhalb des Bereichs Machine Learning (vgl. Shalev-Shwartz und Ben-David, 2014). Machine Learning hat Überschneidungen mit Mathematik, Ingenieurwissenschaften, Sozialwissenschaften, Wirtschaftswissenschaften und Geisteswissenschaften. Diese Breite ermöglicht vielfältige Forschung, bringt allerdings auch Verunsicherung mit sich. Tedre (2011) identifiziert eine Sorge oder auch Hoffnung, in der Informatik über ihre Zukunft als Disziplin: Obwohl der Computer komplexer als beispielsweise ein Mikro-

skop ist (vgl. Newell et al., 1967), wäre es möglich, dass die *Computerwissenschaft* sich in ihre angrenzenden Disziplinen auflöst und der Computer selbst lediglich als Werkzeug für andere Fächer verbleibt (vgl. Smith, 1996). Weniger pessimistische Annahmen sehen jedoch auch eine inhaltliche Ausdifferenzierung des Faches (vgl. Freitas, 2009).

Auch die Kommunikationswissenschaft wird als interdisziplinäre Wissenschaft verstanden (vgl. Karmasin, 2008; Strippel et al., 2018). Spätestens mit dem Aufkommen sozialer Medien verstärken sich interdisziplinäre Bestrebungen in Richtung Informatik: anfangs unter dem Begriff Big Data (vgl. Boyd und Crawford, 2012), mittlerweile als Computational Communication Science (vgl. Atteveldt und Peng, 2018) mit eigens gegründeter Fachzeitschrift (vgl. Atteveldt et al., 2019a). Die plötzlich verfügbare wachsende Menge an Kommunikationsdaten führte zu durchaus steilen Thesen, die beispielsweise das Ende der Theorie vorhersagen (vgl. Anderson, 2008). Die Forschung der letzten Jahre zeigt jedoch, dass die Kommunikationswissenschaft sich nicht von Theorien abgewendet hat. Dennoch sieht Kitchin signifikante epistemologische Veränderungen auf das Fach zukommen, wenn es datengesteuert forschen will (vgl. Kitchin, 2014). Das Fach sorgt sich also, anders als die Informatik, weniger um seine Auflösung, diskutiert aber durchaus, ob sich durch die verstärkte Integration von computergestützten Verfahren ein neuer Forschungszweig abspalten wird (vgl. Von Welker und Matzat, 2009).

Die Tatsache, dass beide Disziplinen fachübergreifenden Kooperationen gegenüber offen sind, ist eine solide Grundlage für die Integration von Verfahren des maschinellen Lernens in die Kommunikationswissenschaft. Ein gemeinsames Interesse an bestimmten Forschungsgegenständen oder Forschungszielen bietet einen guten Motivator, um zusammen zu kommen und die Expertise aus beiden Bereichen zu verbinden. Dennoch dürfen grundlegende Fragen darüber, wie tatsächlich gemeinsam Wissen produziert, eingeordnet und kritisch geprüft werden kann, nicht außen vor gelassen werden. Denn die unterschiedlichen epistemologischen Standpunkte haben zur Folge, dass der Umgang mit Theorie und Theoriebildung, die Funktion von Qualitäts- und Gütekriterien als auch das Vorgehen an sich unterschiedlich sind. Aus kommunikationswissenschaftlicher Perspektive müssen eine stabile theoretische Grundlage und ein nachvollziehbarer, möglichst standardisierter Forschungsprozess gegeben sein. Validität und Reliabilität müssen ebenfalls gesichert werden. Aus der ingenieurwissenschaftlichen Perspektive des maschinellen Lernens muss die technische Entwicklung iterativ möglich sein, um bessere Ergebnisse zu ermöglichen. Aus naturwissenschaftlich-experimenteller Perspektive des maschinellen Lernens müssen diese Iterationen umfassend dokumentiert und abstrahiert werden, um daraus Erkenntnisse zu synthetisieren. In den folgenden zwei Kapiteln wird die Frage nach der Qualitätssicherung innerhalb eines potenziell gemeinsamen Forschungsprozesses

konkretisiert. Im Zentrum stehen dabei die sozialwissenschaftlichen Anforderungen an die Validität der eingesetzten Verfahren des maschinellen Lernens.

Qualitätssicherung automatischer Inhaltsanalyse

Aus den in Kapitel 2 dargelegten epistemologischen Positionen ergeben sich unterschiedliche Konventionen für den Forschungsprozess, um die Qualität wissenschaftlicher Arbeiten zu sichern. Im Folgenden werde ich anhand der manuellen Inhaltsanalyse sowie der automatischen Textklassifikation gegenüberstellen, wie die Forschungsprozesse angelegt sind und zu welchen Zeitpunkten welche Kriterien erarbeitet oder geprüft werden. Die Betrachtungsrichtung geht dabei von der Kommunikationswissenschaft zum maschinellen Lernen, da in der Kommunikationswissenschaft ein stärkerer Konsens über die entsprechenden Kriterien besteht (vgl. 2.3). Hier wird sich zeigen, dass die Strategien zur Sicherstellung von Validität in der Kommunikationswissenschaft vor allem zum Beginn des Forschungsprozesses, in der Konstruktion des Forschungsinstruments, Anwendung finden. Im Machine Learning hingegen wird die Validität zum Ende des Forschungsprozesses, also *ex post*, festgestellt und damit eher empirisch begründet. Zuletzt werde ich in Kapitel 3.3 zeigen, welche Konsequenzen sich für eine Integration der automatischen Textklassifikation in die kommunikationswissenschaftliche Inhaltsanalyse ergeben.

3.1 Manuelle Inhaltsanalyse: Qualitätskriterien

Ziel der quantitativen Inhaltsanalyse ist es, zentrale Muster in Medieninhalten herauszuarbeiten (vgl. Rössler, 2010, S. 19). Der Methode kommt in der Kommunikationswissenschaft eine besondere Rolle zu, da sie „in der Kommunikationsforschung selbst entwickelt wurde“ (Springer et al., 2015, S. 76). Eine der ersten Definitionen lieferte bereits 1952 Bernard Berelson:

„Content analysis is a research technique for the objective, systematic and quantitative description of the manifest content of communication“ (Berelson, 1952, S. 18; zit. nach Rössler, 2010, S. 20).

Rössler greift diese sowie zwei weitere Definitionen der Inhaltsanalyse auf und zeigt, wie sich im Laufe der Jahrzehnte unterschiedliche auch erkenntnistheoretische Standpunkte darin zeigen. Ging Berelson in den Fünfzigerjahren noch von Objektivität

der Methode aus, wählte Früh das Konzept der intersubjektiven Nachvollziehbarkeit, da „Objektivität‘ im Sinne einer vom jeweiligen Beobachter unabhängigen, stets gleichartig ausfallenden Wahrnehmung nicht erreichbar ist“ (Rössler, 2010, S. 22). Mit dieser Formulierung denkt Früh ein erkennendes Subjekt mit und fordert ein möglichst nachvollziehbares Vorgehen, um Unterschiede in der Wahrnehmung einzelner Subjekte auszugleichen. Ferner bezog Berelson sich ausschließlich auf die manifesten Inhalte von Medien(text), wohingegen Merten dafür plädiert, „von Merkmalen eines manifesten Textes auf Merkmale eines nichtmanifesten Kontextes“ (Merten, 1983, S. 59) zu schließen. Diese Definition schafft die Möglichkeit für komplexere Fragestellungen, weil sie anerkennt, dass Menschen aufgrund ihres Vorwissens „die ungeheuer vielfältigen Ausdrucksformen in (Medien-)Botschaften [...] deutlich sinnvoller und effektiver auf bestimmte Kategorien“ (Rössler, 2010, S. 23) zurückführen können.

Der Forschungsprozess, der die manuelle quantitative Inhaltsanalyse rahmt, orientiert sich am generellen quantitativ-empirischen Forschungsprozess. Zuerst teilt sich dieser in den Entdeckungs-, Begründungs- und Verwertungszusammenhang (vgl. Früh, 2017, S. 96; Rössler, 2010, S. 38). Im Entdeckungszusammenhang bildet sich das Forschungsinteresse aus, im Verwertungszusammenhang endet die Forschung mit Publikationen und weiterer Wissensvermittlung. Der Begründungszusammenhang beschreibt die wichtigste Phase des wissenschaftlichen Arbeitens. Hier werden die theoretischen Grundlagen ausgearbeitet und Hypothesen gebildet, das Messinstrument entwickelt und getestet, Daten erhoben und ausgewertet. Abbildung 3.1 zeigt den Ablauf schematisch, wobei an verschiedenen Stellen Schleifen möglich sind. So greifen zum Beispiel Test- und Entwicklungsphase durchaus ineinander, wenn sich in Codierschulungen zeigt, dass das Codebuch Korrekturen und Ergänzungen benötigt.

Elementare Merkmale der quantitativen Inhaltsanalyse sind nach Springer et al. Nicht-Reaktivität, deduktives Vorgehen, intersubjektive Nachvollziehbarkeit, Systematik und Quantifizierung (vgl. Springer et al., 2015, 79ff). Darüber hinaus stellt sich die Frage, „[o]b Instrument und Messung nun tatsächlich geeignet sind, Antworten auf die Forschungsfragen zu geben“ (vgl. Rössler, 2010, S. 195). Diese Frage nach *Qualität* wird in zwei Kriterien festgehalten: Validität und Reliabilität. Sie sind die zentralen Forderungen an eine Messung in der Inhaltsanalyse (vgl. Brosius et al., 2009, S. 63).

3.1.1 Validität

Validität ist die übergeordnete Anforderung an eine Messung (vgl. Rössler, 2010, S. 195), aber auch diejenige, die weniger klar messbar ist. Sie beschreibt „die Korre-

Abb. 3.1: Forschungsprozess für die manuelle Inhaltsanalyse, orientiert an Rössler, 2010, S. 38; Früh, 2017, S. 96

spondenz von Analyseergebnis und sozialer Wirklichkeit“ (Merten, 1983, S. 302), allgemein bekannt als die *Gültigkeit* der Messung und „stellt sicher, dass man das misst, was man messen will“ (Brosius et al., 2009, S. 63). Im Sinne des kritischen Rationalismus erklärt Merten, dass diese externe Gültigkeit einer Messung, wie jegliche Gültigkeitsprüfungen, nicht verifizierbar, sondern lediglich falsifizierbar seien. Gültigkeit könne vorläufig angenommen werden, wenn sich keine belastbaren Einwände gegen die Anwendung eines Instruments präzisieren ließe (vgl. Merten, 1983, S. 310). Für ihre Prüfung existieren deswegen, anders als für die Prüfung der Reliabilität, wenige eindeutige Kennzahlen. Transparenz der einzelnen Schritte im Forschungsprozess ermöglicht die Prüfung der Validität überhaupt erst und kon-

krete Strategien zur Sicherung von Validität werden in verschiedenen Phasen des Forschungsprozesses angewandt. Hier zeigen sich zwei verschiedene Dimensionen von Validität: Die Strategien, mit denen sie umgesetzt werden soll und die spätere Prüfung, ob jene Strategien erfolgreich waren (vgl. Früh, 2017, S. 187). Jedoch

„enden Validitätsprüfungen [...] nicht mit einem Koeffizienten zwischen 0 und 1, sondern mit Verbesserungsvorschlägen für das Untersuchungsdesign und das Erhebungsinstrument“ (Brosius et al., 2009, S. 69).

Welche Maßnahmen sollen also ergriffen werden, um die Validität einer Messung zu erhöhen? Rössler versteht Inhaltsvalidität, Analysevalidität, Kriteriumsvalidität und Inferenzvalidität als die wichtigsten Aspekte (vgl. Rössler, 2010), weshalb diese im Folgenden zusammengefasst werden.

Im Schritt der Kategorienbildung und Definition von Ausprägungen wird die *Inhaltsvalidität* gestärkt. Sie erfasst „die Beziehung zwischen Theorie und dem gebildeten Erhebungsinstrument“ (Merten, 1983, S. 311) und beinhaltet auch eine logische oder semantische Gültigkeit, die sich auf die Übersetzung von Hypothesen hin zur Operationalisierung beziehe (vgl. Merten, 1983, S. 311). Diese Vollständigkeit der Konstrukte kann durch eine umfassende Aufarbeitung des Forschungsstandes, „externer Quellen (wie z. B. Expertenmeinungen) und einer ersten Sichtung des Untersuchungsmaterials“ (Rössler, 2010, S. 206) gesichert werden. Geprüft wird das Erreichen der Inhaltsvalidität später beispielsweise daran, wie häufig die Codierenden Auffangwerte oder offene Kategorien nutzten und welche Schwierigkeiten sie in der Codierung berichten (vgl. Rössler, 2010, S. 207). Davon ausgehend, dass die Inhaltsvalidität zufriedenstellend ist, zeigt sich in dem Kriterium der *Analysevalidität* dann, ob das Instrument für die Codierenden nachvollziehbar ist und dessen Ausformulierung sich mit den Ideen der Forschenden deckt. Eine gute Analysevalidität ist wichtig, da die Interpretationen zum Ende der Forschung natürlich von den Forschenden getroffen werden und eine Diskrepanz zwischen deren Verständnis der Konstrukte und dem der Codierenden das Material für die Interpretation unbrauchbar machen würde. Die Forschenden müssen den Codierenden ihre Interpretationsweise angemessen vermitteln, denn nur so „sind auch die produzierten Daten in diesem Sinne valide“ (Früh, 2017, S. 188). Dieses Kriterium kann durch solide Codierschulungen sichergestellt werden. Wenn sowohl Forschende als auch Codierende einen Teil der Daten codieren, lässt sich die Prüfung relativ leicht umsetzen und sogar in einen Zahlwert übersetzen: Die Forscher-Codierer-Reliabilität drückt aus, inwieweit die als korrekt angenommenen Codierungen der Forschenden mit denen der Codierenden übereinstimmen (vgl. Rössler, 2010). Im Rückschluss der Ergebnisse auf die Theorie zeigt sich die *Kriteriumsvalidität*, bestehend aus Vorhersagevalidität und Konkurrenzvalidität (vgl. Früh, 2017, S. 188). Sie „prüft,

ob die erhobenen Daten im Vergleich zu einem anderen Messkriterium gültig sind“ (Brosius et al., 2009, S. 69). Durch den Vergleich der eigenen Ergebnisse mit denen anderer Inhaltsanalysen, gegebenenfalls auch qualitativer Ergebnisse, kann eine hohe Kriteriumsvalidität gesichert werden (vgl. Rössler, 2010, S. 207). Zuletzt greift Rössler die *Inferenzvalidität* auf, die ebenfalls in der Phase der Interpretation relevant wird. Sie prüft, ob Schlüsse, die über den konkreten Text hinaus gehen, mit den Ergebnissen von anderen Forschungen vereinbar sind (vgl. Rössler, 2010, S. 208). Diese externe Validierung erfolgt vor allem durch den Vergleich mit Studien, die andere Methoden nutzten (vgl. Rössler, 2010, S. 208).

Für die Dokumentation der Validität einer Studie finden sich in Lehrbüchern mehr oder weniger konkrete Hinweise (vgl. Krippendorff, 2004; Rössler, 2010; Wirth et al., 2015) und vereinzelte kritische Analysen, welche beispielsweise die Validität von Zusammenhängen in Frage stellen, die auf unterschiedlichen hierarchischen Ebenen der Untersuchungseinheit gefunden werden (vgl. Lauf und Berens, 2003). Eine umfassende Reflexion der Praxis im Wissenschaftsbetrieb analog zu Vogelgesang und Scharkow (2012) für die Reliabilität scheint mit Bezug auf Validität eine Leerstelle zu sein.

3.1.2 Reliabilität

Reliabilität kann definiert werden als „das Ausmaß der Reproduzierbarkeit von Meßergebnissen bei wiederholter Anwendung des gleichen Instruments auf das gleiche Meßobjekt“ (Merten, 1983, S. 302). Während sich die Validität also auf die Gültigkeit des theoretischen Konstruktes bezieht, beschreibt die Reliabilität vielmehr die Verlässlichkeit des Instrumentariums selbst sowie dessen Anwendung (vgl. Früh, 2017, S. 112). Die Prüfung der Reliabilität ist logisch jener der Validität untergeordnet, kann man doch „mit großer Präzision und Verlässlichkeit immer wieder die falschen oder nur einen Teil der gesuchten Textmerkmale erfassen“ (Früh, 2017, S. 180). Reliabilität und Validität stehen jedoch auch in einer Art von Konkurrenz. Je enger ein Konstrukt gefasst ist, indem es zum Beispiel latente Bedeutungen eines Textes nicht erfasst, desto zuverlässiger lässt sich dies messen. Je umfassender und detaillierter ein Konstrukt die Realität abbilden soll, desto schwieriger wird es, diese Messung zu reproduzieren (vgl. Brosius et al., 2009, S. 162). Die Abwägung zwischen beiden Qualitätskriterien, zwischen Gültigkeit und Verlässlichkeit, muss also immer wieder neu verhandelt werden.

Für die Prüfung der Reliabilität existieren diverse Kenngrößen. Dabei beschreibt die *Intercoderreliabilität*, inwieweit verschiedene Codierer:innen in ihrer Codierung übereinstimmen und die *Intracoderreliabilität*, inwieweit ein:e einzelne:r Codierer:in im Laufe der Zeit in den eigenen Codierungen konsistent bleibt. Die Prüfung der

Intercoderreliabilität wird allgemein als erforderlich angesehen, während die Prüfung der Intracoderreliabilität lediglich bei längeren Projekten und Codierphasen empfohlen ist (vgl. Früh, 2017, S. 179). Es stellen sich vor der Reliabilitätsprüfung aber drei Fragen (vgl. Rössler, 2010, S. 198):

1. Was gilt überhaupt als Übereinstimmung oder Abweichung? Dies scheint zunächst klar definiert: geben Codierende den gleichen Code an, stimmen sie überein, geben sie unterschiedliche Codes an, weichen sie voneinander ab. Komplexer wird es jedoch bei Werten, die kontinuierliche, gestufte oder hierarchische Variablen messen (vgl. Rössler, 2010, S. 199). Hier können die Codierungen vor der Prüfung in Toleranzintervalle, Tendenzen oder hierarchisch höher angelegte Variablen überführt werden, was sich jedoch auch auf die Messgenauigkeit auswirkt (vgl. Rössler, 2010, S. 199). Auch Mehrfachnennungen von Variablen sind schwierig, wie Früh für eine Themen-Frequenzanalyse zeigt: Sind beispielsweise mehrere Themen für einen Text zu vergeben, dann können sowohl Reihenfolge als auch Anzahl variieren. Er empfiehlt dann eine kleinere Analyseeinheit (vgl. Früh, 2017, S. 183), die durch textnahe Annotationen erreicht werden könnten. Eine weitere Herausforderung für die Entscheidung, was als Übereinstimmung oder Abweichung zählt, wird unter dem Begriff der Identifikationsreliabilität gefasst. Ob zwei Codierende die gleiche Analyseeinheit im Text überhaupt identifiziert haben, sollte getrennt von und vor der Übereinstimmung der codierten Variablen erfolgen (vgl. Rössler, 2010, S. 199).
2. Wie ist der Grad der Übereinstimmung zu berechnen? Merten unterscheidet drei Typen der Intercoderreliabilität: Typ 1 prüft die Übereinstimmung von jeweils zwei Codierenden und gibt somit keine Informationen über die Übereinstimmung aller Codierenden miteinander aus, kann aber einzelne abweichende Codierende gut identifizieren. Typ 2 prüft die Übereinstimmung aller Codierenden miteinander und gibt somit einen umfassenderen Einblick, jedoch als Konsequenz auch niedrigere Werte, da Typ 2 nie besser ausfallen kann als die niedrigste paarweise Übereinstimmung zwischen zwei Codierenden. Typ 3 prüft die höchste Übereinstimmung von beliebig vielen der Gesamtzahl an Codierenden (vgl. Merten, 1983, S. 304). Merten empfiehlt sie vor allem für Codierungen, bei denen bereits im Vorhinein eine breite Streuung erwartet wird. Als Ergänzung oder Sonderform von Typ 3 ist außerdem RICO (Reliability If Coder Omitted) zu erwähnen, bei dessen Berechnung immer nur ein:e Codierer:in ausgelassen wird, um potenziell schlechte Codierer:innen zu identifizieren (vgl. Scharkow, 2012, 129f.). Typ 1 und Typ 2 lassen sich beispielsweise mit dem Reliabilitätsmaß nach Holsti (1969) berechnen:

$$\text{Codiererreliabilität} = \frac{2 \times \text{Anzahl der übereinstimmenden Codierungen}}{\text{Anz. Codierungen Codierer 1} + \text{Anz. Codierungen Codierer 2}} \quad (3.1)$$

Abb. 3.2: Reliabilitätsmaß (Holsti, 1969)

Diese Formel ergibt eine Zahl zwischen 0 und 1, wobei 0 die geringste Übereinstimmung und 1 die höchste Übereinstimmung repräsentiert. Bei mehr als zwei Codierenden kann dann der Mittelwert der einzelnen Übereinstimmungen errechnet werden (vgl. Früh, 2017, S. 181), was Typ 2 entspricht. Welcher Wert als zulässig einzustufen ist, hängt davon ab, was eine einzelne Variable erfasst (vgl. Lacy et al., 2015). Formale Variablen wie beispielsweise das Veröffentlichungsdatum eines Textes sollten bei 1 liegen, für komplexere Variablen können aber auch Werte um .75 (z.B. für Themen) ein gutes Ergebnis sein (vgl. Früh, 2017, S. 184). Alternative Reliabilitätsmaße wie Scotts Pi, Cohens Kappa und Krippendorfs Alpha greifen unter anderem oben genannte Herausforderungen sowie schief verteilte Kategorien und zufällige Übereinstimmungen auf oder erlauben den Umgang mit unterschiedlichen Skalenniveaus (vgl. Lacy et al., 2015). Krippendorfs Alpha wird als aussagekräftigster Koeffizient verteidigt (Lacy et al., 2015), ist jedoch vergleichsweise komplex zu berechnen, weshalb er im Rahmen dieser Arbeit lediglich erwähnt wird, ebenso wie der vergleichsweise neue Koeffizient Lotus (vgl. Fretwurst, 2015). In einem Experiment vergleicht Wettstein die Ergebnisse verschiedener Reliabilitätsmaße unter unterschiedlichen Bedingungen und zeigt, wie diese auf „unterschiedliche Fehlerarten und auf die Randbedingungen der Inhaltsanalysen reagieren“ (Wettstein, 2016, S. 57). Er stellt fest, dass die Anzahl der Codierenden keinen Einfluss auf die Werte hat. Unterschiedliche Ergebnisse zeigen sich jedoch, wenn einzelne Codierende besonders schlecht codieren (vgl. Wettstein, 2016, S. 68). Fehlentscheidungen in schief verteilten Kategorien können bei Pi, Alpha und Kappa zu stark ins Gewicht fallen, wohingegen Lotus die Codierqualität überschätzt (vgl. Wettstein, 2016, S. 68). Wettstein empfiehlt deswegen, wenn möglich unterschiedliche (aber mindestens einen zufallsbereinigten) Koeffizienten zu berechnen, die prozentuale Übereinstimmung anzugeben und zu dokumentieren, welche Kategorien auf Basis dieser Werte eher unzuverlässig erhoben wurden.

3. Wie viel Testmaterial ist nötig, um statistisch sichere Werte zu produzieren? Früh adressiert die Menge des Materials, anhand dessen die Reliabilität berechnet wird und empfiehlt beispielsweise, dass mindestens 30-50, besser jedoch 200-300 Nennungen jeder Variable im Testmaterial vorhanden sein sollten, um einen sicheren statistischen Vergleich zu bieten (vgl. Früh, 2017, S. 180).

Im Forschungsprozess wird die Reliabilität mindestens ein Mal im Rahmen der Codierschulung berechnet. Früh empfiehlt, dass die Werte vor allem auch für einzelne Codierende, Variablen und Textsorten geprüft werden, um rechtzeitig Mängel in den Codierfähigkeiten, den Codieranweisungen, der Trennschärfe des Kategoriensystems und dessen operationalen Definitionen sowie den Daten selbst zu identifizieren. Daraufhin können Korrekturen erfolgen, die durch einen erneuten Reliabilitätstest geprüft werden (vgl. Früh, 2017, S. 185).

Zur Dokumentation der Werte zeigen sich durchaus unterschiedliche Meinungen. Rössler fordert beispielsweise, dass die Ergebnisse des letzten Reliabilitätstests als Matrix dargestellt werden, die die Übereinstimmung nach Kategorien und Codierenden aufschlüsselt (vgl. Rössler, 2010, S. 204). Vogelgesang und Scharkow sprechen sich dafür aus, weitere Angaben zu Identifikationsreliabilität, Forscher-Coder-Reliabilität, Anzahl der getesteten Codierer und zur Stichprobe des Reliabilitätstests zu machen, da diese ebenfalls die Möglichkeit erhöhen, die Reproduzierbarkeit einer Studie einzuschätzen (vgl. Vogelgesang und Scharkow, 2012). Im Jahr 2012 wiederholten Vogelgesang und Scharkow eine Studie von Edmund Lauf (2001), in welcher dieser die Dokumentationspraxis von Reliabilitätswerten in kommunikationswissenschaftlichen Fachzeitschriften analysierte. Vogelgesang und Scharkow finden insgesamt eine bessere Dokumentationpraxis vor, sehen jedoch weiterhin eine Abweichung kommunikationswissenschaftlicher Veröffentlichungen zu den in Lehrbüchern empfohlenen Standards. Die Autoren begründen dieses Verhalten unter anderem mit der Kürze eines wissenschaftlichen Aufsatzes, aus der folglich die Dokumentation der Reliabilität knapp gehalten werden müsse (vgl. Vogelgesang und Scharkow, 2012, S. 341). Für Studien mit überwachter Textklassifikation, in der manuelle und automatische Verfahren kombiniert werden, ist die umfassende Dokumentation ihrer Reliabilität demnach eine noch größere Herausforderung.

3.2 Automatische Textklassifikation: Gütekriterien

Für die automatische Textklassifikation können im überwachten maschinellen Lernen diverse Verfahren verwendet werden. Die in den letzten Jahren wichtigste Unterscheidung besteht zwischen klassischen statistischen Klassifikatoren (beispielsweise logistische Regression, Naive Bayes, Support Vector Machines (SVM)), die auch in der Kommunikationswissenschaft bereits genutzt werden (z.B. Ziegele et al., 2018; Stoll, 2020) und sogenannten *neuronalen Netzen* (beispielsweise Convolutional Neural Networks, Recurrent Neural Networks, Long Short-Term Neural Networks)¹.

¹Eine Beschreibung der einzelnen Verfahren ist nicht Gegenstand dieser Arbeit, für interessierte Leser:innen sei aber auf die Überblicksarbeiten von Kotsiantis (2007) und Shalev-Shwartz und Ben-David (2014) verwiesen, sowie auf Minaee et al. (2020) für einen spezifischen Überblick über Textklassifikation mit Deep Learning.

Es ist außerdem zu differenzieren zwischen den Begriffen Algorithmus und Modell. Ein Algorithmus ist lediglich ein logisches Konstrukt aus Anweisungen und kann unterschiedlich komplex sein. Statistische Verfahren nutzen klassischerweise einen einzelnen Algorithmus, während in neuronalen Netzen mehrere Algorithmen verknüpft werden können. Die einzelnen oder kombinierten Algorithmen werden auf einen Großteil der zu untersuchenden Daten angewandt, um für diese ein spezifisches Modell (Klassifikator) zu entwickeln. Daraufhin werden sie an einem kleineren Teil der Daten getestet, um das Modell zu prüfen. Das Modell besteht dann aus Gewichten oder anderen konkreten Werten und kann, wenn es neue Daten erhält, einzelne Klassifikationen dafür generieren. Im Folgenden liegt der Fokus nun auf der generellen Funktionsweise von Textklassifikation und deren Gütekriterien.

„Ziel der Textklassifikation im Supervised Machine Learning (SML) ist es, einen statistischen Zusammenhang zwischen [...] Mustern in einem Text (unabhängigen Variablen) und der Textkategorie (abhängiger Variable) zu modellieren, um die Textkategorie auf Basis der Textmuster vorherzusagen zu können“ (Stoll, 2020, S. 3).

Es sollen also vordefinierte Labels an neue Texte oder Textabschnitte vergeben werden; und zwar möglichst in Übereinstimmung mit einem vorgegebenen Referenzdatensatz, dessen Codierungen als korrekt vorausgesetzt werden. Das Modell nutzt üblicherweise für die Klassifikation keine für Menschen nachvollziehbaren Abfolgen von Regeln mehr, sondern folgt dem Prinzip der Optimierung - es ist sozusagen indirekt programmiert. Das überwachte Lernen funktioniert also anders als diktions- oder regelbasierte Ansätze, die deduktiv vorgehen und möglichst jeden Fall vordefinieren müssen, damit ein Algorithmus, der üblicherweise keinen Kontext erfassen kann, gut funktioniert. Das überwachte Lernen geht vielmehr induktiv vor und leitet die Regeln zur Codierung aus Eigenschaften des vorgegebenen Datensatzes selbst ab (vgl. Scharkow, 2013, S. 763).

Der Forschungsprozess, der die überwachte automatische Textklassifikation rahmt, ist weniger standardisiert als jener für die manuelle Inhaltsanalyse. Vereinzelt werden Phasen des Prozesses beschrieben (vgl. Kandel et al., 2012) oder Workflows schematisch dargestellt (vgl. Kotsiantis, 2007; Gill et al., 2020; Chaoji et al., 2016; Amershi et al., 2019), wobei sich diese Ansätze eher unter dem Begriff „Data Science“ sammeln und nicht „Machine Learning“ oder „Textklassifikation“. Die Anzahl der Schritte sowie deren Inhalt variiert. Ein grundlegender Unterschied zeigt sich in der Frage, ob an den Anfang des Prozesses eine Fragestellung / ein Problem gesetzt wird, oder Datenexploration. Neumann legt in ihrem Vorschlag einen stärkeren Fokus auf die Übersetzung eines wissenschaftlichen, sozialen oder unternehmerischen Problems in eine Frage an Daten, verkürzt aber gleichzeitig den Schritt der Entwicklung

und Auswertung des Klassifikators (vgl. Neumann, 2019). Mit diesem Schwerpunkt auf den Anfang des Prozesses stellt ihr Prozess jedoch eher eine Ausnahme dar. Der in Abbildung 3.3 dargestellte Prozess ist daher eine Synthese verschiedener Schemata und gleichzeitig der Versuch, einen mit dem in Abschnitt 3.1 gezeigten Forschungsprozess vergleichbaren Ablauf darzustellen.

Teilt man den Workflow für überwachtes Lernen in die gleichen drei Zusammenhänge ein wie beim kommunikationswissenschaftlichen Forschungsprozess, dann ist die Frage- oder Problemstellung im Entdeckungszusammenhang angesiedelt, da diese nicht als wissenschaftliche Fragestellung, sondern als unternehmerisches Problem verstanden (vgl. Chaoji et al., 2016; Amershi et al., 2019), als bereits vorhanden vorausgesetzt (vgl. Domingos, 2012) wird oder aus einem explorativen Umgang mit Daten entsteht, für den ebenfalls keine theoretische Rahmung gefordert ist (vgl. Gill et al., 2020). Ich habe mich dazu entschieden, die Datenerhebung ebenfalls bereits an diesem Punkt zu verankern, aus verschiedenen Gründen: Erstens erfordert die explorative Datenanalyse bereits Daten (vgl. Larose und Larose, 2014), zweitens wird die Erhebung von Daten im Machine Learning Prozess zwar erwähnt, jedoch ist damit oft der Rückgriff auf bestehende aufbereitete Datensets (vgl. Amershi et al., 2019) oder jene Daten, die im Unternehmensalltag bereits mitgeschnitten werden (vgl. Chaoji et al., 2016), gemeint. Für die automatische Textklassifikation bedeutet dieser Schritt das Sammeln von Daten, die bestenfalls bereits mit den Ziel-Labels versehen sind.

In der Anforderung an das zu entwickelnde Modell wird festgelegt, wie die Fragestellung oder das Problem gelöst werden soll: Kann das Ergebnis der Klassifikation durch binäre Labels repräsentiert sein, durch reale Werte oder durch ein bzw. mehrere Labels aus einem begrenzten Set (vgl. Shalev-Shwartz und Ben-David, 2014, S. 45)? Daraufhin wird der Algorithmus gewählt, der das Modell trainieren soll. Die Daten werden bereinigt und vorbereitet, Labels können entsprechend der Ziele angepasst werden. Dieser Schritt beinhaltet außerdem Tokenisierung (Aufsplitten des Textes in kleine Einheiten wie Wörter oder einzelne Buchstaben), Textfilterung (Entfernen von Satzzeichen und Stopwörtern), Überführung in Kleinbuchstaben, Stemming (Reduktion von Wörtern auf ihren Wortstamm), Lemmatisierung (Ersetzen von gebeugten Wörtern durch ihre Grundform), Disambiguierung (Auflösung von Homonymen und Polysemen), Anaphoren-Auflösung (Ersetzung von Pronomen durch ihr Bezugswort) und Part-of-Speech-Tagging (Zuordnung der Wortart zu jedem Wort) (vgl. Scharow 2012; Uysal und Gunal 2014). Obwohl dieser Schritt bereits Auswirkung auf die Ergebnisse hat, besteht kein Konsens darüber, welche Schritte in welchem Kontext notwendig und angemessen sind (vgl. Uysal und Gunal, 2014). Daraufhin folgt das Feature Engineering, in dem die Einheit ausgewählt wird, anhand derer das Modell lernen soll (vgl. Domingos, 2012). Für tabellarische Daten sind das einzelne Variablen wie beispielsweise Alter, Einkommen oder Körpergröße.

Abb. 3.3: Workflow für die automatische Textklassifikation, orientiert an Kotsiantis, 2007, S. 250; Amershi et al., 2019, S. 292; Gill et al., 2020, S. 137

Da bei der Textklassifikation jedoch der Input lediglich Text ist, sind die Features hier Zeichen – also Folgen von einzelnen Buchstaben, Buchstabengruppen, einzelnen Worten oder auch Wortgruppen. An dieser Stelle wird entschieden, wie groß der Kontext ist, den ein Modell betrachten soll: Werden einzelne Worte als Features verwendet oder Gruppen von beispielsweise drei aufeinander folgenden Worten? Sollen mehrere Algorithmen aufeinander aufbauen oder miteinander verbunden werden? Anschließend wird der Algorithmus auf die Daten angewendet, um ein Modell zu trainieren und zu erstellen. Der zugrunde liegende Datensatz wird dafür in ein Trainings- und ein Testset aufgeteilt. Um Unausgewogenheiten in den Trainings- und Testdaten auszugleichen, wird standardmäßig die k -fache-Kreuzvalidierung (k -fold cross validation) genutzt, in der der Datensatz zufällig in k Gruppen aufgeteilt wird. Mit allen bis auf eine Gruppe wird dann das Modell trainiert und mit der letzten Gruppe wird es getestet. Dieser Prozess wird wiederholt, bis alle Gruppen

ein Mal zum Testen verwendet wurden und die Gewichte für die Features werden daraufhin gemittelt (vgl. Scharkow, 2012). So sind Klassifikatoren weniger verzerrt und die folgenden Werte haben mehr Aussagekraft. Mit den bestehenden Daten wird also sowohl das Modell trainiert, in dem Gewichte für einzelne Features berechnet werden, als auch getestet, indem geprüft wird, wie gut das Modell die ursprünglichen Klassifikationen anhand der gelernten Gewichte reproduzieren kann. Dies geschieht iterativ und zusätzlich kann das Modell auch auf einen weiteren, bis dato unbekannten Datensatz angewandt werden. An dieser Stelle unterscheiden sich die Begriffe für verschiedene Phasen. Training, Optimierung, Validierung, Testing und Evaluation sind unscharf definiert und beschreiben zusätzlich keine abgeschlossenen, aufeinanderfolgenden Phasen, sondern einen iterativen Verlauf, in dem das Modell durch Testen und darauffolgendes Anpassen von Hyperparametern möglichst gut trainiert wird. Sind die Ergebnisse (siehe 3.2.1) zufriedenstellend, wird das Modell ein letztes Mal anhand der relevanten Gütekriterien evaluiert. Es folgt die Auswertung in Form eines Berichtes und dann gegebenenfalls die Anwendung des Modells im Live-Betrieb.

Der Machine Learning Workflow ist also weniger klar definiert als der sozialwissenschaftliche Forschungsprozess, was sich eben auch aus den drei verschiedenen epistemologischen Positionen ergibt. Ein Machine Learning Prozess aus mathematischer Perspektive würde mit dem Beweis enden, dass ein Algorithmus das jeweilige Klassifikationsproblem lösen kann. Ein ingenieurwissenschaftlicher Ansatz ist enger in unternehmerische Prozesse eingebunden oder daran orientiert und weniger an Hypothesenbildung, sondern an Problemlösung und der Entwicklung einer zufriedenstellend funktionierenden Software-Anwendung interessiert. Die naturwissenschaftliche Perspektive hingegen legt Wert auf einen systematischen Vergleich und eine systematische Optimierung des Modells. Trotz dieser unterschiedlichen Ansätze können, wie oben beschrieben, elementare Phasen des Machine Learning Workflows ausgemacht werden. Entsprechend dieser mangelnden Standardisierung in der Forschungspraxis ergeben sich jedoch Konsequenzen für die Prüfung der Qualität oder der Güte des gelernten Modells. 1989 schlagen Cohen und Howe eine Reihe von evaluativen Fragen vor, die sie an unterschiedliche Phasen empirischer KI-Forschung ansiedeln (vgl. Cohen und Howe, 1989). Dieser Fahrplan hat sich jedoch nicht durchgesetzt. Üblich sind heute konkrete Werte wie Accuracy, Precision und Recall (vgl. Shalev-Shwartz und Ben-David, 2014, 244f), die sehr direkt die Ergebnisse des Klassifikators messen. Darüber hinaus existieren auch Bestrebungen nach der Prüfung von Reproduzierbarkeit, Performanz, Transparenz, Sicherheit und ethischen Standards. Diese werden im Folgenden genauer dargestellt.

3.2.1 Accuracy, Precision und Recall

Die Güte eines Modells im überwachten Lernen wird generell daran gemessen, inwieweit die Ergebnisse der automatischen Klassifikation mit den Ergebnissen des zugrunde liegenden Datensatzes übereinstimmen, auf dem das Modell trainiert wurde. Diese Art der Messung bezeichnet Scharkow als *Expertenvalidität*, die bedeutet, dass „die Codierung des Computers [...] mit einem vorgegebenen Goldstandard verglichen“ (vgl. Scharkow, 2012, S. 134) wird. Während in der manuellen Inhaltsanalyse also die Übereinstimmung gleichberechtigter Codierer:innen mit möglicherweise verschiedenen Interpretationsrahmen in Form von Reliabilitätskoeffizienten quantifiziert wird, berechnet die Informatik stattdessen die Übereinstimmung der Codierungen im Testset durch einen potenziell fehlbaren algorithmischen Codierer mit den als korrekt verstandenen menschlichen Codierungen. Der simpelste Wert für binäre Klassifikatoren, die *Accuracy*, beschreibt dabei lediglich den Anteil der in der automatischen Textklassifikation korrekt klassifizierten Einheiten (TP und TN)². Dieser Wert sagt jedoch wenig darüber aus, welche Art von Abweichungen vom „Goldstandard“ das Modell produziert (vgl. Scharkow, 2012, S. 138). Je nach Anwendungsfall ist eine falsch positiv codierte Einheit (FP)³ akzeptabler als eine falsch negativ codierte Einheit (FN)⁴. Ein Klassifikator für Hate Speech in Nutzerkommentaren hat beispielsweise meist mehr Wert, wenn er zu viel Hate Speech vermutet, sodass einige Klassifikationen daraufhin manuell korrigiert und Kommentare freigegeben werden müssen. Problematischer ist es, wenn der gleiche Klassifikator zu wenig Hate Speech erkennt und Kommentare zur Veröffentlichung frei gibt, die eine Diskussion eskalieren lassen oder verunmöglichen. Daher helfen drei weitere Werte, die Güte des Modells und die Art der Fehler zu berechnen, wie die Tabelle 3.1 illustriert.

Kennwert	Formel	Bedeutung
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Wahrscheinlichkeit einer korrekten Codierung
Precision	$\frac{TP}{TP+FP}$	Wahrscheinlichkeit, mit der ein codiertes Dokument relevant ist
Recall	$\frac{TP}{TP+FN}$	Wahrscheinlichkeit, mit der ein relevantes Dokument codiert wird
F-Measure	$\frac{2*Prec*Rec}{Prec+Rec}$	harmonisches Mittel aus Precision und Recall

Tab. 3.1: Maße für die Klassifikatorevaluation nach Scharkow (2012)

²True positive: Der Klassifikator vergibt ein Label für eine Einheit, die im Testset das gleiche Label hat. True negative: Der Klassifikator vergibt kein Label für eine Einheit, die auch im Testset dieses Label nicht hat.

³False positive: Der Klassifikator vergibt ein Label für eine Einheit, obwohl diese im Testset das Label nicht hat

⁴False negative: Klassifikator vergibt kein Label für eine Einheit, obwohl der Testset für diese Einheit ein Label hat

Eine hohe *Precision* beschreibt damit den Anteil der identifizierten Klassifizierungen an jenen, die gefunden werden sollten. Ein hoher Wert heißt, dass viele Klassifizierungen, die durch den Klassifikator getroffen wurden, auch so im Testset existieren (wenige *false positives*). Hingegen beschreibt ein hoher *Recall* den Anteil der gesamten korrekt identifizierten Klassifizierungen an allen, die getroffen werden sollten. Ein hoher Recall heißt, dass der Klassifikator wenige Klassifizierungen unterlassen hat, die im Testset jedoch bestehen (wenige *false negatives*). Er beschreibt damit auch eine Form von Identifikationsreliabilität – davon ausgehend, dass der Klassifikator alle positiven Klassifikationen aus dem „Goldstandard“ identifizieren könnte, wird geprüft, wie viele tatsächlich identifiziert wurden. Die Werte Precision und Recall stehen in der praktischen Umsetzung oft konträr zueinander. Deswegen müssen Entwickler:innen meist entscheiden, welcher der beiden Werte für ihre Anwendung wichtiger ist. Eine Kombination aus *Precision* und *Recall* stellt der letzte Kennwert dar – der *F-Measure* oder auch der *F1-score*.

Die Kritik an Accuracy, Precision, Recall und F-Measure findet auf zwei Ebenen statt. Erstens kritisieren beispielsweise Henderson et al., dass diese Werte unzuverlässig seien, weil Algorithmen in ihrer Anwendung eben nicht reliabel sind, sondern abhängig von – unter anderem – Hyperparametern, Implementierungsdetails, experimentellem Setup und Evaluationsmethoden äußerst verschiedene Ergebnisse liefern können (vgl. Henderson et al., 2018). Ähnlich der in 3.1.2 genannten kommunikationswissenschaftlichen Kritik an der Dokumentation von Reliabilitätswerten wird also auch in der Informatik eine umfassendere Dokumentation und Reflexion gefordert, um Reproduzierbarkeit zu ermöglichen. Zweitens setzen die Werte konzeptionell voraus, dass der Trainingsdatensatz korrekt codiert wurde. Das ist aber weder bei geschulten Codierer:innen noch beim Crowdsourcing der Fall, was Reliabilitätswerte illustrieren sollen. Bachl und Scharkow (2017) bestehen darauf, dass diese Werte nicht lediglich dokumentiert, sondern auch in die Beurteilungen der Ergebnisse miteinbezogen werden müssen und schlagen verschiedene Strategien dafür vor. Eine Diskussion darüber, inwiefern diese Werte auch im Training eines Modells für die automatische Textklassifikation einbezogen werden könnten, ist nach meiner Recherche noch nicht entstanden.

3.2.2 Performanz, Sicherheit und ethische Standards

Je nach Forschungsinteresse werden in Studien Kriterien wie Performanz, Sicherheit und ethische Standards angeführt, um die Güte eines Modells zu bewerten. Die Performanz kann sich auf die Komplexität des Algorithmus selbst beziehen (vgl. Lipton, 2018) und stellt dann die Frage, inwiefern der Algorithmus von Menschen in seiner Funktionsweise gänzlich verstanden werden kann. Hierfür gibt es keine Kenngrößen, die Messung muss im Abgleich mit der Zielgruppe geschehen, die die

Komplexität des Algorithmus verstehen soll. Performanz kann sich aber auch auf den Umgang mit lokalen vs. globalen Optima beziehen, oder auf die *computational cost*, die Kosten (beispielweise Dauer, Energie) des Berechnens (vgl. Domingos, 2012). Auch hier gibt es keine eindeutigen Kennzahlen und die Ansprüche an diese Elemente von Performanz sind klar anhängig davon, was im Rahmen des Forschungsprojektes möglich ist.

Verzerrungen in den Daten und Verzerrungen durch das maschinelle Lernen an sich haben mindestens zwei kritische Konsequenzen. Aus der Perspektive der Nutzer:innenfreundlichkeit ist es nicht zuträglich, wenn bestimmte Gruppen durch algorithmische Entscheidungen benachteiligt werden. Dass dies dennoch geschieht, lässt sich durch unzählige Beispiele der letzten Jahre aufzeigen. Erwähnt sei hier beispielsweise Google, das weniger Werbung für hochbezahlte Jobausschreibungen zeigte, wenn es davon ausging, dass die Userin eine Frau ist (vgl. Datta et al., 2014) und in der Bildersuche für verschiedene Berufsbezeichnungen eine systematische Überrepräsentation entsprechend des vorherstehenden Stereotyps (bezüglich Geschlecht und *race*) aufwies (vgl. Kay et al., 2015). Diese Beispiele sind langfristig gesehen problematisch, wirken aber zu latent, um als Sicherheitsrisiko eingestuft zu werden. Andere algorithmische Verzerrungen hingegen benachteiligen marginalisierte und Risikogruppen direkt, systematisch und setzen ihre Sicherheit aufs Spiel, wie Studien zu Kreditvergabe, Gerichtsurteilen und Gesichtserkennung (vgl. Rudin, 2019) zeigen. Diese Verzerrungen werden oft als ethisches Problem gerahmt, als könnte man darüber diskutieren, ob die Fortführung bestehender Diskriminierungen durch Algorithmen akzeptabel sei. Aktuellere Ansätze verstehen dieses Problem aber als eines der Sicherheit. Ethische Standards hingegen beziehen sich auf die Frage nach dem Ursprung der Daten, auf denen gelernt wird, ob für deren Erhebung und Weiterverarbeitung eigentlich informierter Konsens mit den Datenproduzierenden bestand (vgl. Boyd und Crawford, 2012).

Um diese unterschiedlichen Aspekte standardisiert in den Entwicklungsprozess des maschinellen Lernens einzubringen, haben Mitchell et al. sogenannte *model cards* entwickelt. Diese listen auf, welche Information über und um die Entwicklung eines Modells geteilt werden sollten:

„Model cards serve to disclose information about a trained machine learning model. This includes how it was built, what assumptions were made during its development, what type of model behavior different cultural, demographic, or phenotypic population groups may experience, and an evaluation of how well the model performs with respect to those groups“ (Mitchell et al., 2019, S. 222).

Neben Details zum Modell selbst, wie beispielsweise Trainingprozess und Lizenz, sollen auch der beabsichtigte Zweck und die Zielgruppe definiert werden. Eine Reflexion darüber, welchen Einfluss Gruppenmerkmale, Hardware und Umweltfaktoren auf das Modell haben, soll ebenso veröffentlicht werden wie die Metriken, mit denen das Modell evaluiert wurde. Die Daten, die für Training und Evaluation verwendet wurden, sollen, soweit möglich, zur Verfügung gestellt werden. Und auch eine quantitative Analyse, ethische Abwägungen und darüber hinausgehende Vorbehalte oder Empfehlungen, sollen zum Modell veröffentlicht werden. Die Autor:innen zeigen exemplarisch an dem Klassifikator zu toxischer Sprache in User:innenkommentaren von Perspective API⁵, wie diese model card auf einer DIN A4 Seite umgesetzt werden kann (vgl. Mitchell et al., 2019, S. 227). Die geforderten Angaben klingen äußerst umfangreich, können jedoch, wenn sie dicht formuliert sind, einen wertvollen Überblick über ein Modell geben, der auch für das folgende Qualitätskriterium nützlich ist, welches immer wieder gefordert wird: die Reproduzierbarkeit von Modellen.

3.2.3 Reproduzierbarkeit

„The Computing Science literature is littered with experimental papers that are irrelevant even before they are published because of the careless fashion in which the experiments were conducted and reported. Often the authors themselves could not reproduce the experiments only a few weeks after they rushed to obtain the experimental results for a mid-night conference deadline“ (Amaral et al., 2007, S. 3).

Dieses strenge Urteil wird bekräftigt von der bereits in 3.2.1 erwähnten Studie von Henderson et al., in welcher Standardlösungen (sogenannte „baseline algorithms“, die zum Vergleich mit neuen Lösungen herangezogen werden) für Machine Learning Aufgaben reproduziert werden sollten. Sie stellen fest, dass die Diversität von berichteten Metriken sowie ein Mangel an Signifikanztestungen zu potenziell irreführenden Ergebnissen führen kann (vgl. Henderson et al., 2018, S. 3207). Um dem entgegen zu wirken, fordern sie das Veröffentlichen jeglicher Entscheidungen im experimentellen Entwicklungsprozess des maschinellen Lernens. Dazu gehören intrinsische Faktoren wie Anfangswerte für Zufallsgeneratoren oder die Entwicklungsumgebung genauso wie extrinsische Faktoren wie Hyperparameter oder die Code-Implementierung eines Algorithmus.

Auch die von Pineau 2018 erstmals entwickelte *Machine Learning Reproducibility Checklist* regt zu detailliertem Reporting an und zwar auf den Ebenen von Modell/Algorithmus, Theorie/Hypothese, Daten, Code und experimenteller Ergebnisse: Für den Algorithmus sollen das mathematische Setting, Erklärungen zu Vorannahmen

⁵<https://github.com/conversationai/perspectiveapi/tree/master/1-get-started>

und eine Analyse der Komplexität bereitgestellt werden. Theoretische Behauptungen benötigen Klarheit und umfassende Beweise. Für Datensets sollen deskriptive Statistiken sowie die Größe der Sets für Training, Validierung und Test angegeben werden. Auch eine Beschreibung der Schritte im Preprocessing, der Datenerhebung, Annotationsanweisungen und Methoden zur Qualitätssicherung sollen vermerkt werden. Das Datenset an sich soll bereitgestellt werden, ebenso wie der Code für Training und Evaluation. Für die experimentellen Schritte sollen die unterschiedlichen Hyperparameter berichtet werden und außerdem die Anzahl der Trainings- und Evaluationsdurchläufe, eine Definition der Maße oder Statistiken, mit denen die Ergebnisse dargestellt werden, die durchschnittliche Laufzeit oder geschätzten Energiekosten für das Experiment sowie eine Beschreibung der genutzten IT-Infrastruktur (vgl. Pineau, 2020). Diese Merkmale sollen es ermöglichen, einen Machine Learning Workflow zu wiederholen⁶.

3.2.4 Transparenz

Transparenz kann im Rahmen des Machine Learning Workflows auf verschiedenen Ebenen stattfinden: Bezüglich der im Entdeckungs- und Verwertungszusammenhang liegenden Faktoren für Forschungsentscheidungen, der Veröffentlichung aller relevanten Forschungsentscheidungen im Begründungszusammenhang oder auch der Einsehbarkeit der Klassifikatoren selbst. Ersteres wurde bereits unter Punkt 3.2.2 behandelt, zweites in Abschnitt 3.2.3. In diesem Abschnitt ist Transparenz also als die Nachvollziehbarkeit des Modells und seiner Klassifikationentscheidungen zu verstehen.

Dass Algorithmen sowie Modelle und deren Entscheidungen nachvollziehbar sein müssen, wird spätestens seit dem breiteren Einsatz von automatisierten Entscheidungen in Wirtschaft, Bankenwesen und dem öffentlichen Sektor gefordert (vgl. Mittelstadt et al., 2016). Mit neueren Gesetzen wie der Datenschutzgrundverordnung ist diese Forderung sogar legislativ verankert (vgl. Guidotti et al., 2018). Gründe für das Erklären von Modellen des maschinellen Lernens sind vielfältig. Ihnen übergeordnet ist, dass die Formalisierung des Problems sehr wahrscheinlich unvollständig ist (vgl. Molnar, 2020). Für bestimmte Probleme sei es nicht genug

„to get the prediction (the **what**). The model must also explain how it came to the prediction (the **why**), because a correct prediction only partially solves your original problem“ (Molnar, 2020, S. 19, Hervorh. im Original).

⁶Aus einer sozialwissenschaftlichen Perspektive betrachtet Stodden die Praxis des Veröffentlichens von Studiendetails in der Informatik und stellt fest, dass Faktoren wie Zeitmangel, Platzmangel und Konkurrenzdruck dazu führen können, dass dieser Forderung nach Offenlegung dennoch nicht nachgegangen wird (vgl. Stodden, 2010).

Das Etablieren von Vertrauen, das Identifizieren von Verzerrungen oder Fehlern, das Vorbeugen von Schaden, aber auch ein menschliches Bedürfnis nach Verstehen und Bedeutung sind maßgeblich für die Problemlösung durch Algorithmen (Molnar, 2020, S. 19). Ein großer Teil der Forderungen nach Transparenz richtet sich an Unternehmen, die ihren Konsument:innen algorithmische Entscheidungen vor dem Hintergrund der Nutzer:innenfreundlichkeit erklären sollen oder an Institutionen, die ihre algorithmischen Entscheidungen vor dem Hintergrund von Anti-Diskriminierungsgesetzen verteidigen müssen. Molnar erwähnt als Motivation für Transparenz von maschinellem Lernen aber auch das Streben nach Erkenntnisgewinn in der Wissenschaft:

„The goal of science is to gain knowledge, but many problems are solved with big datasets and black box machine learning models. The model itself becomes the source of knowledge instead of the data“ (Molnar, 2020, S. 20).

Modelle des maschinellen Lernens funktionieren nicht als Werkzeug, mit dem Daten abstrahiert werden, sondern werden selbst zur Quelle des Wissens und produzieren neues Wissen. Lapuschkin et al. weisen auf den „Kluger-Hans-Effekt“ aus der Psychologie hin, nach dem in Experimenten Verhaltenweisen (unbeabsichtigt) so beeinflusst werden können, dass zwar das richtige oder erwartete Ergebnis produziert wird, aber das Versuchssubjekt nur das entsprechende Verhalten erlernt hat, nicht die Gründe dafür. Gerade beim maschinellen Lernen bestünde die Gefahr, dass gelernte Features lediglich auf Scheinkorrelationen (vgl. Lapuschkin et al., 2019, S. 2) oder Artefakten (vgl. Guidotti et al., 2019) beruhen und damit für die wissenschaftliche Anwendung nutzlos sind. Ihre Lösung für dieses Problem ist die Transparenz von maschinellem Lernen:

„[W]e posit that the ability to explain decisions made by learning machines allows us to judge and gain a deeper understanding of whether or not the machine is embodying a particular strategic decision making. Without this understanding we can merely monitor behavior and apply performance measures without possibility to reason deeper about the underlying learned representation“ (Lapuschkin et al., 2019, S. 6).

Diese Unterscheidung zwischen bloßen, messbaren Gütekriterien und den tatsächlichen Inhalten des Gelernten ist, wenn auch nicht so formuliert, eine Unterscheidung zwischen Reliabilität und Validität. An dieser Stelle zeigt sich also, dass ein neuralgischer Punkt in der Qualitätssicherung maschinellen Lernens vor allem ein Validitätsproblem ist, das selten so benannt wird. Das mag auch daran liegen, dass Validität weniger leicht in Zahlen zu fassen ist (vgl. Abschnitt 3.1.1). Für eine wissen-

schaftliche Perspektive auf Machine Learning sollte das Streben nach Validität nun auf der Hand liegen. In Kapitel 4 werde ich deswegen die Frage nach der Transparenz von maschinellem Lernen im integrativen Forschungsprozess der automatischen Inhaltsanalyse als elementares Kriterium für die Validität ausarbeiten. Davor werden noch Konsequenzen zusammengefasst, die sich aus diesen Unterschieden für die automatische Inhaltsanalyse ergeben.

3.3 Herausforderungen für die automatische Inhaltsanalyse

Während computergestützte Verfahren schon früh in der Kommunikationswissenschaft eingesetzt wurden (vgl. Merten, 1983, 35ff und Conway, 2006, S. 187), war dies eher auf die Erhebung von Daten (vgl. Wettstein, 2016) oder auf explorative Ansätze (vgl. Papacharissi und De Fatima Oliveira, 2012) begrenzt. Auch regelbasierte Ansätze, in denen Forschende genau definieren mussten, was der Algorithmus codieren soll, wurden ausprobiert. So stellte Conway (2006) menschliche und algorithmische Codierung gegenüber, indem er den gleichen Datensatz ein Mal menschlich und ein Mal algorithmisch codierte, um Unterschiede in den Forschungsprozessen darzustellen. Auch Zamith und Lewis (2015) und Lacy et al. (2015) beziehen sich auf die Idee eines „algorithmic coder“ der unabhängig von menschlicher Codierung arbeitet. Er produziert grundlegend andere Ergebnisse und wird für andere Aufgaben eingesetzt. Durchgesetzt haben sich regelbasierte Ansätze nicht, dennoch sind einzelne Diskussionspunkte der Autor:innen auch auf das maschinelle Lernen übertragbar und werden im Folgenden aufgegriffen.

Die automatische Inhaltsanalyse bietet großes Potenzial für eine methodologische Weiterentwicklung der Kommunikationswissenschaft. Durch den Einsatz computergestützter Verfahren kann die Menge der codierten Daten gesteigert werden, was wiederum die Zuverlässigkeit und Validität der Inferenz steigert (vgl. Scharkow, 2012):

1. Die Inferenz gewinnt an Reichweite, weil die Stichprobe beispielsweise einen größeren Zeitraum oder mehr Quellen umfassen kann.
2. Stichprobenfehler werden durch eine größere Anzahl an Codierungen ausgeglichen und die Inferenz wird präziser.
3. Messfehler können durch eine Mehrfachcodierung mit verschiedenen oder gleichen Instrumenten ausgeglichen werden.

In den in den letzten Jahren entstehenden *Computational Communication Studies* wird ebenfalls argumentiert, dass durch das Messen in Echtzeit anstelle von Querschnittsstudien die interne Validität von Ergebnissen erhöht wird und eine generell größere Menge an Daten höhere statistische Erklärkraft hat (vgl. Atteveldt et al., 2019a). Gleichzeitig müsse man bedenken, dass bei einer ausreichend großen Menge an Daten jeder Zusammenhang signifikant werden könne (vgl. Atteveldt et al., 2019a). Ebenfalls besteht ein Validitätsproblem, wenn Daten nicht explizit im Hinblick auf eine Fragestellung erhoben, sondern quasi als Nebenprodukt miterhoben (vgl. Atteveldt et al., 2019a) werden. Dieser Diskussionsausschnitt spiegelt eine generellere Diskussion um die Nutzung großer Datenmengen in den Sozialwissenschaften wieder (vgl. Abschnitt 2.1).

Konkreter auf die automatische Inhaltsanalyse bezogen stellt Scharkow die Frage nach der Qualitätssicherung: Er fordert, dass Forschende „für eine gleichermaßen aussagekräftige und anschlussfähige Evaluation von sozialwissenschaftlich relevanten Textklassifikationen [...] eine Vielzahl von Koeffizienten“ (Scharkow, 2012, S. 138) berichten. Die Veröffentlichung von Validitäts- und Reliabilitätswerten in beiden Verfahren sollte insoweit keine große Schwierigkeit darstellen, als dass sie in der Logik des Verfahrens jeweils als notwendig anerkannt sind und Messgrößen sowie Reflektionen über diese existieren. Schwieriger wird es jedoch für Scharkows zweite Forderung, nämlich, dass ein tieferer Blick in die Auswirkungen des Trainings- und Testmaterials sowie des Testens selbst auf die Gütekriterien geworfen wird (vgl. Scharkow, 2012, S. 138). Und keine Erwähnung findet die Frage, ob denn die Gütekriterien der automatischen Textklassifikation allein ausreichen, um das Modell und die Klassifikationen für die Kommunikationswissenschaft als valide zu verstehen. Die in 3.2.1 erwähnte Kritik zeigt, dass Modelle durchaus Scheinkorrelationen lernen können, die für einen ingenieurwissenschaftlichen Ansatz irrelevant sein können, solange die Gütekriterien erreicht sind. Für einen sozialwissenschaftlichen und auch einen naturwissenschaftlichen Ansatz, der aber sicher gehen muss, dass er tatsächlich misst, was er messen will, sind diese Gütekriterien lediglich ein Anfang. Zur Strukturierung werde ich nun erst den Prozess der automatischen Inhaltsanalyse beschreiben und an diesem existierende Strategien, die über die in Abschnitt 3.1 und 3.2 erwähnten Punkte hinausgehen und Lücken in der Qualitätssicherung aufzeigen, die für eine Integration der automatischen Textklassifikation in die kommunikationswissenschaftliche Inhaltsanalyse relevant sind.

Abbildung 3.4 zeigt einen schematischen Ablauf für die automatische Inhaltsanalyse in der Kommunikationswissenschaft, wie sie aktuell umgesetzt wird, wenn Verfahren des überwachten Lernens für die Textklassifikation genutzt werden. Hierbei handelt es sich um eine Kombination der Prozesse in den Abbildungen 3.1 und 3.3 sowie der Praxis, die vereinzelt in Studien beschrieben wird (vgl. Scharkow 2012, S. 51; Stoll et al. 2020). Die automatische Inhaltsanalyse kombiniert also Verfahren der

Abb. 3.4: Forschungsprozess automatische Inhaltsanalyse, IST-Zustand

manuellen Inhaltsanalyse mit Verfahren der automatischen Textklassifikation im überwachten Lernen. Sie verbindet die Entwicklung eines Codebuchs für eine manuelle Inhaltsanalyse und die daraufhin entstandenen codierten Daten mit Verfahren

des maschinellen Lernens, in denen Algorithmen aus jenen codierten Daten Modelle generieren, die auf weitere, größere Datensätze angewandt werden. Die Schritte der manuellen Inhaltsanalyse und der automatischen Textklassifikation folgen üblicherweise aufeinander und haben als Überschneidungspunkt lediglich die gemeinsamen Daten, die von einer Methode zur nächsten weitergereicht werden.

Aufgrund der kommunikationswissenschaftlichen Orientierung beginnt der Forschungsprozess für die automatische Inhaltsanalyse mit einem Erkenntnisinteresse im Entdeckungszusammenhang. Darauf folgen die theoretische Fundierung sowie die Entscheidung für eine Inhaltsanalyse und die weitere Planung des Projektes. Die Datenerhebung und die Entwicklung des Instruments können parallel verlaufen. Die weitere Entwicklung, Anwendung und Auswertung des Erhebungsinstruments unterscheiden sich nicht merklich von der manuellen Inhaltsanalyse. Darauf folgt dann jedoch – anstelle der finalen Interpretation – die automatische Textklassifikation mit Planung, Entwicklung und Test. Erst dann können die Auswertung und Darstellung aller Ergebnisse gemeinsam stattfinden. Zu guter Letzt folgen, je nach Fokus des Projektes, die Publikation und Wissensvermittlung und gegebenenfalls der Einsatz des Modells in einer Anwendung.

Neben der quantitativen Inhaltsanalyse nutzt die Kommunikationswissenschaft auch qualitative Verfahren, die sich jedoch unter anderem in Zielen, Forschungsprozess und Qualitätskriterien von der standardisierten Inhaltsanalyse unterscheiden. Interessant sind qualitative Verfahren für diese Arbeit jedoch, weil sie offener für induktives Vorgehen sind und somit einen Anknüpfungspunkt für die im folgenden Abschnitt beschriebene automatische Textklassifikation bieten können. Lamnek zeigt auf, dass der Forschungsprozess mit qualitativer Inhaltsanalyse offener ist und eine Revision des Kategoriensystems oder einzelner Definitionen auf Basis des Forschungsmaterials vorsieht (vgl. Lamnek, 2010). Diese induktive Ergänzung sowie die explorative, hypothesengenerierende Herangehensweise (vgl. Fürst et al., 2014) kann auch für die Integration manueller Inhaltsanalyse und automatischer Textklassifikation nützlich sein. Vor allem aber das Interesse daran, statt Repräsentativität durch Abstraktion die Vielfältigkeit eines Phänomens zu greifen und auch Grenzfälle zu betrachten kann für die automatische Inhaltsanalyse vorteilhaft sein. Denn eine zu frühe Abstraktion, die bereits in dem manuellen Schritt stattfindet, kann dazu führen, dass letztendlich zu sehr abstrahiert wird, da der automatische Schritt ebenfalls weniger stark vertretene Phänomene an den Rand drängt. Da sich für ein qualitatives Vorgehen jedoch auch andere Qualitätskriterien (vgl. Flick, 2014; Strübing et al., 2018) entwickelt haben und diese im Rahmen der Arbeit nicht weiter verfolgt werden können, sei an dieser Stelle lediglich darauf verwiesen, dass auch diese Perspektive in folgenden Arbeiten gewinnbringend integriert werden könnte, um „eine klar theoriegeleitete, systematische und intersubjektiv nachvollziehbare Analyse mit

einer induktiven Vorgehensweise und vertiefenden Analyse zu kombinieren“ (Fürst et al., 2014, S. 10).

Im Folgenden werden existierende Strategien und Lücken in der Qualitätssicherung dieses Prozesses aufgezeigt. Exemplarisch werden kommunikationswissenschaftliche Studien herangezogen, die sich mit einzelnen Aspekten bereits auseinandergesetzt und damit einen wichtigen Beitrag zu der Reflektion der Methode durch das Fach geleistet haben. Es wird sich auch zeigen, dass der IST-Zustand des Prozesses nur begrenzt funktioniert und eine engere Verzahnung der manuellen Inhaltsanalyse mit der automatischen Textklassifikation einer holistischeren Qualitätssicherung zuträglich wäre.

Theoretische Fundierung

Klassische Verfahren des maschinellen Lernens können nur textimmanentes Wissen nutzen (vgl. Scharnow, 2011, S. 550), Kontextwissen ist nicht abrufbar. Was im Trainingsdatensatz nicht vorhanden ist, kann von dem Modell nicht gelernt werden. Das bedeutet auch, dass in der Zukunft nur das erkannt werden kann, was aus der Vergangenheit schon bekannt ist. Die Prognosen eines Klassifikators sind also begrenzt auf die Erfahrung aus der Vergangenheit. Modernere Verfahren wie neuronale Netze und Transfer Learning haben das Potenzial, über den Text hinaus Informationen einzubinden (vgl. Aken et al., 2020), jedoch ist hier noch unklarer, wie diese Informationen gelernt werden. Die theoretische Fundierung des Forschungsvorhabens muss also berücksichtigen, dass sowohl die Zeit als auch der Kontext Faktoren sind, die besondere Beachtung in der Ausarbeitung des Forschungsstandes und der zentralen Begriffe erfordern.

Für die Hypothesenbildung in der automatischen Inhaltsanalyse muss beachtet werden, dass sozusagen eine Hypothese innerhalb der Hypothese gebildet wird. Die erste Hypothese ist jene, die wir aus dem Erkenntnisinteresse erarbeiten und die den kommunikationswissenschaftlichen Erkenntnisgewinn fördern soll. Die zweite Hypothese ist die, mit der das Modell arbeitet: Dass aus den Features und Mustern im Text (unabhängige Variable) eine bestimmte Kategorie abzuleiten ist (abhängige Variable) (vgl. Stoll, 2020). Die Formulierung beider Hypothesen muss miteinander abgestimmt werden und die Beantwortung der ersten Hypothese muss im späteren Verlauf auch logisch beeinflusst sein von den Ergebnissen der zweiten Hypothese.

Entwicklung

Die manuelle Inhaltsanalyse kennt sowohl direkte als auch indirekte, über einen Index gebildete Messungen (Merten, 1983, S. 318). Die automatische Textklassifikation ist zwar in der Lage, unterschiedliche Formen von Text⁷ entgegen zu nehmen, kann aber nur die Labels vorhersagen, die direkt vorgegeben sind. Wollen wir also beispielsweise Hassrede darüber klassifizieren, dass sie negative Bewertungen von Menschen, Verallgemeinerungen, Entmenschlichungen oder Beschimpfungen enthält, muss ein Klassifikator erst lernen, diese einzelnen Kategorien vorherzusagen, bevor daraus dann die Existenz von Hassrede abgeleitet werden kann. Wird stattdessen nur das Ziel-Label „Hassrede“ für das Training genutzt, sind die Indexvariablen Informationen, die in komplexeren Klassifikatoren für das Lernen nutzbar gemacht werden können, aber in der Anwendung dann nicht explizit identifiziert werden. In diesem Falle könnten wir die Indexvariablen als Features verstehen und es wäre äußerst interessant zu sehen, ob die automatische Textklassifikation diese Indexvariablen als Features identifiziert.

Auch die Größe der Codiereinheit kann variieren. Üblicherweise werden vor allem bei nutzer:innengenerierten Daten ganze Kommentare codiert und klassifiziert, jedoch ist es auch möglich, textnaher einzelne Sätze oder Worte zu annotieren (vgl. Stenetorp et al., 2012). Hier stellt die Identifikationsreliabilität eine besondere Herausforderung. Je kleiner die Einheit wird, desto schwieriger wird es, Einigkeit über die zu identifizierende Kategorie herzustellen. Ein Modell wird letztendlich nur die Größe identifizieren können, die durch die Featuregröße und die Klassifikationseinheit vorgegeben ist. Was bedeutet es für die Identifikationsreliabilität und den Recall, wenn die menschlichen Codierenden eine andere Größe codieren als später die Maschine?

Das Codebuch gibt Regeln und Beispiele für die menschlichen Codierenden vor, mit denen sie die Codierung möglichst gleich durchführen können. Da sich für die automatische Klassifikation regelbasierte Modelle nicht durchgesetzt haben, trägt das Codebuch nicht direkt zu der automatischen Textklassifikation bei. An dieser Stelle gehen wichtige und bereits sorgsam ausgearbeitete Informationen verloren, die zu präziseren Ergebnissen beitragen könnten. Stattdessen „entdeckt“ das Modell eigene Regeln, die wiederum für die manuelle Codierung keine Rolle spielen. Wie können diese Informationen für den jeweils anderen Codier-Prozess nutzbar gemacht werden? Können Features, die das Modell identifiziert und die mit Erkenntnissen, die das Codebuch leiten, beispielsweise eine stärkere Gewichtung erfahren?

⁷zu Bag of Words und Word Embeddings siehe beispielsweise MacAvaney et al., 2019, S. 7.

An dieser Stelle zeigt sich, dass es keine Strategie für die Prüfung der Inhaltsvalidität im maschinellen Lernen gibt. Für die automatische Textklassifikation, die üblicherweise keine starke theoretische Fundierung hat, ist der Abgleich mit dem Forschungsinstrument schwierig. Die Kommunikationswissenschaft bringt diese theoretische Fundierung mit, hat aber keine Werkzeuge, mit denen die technische Dimension des Instruments geprüft werden kann. Zur Prüfung Inhaltsvalidität in der automatischen Inhaltsanalyse können die oben genannten Punkte einen ersten Beitrag leisten.

Reliabilität(en)

Machine Learning geht konzeptionell davon aus, dass der Trainingsdatensatz korrekt codiert wurde, also eine Art Wahrheit darstellt. Sozialwissenschaftliche Forschung hingegen versteht diese Datensätze als eine Annäherung an eine intersubjektiv nachvollziehbare Realität, also eine Überschneidung von unterschiedlichen Wahrheiten. Die Reliabilitätswerte geben an, wie gut diese Überschneidung gelungen ist. Bachl und Scharkow bemängeln bereits für die manuelle Inhaltsanalyse, dass Werte zwar berichtet, aber nicht logisch in die Auswertung und Interpretation der Ergebnisse miteinbezogen werden (vgl. Bachl und Scharkow, 2017). Dies muss umso mehr für die maschinelle Weiterverwertung solcher Daten gelten. Die Zuverlässigkeit der Messung einzelner Konzepte muss widerspiegelt werden in den Erwartungen an die Zuverlässigkeit ihrer maschinellen Identifikation; vielleicht könnten sogar einzelne Kategorien unterschiedlich behandelt werden, je nach Trennschärfe der Kategorien, die sich in den Reliabilitätswerten zeigt.

Es stellt sich auch die Frage, welche Rolle dem Klassifikator zugeschrieben wird (vgl. Conway, 2006). Verstehen wir das Modell als weitere:n Codierer:in, könnte eine Prüfung der Analysevalidität stattfinden: eine Forscher-Klassifikator-Reliabilität würde Auskunft über die Schere zwischen der „korrektesten“ und der maschinellen Codierung geben. Allerdings stellt sich für deren Berechnung eine Herausforderung, die in der manuellen Inhaltsanalyse kein größeres Problem darstellt, für die automatische Textklassifikation jedoch deutlich schwieriger zu beantworten ist: Die Datengrundlage für die Berechnung der Forscher-Klassifikator-Reliabilität. Da der Klassifikator auf einem begrenzten Set an Daten trainiert wurde und üblicherweise über kein Kontextwissen verfügt, hat er einen direkten Nachteil, falls die Daten aus einem zusätzlichen, ihm unbekannten Set stammen. Denn während die menschlichen Codierer:innen Worte, die vorher noch nicht aufgetaucht sind, aufgrund von Kontextwissen in ihr mentales Modell integrieren können, ist das dem Klassifikator nicht möglich⁸. Umgekehrt ist es aber auch problematisch, wenn die Daten dem

⁸Transfer Learning, in dem Modelle erst generell sprachliche Zusammenhänge lernen, bevor sie auf eine bestimmte Klassifikation trainiert werden, können hier interessante Fortschritte bieten. Es

Klassifikator bereits bekannt sind, denn dann wurde er darauf trainiert und hat die manuell vorgegebenen Labels bereits als korrekt angenommen. Eine vom Forscher „unabhängige“ Entscheidung des Klassifikators ist damit nicht gegeben und die Berechnung der Forscher-Klassifikator-Reliabilität nutzlos. Gleiches gilt für die Berechnung von InterCoderreliabilität.

Interessant für die These, dass mehr Trainingsdaten zu einem zuverlässigeren Modell führen, wäre die Berechnung einer Intracoderreliabilität. Das Modell könnte nach und nach mit mehr Daten trainiert und zwischendrin immer wieder am gleichen Validierungsset geprüft werden. Das Ergebnis würde jedoch nichts über die Zuverlässigkeit des Codierers über die Zeit aussagen, sondern über den Einfluss der Datenmenge und gegebenenfalls -breite auf das Training.

Training

Unbalancierte Trainingsdaten stellen für das Training von Modellen ein Problem dar, weil sie lernen, die dominante Klasse häufiger zu schätzen als eine marginal repräsentierte Klasse. Statistische Verfahren laufen eben Gefahr, dass „viele potenziell nützliche Informationen [...] statistisch unsichtbar“ (Stoll, 2020, S. 17) bleiben. So können systematische Verzerrungen entstehen, was die Klassifikation „unsicher und unzuverlässig“ (Stoll, 2020, S. 2) macht, zumal die Verteilung für neue Datensätze nicht gleich bleiben muss. In dem Gütekriterium der Accuracy wird dies jedoch nicht sichtbar. Sie kann beispielsweise 90% erreichen, wenn bei einer 9:1-Verteilung ausschließlich die dominante Klasse geschätzt wird (vgl. Stoll, 2020, S. 2). Anke Stoll vergleicht deswegen zwei Verfahren des Oversampling zum Ausbalancieren der Trainingsdaten und deren Effekt auf die Schätzgenauigkeit für unterrepräsentierte Kategorien. Eine Möglichkeit, mehr Daten für die unterrepräsentierte Kategorie zu erlangen, ist die Produktion weiterer künstlicher Beispiele auf Basis der realen, bereits codierten Beispiele (Synthetic Minority Over-Sampling Technique, SMOTE). Simpler ist die Vervielfachung bereits vorhandener Beispiele, um diese stärker zu gewichten (ROS) (vgl. Stoll, 2020, S. 6). Es zeigt sich, dass beide Verfahren die Schätzgenauigkeit für die unterrepräsentierte Kategorie verbessern, jedoch auch, dass die überrepräsentierte Gruppe nach Anwendung dieser Verfahren seltener erkannt wird. Stoll erklärt, dass damit statistische Fehler ausgeglichen werden und „gute“ Features, die in den unterrepräsentierten Daten bereits vorhanden sind, mehr Aufmerksamkeit bekommen. Sie weist jedoch auch darauf hin, dass inhaltlich durch dieses Verfahren keine Verbesserung erzielt werden kann:

bringt aber auch neue Ungewissheiten in den Forschungsprozess, weil wir verstehen müssen, *wie beziehungsweise welche Merkmale* der Sprache „gelernt“ werden.

„Wo entscheidende Informationen für die Erkennung von Textkategorien fehlen, kann Oversampling die Klassifikation jedoch nicht verbessern, da weder SMOTE noch ROS neue, „bessere“ Features (hier Wörter und Wortkombinationen) generieren kann“ (Stoll, 2020, S. 15).

Oversampling ist keine Anreicherung der Daten, die genuin neue Beispiele einbringt; der Kontext des Modells wird also nicht erweitert. Außerdem können bestehende, uneindeutige Beispiele durch Oversampling mehr Aufmerksamkeit erhalten, sodass hier wiederum mehr Fehlklassifikationen möglich sind (vgl. Stoll, 2020, S. 15). Oversampling gleicht also die Ungleichverteilung von Kategorien aus und kann den Recall für marginalisiert repräsentierte Kategorien verbessern, führt aber auch neue Probleme ein, wie die stärkere Gewichtung von „schlechten“ Features. Ob Oversampling sinnvoll ist, ist also keine Frage, die lediglich im Machine Learning Prozess geklärt werden kann. Die aufgezeigten Vor- und Nachteile sollten je nach kommunikationswissenschaftlicher Fragestellung und Operationalisierungsmöglichkeiten abgewogen werden. Bei klar abgegrenzten Kategorien könnten die Vorteile überwiegen, bei undeutlicheren Kategorien hingegen könnten mehr Fehler produziert werden.

Validierung

Auf die Eindeutigkeit von Kategorien bezieht sich implizit auch die Studie von Song et al., in welcher der Einfluss unvollkommener menschlicher Annotationen in Form eines Goldstandards auf die Validierung automatischer Inhaltsanalysen gemessen wird. Ihr Fokus liegt jedoch nicht auf dem Datensatz, der für Training und Test verwendet wird, sondern auf dem, der für die Validierung oder auch Evaluation Verwendung findet. In der Nutzung manueller Codierungen für die Validierung automatischer Testklassifikation wird selten berücksichtigt, dass die Daten, gegen die validiert wird, bereits nicht perfekt sind (vgl. Song et al., 2020, S. 6), was zu zwei systematischen Fehlern führt:

„First, although automated methods themselves are perfectly reliable, imperfect human judgments of validation materials essentially make the ultimate “target” of such reliable measurements radically deviate from the true yet unknown target of inference, rendering them “reliably wrong” on-target. Second, and relatedly, systematically flawed human judgments of validation materials can introduce unknown bias when evaluating algorithms’ performance *vis-à-vis* true performance“ (Song et al., 2020, S. 6).

Der Automatisierungsschritt macht die Inferenz auf eine wahre, aber unbekannte Zielklassifikation zuverlässig und gleichzeitig fehlerhafter, wenn sie auf bereits feh-

lerbehafteten, menschlich codierten Daten geschieht. Daraus folgt, dass auch in der Messung der Güte eines Modells Verzerrungen vorprogrammiert sind. Song et al. prüfen deswegen mithilfe einer Monte-Carlo-Simulation den Einfluss unterschiedlicher Merkmale des Validierungsset auf die Güte des Modells. Sie stellen fest, dass die Intercoderreliabilität (Krippendorfs Alpha) sowie die Größe und eine passende Randomisierung des Validierungsdatensets den größten systematischen Einfluss auf die Evaluation der Accuracy eines Modells haben, während Mehrfachcodierungen und die Anzahl der Codierenden keinen signifikanten Einfluss nehmen (vgl. Song et al., 2020, S. 14).

Diese Ergebnisse überschneiden sich vor allem im Bezug auf die klare Abgrenzung der Kategorien mit Stolls Studie, lässt doch ein hoher Wert von Krippendorfs Alpha den Schluss zu, dass die Kategorien intersubjektiv nachvollziehbar und eindeutig operationalisiert sind. Auch die Ergebnisse bezüglich der Kategorienverteilung schließen einander nicht aus: Während im Training systematische Ungleichgewichte zwischen den Kategorien ausgeglichen werden können, sollte die Verteilung in der Validierungsphase wieder der ursprünglichen Verteilung entsprechen. Zu guter Letzt zeigen beide Studien, dass es eine engere Verknüpfung braucht zwischen dem Verständnis davon, was ein Trainingsdatensatz bietet, was ein Klassifikator daraus lernen kann und wie diese Ergebnisse einzuordnen sind.

Evaluation

Die Ergebnisse des Klassifikators können umfangreicher evaluiert werden, als es die Gütekriterien Accuracy, Precision und Recall festhalten. Beschränkt die automatische Inhaltsanalyse sich auf diese Werte, gehen auch hier möglicherweise wichtige Informationen verloren. Denn ‚inkorrekt‘ klassifizierte Fälle geben Auskunft darüber, welche Texte schwer automatisch zu klassifizieren sind, während die korrekt klassifizierten Fälle auch zeigen, welche Klassifikationen bereits gut automatisierbar sind. Diese Informationen können natürlich zu einer Anpassung des Instruments führen. Sie könnten eine Grundlage für die Entscheidung bieten, beispielsweise gut funktionierende Klassifikationen an Klassifikatoren auszulagern, um menschlichen Codierer:innen mehr Zeit für die Optimierung der Codierungen zu schaffen, deren technische Umsetzung schwieriger wäre.

Neben der Bewertung anhand von Gütekriterien erfolgen vereinzelt auch Fehleranalysen, die spezifische Fälle betrachten, in denen Klassifikator und die Codierung im Basisdatensatz nicht übereinstimmen (vgl. Aken et al., 2018). Hier wird nicht nur analysiert, welche Art von Klassifikationen für das Modell schwierig sind, sondern auch, welche ursprünglichen menschlichen Codierungen möglicherweise falsch waren. Unter dem Begriff „doubtful labels“ (vgl. Aken et al., 2018) wird also auch

grundsätzlich die „Wahrheit“ des Goldstandards in Frage gestellt, sodass sich das maschinelle Lernen hier erkenntnistheoretisch einer sozialwissenschaftlichen Perspektive annähert. Eine Konsequenz daraus muss allerdings noch gezogen werden. Sollen fragwürdige Klassifikationen aus dem Basisdatensatz geprüft und gegebenenfalls korrigiert oder gar aus dem Training ausgeschlossen werden?

Insgesamt bündelt dieses Kapitel mehr Lücken, als es schließen kann. So zeigt sich das Potenzial, aber auch die Notwendigkeit, einer substanziellen Weiterentwicklung der automatischen Inhaltsanalyse. Denn zuverlässigere Inferenzen durch größere Datenmengen sind für die Kommunikationswissenschaft vorteilhaft, jedoch haben sie nur Wert, solange auch ihre Gültigkeit gegeben ist. Eine automatische Inhaltsanalyse, in der die Schnittstelle zwischen menschlicher und maschineller Auswertung lediglich die Daten an sich sind, reicht offensichtlich nicht aus.

Gezeigt wurde in diesem Kapitel, dass die Balance der Trainingsdaten Einfluss auf das Modell hat und Reliabilität sowie Größe und Randomisierung der Daten für die Validierung eine wichtige Rolle spielen. In einem enger verzahnten, iterativen Forschungsprozess, sollten darüber hinaus folgende Aspekte ausgearbeitet werden:

- Theoretische Grundlagen und der Forschungsstand müssen sowohl für die Kommunikationswissenschaft als auch die Informatik aufbereitet und miteinander verbunden werden.
- Hypothesen müssen sowohl für die sozialwissenschaftlichen als auch die technologischen Aspekte der Fragestellung formuliert werden.
- Durch theoretische und empirische Auseinandersetzungen muss geklärt werden, inwieweit sich technische Rahmenbedingungen in Operationalisierung und Codebuch wiederfinden, welcher Zusammenhang zwischen Features und Zielvariablen besteht oder welchen Einfluss die Größe der Codiereinheit hat.
- Es müssen Strategien entwickelt werden, um Reliabilitätswerte der manuellen Codierung in die darauffolgenden Schritte miteinzubeziehen.
- Letztlich muss ein Umgang gefunden werden für Fälle, in denen manuelle Codierung und automatische Klassifikation voneinander abweichen.

Im folgenden Kapitel werden nun Strategien ausgearbeitet, die ebenfalls der Evaluation einer automatischen Inhaltsanalyse dienen, jedoch mit Fokus auf die Erklärung der automatisierten Entscheidungen, welche zur Prüfung von Inhaltsvalidität genutzt werden könnte.

Erklärbarkeit als Validitätskriterium

Ein wiederkehrendes Thema im maschinellen Lernen ist die Frage danach, *was* ein Modell genau gelernt hat und *warum* es zu einer Entscheidung gekommen ist. Noch besteht kein Konsens darüber, welche Informationen über ein Modell notwendig sind, damit Wissenschaftler:innen automatische Klassifizierungen nachvollziehen können (vgl. Vilone und Longo, 2020, S. 6). So kann ein Modell beispielsweise mathematisch vollständig beschrieben werden; eine menschlich nachvollziehbare, inhaltliche Begründung für eine einzelne Klassifikationsentscheidung liefert diese Information jedoch nicht. Die Begriffsvielfalt zur Erklärbarkeit von Modellen (vgl. Abschnitt 3.2.4 und Vilone und Longo, 2020, S. 9) zeigt, wie je nach Perspektive leicht andere Aspekte in den Fokus rücken. Mit Blick auf Verantwortung und politische Leitlinien fragen beispielsweise die Critical Algorithm Studies nach der Transparenz oder Undurchsichtigkeit von Algorithmen (vgl. Burrell, 2016). Verfechter:innen von Open Science fordern Transparenz und die Veröffentlichung von Algorithmen als Voraussetzung für deren wissenschaftliche Wiederverwendung und Weiterentwicklung (vgl. Stodden und Miguez, 2013; Atteveldt et al., 2019b). Erklärbarkeit sei essentiell für gute Wissenschaft, denn

„the use of machine-learning models in scientific research, for example, in medicine, biology, socio-economic sciences, requires an explanation not only for trust and acceptance of results but also for the sake of the openness of scientific discovery and the progress of research“ (Guidotti et al., 2019).

Die Begriffe Interpretierbarkeit und Erklärbarkeit von Modellen gewinnen in der Usability-Forschung an Relevanz (vgl. Kulesza et al., 2011; Doshi-Velez und Kim, 2017), um die Interaktionen von Mensch und Maschine leichter zu gestalten. Unter dem Begriff der Begründung (justification) sollen Modelle entwickelt werden, die durch den Einblick in ihr Funktionieren Vertrauen schaffen und überzeugend sind (vgl. Biran und Cotton, 2017). Und zu guter Letzt weisen Mueller et al. (2019) darauf hin, dass Entwickler:innen selbst oft nicht verstehen, wie oder warum ihr System funktioniert, sodass Undurchsichtigkeit wiederum die strategische Verbesserung der Systeme verhindert.

Die Forderung danach, eine Klassifikation erklärbar zu machen, ist auch der manuellen Inhaltsanalyse nicht unbekannt. Eine hohe Validität wird vor allem daran gemessen, ob die theoretischen Konzepte in dem Erhebungsinstrument zufriedenstellend abgebildet sind (vgl. Abschnitt 3.1.1). Eine Strategie, um dies einsehbarer zu machen, ist die Veröffentlichung des Codebuchs (z.B. Eilders, 1996; Rössler, 2010) und von Details zum Training der Codierenden. Das Codebuch kann zwar keine Erklärung für jede einzelne Entscheidung einer Codierung liefern, ist aber das Dokument, welches das mentale Modell der Codierenden prägen sollte. Es ist gewissermaßen vergleichbar mit dem Algorithmus, der im Machine Learning Prozess genutzt wird, um einen Klassifikator zu trainieren. Da der Algorithmus für die automatische Textklassifikation jedoch keinen inhaltlichen Bezug zur kommunikationswissenschaftlichen Theorie hat, kann seine Validität erst an der Stelle gemessen werden, an der die Inhalte relevant werden: im Modell und den einzelnen Klassifikationen.

Erklärbarkeit erfüllt damit diverse Funktionen. Sie kann dabei helfen, Verzerrungen in der Forschung aufzudecken, trägt zu einem offenen und nachvollziehbaren wissenschaftlichen Prozess bei, schafft Vertrauen in und Verständnis von den Entscheidungen eines Modells bei einer recht spezifischen Gruppe von User:innen, nämlich Forschenden und Wissenschaftler:innen. Und sie kann das Lernen über und Verstehen von den einem Modell zugrundeliegenden Annahmen unterstützen (vgl. Lombrozo, 2006). Vor allem aber erfüllen Erklärungen auch einen wichtigen erkenntnistheoretischen Zweck: Sie sind das Werkzeug, mit dem Forschende ihre epistemische Verantwortung einlösen (vgl. Miller und Record, 2013). Mit wachsenden technologischen Möglichkeiten sehen Miller und Record Subjekte in der Pflicht, sich so weit wie möglich über die Ursprünge der Annahmen zu informieren, die ihre Erkenntnis prägen:

„When a technology enables subjects to conduct certain inquiries or make determinations that would be impracticable otherwise, they may reasonably be expected to perform them in order to fulfill their epistemic responsibilities“ (Miller und Record, 2013, S. 130).

Diese Prüfung sei nicht nur eine moralische oder vernünftige Verpflichtung, sondern auch epistemisch, da sie zur Sicherung und Vertiefung des Erkenntnisgewinns beiträgt (vgl. Miller und Record, 2013, S. 123). In diesem Kapitel werden wir die Erklärbarkeit des Modells demnach als Voraussetzung für valide wissenschaftliche Erkenntnisse verstehen. Im Folgenden wird erst ausgearbeitet, was Erklärungen sind und welche Formen von Erklärungen Menschen gut verstehen. Daraufhin wird beleuchtet, welche technischen Möglichkeiten bestehen, um die Erklärbarkeit von Modellen zu erhöhen.

4.1 Struktur von Erklärungen

Für die Wissenschaft haben Erklärungen mindestens zwei Dimensionen: als Ziel von Wissenschaft sollen Erklärungen Auskunft über Phänomene dieser Welt geben, als Qualitätskriterium für Wissenschaft sollen Erklärungen Auskunft über das Zustandekommen wissenschaftlicher Ergebnisse geben. Ersteres findet sich sowohl in den Sozialwissenschaften (vgl. Popper, 1957; Harding, 1986), als auch in der Informatik (vgl. Eden, 2007, S. 8) und in aktueller Forschung zu Erklärbarkeit von künstlicher Intelligenz (vgl. Mueller et al., 2018). Eine grundlegende Beschreibung davon, wie Erklärungen aufgebaut sind, bietet Popper:

„Mit einer (kausalen) *Erklärung* ist eine Klasse von Sätzen gemeint, von denen einer den Sachverhalt beschreibt, der erklärt werden soll (das *explicandum*), während die anderen, die erklärenden Aussagen, die 'Erklärung' im engeren Sinne des Wortes bilden (das *explicans* des *explicandums*)“ (Herv. i. O. Popper, 1957, S. 144).

Dabei soll das *explicandum* als wahr bzw. bekannt vorausgesetzt werden, während das *explicans* unbekannt ist, also deduktiv hergeleitet (entdeckt) und empirisch geprüft werden muss. Als kognitive Prozesse stehen Erklärungen in Verbindung mit abduktiver Inferenz und mit Retrospektion (vgl. Mueller et al., 2018, S. 21): So beschreiben sowohl Miller (2019, S.13) als auch Overton (2011) den Prozess des Erklärens als abduktiv. Mit Bezug auf Aristoteles beschreibt Miller vier unterschiedliche Ebenen von Erklärungen: Eine materielle Erklärung bezieht sich auf das Material, aus dem etwas gemacht ist. Beispielweise ist die Hardware, auf der ein Klassifikator ausgeführt wird, eine materielle Ursache für eine Klassifikation. Eine formale oder auch kategorische Erklärung bezieht sich auf die Form oder Eigenschaft einer Sache, beispielsweise ist das Modell eine formale Ursache für eine Klassifikation. Eine effiziente oder auch mechanistische Erklärung bezieht sich auf die Ursache einer Veränderung. Zum Beispiel ist der Algorithmus, der bestimmt, wie Features identifiziert werden, ein effizienter Grund für eine Klassifikation. Und die finale (auch funktionale oder teleologische) Erklärung bezieht sich auf das Ende oder Ziel einer Sache, beispielsweise ist die Automatisierung einer Inhaltsanalyse eine finale Erklärung für eine Klassifikation (vgl. Miller, 2019, 17ff). Vor allem die formale und die mechanistische Erklärung werden für die Kommunikationswissenschaft relevant, wenn sie Ergebnisse des maschinellen Lernens erklären möchte.

Overton beschreibt Erklärungen als die Antwort auf eine Warum-Frage. Diese Art von Fragen beinhaltet eine Annahme, eine innere ja/nein-Frage dazu, ob diese Annahme zutrifft und eine äußere Warum-Frage, die nach einem Grund für das Zutreffen der Annahme fragt. Für eine Anwendung, die automatisiert Hassrede erkennen

soll, würde die Frage „Warum beinhaltet dieser Kommentar Hassrede?“ also die innere Frage beinhalten, ob jener Kommentar Hassrede enthält und die Annahme, dass jener Kommentar Hassrede beinhaltet. Die äußere Warum-Frage will dann den Grund dafür wissen, weshalb die Annahme zutrifft. Diese Erklärung besteht für Overton aus dem Paar von explicans und explicandum, oder auch Annahme und Antwort - im Falle des oben genannten Beispiels also der Annahme, dass der Kommentar Hassrede beinhaltet und einer Antwort, die erklärt, warum das so ist.

Abb. 4.1: Warum-Frage nach Overton, 2011

Wie diese Warum-Fragen für Menschen zufriedenstellend beantwortet werden können, ist ausführlich erforscht. Miller fasst den Forschungsstand wie folgt zusammen: Kausale Erklärungen sind am zugänglichsten (vgl. Miller, 2019, S. 9) und funktionieren am besten, wenn sie im Gegensatz zu einer anderen Möglichkeit oder Erklärung stehen. Menschen fragen eher „Warum ist nicht Q passiert?“ und nicht „Warum ist P passiert?“ (Miller, 2019, S. 6). Sie wählen Erklärungen aus, indem sie erst kausale Verbindungen identifizieren und dann eine Teilmenge dieser Gründe als Erklärung auswählen. Daraufhin folgt eine Evaluation, in der die Person, der etwas erklärt werden soll, die Qualität der Erklärung bewertet (vgl. Miller, 2019, S. 32). Erklärungen sind somit auch in soziale Kontexte eingebettet. Als Form von Wissenstransfer finden sie in Unterhaltungen oder Interaktionen statt und entstehen aus den Annahmen der erklärenden Person über das Informationsbedürfnis der Person, für die die Erklärung gedacht ist (vgl. Miller, 2019, S. 50). Zu guter Letzt sind Erklärungen, die auf statistischen Korrelationen oder Kookurrenzen beruhen, selten zufriedenstellend. Menschen haben ein Bedürfnis nach Gründen, nicht Wahrscheinlichkeiten. Zwar sind Wahrscheinlichkeiten für eine wissenschaftliche Annäherung an Wahrheit relevant, aber um Instanzen mit statistischen Generalisierungen zu erklären, müssen die Ursachen dieser Generalisierungen erklärt werden, nicht nur die Generalisierung an sich (vgl. Miller, 2019, S. 44).

Erklärungen sind also besonders zufriedenstellend, wenn Alternativen miteinbezogen werden; sie werden aktiv und in Interaktion ausgewählt und der Grad ihrer Wahrscheinlichkeit ist nur begrenzt relevant. Diese Erkenntnisse beziehen sich auf Alltagserklärungen und sind auch für die Erklärungen von Klassifikationen nützlich. Betrachten wir jedoch Wissenschaftler:innen – die Zielgruppe dieser Arbeit – dann zeigen sich Widersprüche. Vor allem der letzte Punkt steht in direktem Gegensatz zu der in quantitativen Studien üblichen Praxis, Wahrscheinlichkeiten für die eige-

nen Ergebnisse anzugeben. Dass Wahrscheinlichkeiten allein nicht ausreichen, um überzeugend zu wirken, ist eine interessante Erkenntnis. Überzeugung ist hier aber nicht das Ziel, sondern Transparenz. Es zeigt sich also eine Diskrepanz zwischen den alltäglichen Ansprüchen an Erklärungen und jenen, die für Wissenschaftler:innen funktionieren können. Im Folgenden verstehe ich Erklärungen also im eingangs erwähnten zweiten Sinne: die Erklärung einer wissenschaftlichen Erkenntnis als Sicherung ihrer Qualität, genauer als Sicherung ihrer Validität. Erklärbarkeit kann hier also nicht nur aus einer Perspektive der User:innenfreundlichkeit (für weitere Arbeiten zu diesem Thema vgl. Lage et al., 2019) verstanden werden. Denn es geht nicht ausschließlich darum, dass Forschende dem Klassifikator mehr Vertrauen schenken oder weniger zurückhaltend in der Nutzung sind. Erklärbarkeit für wissenschaftliche Zwecke will prüfen, ob die theoretischen Konzepte auch vom Klassifikator gelernt wurden, ob dieser also valide Ergebnisse produziert.

4.2 Erklärbare Textklassifikation

Als *ex ante* Strategie für Validität sind eine umfassende theoretische Ausarbeitung der Konzepte, eine treffende Operationalisierung, ein gutes Codebuch und eine intensive Schulung mit den Codierenden unumgänglich. *Ex post* wird zur Prüfung der Validität aber auch eine Antwort auf die Frage benötigt, ob das, was so ausführlich erarbeitet und Menschen beigebracht wurde, auch maschinell gelernt wurde. Seit den Siebzigerjahren (vgl. Rudin, 2019) gibt es in der Informatik Bestrebungen, Klassifikationsentscheidungen und die dazugehörigen Modelle nachvollziehbarer zu gestalten, da diese oft als black boxes erscheinen: Entweder können sie aufgrund von proprietärer Software nicht eingesehen werden oder die Mechanismen sind zu kompliziert und damit unverständlich. Es wurden daher unterschiedliche Strategien entwickelt, um mit der Intransparenz von Klassifikatoren umzugehen. Zu beachten ist allerdings, dass eine Erklärung letztendlich nur so gut sein kann wie das Modell selbst - macht das Modell viele Fehler, werden auch die Erklärungen mangelhaft sein. Die Unterscheidung ist jedoch eine Herausforderung: „[I]t is hard to tease apart errors that stem from the misbehavior of the model versus the misbehavior of the attribution method“ (Sundararajan et al., 2017, S. 5109). Um es positiv zu formulieren: Mangelhafte Erklärungen können unter anderem darauf hinweisen, dass das Modell selbst problematisch ist und sollten einen Anstoß zu dessen weiterer Prüfung geben.

Die jeweilige Strategie, mit der Klassifikationsentscheidungen erklärt werden können, ist abhängig von verschiedenen Faktoren, die von Guidotti et al. wie folgt strukturiert werden:

- *Problemtyp*: Entsprechend der in 4.1 erwähnten Ebenen sind vor allem formale und mechanistische Erklärungen für die Kommunikationswissenschaft interessant. Erstere nehmen das Modell als Ursache, letztere hingegen sehen Prozesse im Modell als Ursache. Guidotti et al. nennen erstere eine globale Erklärung, die das Modell als Ganzes beschreiben will (Model Explanation) (vgl. Guidotti et al., 2019, S. 13). Auch ist es möglich, einen direkt interpretierbaren Klassifikator zu nutzen, der von sich aus transparent ist (Transparent Box Design). Die Erklärung von Prozessen im Modell hingegen sind lokale Erklärungen, die einzelne Klassifikationen (Outcome Explanation) oder aber einzelne Merkmale des Klassifikators (Model Inspection) beschreiben.
- *Art der Erklärung*: beinhaltet wiederum zwei Dimensionen (vgl. Mueller et al., 2018). Die inhaltliche Referenz entscheidet, *was* erklärt werden soll – einzelne Beispiele (Prototypen, Fehlklassifikationen, Sonderfälle), Muster, Features oder Ähnliches. Das Format hingegen entscheidet, *wie* erklärt werden soll: mit Visualisierungen (zum Beispiel Heatmaps), Text, Formeln, Diagrammen, Hierarchien und anderen Möglichkeiten. Die Kombination aus inhaltlicher Referenz und dem Format führt dann konkret zu der Wahl von beispielsweise Entscheidungsbäumen, Regeln, Feature Importance, Saliency Masks oder Sensitivitätsanalysen (vgl. Guidotti et al., 2019).
- *Art von Modell*, das erklärbar gemacht werden soll: Modellunabhängige Erklärungen sind möglich. Aber je nach dem, ob beispielsweise ein Neuronales Netz, eine Support Vector Machine, Entscheidungsbäume oder nichtlineare Modelle genutzt werden, können auch spezifischere, modellbasierte Erklärungen entwickelt werden.
- *Art von Daten*, die verarbeitet werden sollen: Tabellarische Daten ermöglichen andere Strategien als Bilder, Text oder Zeitreihen.

Vilone und Longo erwähnen außerdem auch die *Domäne* und die *Zielgruppe* als wichtige Faktoren dafür, welche Strategie für eine Erklärung gewählt werden sollte (vgl. Vilone und Longo, 2020). Zu guter Letzt können die Erklärungen einer *Evaluation* unterzogen werden, sowohl technisch als auch von Menschen.

Der Forschungsstand für Erklärungen von Textklassifikationen ist im Vergleich zu Erklärungen von Computer Vision oder tabellarischen Daten relativ klein. Vilone und Longo zeigen in einer Review, dass lediglich elf Prozent der Methoden für Erklärbarkeit textbezogen sind (vgl. Vilone und Longo, 2020, S. 16). Textklassifikation und deren Erklärungen haben eine besondere Form der Komplexität. So können etwa kontrastierende Beispiele nicht sinnvoll durch kleine Störungen in den Original-

Daten generiert werden (vgl. Liu und Avci, 2019). Wenn wir ein Wort extrahieren wollen, das erklärt, warum eine Klassifikation getroffen wurde und nicht eine andere, reicht die Variation eines Buchstabens nicht aus, um ein sinnvolles Gegenbeispiel zu produzieren: „häflich“ ist kein Gegenteil zu „hässlich“ – die Störung muss Semantik miteinbeziehen. Im Folgenden werden dennoch nur Erklärungsmöglichkeiten aufgezeigt, die spezifisch für Text funktionieren oder generell auf Text anwendbar sind (vgl. Guidotti et al., 2019), um den Fokus auf für die automatische Inhaltsanalyse von Text relevante Strategien zu wahren. Da dieses Forschungsfeld im Machine Learning aktuell sehr aktiv ist und teilweise sehr spezifische Lösungen entwickelt, kann kein Anspruch auf Vollständigkeit erhoben werden. Vielmehr handelt es sich um das Aufzeigen möglichst unterschiedlicher Ansätze als Orientierung dafür, was überhaupt möglich ist. Zum besseren Überblick stellt Tabelle 4.1 die hier gebündelten Ansätze entlang der von Guidotti et al., 2019 entwickelten Faktoren dar. In den nächsten Abschnitten werden diese Ansätze ausführlicher beschrieben.

Methode	Problemtyp	Erklärung	Modell	Daten
PALM	Model Expl.	Entscheidungsbaum	unabhängig	beliebig
LIME	Outcome Expl.	Gewichte	unabhängig	beliebig
Anchors	Outcome Expl.	Wenn-Dann-Regeln	unabhängig	beliebig
MES	Outcome Expl.	Entscheidungsregeln	unabhängig	beliebig
Shapley-Werte	Model Insp.	Gewichte	unabhängig	beliebig
rationales	Model Insp.	Saliency Mask	Neuron. Netze	Text
DeepLIFT	Outcome Expl.	Saliency Mask	Neuron. Netze	beliebig
IG	Model Insp.	Sensitivität	Neuron. Netze	beliebig
VisBERT	Model Insp.	Distanzen	Transformer	Text
Distillation	Model Expl.	Ersatz	Neuron. Netze	beliebig
TCAV	Model Insp.	Konzept-Einfluss	Neuron. Netze	beliebig
Naive Bayes	Transp. Box	Gewichte	Naive Bayes	beliebig
PS	Transp. Box	Prototypen	Data distillation	beliebig
BCM	Transp. Box	Prototypen	Naive Bayes	beliebig

Tab. 4.1: Übersicht über die Methoden für Erklärbarkeit und Interpretierbarkeit, die im Folgenden erklärt werden

4.2.1 Modellunabhängige Erklärungen

Modellunabhängige Erklärungen können sowohl für proprietäre als auch für offene, aber komplexe Systeme¹ verwendet werden. Sie bieten *post hoc* Erklärungen, indem sie ein weiteres Modell trainieren, das unterschiedliche Inputs an die black box reicht und die dazu generierten Outputs betrachtet. Auf dem zu erklärenden Modell sitzt also ein weiteres Modell, das darauf trainiert wurde, ersteres zu erklären. Dieser Ansatz ist eine Form des Reverse Engineering (vgl. Guidotti et al., 2019, S. 19).

PALM (Partition Aware Local Model) bietet eine daten- und modellunabhängige Methode zur Generierung von Erklärungen (vgl. Krishnan und Wu, 2017). Gedacht für das Debugging von Anwendungen, besteht PALM aus zwei Modellen: Ein Meta-Modell partitioniert das Trainingsdatenset und einzelne Sub-Modelle nähern sich dann den lokalen Mustern der Partitionen. Während die Sub-Modelle beliebig komplex sein können, ist das Meta-Modell ein Entscheidungsbaum, anhand dessen Entwickler:innen prüfen können, ob die Entscheidungsregeln mit der Intuition übereinstimmen und einzelne Fehlklassifikationen mit den relevanten Trainingsdaten abgleichen können (vgl. Krishnan und Wu, 2017). Mit diesem Entscheidungsbaum wird also ein reduziertes Modell für das komplexe Modell generiert, das jedoch das gesamte Modell erklären können soll. Somit handelt es sich hier um eine globale Erklärung oder – in Guidottis Terminologie – eine Model Explanation. Für dieses Verfahren muss der Trainingsdatensatz zur Verfügung stehen, was bei der gedachten Anwendung als Debugging-Werkzeug für Entwickler:innen der Fall sein sollte, nicht jedoch bei proprietären Anwendungen, die von Externen genutzt werden.

Eine andere, mittlerweile gut etablierte Methode für die Produktion von modellunabhängigen Erklärungen ist LIME (Local Interpretable Model-agnostic Explanations) von Ribeiro et al. (2016), welche ebenfalls datenunabhängig ist. LIME trainiert einen linearen und somit interpretierbaren Klassifikator für ein einzelnes Beispiel, indem dieser sich weiteren Fällen aus der direkten Nachbarschaft des Beispiels annähert. Damit werden komplexe Modelle in einzelne lokal erklärbare Modelle heruntergebrochen. Abbildung 4.2 zeigt, wie die Ergebnisse grafisch dargestellt werden können. LIME kann so natürlich nicht das Modell im Ganzen erklären. Es handelt sich vielmehr um lokale Erklärungen oder auch eine Antwort auf das Outcome Explanation Problem – wie auch die folgenden Beispiele für modellunabhängige Erklärungen. Für nichtlineare Modelle sind lokale Erklärungen durchaus eine legitime Option.

Anchors ist eine Erweiterung von LIME, die wenn-dann-Regeln zur Erklärung nutzt – genannt Anker. Diese Erklärungen verankern eine Regel hinreichend lokal, sodass

¹ Komplexe Systeme sind beispielsweise nichtlineare Regressionen, Deep Neural Networks oder andere Modelle, die viele Features oder Dimensionen beinhalten.

Abb. 4.2: Beispiel dafür, wie Ergebnisse von LIME dafür verwendet werden können, um Erklärungen für einzelne Klassifikationen von zwei unterschiedlichen Modelle zu vergleichen. In diesem Fall sollen die Klassifikatoren einen Text mit „Christentum“ (grün) oder „Atheismus“ (pink) klassifizieren. Die Balken repräsentieren die Wichtigkeit der relevantesten Worte, die auch im Text darunter hervorgehoben werden (vgl. Ribeiro et al., 2016, S. 1136)

Änderungen an anderen Featurewerten der Instanz keinen Unterschied machen. Ähnliche Instanzen, die mit dem gleichen Anker erklärt werden, haben also auch fast immer die gleiche Klassifikation (vgl. Ribeiro et al., 2018). Anchors ist ebenfalls datenunabhängig und bietet lokale Erklärungen.

Das Model Explanation System (MES) bietet, wie LIME, lokale Erklärungen und ist datenunabhängig. MES geht davon aus, dass Erklärungen einfache logische Aussagen (Entscheidungsregeln) sind und nutzt einen Monte-Carlo-Algorithmus, um mit einem Scoring-System die beste Erklärung für eine einzelne Klassifikation zu finden (vgl. Turner, 2016). Obwohl es auch für Text funktionieren soll, hat Turner lediglich Beispiele für Computer Vision und Credit Scoring (tabellarische Daten) zur Verfügung gestellt. Ob sinnvolle automatisch generierte Erklärungen für Text mit MES umgesetzt werden können, ist offen.

Ein Ansatz, der von der kooperativen Spieltheorie inspiriert ist, orientiert sich an sogenannten Shapley-Werten. Die Idee ist, dass jedes Feature ein:e Spieler:in in einem Spiel ist, bei dem die Klassifikation eine Art Gewinnausschüttung darstellt. Shapley-Werte geben an, wie dieser Gewinn fair zwischen den Features verteilt werden muss. Dafür werden jede mögliche Feature-Kombination und deren Auswirkung auf die Klassifikation verglichen. Somit kann – wenn alle Klassifikationen betrachtet werden – tatsächlich eine Aussage über die globale Relevanz jedes Features und auch über die lokale Relevanz des Features in einer einzelnen Klassifikation angegeben werden (vgl. Molnar, 2020). Praktisch sind diese Werte aufgrund der großen Menge von Features

und deren Kombinationsmöglichkeiten jedoch nicht zu berechnen, dennoch wird die Intuition für verschiedene Methoden genutzt, um Modelle zu erklären. Exemplarisch seien Chen et al. genannt, die in ihrem Experiment Textdaten als lineare Graphen verstehen und so die Distanz zwischen Features (einzelnen Worten) nutzen, um sich Shapley-Werten anzunähern (vgl. Chen et al., 2019). Diese Methoden sind datenunabhängig und je nach Implementierung können lokale oder auch globale Erklärungen für Feature gegeben werden.

Method	Explanation									
Shapley	It	is	not	heartwarming	or	entertaining	.	It	just	sucks
C-Shapley	It	is	not	heartwarming	or	entertaining	.	It	just	sucks
L-Shapley	It	is	not	heartwarming	or	entertaining	.	It	just	sucks
KernelSHAP	It	is	not	heartwarming	or	entertaining	.	It	just	sucks
SampleShapley	It	is	not	heartwarming	or	entertaining	.	It	just	sucks

Abb. 4.3: Beispiel dafür, wie Ergebnisse unterschiedlicher Shapley-Implementierungen dafür verwendet werden können, um Erklärungen für einzelne Klassifikationen zu visualisieren. In diesem Fall wird gezeigt, welche Relevanz die einzelnen Worte für die Klassifikation „negative Filmbewertung“ haben. Positive Worte wurden zwischen weiß und rot markiert, negative Worte zwischen weiß und blau (vgl. Chen et al., 2019, S. 9)

Modellunabhängige Erklärungen bieten den großen Vorteil, dass sie einen beliebigen, unzugänglichen Klassifikator erklären können und bis auf PALM auch keinen Zugriff auf die ursprünglichen Trainingsdaten fordern. Für Forschungsarbeiten, die von solchen Umständen geprägt sind, können modellunabhängige Erklärungen die einzige Möglichkeit sein, um sich einem Verständnis für den Klassifikator anzunähern. Allerdings finden sich auch einige Nachteile in der Tatsache, dass hier nicht der Klassifikator selbst „durchleuchtet“ sondern auf Basis seiner In- und Outputs ein weiteres Modell gebaut wird. Das Modell eines Modells ist eine weitere Vereinfachung, die die Gefahr birgt, Fehler des ursprünglichen Modells zu übergehen oder zu amplifizieren. Rudin plädiert deswegen dafür, diese Art von Erklärungen lediglich Annäherungen oder Trends zu nennen, da sie nicht die Mechanismen des ursprünglichen Modells erklären, sondern Tendenzen darin, wie Features und Vorhersagen miteinander in Verbindung stehen (vgl. Rudin, 2019, S. 208). Sie zeigen nicht, was das Modell gelernt hat und sind somit für eine Validitätsprüfung unzureichend.

4.2.2 Modellbasierte Erklärungen

Modellbasierte Erklärungen benötigen Zugriff auf das Modell selbst, da sie mit Informationen über die Struktur des Systems arbeiten. Sie sind damit spezifisch für eine bestimmte Art von Modell - seien es Neuronale Netze, SVMs oder nicht-lineare Modelle. Anders als modellunabhängige Erklärungen können sie somit auf das Model Inspection Problem eingehen, beispielsweise indem sie die Rolle einzelner Schichten

in Neuronalen Netzen aufgreifen. Aber auch Lösungen für das Explanation Outcome Problem sowie das Transparent Box Design Problem finden sich hier.

Als Lösungsvorschlag für das Explanation Outcome Problem entwickeln Lei et al. einen Ansatz für Neuronale Netze, in dem parallel zu dem Klassifikator auch ein *rational generator* trainiert wird. Dieser selektiert eine Untermenge der Inputsequenz als Begründung für eine Klassifikation. Die Begründung enthält wenige, aber aussagekräftige Worte und soll, wenn sie so reduziert an den Klassifikator gereicht wird, in der gleichen Klassifikation resultieren (vgl. Lei et al., 2016). Diese Art von Erklärung wird auch Saliency Mask genannt und wird vor allem für komplexere Neuronale Netze genutzt, die Bilder verarbeiten; Lei et al.s Ansatz jedoch funktioniert für Text.

a beer that is not sold in my neck of the woods , but managed to get while on a roadtrip . poured into an imperial pint glass with a **generous head that sustained life throughout** . nothing out of the ordinary here , but a good brew still . body **was kind of heavy , but not thick** . the **hop smell was excellent and enticing . very drinkable**

very dark beer . pours **a nice finger and a half of creamy foam and stays** throughout the beer . **smells of coffee and roasted malt . has a major coffee-like taste with hints** of chocolate . if you like black coffee , you will love **this porter . creamy smooth mouthfeel and definitely gets smoother on** the palate once it warms . it 's an ok porter but i feel there are much better one 's out there .

i really did not like this . it just **seemed extremely watery** . i dont ' think this had any **carbonation whatsoever** . maybe it was flat , who knows ? but even if i got a bad brew i do n't see how this would possibly be something i 'd get time and time again . i could taste the hops towards the middle , but the beer got pretty **nasty** towards the bottom . i would never drink this again , unless it was free . i 'm kind of upset i bought this .

a : poured a **nice dark brown with a tan colored head about half an inch thick , nice red/garnet accents when held to the light . little clumps of lacing all around** the glass , not too shabby . not terribly impressive though s : smells **like a more guinness-y guinness really** , there are some roasted malts there , signature guinness smells , less burnt though , a little bit of chocolate ... m : **relatively thick , it is n't an export stout or imperial stout , but still is pretty hefty in the mouth , very smooth , not much carbonation . not too shabby** d : not quite as drinkable as the draught , but still not too bad . i could easily see drinking a few of these .

Abb. 4.4: Beispiel dafür, wie Ergebnisse des rationale generator dafür verwendet werden können, um Erklärungen für einzelne Klassifikationen zu erstellen. In diesem Fall werden Worte hervorgehoben, die eine Stimmung bezüglich Erscheinung (rot), Geruch (blau) und Geschmackserlebnis (grün) klassifizieren (vgl. Lei et al., 2016, S. 7)

DeepLIFT (Deep Learning Important FeaTures) arbeitet ebenfalls mit Saliency Masks und nutzt das Prinzip der Layer-wise Relevance Propagation (LRP), mit der eine einzelne Klassifikationsentscheidung eines Neuronalen Netzes rückwärts durch das Netz verfolgt und je Schicht betrachtet wird. DeepLIFT ist datenunabhängig und analysiert die Differenz der Aktivierungswerte einzelner Neuronen für ein Input-Output-Paar im Vergleich zu einem Referenz-Input-Output-Paar (vgl. Shrikumar et al., 2017). Diese Referenz soll mit Blick auf die spezifische Anwendung gewählt werden, für die Anwendung auf Text liegt es scheinbar nahe, Null zu wählen (vgl. Lertvittayakumjorn und Toni, 2019; Sundararajan et al., 2017). DeepLIFT gibt dann an, welche Features (also zum Beispiel einzelne Worte) am meisten für eine Klassifikation oder für ihr Gegenteil gesprochen haben.

Die Integrated Gradients (IG) Methode geht das Model Inspection Problem für Neuronale Netze an, indem sie die Sensitivität zu Veränderungen des Inputs betrachtet.

Zuerst definiert sie jedoch zwei Eigenschaften, die klassifikationserklärende Methoden allgemein haben sollten. Mit Empfindlichkeit (sensitivity) wird erwartet, dass eine Relevanz von nicht null zugewiesen wird, wenn Input und Baseline sich in einem Feature unterscheiden und unterschiedliche Klassifikationen daraus resultieren. Mit Implementierungsbeständigkeit (implementation invariance) wird erwartet, dass die Attribuierung für funktional äquivalente Neuronale Netze identisch ist – wenn also zwei NNs den gleichen Input gleich klassifizieren, muss auch die Erklärung dafür gleich sein (vgl. Sundararajan et al., 2017). Integrated Gradients erfüllen beide Eigenschaften, indem sie eine Sequenz von Gradienten erstellen, die von der Baseline zum Input führt und deren Durchschnitt berechnet wird. Das neuronale Netz selbst muss dafür nicht manipuliert werden. Diese Art von Erklärung ist eine Sensitivitätsanalyse, weil hier die Korrelation zwischen der Unsicherheit im Output eines Klassifikators und dessen Input analysiert wird. Integrated Gradients beeinflussen das Neuronale Netz nicht und sind relativ simpel zu implementieren (vgl. Sundararajan et al., 2017).

Method	Sentence	Probability
Baseline	I am gay	0.915
	I am straight	0.085
Our Method	I am gay	0.141
	I am straight	0.144

Abb. 4.5: Beispiel dafür, wie Ergebnisse der Integrated Gradients dafür verwendet werden können, um Erklärungen für einzelne Klassifikationen zu erstellen. In diesem Fall werden zwei unterschiedliche NNs zur Klassifikation von toxischer Sprache verglichen. Der zweite Klassifikator wurde entwickelt, um die Verzerrung im ersten zu korrigieren, welcher Identitätsbegriffe generell als Indikator für toxische Sprache identifiziert. Es werden die Worte hervorgehoben, zur Klassifikation positiv beitragen (vgl. Liu und Avci, 2019, S. 6274)

Mit dem jüngsten und äußerst schnellen Erfolg von sogenannten Transformermodellen haben die Methoden zur Visualisierung von Attention (Aufmerksamkeit) und Hidden States (verborgene Zustände) für Erklärungen an Popularität gewonnen (vgl. Aken et al., 2020). Transformer Modelle sind eine besondere Form von tiefen Neuronalen Netzen, die erst grundlegende Sprachstrukturen „lernen“ bevor sie auf spezifische Aufgaben trainiert werden (vgl. Minaee et al., 2020). Aufmerksamkeitsmechanismen sollen in Textverarbeitung den einzelnen Features je nach Kontext unterschiedliche Gewichte geben. Diese Gewichte werden oft als Feature-Relevanz verstanden und sollen so einzelne Klassifikationen erklären, jedoch zeigen Jain und Wallace (2019), dass die Korrelation zwischen anderen, beispielsweise gradientenbasierten, Methoden und den Aufmerksamkeitsgewichten schwach ist, was nahelegt, dass sie keine nützlichen Erklärungen liefern können. Stattdessen schlagen Aken et al. eine Betrachtung der Feature-Embeddings einzelner Schichten (Hidden States) vor, um den Lernprozess von Transformermodellen für einzelne Klassifikationen zu visualisieren. Dafür werden die Vektoren, mit denen die einzelnen Inputs technisch

repräsentiert sind, nach jeder Schicht mit PCA (Principal Component Analysis) in ihrer Dimension reduziert und auf einer zweidimensionalen Fläche abgebildet. Über die Schichten hinweg zeigt sich die Nähe unterschiedlicher Begriffe (vgl. Aken et al., 2020). Anhand einer Question Answering Aufgabe zeigen die Entwickler:innen von VisBERT diese Distanzen in einer Webanwendung². Abbildung 4.6 zeigt, wie dies für ein bereits bekanntes Beispiel aus Abbildung 4.5 aussehen kann. Auch dieser Ansatz erklärt also einzelne Klassifikationen, orientiert sich durch die Analyse einzelner Schichten aber eher an dem Model Inspection Problem. Es ist nicht datenunabhängig, sondern zumindest in der hier erwähnten Implementierung auf Textdaten beschränkt. Das Zeigen von Wortdistanzen legt innere Strukturen des neuronalen Netzes dar, es bedarf jedoch einer weiteren menschlichen Interpretation dessen, um daraus eine Erklärung werden zu lassen.

Abb. 4.6: Beispiel dafür, wie VisBERT Distanzen zwischen gelernten Wortvektoren für eine Klassifikation präsentiert. Der Text ist selbst hinzugefügt und dem Beispiel in Abbildung 4.4 entnommen.

Molnar (2020) erwähnt außerdem die Methode der Model Distillation, die komplexe Modelle und vor allem solche, die aus mehreren Unter-Modellen bestehen (Ensemble Learning), durch ein simpleres Modell ersetzen, das dadurch bestenfalls auch mehr

²<https://visbert.demo.dataxis.com/>

Erklärbarkeit bietet (vgl. Hinton et al., 2015). DistillBERT ist solch ein Modell, das im Vergleich zu BERT 40% kleiner und 60% schneller ist, dabei aber 97% des Sprachverstehens beibehält (vgl. Sanh et al., 2019). Model Distillation wird auch genutzt, um black box Modelle bezüglich ihrer Verzerrungen zu inspizieren (vgl. Tan et al., 2018). Es ist jedoch fragwürdig, inwiefern ein Algorithmus, der einen anderen Algorithmus imitieren soll, tatsächlich Einblicke in letzteren bietet.

Eine erst kürzlich vorgeschlagene Methode zur Interpretierbarkeit von tiefen Neuronalen Netzen sind konzeptbasierte Erklärungen. Testing with Concept Activation Vectors (TCAV) involviert Menschen in die Generierung von Erklärungen, indem diese nach dem Training Beispiele und Gegenbeispiele für bestimmte Konzepte (beispielweise Streifen in Bilderkennung) auswählen, auf denen dann ein linearer Klassifikator trainiert wird, der zwischen den Aktivierungen für beide Sets unterscheiden kann (vgl. Kim et al., 2018). Mit TCAV können globale Erklärungen für den Einfluss der Konzepte auf die Klassen generiert werden. Allerdings sind diese Konzepte abhängig davon, was die Forschenden inhaltlich auswählen und es ist unklar, ob damit alle für das Modell relevanten Konzepte abgedeckt werden. Deshalb fügt eine Weiterentwicklung dieser Methode einen Schritt des *unüberwachten* Lernens hinzu, in dem Konzepte, die ausreichend prädikativ für eine Klassifikation sind, automatisch extrahiert werden (vgl. Yeh et al., 2019). Diese Strategie ähnelt dem topic modeling (vgl. Chaudhary et al., 2020), welches Themen aus Texten extrahiert, allerdings fügen Yeh et al. einen zusätzlichen Schritt ein, der auf räumliche Abhängigkeiten und somit Kohärenz der Konzepte achtet. Für die Berechnung der Wichtigkeit der einzelnen Konzepte nutzen die Forschenden Shapley-Werte. Abbildung 4.7 zeigt exemplarisch an Filmbewertungen, wie diese induktiv erstellten Konzepte durch häufige und nahe Worte dargestellt werden können. Der Wert ConceptSHAP gibt die Wichtigkeit des Konzeptes für alle Klassifikationen an. Die Auswertung der Bedeutung dieser Konzepte liegt nun wieder in der Verantwortung der Forschenden, die beispielsweise zwei eher negative (1&2) und zwei eher positive (3&4) Konzepte erkennen können (vgl. Yeh et al., 2019).

Concept	Nearest Neighbors	Frequent words	ConceptSHAP
1	poorly constructed what comes across as interesting is the wasting my time with a comment but this movie awful in my opinion there were <UNK> and the	worst (168) ever (69) movie (61) seen (55) film (50) awful (42) time(40) waste (34) poorly (26) movies (24) films (18) long (17)	0.280
2	normally it would earn at least 2 or 3 <UNK> <UNK> is just too dumb to be called i feel like i was ripped off and hollywood	not (58) movie (39) make (25) too (23) film (22) even (19) like (18) 2 (16) never (14) minutes (13) 1 (12) doesn't (11)	0.306
3	remember awaiting return of the jedi with almost <UNK> better than most sequels for tv movies i hate male because marie has a crush on her attractive	movies (19) like (18) see (16) movie (15) love (15) good (12) character (11) life (11) little (10) ever (9) watch (9) first (9)	0.174
4	new <UNK> <UNK> via <UNK> <UNK> with absolutely hilarious homosexual and an italian clown <UNK> is an entertaining stephen <UNK> on the vampire <UNK> as a masterpiece	excellent (50) film (25) perfectly (19) wonderful (19) perfect (16) hilarious (15) best (13) fun (12) highly (11) movie (11) brilliant (9) old (9)	0.141

Abb. 4.7: Beispiel dafür, wie Ergebnisse von konzeptbasierten Erklärungen dargestellt werden können (vgl. Yeh et al., 2019, S. 8)

Für kommunikationswissenschaftliche Analysen wäre möglicherweise eine Kombination aus der Analyse unterschiedlicher Schichten in neuronalen Netzen, in der auch festgestellt wird, dass unterschiedliche sprachliche Konzepte in unterschiedlichen Schichten erkannt werden (Aken et al., 2019), und der konzeptbasierten Analyse von Yeh et al., (2019) interessant. So könnten die unterschiedlichen sprachlichen Ebenen für einzelne Konzepte betrachtet werden. Beispielsweise wäre es für die Klassifikation von Hate Speech möglich, dass Beschimpfungen, die vergleichsweise manifest sind, schon auf frühen Schichten identifiziert werden, während latente Konzepte wie beispielsweise Entmenschlichungen erst in späteren Schichten zu finden sind. Dies ist jedoch lediglich eine Vermutung, dessen empirische Prüfung äußerst interessant sein könnte.

Modellbasierte Erklärungen greifen auf das Modell selbst zu, um Erklärungen zu generieren. Dennoch wird selten das Modell im Ganzen erklärt, vielmehr überwiegen auch hier lokale Erklärungen für einzelne Klassifikationsentscheidungen oder einzelne Elemente der Modells. Von diesen Erklärungen kann jedoch nicht auf die Funktionsweise des Modells im Ganzen geschlossen werden (vgl. Mittelstadt et al., 2019) – das wäre ein induktiver Fehlschluss. Tatsächlich tragen diese Methoden auch nicht dazu bei, User:innen ein umfassenderes Verständnis für das Modellverhalten zu geben (vgl. Lertvittayakumjorn und Toni, 2019). Weitere Kritikpunkte an erklärbaren Modellen bringt Rudin an. Sie kritisiert, dass es in dieser Forschung einen „blind belief in the myth of the accuracy-interpretability trade-off“ (vgl. Rudin, 2019, S. 207) gebe. Je besser das Modell klassifiziere, desto komplexer müsse es sein und desto schwieriger sei es zu interpretieren. Als Gegenbeispiel führt sie die Explainable Machine Learning Challenge aus dem Jahre 2018 an, für die ihr Team als einziges ein interpretierbares Modell anstelle von Erklärungen entwickelte, das genauso zuverlässig war wie die Neuronalen Netze, mit denen es in Konkurrenz stand und damit gewann (vgl. Rudin und Radin, 2019). Darüber hinaus zeigen Erklärungen nicht automatisch, was die black box tatsächlich macht. Als Beispiel (siehe Abbildung 4.8) führt sie Salienz in Computer Vision an, die lediglich erklärt, wo das Neuronale Netz besonders hinschaut, aber nicht, was es dort sieht. Auch sind übermäßig komplizierte Modelle fehleranfälliger und schwieriger zu debuggen (vgl. Rudin, 2019, S. 209). Rudin plädiert deswegen dafür, auf black boxes zu verzichten und stattdessen mehr Zeit in die Verbesserung interpretierbarer Modelle zu investieren.

4.2.3 Interpretierbare Modelle

Anstatt komplexe oder geschlossene Systeme für die Textklassifikation zu nutzen, können Forschende also auch auf Modelle zurückgreifen, die von sich aus interpretierbar sind. *Ante hoc* Methoden sollen Modelle von Anfang an erklärbar halten und

	Test Image	Evidence for Animal Being a Siberian Husky	Evidence for Animal Being a Transverse Flute
Explanations Using Attention Maps

Abb. 4.8: Irreführende Erklärung eines Neuronalen Netzes (Rudin, 2019)

werden deswegen auch *white box models* genannt (vgl. Rosenfeld und Richardson, 2019). Sie sind meist modell- und datenspezifisch und können durch Entscheidungsbäume und Entscheidungsregeln oder K-Nearest-Neighbours (kNN) implementiert werden sowie mit Hilfe von diskreten Wahrscheinlichkeiten, wie logistischer Regression oder Naive Bayes (vgl. Molnar, 2020, S. 48). Allerdings sind diese Modelle üblicherweise domänenspezifischer und können, wenn die Daten nicht gut strukturiert und klar sind, enorm viel Rechenaufwand erfordern:

„[I]nterpretability usually translates in practice to a set of application-specific constraints on the model. Solving constrained problems is generally harder than solving unconstrained problems. Domain expertise is needed to construct the definition of interpretability for the domain, and the features for machine learning. For data that are unconfounded, complete, and clean, it is much easier to use a black box machine learning method than to troubleshoot and solve computationally hard problems“ (Rudin, 2019, S. 210).

Der Kontext, in dem das interpretierbare Modell entwickelt wird, spielt also eine wichtige Rolle. Rudin verweist darauf, dass sich Entwickler:innen risikoreicher Anwendungen, in denen die Auswirkungen von Fehlklassifikationen schädliche Konsequenzen haben können, die Zeit nehmen müssen, die komplexeren Probleme zu lösen, um sie gut zu lösen (vgl. Rudin, 2019). Aber auch für die Forschung kann dieses Argument gelten, wenn wir zurück auf das Konzept der epistemischen Verantwortung blicken. Forschende haben darüber hinaus den immensen Vorteil, dass sie bereits über ein fundiertes Domänenwissen verfügen, welches zu einer besser verzahnten Machine Learning Anwendung beiträgt. Auch ist das Streben nach Verstehen eine grundlegende Motivation für Forschung – eine weitere Stärke, die Wissenschaftler:innen in den Prozess des maschinellen Lernens einbringen.

Interpretierbare Modelle für Text lösen das Transparent Box Problem, indem sie an sich lokal oder global interpretierbar sind (vgl. Guidotti et al., 2019, S. 33). Stoll et al. nutzen beispielsweise Naive Bayes für globale Erklärungen, die einzelnen

Worten Gewichte geben: Das Gewicht eines Wortes (oder Features) wird durch seine Wahrscheinlichkeit berechnet, in einer bestimmten Klassifikation aufzutauchen (vgl. Stoll et al., 2020). Die globale Erklärung gibt also eine Liste an Features aus, die relevant für eine Klasse sind. Auch Risch et al. nutzen diese Strategie in einem Vergleich mit erklärbaren Modellen und stellen fest, dass Naive Bayes die niedrigste Leistung zeigt (vgl. Risch et al., 2020). Anders als Stoll et al., die verschiedene Preprocessing-Methoden ausprobieren, zeigen Risch et al. jedoch nicht, ob sie weitere Schritte unternommen haben, um den Klassifikator zu verbessern, was nach Rudin für interpretierbare Modelle jedoch notwendig ist.

Battle of Sarsa It was not won by anyone. Sikh forces managed to cross river and Mughals were retreated

Battle of Sirhind .Sikhs killed wazir khan and established sikh rule in punjab .Fateh burj commemorates the great victory of sikhs over Sirhind. now tell me why you are ignorant fool

Abb. 4.9: Beispiel dafür, wie Ergebnisse von Naive Bayes dafür verwendet werden können, um Erklärungen für einzelne Klassifikationen zu erstellen. In diesem Fall sind die Worte hervorgehoben, die zur Klassifikation als toxische Sprache beitragen (vgl. Risch et al., 2020, S. 10)

Anstelle von gewichteten Features können mit interpretierbaren Modellen auch Prototypen erstellt werden. Eine mögliche Methode wurde von Bien und Tibshirani entwickelt. Der Prototype Selection (PS) Ansatz versucht nicht, die Anzahl der Features so zu reduzieren, dass eine überschaubare Menge dabei rauskommt. Stattdessen sollen die Daten an sich gebündelt werden, indem für jede Instanz ein dazugehöriger Prototyp aus der Nachbarschaft gesucht wird, der das gleiche Label hat. Der Anspruch der Autor:innen ist es, dass insgesamt so wenig Prototypen wie möglich existieren und keine Instanz einen Prototypen hat, der ein anderes Label besitzt (vgl. Bien und Tibshirani, 2011). Besonders für Text überlässt jedoch diese Methode den Forschenden einen großen Inferenzschritt, da nicht markiert wird, welche konkreten Eigenschaften des Prototyps für seine Auswahl relevant waren.

Den Mangel an konkreten Feature-Bedeutungen greift das Bayesian Case Model (BCM) auf, indem es erst die Daten clustert und dann Prototypen sowie Feature-Gewichte für diese Cluster generiert (vgl. Kim et al., 2015b). Damit liefert es globale Erklärungen, allerdings kann es passieren, dass zu viele Cluster gebildet werden und so sowohl die Rechenzeit zu hoch ist als auch die Menge an Erklärungen, was wiederum keine Interpretierbarkeit mehr ermöglicht. Guidotti et al. weisen darauf hin, dass mehrere Weiterentwicklungen für BCM existieren: Menschen können mit dem Modell interagieren, um die Prototypen zu verbessern (vgl. Kim et al., 2015a) oder es werden Instanzen ausgegeben, die nicht gut in das Modell passen, um Gegenbeispiele zu liefern (vgl. Kim et al., 2016).

Interpretierbare Modelle bieten vor allem für jene Forschenden, die bereits Kompetenzen in Statistik haben, eine große Chance. Wenn sie sorgfältig trainiert werden, sind sie von sich aus verständlich, sodass ihre Erklärungen tatsächlich als Validitätsprüfung fungieren können. Allerdings benötigen präzise interpretierbare Modelle eben auch ein wenig mehr Entwicklungszeit und -aufwand, sodass Kosten und Nutzen für das jeweilige Forschungsvorhaben abgewogen werden müssen. Aufgrund der Datenstruktur und Mehrdeutigkeit von Text scheinen interpretierbare Modelle für Textklassifikation jedoch aktuell wenig weiterentwickelt zu werden. Selbst Kim et al., die sich 2015 klar für interpretierbare Modelle aussprechen, entwickeln 2020 die im Abschnitt 4.2.2 dargestellte konzeptbasierte Methode, die zwar Erklärungen auf Basis linearer Klassifikatoren entwickelt, aber nicht das Modell im Ganzen als interpretierbar konzipiert. Das Bestreben im Machine Learning, möglichst generelle Lösungen zu entwickeln, die auf viele verschiedene Probleme angewandt werden, überschneidet sich nicht unbedingt mit dem der Sozialwissenschaften, Probleme im Kontext zu betrachten. Es bleibt zu beobachten, ob vielleicht auch aufgrund neuer Anforderungen an Erklärungen neue Denkanstöße in die Forschung und Entwicklung einfließen.

4.3 Eignung für die Kommunikationswissenschaft

Erklärbarkeit und Interpretierbarkeit bieten die Chance, automatische Textklassifikation auf Validität zu untersuchen. Auch ermöglichen sie den Forschenden ein besseres Verständnis davon, wie und warum ein Modell funktioniert und wo es möglicherweise Fehler macht oder Verzerrungen amplifiziert. Potenziell können Forschende überprüfen, ob die Konzepte aus der Theorie nicht nur von menschlichen, sondern auch von dem algorithmischen Codierer erlernt wurden. Drei Herangehensweisen wurden in diesem Kapitel dargestellt: Interpretierbare Modelle bieten direkte Einblicke in das Modell, jedoch ist der Forschungsstand gerade für die Textklassifikation klein. Modellbasierte Erklärungen können größtenteils lokale Erklärungen liefern oder die Funktion einzelner Schichten eines Modells offenlegen. Allerdings besteht hier eine größere Gefahr der Fehlinterpretation von Erklärungen. Ähnliches gilt für modellunabhängige Erklärungen. Sie bieten zwar den Vorteil, dass für die Erklärungen kein Zugriff auf das ursprüngliche Modell nötig ist und damit auch eine Vergleichbarkeit der Methoden über mehrere Anwendungen oder Forschungsprojekte möglich ist. Allerdings geben sie keinen direkten Einblick in das ursprüngliche Modell, sondern lediglich eine Annäherung, die die Ergebnisse des Modells erklären kann. Für die kommunikationswissenschaftliche Nutzung müssen diese Methoden also noch weiter ausgearbeitet und auf wissenschaftliche Erwartungen angepasst werden.

Die Evaluation von Methoden zur Erklärbarkeit und Interpretierbarkeit kann sowohl technisch als auch aus der Perspektive menschlicher Nutzer:innen erfolgen³ und ist abhängig von den Zielen des Modells, aber auch von den finanziellen und zeitlichen Rahmenbedingungen der Forschung und den Kompetenzen der Forschenden. Die Anforderungen an solch eine Evaluation sollten unter anderem auf Basis des Forschungsstands zur Funktionsweise von Erklärungen (4.1) gewählt werden. *De facto* überschneiden sich die Art der Erklärungen, die aktuelle Erklärbarkeitsstrategien liefern, jedoch wenig mit diesem Forschungsstand. Eine kausale Erklärung mag mit dem Hervorheben oder einer farblichen Markierung von Text nahegelegt werden, aber keine kontrafaktische. Bedenklich ist auch, wenn für das Training eines Erklärers als Baseline schlicht Nullen verwendet werden, denn „Warum wurde diese Klassifikation getroffen anstatt keiner?“ resultiert nicht in einer formalen oder mechanischen Erklärung, sondern in einer funktionalen. Schlussendlich sind keine Evaluationsmethoden bekannt, die Erklärungen darauf prüfen, ob sie für wissenschaftliche Anwender:innen geeignet sind, die damit die Validität eines Modells oder einer Klassifikationen prüfen wollen. Es besteht eine Forschungslücke, die aus mehreren Perspektiven heraus befüllt werden muss: Usability-Forscher:innen finden eine besondere Zielgruppe vor, die neue Qualitätsanforderungen stellt und ein anderes Vorwissen hat als es bei Standardnutzer:innen der Fall ist. Forschende aus dem Bereich des Machine Learnings können die eigenen Methoden in Kooperation mit Wissenschaftler:innen aus anderen Disziplinen absichern. Und Forschenden aus der Kommunikationswissenschaft kann die Entwicklung von Evaluationsmethoden für erklärbare Textklassifikation dabei helfen, klarere Anforderungen an diese Systeme zu formulieren.

Dieses Kapitel hat also gezeigt, dass sich die (kommunikations-)wissenschaftlichen Anforderungen an Erklärungen nicht mit jenen decken, die für andere Gruppen von Nutzer:innen formuliert wurden. Welche Anforderungen genau Erklärbarkeit und Interpretierbarkeit erfüllen müssen, um der *wissenschaftlichen* Perspektive gerecht zu werden, ist scheinbar noch nicht ausreichend erforscht. Wenn maschinelles Lernen gewinnbringend in sozialwissenschaftlicher Forschung eingesetzt werden soll, müssen folgende Forschungslücken geschlossen werden: Sind kausale, kontrastive Erklärungen wissenschaftlich die nachvollziehbarsten? Welche Rolle dürfen Wahrscheinlichkeiten spielen? In welchem Verhältnis stehen individuelle Bedürfnisse zu wissenschaftlichem Anspruch in der Auswahl von Erklärungen? Mit Antworten auf diese Fragen ermöglichen Erklärbarkeit und Interpretierbarkeit nicht nur eine erhöhte Qualitätssicherung, sondern geben Forschenden auch die Möglichkeit, ihre epistemische Verantwortung einzulösen.

³Für einen Überblick über verschiedene Methoden der Evaluation von Modellen für Textklassifikation, siehe Nguyen, 2018.

Darüber hinaus gestatten Erklärbarkeit und Interpretierbarkeit eine Feedbackschleife, in der das Verständnis für ein Modell wiederum positive Auswirkungen auf folgende Operationalisierungen für eine manuelle Inhaltsanalyse haben kann. Denn aktuell führen die Ergebnisse solcher Erklärungen höchstens dazu, dass mehr Daten in das Training gesteckt werden, was jedoch teuer ist und das ursprüngliche Problem nicht angeht (vgl. Liu und Avci, 2019). Möglicherweise sollte eine Iteration dieser Schleife generell in den Forschungsprozess mit einer automatischen Inhaltsanalyse integriert werden. Im folgenden Kapitel werden die bis hierhin erarbeiteten Erkenntnisse genutzt, um einen neuen, enger verzahnten Forschungsprozess für die Integration von manueller Inhaltsanalyse und automatischer Textklassifikation zu entwickeln.

Empfehlungen für die Integration automatischer Textklassifikation in kommunikationswissenschaftliche Forschung

Ein kritisch-rationales Wissenschaftsverständnis, wie es die Kommunikationswissenschaft prägt, fordert falsifizierbare, intersubjektiv nachvollziehbare und empirisch überprüfbare Forschung, die es zum Ziel hat, falsche Theorien zu verwerfen. Mit der stetig wachsenden Menge an Kommunikationsdaten sowie dem zunehmend einfacheren Zugang dazu steigt die Möglichkeit für einen Paradigmenwechsel, der in einem stärker datengläubigen Empirismus resultiert, in dem Theorien nachrangig sind. Vor allem der ingenieurwissenschaftliche Zweig der Informatik knüpft an dieses Vorgehen an. Interdisziplinäre Kooperationen mit diesem Zweig könnten jenen Paradigmenwechsel befeuern. Allerdings streben sowohl in der Kommunikationswissenschaft als auch in der Informatik konstruktiv-kritische Forschende nach intensiver Reflexion der eigenen Arbeit, um Wissenschaft präziser, gültiger, verlässlicher, transparenter, fairer und fehlerresistenter zu gestalten. Anstatt eine Spaltung der Forschungsgemeinschaft in unvereinbare Lager vorauszusagen, ist also auch eine gemeinsame kritische Weiterentwicklung denkbar.

Als Positivbeispiel dafür, wie eine umfassend reflektierte Integration automatischer Textklassifikation durchgeführt und berichtet werden kann, bietet sich „Detecting Impoliteness and Incivility in Online Discussion“ von Stoll, Ziegele und Quiring, (2020) an. Die Autor:innen der Studie beginnen mit einer theoretischen Ausarbeitung der Konzepte „incivility“ und „impoliteness“ und diskutieren dann, wie diese technisch erfasst werden könnten. Sie vergleichen die Stärken und Schwächen diktionsbasierter und maschinell lernender Verfahren und rufen erneut in Erinnerung, dass Algorithmen nicht über den direkten sprachlichen Kontext hinaus, von dem sie lernen, abstrahieren können. Daraufhin beschreiben sie sowohl die sozialwissenschaftliche Operationalisierung und die Reliabilitätswerte ihrer manuellen Inhaltsanalyse nach Krippendorfs Alpha je Kategorie als auch die quasi-Operationalisierung (Feature-Bildung) für die automatische Textklassifikation. Daraufhin werden die Klassifikationsmodelle gewählt und der Trainingsprozess beschrieben. Auch hier finden Vor- und Nachteile sowie die Entscheidung für eine Balancierung der Datensets

aufgrund schiefer Kategorienverteilungen Erwähnung. Die Auswertung der Textklassifikation erfolgt getrennt für jede inhaltliche Kategorie und die unterschiedlichen Algorithmen, deren Performanz verglichen werden sollte. Accuracy, Precision, Recall und F-Measure werden kritisch betrachtet. Besonders hervorzuheben ist außerdem, dass die Autor:innen einzelne informative Features herausgreifen und konkrete Worte sowie Wortkombinationen zeigen, die für eine Klassifikation bedeutsam waren. Dies ist eine mögliche Strategie um zu prüfen, ob die Modelle valide Ergebnisse produzieren: Wenn die Begriffe auch unter der Perspektive der kommunikationswissenschaftlichen Operationalisierungen sinnvoll erscheinen, könnte das ein Hinweis darauf sein, dass aussagekräftige Features gelernt wurden. Zusätzlich ziehen die Autor:innen ein weiteres Datenset für eine externe Validierung heran. Im Fazit fassen sie zusammen, dass sowohl menschliche als auch algorithmische Klassifizierungen Fehler machen können, vor allem, wenn die Ziel-Labels komplexe Konzepte beschreiben (vgl. Stoll et al., 2020, S. 129). Sie schließen mit vier Empfehlungen: Menschliche Codierer:innen sollten wortnaher codieren, wenn Algorithmen von ihnen lernen sollen. Eine Reduktion in der Varianz der Daten könnte präzisere Klassifikatoren ermöglichen. Neuronale Netze könnten möglicherweise besser abstrahieren und ein anderes Input-Format für den Text (Wortvektoren anstelle des Bag of Words-Prinzips) könnte den einzelnen Worten mehr Kontext innerhalb des Satzes geben (vgl. Stoll et al., 2020, S. 130). Diese umfassende und kritische Studie von Stoll et al. liefert eine solide Vorlage für weitere Forschung und wird nun ergänzt durch die Aspekte, die im Rahmen dieser Masterarbeit ausgearbeitet wurden.

Diese Arbeit hat gezeigt, dass unterschiedliche epistemologische Standpunkte in der Kommunikationswissenschaft und dem maschinellen Lernen zu unterschiedlichen Forschungsprozessen und -herangehensweisen führen. Reibungspunkte, mit denen Forschende sich auseinandersetzen sollten, finden sich im eigentlichen Ziel einer Forschung, in der Bedeutung von Theorie, der Standardisierung des Prozesses, der Wahl der Methoden, der Qualitätssicherung der Forschung bis hin zur Dokumentation der Erkenntnisse. Im Folgenden gruppiere ich diese Punkte in drei Bereiche, die vom Konkreten und sofort Umsetzbaren hin zum Abstrakten führen, das für zukünftige Arbeiten auszubauen ist. In 5.1 fasse ich zusammen, welche Werte zur Qualitätssicherung einer Studie berichtet und welche Strategien durchgeführt werden sollten; in 5.2 zeige ich eine mögliche Version des Forschungsprozesses für eine tatsächlich integrierte Inhaltsanalyse und in 5.3 zeige ich auf, welche theoretischen und methodologischen Auseinandersetzungen für eine fundierte interdisziplinäre Weiterentwicklung nötig sind.

5.1 Reporting

Für die Nutzung von Verfahren des maschinellen Lernens in der Kommunikationswissenschaft haben sich noch keine spezifischen Anforderungen bezüglich der Qualitätssicherung etabliert. Bereits für die in der Kommunikationswissenschaft übliche Dokumentation von Qualitätssicherungsstrategien und Reliabilitätswerten ist in Veröffentlichungen wenig Platz, sodass eine ausführlichere Variante, die auch Rechenschaft über den Part des maschinellen Lernens ablegt, umso dichter dokumentiert werden muss. Dokumentation muss ein Verständnis ermöglichen, mit dem eine Studie anhand der Strategien und Kennzahlen bewertet und eingeordnet werden kann. Und besonders in einem interdisziplinären Feld, das gerade erst zusammen wächst, sollte eine Auseinandersetzung mit den Wechselwirkungen zwischen unterschiedlichen Herangehensweisen ermöglicht, wenn nicht sogar angeregt, werden. Reflektiert die Dokumentation erfolgreich die unterschiedlichen epistemologischen Standpunkte, kann außerdem ein Bewusstsein dafür geschaffen werden, dass der Einsatz technischer Lösungen nicht notwendigerweise korrektere oder objektivere Ergebnisse produziert.

Zur Sicherung der Validität einer automatischen Inhaltsanalyse sollte daher gewährleistet werden, dass

1. die Theorie angemessen in das Forschungsinstrument überführt wurde,
2. Forschende und Codierende in ihrem Verständnis des Forschungsinstruments möglichst eng übereinstimmen,
3. das gewählte Messkriterien sich an anderen bereits etablierten Messkriterien ausrichten,
4. die Ergebnisse stimmig sind mit Ergebnissen anderer hochwertiger Studien.

Diese Strategien müssen nicht nur auf den Schritt der manuellen Inhaltsanalyse angewandt werden, sondern auch für die automatische Textklassifikation.

Dies ist durchaus möglich, indem für (1) die Fähigkeiten des gewählten ML-Verfahrens reflektiert werden, das theoretische Konzept zu fassen. Diese Inhaltsvalidität ist nach Einschätzung dieser Ausarbeitung die bis dato am stärksten vernachlässigte Qualitätssicherungsstrategie. Hier können auch Formen der Erklärbarkeit von Modellen nützlich sein, wenn diese einen Einblick in die Funktionsweisen unterschiedlicher Verfahren und deren Angemessenheit für die Untersuchung bieten. Wichtig ist vor allem, dass bei der gewählten Erklärung möglichst wenig über das Verhalten des

Modells inferiert werden muss, weshalb vorzugsweise interpretierbare Modelle zu nutzen sind. Modellbasierte Erklärungen, die zumindest Einzelaspekte eines Modells beleuchten, sind auch denkbar. Modellunabhängige Erklärungen können genutzt werden, wenn kein direkter Zugriff auf das Modell möglich ist. Ein Vorteil von modellunabhängigen Erklärungen wäre tatsächlich die Vergleichbarkeit einer Erklärbarkeitsstrategie über verschiedene Forschungsprojekte hinweg. Für (2) ist ein Forscher-Klassifikator-Reliabilitätstest denkbar. Für die Prüfungs der Kriteriumsvalidität (3) wird bereits häufig das Modell auf einen anderen Datensatz angewandt. Und für (4) ist eine solide Aufarbeitung des auf die Forschungsfrage bezogenen Forschungsstandes im maschinellen Lernen notwendig, was aufgrund des schnell wachsenden Feldes zumindest nicht am Mangel von Veröffentlichungen scheitern sollte.

Zur Sicherung der Reliabilität einer automatischen Inhaltsanalyse sollte die Inter-coderreliabilität aufgeschlüsselt nach einzelnen Kategorien und gegebenenfalls mit unterschiedlichen Maßen berechnet werden, um verschiedene Fehler oder Schiefheiten in den Daten abzudecken. Mit den Reliabilitätswerten für einzelne Kategorien können dann auch Accuracy, Precision, Recall und F-Measure besser eingeordnet werden. Auch mit dem Verständnis, dass die Accuracy eine Art Expertenvalidität und der Recall eine Art Identifikationsreliabilität darstellen, können die Kennwerte besser eingeordnet werden. Da Reproduzierbarkeit im Machine Learning eine wichtige Rolle spielt und den Ankerpunkt für Standardisierungsbestrebungen bietet, kann für eine kompakte Dokumentation des Modells auf die Checkliste von Pineau (2018) zurückgegriffen werden, um die bis hierhin noch nicht abgedeckten Elemente (insbesondere bezogen auf Modell, Daten und Code) festzuhalten.

Diese Strategien sind gedacht als ein erster Schritt zur Standardisierung, als eine weitere Iteration in der Auseinandersetzung mit Qualitätssicherung in der automatischen Inhaltsanalyse. Sie können umgehend umgesetzt werden und tragen dazu bei, die für beide Disziplinen relevanten Qualitäts- bzw. Gütekriterien abzudecken. Allerdings schöpfen sie nicht das volle Potenzial interdisziplinärer Forschung aus, weshalb im folgenden Abschnitt tiefergreifende Möglichkeiten gebündelt werden.

5.2 Forschungsprozess

Ein integrierter Forschungsprozess für die automatische Inhaltsanalyse geht davon aus, dass nicht die Daten die Schnittstelle zwischen manueller Inhaltsanalyse und automatischer Textklassifikation sind, sondern das Forschungsinteresse. Nur so kann Interdisziplinarität von Anfang an im Forschungsprozess verankert werden. Während eine Standardisierung des Prozesses anzustreben ist, müssen Forschende jedoch auch

anerkennen, dass auf dem Weg dorthin noch ein intensiverer Austausch zwischen den Disziplinen nötig ist und aktuelle Forschungsprozesse möglicherweise iterativer arbeiten als jener, der langfristig etabliert werden kann. Abbildung 5.1 zeigt eine mögliche Version des zukünftigen integrierten Forschungsprozesses für die automatische Inhaltsanalyse. Die in grün angefügten Fragen sollen den Prozess begleiten und als Diskussionsgrundlage für den interdisziplinären Austausch fungieren.

Abb. 5.1: Forschungsprozess automatische Inhaltsanalyse, möglicher SOLL-Zustand

Entsprechend dem kritisch-rationalen kommunikationswissenschaftlichen Forschungsprozess wird das Erkenntnisinteresse wieder in den Entdeckungszusammenhang gerückt, wenngleich ich Wechselwirkungen zwischen Entdeckungs- und Begründungszusammenhang nicht ausschließen möchte. Für die theoretische Fundierung ist es von grundlegender Bedeutung, dass Theorie und Forschungsstand aus beiden Disziplinen aufgearbeitet werden, denn sowohl die Inhalte als auch die technischen

Möglichkeiten und Grenzen der automatischen Inhaltsanalyse sollten klar abgesteckt sein. Bei der Formulierung der Hypothesen muss dann nicht nur für die kommunikationswissenschaftliche Fragestellung, sondern auch für die technische Umsetzung eine Erwartung ausgearbeitet werden, damit folgend die Anforderungen umfassend gesammelt werden können. Der Schritt der Planung, in dem die Auswahl auf die automatische Inhaltsanalyse fällt, wird hier explizit erwähnt, um der Projektplanung Sichtbarkeit zu verschaffen, die ebenfalls beide Disziplinen involvieren sollte. In der Phase der Anforderungsanalyse wird festgelegt, welches die Analyseeinheit sein soll und auch, wie sich diese in die Anforderungen an ein Modell übersetzen lassen: Welche Art von Klassifikation wird benötigt? Welche Arten von Fehlern sind eher akzeptabel? Innerhalb welches Grenzbereichs können Fehler vertretbar sein? Weitere Anforderungen, beispielsweise an die Komplexität, Rechenzeit und Erklärbarkeit oder Interpretierbarkeit des Modells werden ebenfalls in diesem Schritt festgelegt. Dieser Schritt ist wichtig für das Erwartungsmanagement und die Kommunikation über die Möglichkeiten der automatischen Textklassifikation. Die Anforderungsanalyse findet im IST-Zustand (siehe 3.4) noch keinen Platz, vermutlich auch, weil das Konzept aus dem Software Engineering stammt und der Kommunikationswissenschaft nicht vertraut ist¹.

Die darauffolgende Entwicklungsphase ist jene, in der aktuell die größte Herausforderung liegt. Hier findet sich das größte Potenzial für welche Wechselwirkungen zwischen Fragestellung, inhaltlicher und technischer Operationalisierung sowie der Übertragung in ein Codebuch und das Training eines Modells. Da sich noch nicht gezeigt hat, welche Algorithmen besonders passend für welche Art von kommunikationswissenschaftlicher Fragestellung sind, ist zu erwarten, dass diese Phase iterativ verläuft und kleinere Experimente nutzt, bis eine Standardisierung erfolgen kann. Denn in diesem Schritt operierten bis dato Kommunikationswissenschaft und maschinelles Lernen isoliert voneinander. Interdisziplinäres Arbeiten setzt an dieser Stelle ein geteiltes Vokabular voraus, um ein möglichst gutes Verständnis des jeweils anderen Bereichs zu erlangen. In der Testphase wird das Instrument – technisch sowie inhaltlich – ausprobiert und geprüft. Der zusätzlich eingeführte Punkt des Probetrainings ist idealistisch, benötigen die meisten Algorithmen doch wesentlich größere Datensätze zum Erstellen eines Modells als jene, die in einer Probecodierung anfallen. Dennoch sollte der Versuch unternommen werden, um nicht erst nach der ausführlichen manuellen Codierung festzustellen, dass sich das erarbeitete Instrument nicht automatisieren lässt. In der Testphase werden dann auch die ersten Qualitätsmaße erhoben (5.1), die unter anderem Auskunft über die Trennschärfe der Kategorien geben können.

¹Für interessierte Lesende sei deshalb auf ein Standardwerk der Anforderungsanalyse (vgl. Jünger und Schade, 2018) und eine erkenntnistheoretische Auseinandersetzung damit (vgl. Scheffé, 1999) verwiesen.

Sind die Ergebnisse der Testphase zufriedenstellend, kann die manuelle Codierung wie gewohnt erfolgen. Die statistische Auswertung in der Reflexionsphase hilft dann dabei, Entscheidungen über eine mögliche Balancierung des Datensets und andere Preprocessing-Methoden zu treffen. Darauf folgt das Training des Modells sowie dessen Optimierung, ebenfalls ein iterativer Schritt. Intercoderreliabilität, Größe des Validierungssets und die Randomisierung der Trainingsdaten haben Auswirkung auf diesen Schritt (vgl. 3.3). In der Auswertungsphase wird das Modell evaluiert, unter Berücksichtigung der Qualitäts- und Gütekriterien sowie der eingangs gesammelten Anforderungen. Auch die Erklärbarkeit des Modells findet hier erneut Berücksichtigung. In der Interpretation und Darstellung der Ergebnisse muss nun besonders darauf geachtet werden, dass Inferenzen gerechtfertigt sind. Auch für diesen Schritt bedarf es noch weiterer interdisziplinärer Auseinandersetzung, um die Zusammenhänge zwischen Ergebnissen der manuellen Inhaltsanalyse und automatischer Textklassifikation besser auszuarbeiten. In 5.3 werden Diskussionspunkte für diese Auseinandersetzung gebündelt.

5.3 Kooperative Weiterentwicklung

Gemeinsame Qualitäts- und Gütekriterien sowie ein standardisierter Forschungsprozess sind zwei wichtige Merkmale für die Weiterentwicklung der automatischen Inhaltsanalyse. Es hat sich aber gezeigt, dass beides nicht ohne Unstimmigkeiten umgesetzt werden kann, deren Lösung profundere Diskussionen benötigt, um auch epistemologisch Einigung zu finden.

Mehr Daten schaffen eine gültigere Inferenz und validere Ergebnisse (vgl. Scharnow, 2011), jedoch nur unter der Bedingung, dass die – beispielsweise durch eine automatische Klassifikation – produzierten Daten selbst auch valide sind. Strategien für und Leerstellen um die Sicherung dieser Validität waren ein zentraler Punkt dieser Arbeit. Offen bleibt, ob von der automatischen Textklassifikation überhaupt auf die Theorie geschlossen werden kann oder ob eine Inferenz lediglich auf die Daten der manuellen Inhaltsanalyse berechtigt ist. Davon ausgehend, dass die bisher genutzten Qualitätskriterien für die manuelle Inhaltsanalyse sowie Gütekriterien für die automatische Textklassifikation als Maße nicht ausreichen, um die Validität der produzierten Daten zu prüfen, müssen also Alternativen entwickelt werden. Erklärbarkeit oder Interpretierbarkeit von Modellen bieten eine erste Möglichkeit, allerdings unterscheiden sich die Ansprüche an eine wissenschaftliche Erklärung von solchen, die für Enduser:innen alltagsrelevant sind. Hier besteht das Potenzial für eine Ausarbeitung der Merkmale für wissenschaftliche Erklärungen, sowohl aus einer theoretischen Perspektive, als auch aus der forschungspraktischen Perspektive der User:innen.

Die Wechselwirkungen zwischen manueller Inhaltsanalyse und automatischer Textklassifikation sind ein weiterer Punkt, der gemeinsam analysiert werden muss. Wenn die Operationalisierung für die manuelle Codierung angepasst wird, damit sie auch technisch ähnlich umgesetzt werden kann, stellt sich die Frage, ob Kategorien damit an Tiefe verlieren oder vielleicht auch an Präzision gewinnen. Die Komplexität von Kommunikation und deren Messbarkeit zeigt sich hier besonders deutlich. Inwieweit will und sollte die Kommunikationswissenschaft also eine Anpassung an das technisch Messbare erlauben? Diese Diskussion kann begleitet werden von Experimenten, die konkret vergleichen, inwieweit Operationalisierung und Feature-Auswahl einander beeinflussen. Weitere Experimente sollten im Bereich der Übertragung von Reliabilitätswerten in die automatische Klassifikation gefördert werden.

Eine grundlegende Frage für die Zukunft der automatischen Inhaltsanalyse ist letztlich, in welcher Form Kommunikationswissenschaft und Machine Learning zusammenarbeiten sollen. Denkbar wäre erstens, dass Kommunikationswissenschaftler:innen die Anwendung automatischer Textklassifikation so weit selbst erlernen, dass sie auf Informatiker:innen nicht weiter angewiesen sind. Die Unabhängigkeit mag verlockend sein, jedoch verlieren sie damit auch Expertise von den Forschenden, die ohnehin mit dem aktuellen Forschungsstand und den entsprechenden Technologien vertraut sind. Auch bedeutet die Fähigkeit, ein Modell zu entwickeln noch nicht, dieses auch zu verstehen. Diese Form der Integration würde also nur auf den ersten Blick Arbeit einsparen, aber bei technologischen Innovationen zurückbleiben. Zweitens wäre eine Form der Interdisziplinarität denkbar, in der die einzelnen Schritte der automatischen Inhaltsanalyse in den Händen der jeweiligen Expert:innen liegen. Daraus ergeben sich jedoch Probleme bezüglich der Wechselwirkungen und gemeinsamer Standards (vgl. 3.3). Drittens ist Interdisziplinarität im Sinne der *knowing communities* denkbar, in der Kommunikationswissenschaftler:innen und Machine Learning Expert:innen zusammenkommen und gemeinsam eine automatische Inhaltsanalyse entwickeln, die über die jeweils eigene Disziplin hinaus funktioniert². Auch diese Option ist mit größerem Aufwand verbunden und es stellt sich die Frage, ob Forschende dann lediglich Expert:innen für eine Methode werden, während Inhalte in den Hintergrund rücken.

Eine große Chance besteht jedenfalls darin, dass Forscher:innen beider Bereiche interdisziplinär interessiert sind und sich der eigenen Probleme bewusst werden: Machine Learning erkennt an, dass ein Mangel an Theorie und Theoriebildung, fehlende Standardisierung und opake Modelle problematisch sind. Die Kommunikationswissenschaft muss ihre Qualitätssicherungsstrategien und -prüfung aktiv in der Auswertung ihrer Ergebnisse berücksichtigen und technologische Möglichkeiten vor der Integration kritischer prüfen.

²Und im Sinne der *knowing communities* auch einen kritischen Blick auf Machtverhältnisse in Wissenschaft und der Entwicklung von Technologien beibehält

Fazit und Ausblick

Diese Arbeit hat unterschiedliche epistemologische Standpunkte in der Kommunikationswissenschaft und der Informatik beleuchtet und deren Konsequenzen für Theoriebezug, Methodologie, Forschungsprozess, Standardisierung und Qualitätssicherung sowie interdisziplinäre Bestrebungen herausgearbeitet. Für die sozialwissenschaftliche Seite wurde gezeigt, dass trotz Unterschieden zwischen kritisch-rationalen und feministischen Positionen generell ein Fokus auf Theoriebildung besteht und methodologische Auseinandersetzungen Teil des wissenschaftlichen Diskurses sind. Der Forschungsprozess für die manuelle quantitative Inhaltsanalyse ist bereits standardisiert, ebenso wie die Qualitätssicherung, die vor allem Validität und Reliabilität stärken soll. Für die computerwissenschaftliche Seite wurde gezeigt, dass formalistische, ingenieurwissenschaftliche und naturwissenschaftlich-experimentelle Ansätze das Fach durch unterschiedliche historische Entwicklungen geprägt haben und es bis jetzt keine Konsolidierung der teils äußerst konträren Positionen gibt. Im Bereich des Machine Learning ist ein ingenieurwissenschaftlicher Fokus prominent, der induktiv vorgeht und auf die Entwicklung einer zufriedenstellenden Anwendung hinarbeitet, nicht auf Theoriebildung. Eine Standardisierung des *Forschungsprozesses* mit automatischer Textklassifikation hat sich noch nicht etabliert, für *Entwicklungsprozesse* kommen unterschiedliche Modelle vor. Die Qualitätssicherung zielt hier vor allem auf Reproduzierbarkeit ab, unter dem Begriff der Güte- oder Evaluationskriterien. Sowohl Kommunikationswissenschaft als auch Machine Learning sind offen für interdisziplinäres Arbeiten und Forschen, was eine gute Voraussetzung für die Entwicklung einer automatischen Inhaltsanalyse ist, die die Stärken beider Seiten vereint und jeweilige Schwächen reflektieren kann.

Aktuelle Entwicklungen in beiden Bereichen zeigen Annäherungen: Die Kommunikationswissenschaft arbeitet verstärkt mit größeren Datenmengen und reflektiert, inwiefern ein starker Datenfokus Auswirkungen auf Theoriebildung hat. Sie greift auf Methoden des maschinellen Lernens zurück und bemüht sich um Qualitätssicherung im kommunikationswissenschaftlichen Sinne. Im Machine Learning operieren auch naturwissenschaftlich-experimentelle Perspektiven, die einen stärkeren Fokus auf Theoriebildung legen, Standardisierung des Forschungsprozesses anstreben und neben der Reproduzierbarkeit auch andere Qualitätskriterien anbringen. Das in der Forschung zu künstlicher Intelligenz oft referenzierte Prinzip der Abduktion ähnelt –

von der Begriffswahl abgesehen – stark den Prinzipien des kritischen Rationalismus, sodass diese eine Grundlage für weitere Kooperationen bieten können.

In der Verzahnung einer manuellen Inhaltsanalyse und einer automatischen Textklassifikation zeigen sich derweil jedoch Lücken, die in dieser Arbeit ebenfalls herausgestellt wurden. Theoretische Vorarbeit und Hypothesenbildung beziehen den technischen Rahmen nicht ausreichend mit ein. In der Überführung der theoretischen Konzepte in messbare Einheiten wird der technische Rahmen ebenfalls noch nicht ausreichend miteinbezogen, sodass die automatische Textklassifikation nicht optimal vorbereitet wird. Der Einfluss der Größe von Untersuchungseinheiten auf die Inferenz ist unklar, ebenso wie der Einfluss auf die Genauigkeit der automatischen Textklassifikation. Merkmale des manuell codierten Datensets wie die Verteilung und Trennschärfe von Kategorien sowie Reliabilitätswerte sind bedeutend für die weitere Auswertung, werden aber nicht immer thematisiert. Während die Kommunikationswissenschaft letztlich anerkennt, dass in der Inhaltsanalyse lediglich der Grad der Übereinstimmung zwischen Codierenden gemessen werden kann, übersetzt das Machine Learning mangelnde Übereinstimmung zwischen Mensch und Maschine als Fehler. Das Konzept eines Goldstandards für die Weiterverwendung manuell codierter Daten in der automatischen Textklassifikation muss ernsthaft in Frage gestellt werden. Darüber hinaus muss diskutiert werden, ob eine Inferenz für die automatisch klassifizierten Daten nur bis zu den manuell codierten Daten zulässig ist oder bis zur Hypothese reichen kann. Der Bezug unterschiedlicher Qualitätssicherungsstrategien und -kriterien aufeinander ist unklar und eine Standardisierung dessen, was zu berichten ist, steht noch aus.

Ein Schwerpunkt dieser Arbeit folgte dann aus der Frage, wie sichergestellt werden kann, dass die theoretischen Konzepte in einer automatischen Inhaltsanalyse nicht nur manuell, sondern auch technisch gelernt werden. Strategien zur Erklärbarkeit und Interpretierbarkeit wurden als mögliche Lösungen für dieses Validitätskriterium ausgearbeitet und entsprechend ihrer Treue zum und Reichweite im erklärten Modell geordnet. Es zeigt sich, dass Anforderungen an Erklärbarkeit oder Interpretierbarkeit abhängig von der Zielgruppe sind. Für Forschende ist Interpretierbarkeit das angemessenere Konzept: Sie benötigen globale Erklärungen, die das ursprüngliche Modell nachvollziehbar machen, um dessen Validität beurteilen zu können. Lokale Erklärungen oder Vereinfachungen des ursprünglichen Modells bergen hingegen die Gefahr, dass darüber falsche Einschätzungen getroffen werden und Validität nicht angemessen zu beurteilen ist.

Die Arbeit schließt mit Empfehlungen für eine automatische Inhaltsanalyse, die interdisziplinär ausgearbeitet ist und zeigt Anhaltspunkte für diese Weiterentwicklung auf. Aufeinander abgestimmte Qualitäts- und Gütekriterien können ohne größeren Aufwand eingeführt werden und sollten mindestens Reliabilitätswerte nach zwei

unterschiedlichen Maßen und aufgeteilt nach einzelnen Kategorien sowie Accuracy, Precision und Recall beinhalten. Zu empfehlen sind zusätzlich die in den model cards (vgl. Mitchell et al., 2019) geforderten Informationen über Modell und Daten. Ein vereinter standardisierter Forschungsprozess wird noch einige Iterationen benötigen, um systematische Wechselwirkungen zwischen beiden Bereichen greifbar zu machen und zu entscheiden, wie diese zu handhaben sind. Experimente, die einzelne Wechselwirkungen prüfen und systematisieren können, werden dafür empfohlen. Zu guter Letzt werden grundlegendere Diskussionen angeregt, die bestenfalls voneinander abweichende epistemologische Positionen konsolidieren können. Mit vergleichenden Experimenten und diesen Diskussionen sollte sichergestellt werden, dass sowohl Theorie als auch Praxis der automatischen Inhaltsanalyse neu geprüft werden. Das Konzept der knowing communities wird aufgegriffen, um eine Vision der automatischen Inhaltsanalyse zu schaffen, in der Forschende beider Bereiche zusammenkommen, um an einem gemeinsamen Forschungsziel zu arbeiten und in allen Phasen des Forschungsprozesses ihre eigenen Standpunkte kritisch reflektieren.

Die Grenzen dieser Arbeit zeigen sich an mehreren Stellen. Zuvorderst wurde für die Inhaltsanalyse lediglich Text berücksichtigt, da andere Medienformate ihre Bedeutungen anders codieren und technisch anders zu analysieren sind. Damit war ein großer Part aktueller Forschung zu automatischer Klassifikation in Computer Vision, die wesentlich breiter aufgestellt ist, nicht zu verwenden. Darüber hinaus wurden aus sozialwissenschaftlicher Perspektive qualitative Methoden sowie aus computerwissenschaftlicher Perspektive explorative Methoden für die Analyse von Text lediglich am Rande erwähnt, da die Nutzung von Verfahren des Machine Learnings in der Kommunikationswissenschaft größtenteils auf die Analyse größerer Datenmengen abzielt und quantitative Auswertungen naheliegender sind. Daraus ergibt sich, dass die Offenheit des Forschungsprozesses nur als temporär verstanden wird, bis ein standardisierter Forschungsprozess mit aufeinander folgenden Schritten etabliert werden kann. Der Fokus auf den Forschungsprozess an sich vernachlässigt die Tatsache, dass der Prozess von Individuen gestaltet wird. Welche Ressourcen und Fähigkeiten Forschende konkret benötigen, um den Prozess neu zu gestalten, war nicht Gegenstand der Arbeit. Abschließend konnten in dieser Arbeit weder eine Anforderungsanalyse mit Forschenden bezüglich Erklärbarkeitsstrategien noch eine eigene praktische Entwicklung durchgeführt werden. Beide Punkte sollten in folgenden Arbeiten jedoch durchaus aufgegriffen werden, um für die Forschung angemessene und für Forschende nützliche Methoden der Offenlegung von maschinellern Lernen zu entwickeln.

Glossar

Abduktion Form logischer Inferenz und Begründung. Nach einer Beobachtung werden verschiedene Thesen für deren Existenz aufgestellt und auf ihre Plausibilität geprüft. Daraufhin wird die beste, oder auch wahrscheinlichste, ausgewählt und empirisch geprüft (vgl. Kitchin, 2014).

Algorithmus Reihe von exakt beschriebenen und endlich vielen Handlungsanweisungen, um ein Problem zu lösen.

black box Machine Learning Modell, dessen innere Funktionsweise unklar ist. Black boxes sind entweder proprietär und deswegen nicht einsehbar oder ihre Funktionsweise ist so komplex, dass die zwar einsehbar, aber für Menschen nicht verständlich sind.

Codierung Überführung von Merkmalen und Einheiten eines (Text-)materials in Meta-Daten (vgl. Früh, 2017, S. 148), die üblicherweise vorher festgelegte Variablen sind. Wird in dieser Arbeit verwendet, wenn menschliche Codierende diese Aufgabe ausführen.

Deduktion Form logischer Inferenz und Begründung, mit der von einzelnen Fällen auf eine allgemeine Regel geschlossen wird (vgl. Reichertz, 2003).

Entscheidungsproblem Mathematisches Problem, das auch für die Informatik relevant ist. Sucht nach einem Algorithmus, dessen Ausgabe (ja/nein) universal valide ist, wenn die Eingabe eine Aussage ist (vgl. Turing, 1936).

explainability Offenlegung von zugrundeliegenden Ursachen für die Entscheidung (beispielsweise Klassifikation) eines Modells (vgl. Miller, 2019).

Feature Eigenschaft eines Phänomens und Einheit anhand derer ein Modell trainiert wird. Anders als inhaltliche Variablen in der Kommunikationswissenschaft sind bei der Textklassifikation Features eher struktureller Natur, also einzelne Worte oder kleinere Wortgruppen.

Induktion Form logischer Inferenz und Begründung, mit der von allgemeinen Regeln auf einzelne Fälle geschlossen wird (vgl. Reichertz, 2003).

Klassifikator Modell, das die Aufgabe erfüllt, eine Klassifikation vorzunehmen.

knowing communities Erkenntnistheoretisches Konzept, in dem Wissen als gemeinsames Projekt geschaffen wird (vgl. Pohlhaus, 2002).

Machine Learning siehe maschinelles Lernen.

maschinelles Lernen Verfahren der automatischen Mustererkennung. Algorithmen nutzen statistische Methoden, um aus Daten Modelle zu abstrahieren, die allgemeinere Muster in den Daten beschreiben können. Diese Methoden können einzeln stehen, verknüpft werden oder mit zahlreichen Zwischenschichten implementiert werden.

Modell Abstrahierte Beschreibung von Daten, in der allgemeinere Muster festgehalten werden. Ein Modell im Machine Learning besteht beispielsweise aus Gewichten für einzelne Features..

neuronales Netz Klassifikator, der aus vielen einzelnen Einheiten (Neuronen) besteht, die per gewichtete Verbindungen verknüpft sind und eine große Struktur bilden (vgl. Kubat, 2017).

Textklassifikation Typisierung von Textmaterial mit üblicherweise vorher festgelegten Kategorien. Wird in dieser Arbeit verwendet, wenn maschinelle Codierende (Klassifikator) diese Aufgabe ausführen.

unüberwachtes Lernen Verfahren des maschinellen Lernens, in dem die Zielgrößen bzw. -kategorien nicht vorgegeben sind und beispielsweise Cluster gebildet werden, deren Bedeutung dann von Menschen interpretiert werden müssen.

überwachtes Lernen Verfahren des maschinellen Lernens, in dem die Zielgrößen bzw. -kategorien vorgegeben sind.

Literatur

- Aken, Betty van, Julian Risch, Ralf Krestel und Alexander Löser (2018). „Challenges for Toxic Comment Classification: An In-Depth Error Analysis“. In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Brussels, Belgium.
- Aken, Betty van, Benjamin Winter, Alexander Löser und Felix A Gers (2019). „How does BERT answer questions?“ In: *ACM CIKM 2019*. Beijing, China.
- (2020). „VisBERT: Hidden-State Visualizations for Transformers“. In: *Companion Proceedings of the Web Conference 2020*. New York, NY, USA: ACM, S. 207–211.
- Amaral, José Nelson, Michael Buro, Renee Elio et al. (2007). *About Computing Science Research Methodology*. URL: <https://webdocs.cs.ualberta.ca/~simc603/readings/research-methods.pdf>.
- Amershi, Saleema, Andrew Begel, Christian Bird et al. (2019). „Software Engineering for Machine Learning: A Case Study“. In: *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. Montreal, Canada: IEEE, S. 291–300.
- Anderson, Chris (2008). *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*. URL: <https://www.wired.com/2008/06/pb-theory/> (besucht am 24. Mai 2020).
- Atteveldt, Wouter van und Tai Quan Peng (2018). „When Communication Meets Computation: Opportunities, Challenges, and Pitfalls in Computational Communication Science“. In: *Communication Methods and Measures* 12.2-3, S. 81–92.
- Atteveldt, Wouter van, Drew Margolin, Cuihua Shen, Damian Trilling und René Weber (2019a). „A Roadmap for Computational Communication Research“. In: *Computational Communication Research* 1.1, S. 1–11.
- Atteveldt, Wouter van, Joanna Strycharz, Damian Trilling und Kasper Welbers (2019b). „Toward Open Computational Communication Science: A Practical Road Map for Reusable Data and Code“. In: *International Journal of Communication* 13.0, S. 20.
- Bachl, Marko und Michael Scharkow (2017). „Correcting Measurement Error in Content Analysis“. In: *Communication Methods and Measures* 11.2, S. 87–104.
- Baumann, Eva und Helmut Scherer (2012). „Wider das Ideologische“. In: *Methodenkombinationen in der Kommunikationswissenschaft*. Hrsg. von Wiebke Loosen und Armin Scholl. Köln, Germany: Herbert von Halem.

- Berelson, Bernhard (1952). *Content Analysis in Communication Research*. Glencoe, UK: The Free Press.
- Bien, Jacob und Robert Tibshirani (2011). „Prototype selection for interpretable classification“. In: *Annals of Applied Statistics* 5.4, S. 2403–2424.
- Biran, Or und Courtenay Cotton (2017). „Explanation and Justification in Machine Learning: A Survey“. In: *IJCAI Workshop on Explainable AI (XAI)*. August. Melbourne, Australia, S. 8–14.
- Boumans, Jelle W. und Damian Trilling (2016). „Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars“. In: *Digital Journalism* 4.1, S. 8–23.
- Boyd, danah und Kate Crawford (2012). „Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon“. In: *Information Communication and Society* 15.5, S. 662–679.
- Brosius, Hans-Bernd, Friederike Koschel und Alexander Haas (2009). *Methoden der empirischen Kommunikationsforschung*. 5. Aufl. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Brühl, Rolf (2014). *Wie Wissenschaft Wissen schafft: Wissenschaftstheorie für Sozial- und Wirtschaftswissenschaften*. 2. überarb. Konstanz, Germany: UVK Verlagsgesellschaft.
- Burrell, Jenna (2016). „How the machine ‘thinks’: Understanding opacity in machine learning algorithms“. In: *Big Data and Society* 3.1.
- Chaoji, Vineet, Rajeev Rastogi und Gourav Roy (2016). „Machine learning in the real world“. In: *Proceedings of the VLDB Endowment* 9.13, S. 1597–1600.
- Chaudhary, Yatin, Hinrich Schütze und Pankaj Gupta (2020). „Explainable and Discourse Topic-aware Neural Language Understanding“. In: *arXiv Computer Science*.
- Chauviré, Christiane (2005). „Peirce, Popper, Abduction, and the Idea of a Logic of Discovery“. In: *Semiotica* 4.153, S. 209–221.
- Chen, Jianbo, Le Song, Martin J. Wainwright und Michael I. Jordan (2019). „L-Shapley and C-Shapley: Efficient model interpretation for structured data“. In: *7th International Conference on Learning Representations, ICLR 2019*. New Orleans, United States, S. 1–17.
- Cohen, Paul R. und Adele E. Howe (1988). „How evaluation guides AI research: The message still counts more than the medium“. In: *AI Magazine* 9.4, S. 35–43.
- (1989). „Toward AI Research Methodology: Three Case Studies in Evaluation“. In: *IEEE Transactions on Systems, Man and Cybernetics* 19.3, S. 634–646.
- Colburn, Timothy (2008). „Methodology of Computer Science“. In: *The Blackwell Guide to the Philosophy of Computing and Information*. Oxford, UK: Blackwell Publishing Ltd, S. 318–326.
- Colburn, Timothy und Gary Shute (2010). „Knowledge, Truth, and Values in Computer Science“. In: *Thinking Machines and the Philosophy of Computer Science*. IGI Global, S. 119–131.
- Conway, Mike (2006). „The Subjective Precision of Computers: A Methodological Comparison with Human Coding in Content Analysis“. In: *Journalism & Mass Communication Quarterly* 83.1, S. 186–200.

- Datta, Amit, Michael Carl Tschantz und Anupam Datta (2014). „Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination“. In: *CoRR*, S. 1–26.
- Denning, Peter J. (2005). „Is computer science science?“ In: *Communications of the ACM* 48.4, S. 27.
- Dodge, Jesse, Suchin Gururangan, Dallas Card, Roy Schwartz und Noah A. Smith (2019). „Show Your Work: Improved Reporting of Experimental Results“. In: *Empirical Methods on Natural Language Processing (EMNLP)*. Hong Kong, China. arXiv: 1909.03004.
- Domingos, Pedro (2012). „A few useful things to know about machine learning“. In: *Communications of the ACM* 55.10, S. 78–87.
- Doshi-Velez, Finale und Been Kim (2017). „Towards A Rigorous Science of Interpretable Machine Learning“. In: *ML*, S. 1–13. arXiv: 1702.08608.
- Eden, Amnon H (2007). „Three Paradigms of Computer Science“. In: *Minds and Machines* 17.2, S. 135–167.
- Eilders, Christiane (1996). *Nachrichtenfaktoren und Rezeption*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Flick, Uwe (2014). „Gütekriterien qualitativer Sozialforschung“. In: *Handbuch Methoden der empirischen Sozialforschung*. Wiesbaden: Springer Fachmedien Wiesbaden, S. 411–423.
- Freitas, Ricardo (2009). „Scientific Research Methods and Computer Science“. In: *MAPI Seminars Workshop 2009*. Porto, Portugal.
- Fretwurst, Benjamin (2015). „Reliabilität und Validität von Inhaltsanalysen. Mit Erläuterungen zur Berechnung des Reliabilitätskoeffizienten ›Lotus‹ mit SPSS“. In: *Qualitätskriterien in der Inhaltsanalyse*. Hrsg. von Werner Wirth, Katharina Sommer, Martin Wettstein und Jörg Matthes. Köln, Germany: Herbert von Halem.
- Früh, Werner (2017). *Inhaltsanalyse: Theorie und Praxis*. 9. Aufl. UTB GmbH.
- Fürst, Silke, Constanze Jecker und Philomen Schönhagen (2014). „Die qualitative Inhaltsanalyse in der Kommunikationswissenschaft“. In: *Handbuch nicht standardisierte Methoden in der Kommunikationswissenschaft*. Wiesbaden: Springer Fachmedien Wiesbaden, S. 1–14.
- George, Alexander und Daniel J. Velleman (2018). *Zur Philosophie der Mathematik*. Bd. 8. 1. Berlin, Heidelberg: Springer Berlin Heidelberg, S. 135–145.
- Gill, Navdeep, Patrick Hall, Kim Montgomery und Nicholas Schmidt (2020). „A responsible machine learning workflow with focus on interpretable models, post-hoc explanation, and discrimination testing“. In: *Information (Switzerland)* 11.3, S. 1–32.
- Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri et al. (2019). „A Survey of Methods for Explaining Black Box Models“. In: *ACM Computing Surveys* 51.5, S. 1–42.
- Harding, Sandra G (1986). *The science question in feminism*. New York, NY, USA: Cornell Univ. Press, 271 S.
- Hartmanis, Juris (1993). „Some observations about the nature of computer science“. In: *Foundations of Software Technology and Theoretical Computer Science*. Hrsg. von Rudrapatna K. Shyamasundar. Berlin, Germany: Springer, S. 1–12.
- Henderson, Peter, Riashat Islam, Philip Bachman et al. (2018). „Deep reinforcement learning that matters“. In: *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, S. 3207–3214.

- Hinton, Geoffrey, Oriol Vinyals und Jeff Dean (2015). „Distilling the Knowledge in a Neural Network“. In: *NIPS 2014 Deep Learning Workshop*, S. 1–9. arXiv: 1503.02531.
- Holloway, C. Michael (1995). „Software Engineering and Epistemology“. In: *ACM SIGSOFT Software Engineering Notes* 20.2, S. 20–21.
- Holsti, Ole R. (1969). *Content Analysis for the Social Sciences and Humanities*. Reading, MA, USA: Addison-Wesley.
- Horsten, Leon (2019). „Philosophy of Mathematics“. In: *The Stanford Encyclopedia of Philosophy*. Hrsg. von Edward N Zalta. Spring 201. Metaphysics Research Lab, Stanford University. URL: <https://plato.stanford.edu/archives/spr2019/entries/philosophy-mathematics/>.
- Jain, Sarthak und Byron C. Wallace (2019). „Attention is not Explanation“. In: *Proceedings of the 2019 Conference of the North*. Stroudsburg, PA, USA: Association for Computational Linguistics, S. 3543–3556.
- Jünger, Jakob und Henriette Schade (2018). „Liegt die Zukunft der Kommunikationswissenschaft in der Vergangenheit? Ein Plädoyer für Kontinuität statt Veränderung bei der Analyse von Digitalisierung“. In: *Publizistik* 63.4, S. 497–512.
- Kandel, Sean, Andreas Paepcke, Joseph M. Hellerstein und Jeffrey Heer (2012). „Enterprise Data Analysis and Visualization: An Interview Study“. In: *IEEE Transactions on Visualization and Computer Graphics* 18.12, S. 2917–2926.
- Karmasin, Matthias (2008). „Kommunikations-Kommunikationswissenschaft: Wissenschaftstheoretische Anmerkungen zur Theoriediskussion in den Kommunikationswissenschaften“. In: *Theorien der Kommunikations- und Medienwissenschaft*, S. 229–247.
- Kay, Matthew, Cynthia Matuszek und Sean A. Munson (2015). „Unequal Representation and Gender Stereotypes in Image Search Results for Occupations“. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*. Bd. April. New York, USA: ACM Press, S. 3819–3828.
- Kim, Been, Elena Glassman, Brittney Johnson und Julie Shah (2015a). „iBCM : Interactive Bayesian case model empowering humans via intuitive interaction“. In: *CSAIL-technical report*, S. 1–12. URL: <https://dspace.mit.edu/handle/1721.1/96315>.
- Kim, Been, Cynthia Rudin und Julie Shah (2015b). „The Bayesian Case Model: A Generative Approach for Case-Based Reasoning and Prototype Classification“. In: *Advances in Neural Information Processing Systems* 3. January, S. 1952–1960.
- Kim, Been, Rajiv Khanna und Oluwasanmi O Koyejo (2016). „Examples are not enough, learn to criticize! Criticism for Interpretability“. In: *Advances in Neural Information Processing Systems* 29. Hrsg. von D D Lee, M Sugiyama, U V Luxburg, I Guyon und R Garnett. Barcelona, Spain: Curran Associates, Inc., S. 2280–2288.
- Kim, Been, Martin Wattenberg, Justin Gilmer et al. (2018). „Interpretability beyond feature attribution: Quantitative Testing with Concept Activation Vectors (TCAV)“. In: *35th International Conference on Machine Learning, ICML 2018* 6, S. 4186–4195.
- Kitchin, Rob (2014). „Big Data, new epistemologies and paradigm shifts“. In: *Big Data and Society* April-June, S. 1–12.
- Knuth, Donald E. (2003). *Things a Computer Scientist Rarely Talks About*. Chicago, IL, USA: University of Chicago Press.

- Koch, Thomas, Christina Peter und Philipp Müller (2019). *Das Experiment in der Kommunikations- und Medienwissenschaft*. Studienbücher zur Kommunikations- und Medienwissenschaft. Wiesbaden: Springer Fachmedien Wiesbaden.
- Kotsiantis, S. B. (2007). „Supervised Machine Learning: A Review of Classification Techniques“. In: *Informatica* 31.3, S. 249–268.
- Krippendorff, Klaus (2004). *Content analysis: an introduction to its methodology*. 2. ed. Thousand Oaks, CA, USA: SAGE Publications Ltd.
- Krishnan, Sanjay und Eugene Wu (2017). „PALM“. In: *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics - HILDA'17*. Bd. s3-I. 12. New York, USA: ACM Press, S. 1–6.
- Krotz, Friedrich, Andreas Hepp und Carsten Winter (2008). „Einleitung: Theorien der Kommunikations- und Medienwissenschaft“. In: *Theorien der Kommunikations- und Medienwissenschaft*. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 9–25.
- Kubat, Miroslav (2017). *An Introduction to Machine Learning*, S. 1–348.
- Kuhn, Thomas S. (1977). „Objectivity, Value Judgment, and Theory Choice“. In: *The Essential Tension. Selected Studies in Scientific Tradition and Change*. University of Chicago Press, S. 320–339.
- (1990). „The Road since Structure“. In: *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1990.2, S. 3–13.
- Kulesza, Todd, Margaret Burnett, Simone Stumpf et al. (2011). „Where Are My Intelligent Assistant's Mistakes? A Systematic Testing Approach“. In: *End-User Development*. Hrsg. von Maria Francesca Costabile, Yvonne Dittrich, Gerhard Fischer und Antonio Piccinno. Berlin, Heidelberg: Springer, S. 171–186.
- Lacy, Stephen, Brendan R. Watson, Daniel Riffe und Jennette Lovejoy (2015). „Issues and Best Practices in Content Analysis“. In: *Journalism & Mass Communication Quarterly* 92.4, S. 791–811.
- Lage, Isaac, Emily Chen, Jeffrey He et al. (2019). „An Evaluation of the Human-Interpretability of Explanation“. In: *arXiv preprint*, S. 1–24.
- Lamnek, Siegfried (2010). „Inhaltsanalyse“. In: *Qualitative Sozialforschung*. Weinheim, Germany: Beltz Verlagsgruppe, S. 434–497.
- Lapuschkin, Sebastian, Stephan Wäldchen, Alexander Binder et al. (2019). „Unmasking Clever Hans predictors and assessing what machines really learn“. In: *Nature Communications* 10.1, S. 1–8.
- Larose, Daniel T. und Chantal D. Larose (2014). *Discovering Knowledge in Data*. 2. Aufl. Bd. 39. 5. Hoboken, NJ, USA: John Wiley & Sons, Inc., S. 51–90.
- Lauf, Edmund und Harald Berens (2003). „Das Risiko ökologischer und individualistischer Fehlschlüsse in der Inhaltsanalyse“. In: *Publizistik* 48.4, S. 454–469.
- Lei, Tao, Regina Barzilay und Tommi Jaakkola (2016). „Rationalizing Neural Predictions“. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, S. 107–117.

- Lertvittayakumjorn, Piyawat und Francesca Toni (2019). „Human-grounded Evaluations of Explanation Methods for Text Classification“. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Section 3. Stroudsburg, PA, USA: Association for Computational Linguistics, S. 5194–5204.
- Lipton, Zachary C. (2018). „The mythos of model interpretability“. In: *Communications of the ACM* 61.10, S. 36–43.
- Lipton, Zachary C. und Jacob Steinhardt (2019). „Troubling trends in machine-learning scholarship“. In: *Queue* 17.1, S. 1–15.
- Liu, Frederick und Besim Avci (2019). „Incorporating Priors with Feature Attribution on Text Classification“. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, S. 6274–6283.
- Lombrozo, Tania (2006). „The structure and function of explanations“. In: *Trends in Cognitive Sciences* 10.10, S. 464–470.
- MacAvaney, Sean, Hao-Ren Yao, Eugene Yang et al. (2019). „Hate speech detection: Challenges and solutions“. In: *PLOS ONE* 14.8. Hrsg. von Minlie Huang, S. 1–16.
- Marcos, Esperanza und Alfredo Marcos (1999). „An Aristotelian Approach to the Methodological Research: A Method for Data Model Construcion“. In: *Information Systems - The Next Generation. Proceedings of the 4th UKAIS Conference*. Hrsg. von L. Brooks und C. Kimble. Berkshire, UK: McGraw-Hill, S. 532–543.
- Margolin, Drew B. (2019). „Computational Contributions: A Symbiotic Approach to Integrating Big, Observational Data Studies into the Communication Field“. In: *Communication Methods and Measures* 13.4, S. 229–247.
- McKee, George (1995). „Computer Science or Simply 'Computics'? The open channel“. In: *Computer* 28.12, S. 136.
- Mehler, Alexander und Andy Lücking (2014). „On Covering the Gap between Computation and Humanities“. In: *Dagstuhl Reports*. Bd. 4. 1, S. 91–92.
- Merten, Klaus (1983). *Inhaltsanalyse*. Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.
- Meyen, Michael, Maria Löblich, Senta Pfaff-Rüdiger und Claudia Riesmeyer (2011). *Qualitative Forschung in der Kommunikationswissenschaft*. 2. Aufl. Studienbücher zur Kommunikations- und Medienwissenschaft. Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.
- Miller, Boaz und Isaac Record (2013). „Justified belief in a digital age: On the epistemic implications of secret internet technologies“. In: *Episteme* 10.2, S. 117–134.
- Miller, Tim (2019). „Explanation in artificial intelligence: Insights from the social sciences“. In: *Artificial Intelligence* 267, S. 1–38.
- Minaee, Shervin, Nal Kalchbrenner, Erik Cambria et al. (2020). „Deep Learning Based Text Classification: A Comprehensive Review“. In: *CoRR*, S. 1–42.
- Mitchell, Margaret, Simone Wu, Andrew Zaldivar et al. (2019). „Model Cards for Model Reporting“. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*. New York, New York, USA: ACM Press, S. 220–229.

- Mittelstadt, Brent, Chris Russell und Sandra Wachter (2019). „Explaining Explanations in AI“. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*. New York, New York, USA: ACM Press, S. 279–288.
- Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter und Luciano Floridi (2016). „The ethics of algorithms: Mapping the debate“. In: *Big Data and Society* 3.2, S. 1–21.
- Molnar, Christoph (2020). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Morisville, NC, USA: Lulu. URL: <https://christophm.github.io/interpretable-ml-book>.
- Mueller, Shane T., Robert R. Hoffman, William Clancey, Abigail Emrey und Gary Klein (2018). *Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI*. Techn. Ber. Institute for Human and Machine Cognition.
- Neumann, Marion (2019). „AI education matters“. In: *AI Matters* 5.3, S. 21–24.
- Newell, Allen, Alan J Perlis und Herbert A Simon (1967). „Computer Science“. In: *Science* 157.3795, S. 1373–1374.
- Nguyen, Dong (2018). „Comparing automatic and human evaluation of local explanations for text classification“. In: *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference 1*, S. 1069–1078.
- Overton, James A. (2011). „Scientific Explanation and Computation“. In: *Proceedings of the 6th International Explanation-Aware Computing (Ex-aCt) workshop*. Hrsg. von Thomas Roth-Berghofer, Nava Tintarev und David B. Leake. Barcelona, Spain, S. 41–50.
- Papacharissi, Zizi und Maria De Fatima Oliveira (2012). „Affective News and Networked Publics: The Rhythms of News Storytelling on Egypt“. In: *Journal of Communication* 62.2, S. 266–282.
- Parsons, Charles (2007). *Mathematical Thought and Its Objects*. Cambridge, MA, USA: Cambridge University Press.
- Petzold, Charles (2008). *The annotated Turing: a guided tour through Alan Turing's historic paper on computability and the Turing machine*. Indianapolis, IN, USA: Wiley, XII, 372 S. ;
- Pineau, Joelle (2020). *The Machine Learning Reproducibility Checklist (Version 2.0)*. URL: www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist-v2.0.pdf.
- Pohlhaus, Gaile (2002). „Knowing communities: An investigation of Harding's standpoint epistemology“. In: *Social Epistemology* 16.3, S. 283–293.
- Popper, Karl (1934). „Die wissenschaftliche Methode“. In: *Lesebuch: Ausgewählte Texte zur Erkenntnistheorie, Philosophie der Naturwissenschaften, Metaphysik, Sozialphilosophie*. Hrsg. von David Miller. UTB GmbH.
- (1945). „Die Verteidigung des Rationalismus“. In: *Lesebuch: Ausgewählte Texte zur Erkenntnistheorie, Philosophie der Naturwissenschaften, Metaphysik, Sozialphilosophie*. Hrsg. von David Miller. UTB GmbH.
- (1953). „Das Problem der Induktion“. In: *Lesebuch: Ausgewählte Texte zur Erkenntnistheorie, Philosophie der Naturwissenschaften, Metaphysik, Sozialphilosophie*. Hrsg. von David Miller. UTB GmbH.

- Popper, Karl (1957). „Die Zielsetzung der Erfahrungswissenschaft“. In: *Lesebuch: Ausgewählte Texte zur Erkenntnistheorie, Philosophie der Naturwissenschaften, Metaphysik, Sozialphilosophie*. Hrsg. von David Miller. UTB GmbH.
- (1960). „Erkenntnis ohne Autorität“. In: *Lesebuch: Ausgewählte Texte zur Erkenntnistheorie, Philosophie der Naturwissenschaften, Metaphysik, Sozialphilosophie*. Hrsg. von David Miller. UTB GmbH.
 - (1967). „Subjektive oder objektive Erkenntnis?“. In: *Lesebuch: Ausgewählte Texte zur Erkenntnistheorie, Philosophie der Naturwissenschaften, Metaphysik, Sozialphilosophie*. Hrsg. von David Miller. UTB GmbH.
- Pressesprecher.com (2020). *Zwei Punkte für ein Halleluja*. URL: <https://www.pressesprecher.com/nachrichten/zwei-punkte-fuer-ein-halleluja-491995682>.
- Reichertz, Jo (2003). *Die Abduktion in der qualitativen Sozialforschung*. Bd. 13. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Ribeiro, Marco Tulio, Sameer Singh und Carlos Guestrin (2016). „Why Should I Trust You?“ In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, S. 1135–1144.
- (2018). „Anchors: High-precision model-agnostic explanations“. In: *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, S. 1527–1535.
- Risch, Julian, Robin Ruff und Ralf Krestel (2020). „Explaining Offensive Language Detection“. In: *Journal for Language Technology and Computational Linguistics* 34.1, S. 1–19.
- Rosenfeld, Avi und Ariella Richardson (2019). „Explainability in human-agent systems“. In: *Autonomous Agents and Multi-Agent Systems* 33.6, S. 673–705.
- Rössler, Patrick (2010). *Inhaltsanalyse*. 3. Aufl. Stuttgart, Germany: UTB GmbH.
- Rudin, Cynthia (2019). „Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead“. In: *Nature Machine Intelligence* 1.5, S. 206–215.
- Rudin, Cynthia und Joanna Radin (2019). „Why Are We Using Black Box Models in AI When We Don’t Need To? A Lesson From An Explainable AI Competition“. In: *Harvard Data Science Review* 1.2, S. 1–9.
- Samek, Wojciech und Klaus Robert Müller (2019). „Towards Explainable Artificial Intelligence“. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11700 LNCS, S. 5–22.
- Sanh, Victor, Lysandre Debut, Julien Chaumond und Thomas Wolf (2019). „DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter“. In: *EMC2. The 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing*. Vancouver BC, Canada, S. 2–6.
- Scharkow, Michael (2011). „Zur Verknüpfung manueller und automatischer Inhaltsanalyse durch maschinelles Lernen“. In: *Medien & Kommunikationswissenschaft* 59.4, S. 545–562.
- (2012). „Automatische Inhaltsanalyse und maschinelles Lernen“. Dissertation. Universität der Künste, Berlin.

- Scharkow, Michael (2013). „Thematic content analysis using supervised machine learning: An empirical evaluation using German online news“. In: *Quality & Quantity* 47.2, S. 761–773.
- Scheffe, Peter (1999). „Softwaretechnik und Erkenntnistheorie“. In: *Informatik-Spektrum* 22.2, S. 122–135.
- Scheufele, Bertram und Ines Engelmann (2009). *Empirische Kommunikationsforschung*. Konstanz, Germany: UVK.
- Shalev-Shwartz, Shai und Shai Ben-David (2014). *Understanding Machine Learning: From Theory to Algorithms*. NY, USA: Cambridge University Press.
- Shapere, Dudley (1984). „The Structure of Scientific Revolutions“. In: *The Philosophical Review* 73.3, S. 37–48.
- Shrikumar, Avanti, Peyton Greenside und Anshul Kundaje (2017). „Learning important features through propagating activation differences“. In: *34th International Conference on Machine Learning, ICML 2017*. Bd. 7. Sydney, Australia, S. 4844–4866.
- Simon, Herbert Alexander (1990). *The sciences of the artificial*. 2. Aufl. Cambridge, MA, USA: MIT Press.
- Smith, Brian Cantwell (1996). *On the origin of objects*. Cambridge, MA, USA: MIT Press.
- Song, Hyunjin, Petro Tolochko, Jakob-Moritz Eberl et al. (2020). „In Validations We Trust? The Impact of Imperfect Human Annotations as a Gold Standard on the Quality of Validation of Automated Content Analysis“. In: *Political Communication* 00.00, S. 1–23.
- Springer, Nina, Friederike Koschel, Andreas Fahr und Heinz Pürer (2015). „Techniken empirischer Sozialforschung“. In: *Empirische Methoden der Kommunikationswissenschaft*. UTB GmbH, S. 37–135.
- Stenetorp, Pontus, Sampo Pyysalo, Goran Topic et al. (2012). „BRAT: a Web-based Tool for NLP-Assisted Text Annotation“. In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France: Association for Computational Linguistics, S. 102–107.
- Stodden, Victoria (2010). „The Scientific Method in Practice: Reproducibility in the Computational Sciences“. In: *SSRN Electronic Journal*. MIT Sloan School Working Paper.
- Stodden, Victoria und Sheila Miguez (2013). „Best Practices for Computational Science: Software Infrastructure and Environments for Reproducible and Extensible Research“. In: *SSRN Electronic Journal* August 2012, S. 1–9.
- Stol, Klaas-Jan und Brian Fitzgerald (2015). „Theory-oriented software engineering“. In: *Science of Computer Programming* 101, S. 79–98.
- Stoll, Anke (2020). „Supervised Machine Learning mit Nutzergenerierten Inhalten: Oversampling für nicht balancierte Trainingsdaten“. In: *Publizistik* 65.2, S. 233–251.
- Stoll, Anke, Marc Ziegele und Oliver Quiring (2020). „Detecting Impoliteness and Incivility in Online Discussions: Classification Approaches for German User Comments“. In: *Computational Communication Research* 2.1, S. 109–134.
- Strippel, Christian, Annekatrin Bock, Christian Katzenbach et al. (2018). „Die Zukunft der Kommunikationswissenschaft ist schon da, sie ist nur ungleich verteilt“. In: *Publizistik* 63.1, S. 11–27.

- Strübing, Jörg, Stefan Hirschauer, Ruth Ayaß, Uwe Krähnke und Thomas Scheffer (2018). „Gütekriterien qualitativer Sozialforschung. Ein Diskussionsanstoß“. In: *Zeitschrift für Soziologie* 47.2, S. 83–100.
- Sundararajan, Mukund, Ankur Taly und Qiqi Yan (2017). „Axiomatic attribution for deep networks“. In: *34th International Conference on Machine Learning, ICML 2017* 7, S. 5109–5118.
- Tan, Sarah, Rich Caruana, Giles Hooker und Yin Lou (2018). „Distill-and-Compare“. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: ACM, S. 303–310.
- Tedre, Matti (2011). „Computing as a science: A survey of competing viewpoints“. In: *Minds and Machines* 21.3, S. 361–387.
- Trilling, Damian und Jeroen G F Jonkman (2018). „Scaling up Content Analysis“. In: *Communication Methods and Measures* 12.2-3, S. 158–174.
- Turing, A. (1936). „On Computable Numbers, with an Application to the Entscheidungsproblem“. In: *Proceedings of the London Mathematical Society* 2.42, S. 230–265. URL: <https://www.abelard.org/turpap2/tp2-ie.asp>.
- Turing, Alan (1950). „Computing Machinery and Intelligence“. In: *Mind* LIX.236, S. 433–460.
- Turner, Ryan (2016). „A Model Explanation System: Latest Updates and Extensions“. In: *ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*. New York, NY, USA, S. 1–5.
- Ullman, Jeffrey D. (1995). „The role of theory today“. In: *ACM Computing Surveys (CSUR)* 27.1, S. 43–44.
- Uysal, Alper Kursat und Serkan Gunal (2014). „The impact of preprocessing on text classification“. In: *Information Processing & Management* 50.1, S. 104–112.
- Vilone, Giulia und Luca Longo (2020). „Explainable Artificial Intelligence: a Systematic Review“. In: *Information Fusion*. arXiv: 2006.00093.
- Vogelgesang, Jens und Michael Scharkow (2012). „Reliabilitätstests in Inhaltsanalysen“. In: *Publizistik* 57.3, S. 333–345.
- Von Welker, Martin und Uwe Matzat (2009). „Online-Forschung: Gegenstände, Entwicklung, Institutionalisierung und Ausdifferenzierung eines neuen Forschungszweiges“. In: *Sozialforschung im Internet*. Hrsg. von Nikolaus Jakob, Harald Schoen und Thomas Zerback. Wiesbaden: VS Verlag für Sozialwissenschaften, S. 33–47.
- Warfield, Douglas (2010). „IS/IT research: A research methodologies review“. In: *Journal of Theoretical and Applied Information Technology* 13.1, S. 28–35.
- Waseem, Zeerak und Dirk Hovy (2016). „Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter“. In: *Proceedings of the NAACL Student Research Workshop*. Stroudsburg, PA, USA: Association for Computational Linguistics, S. 88–93.
- Wegner, P (1976). „Research paradigms in computer science“. In: *Proc. 2nd Int'l Conf. Software engineering*. Washington, DC, USA: IEEE Computer Society Press, S. 322–330.
- Wettstein, Martin (2016). „Verfahren zur computerunterstützten Inhaltsanalyse in der Kommunikationswissenschaft“. Dissertation. Universität Zürich, Switzerland.

- Wirth, Werner, Katharina Sommer, Martin Wettstein und Jörg Matthes, Hrsg. (2015). *Qualitätskriterien in der Inhaltsanalyse*. Köln, Germany: Herbert von Halem.
- Yeh, Chih-Kuan, Been Kim, Sercan O. Arik et al. (2019). „On Completeness-aware Concept-Based Explanations in Deep Neural Networks“. In: *CoRR*, S. 1–27.
- Zamith, Rodrigo und Seth C. Lewis (2015). „Content Analysis and the Algorithmic Coder: What Computational Social Science Means for Traditional Modes of Media Analysis“. In: *Annals of the American Academy of Political and Social Science* 659.1, S. 307–318.
- Ziegele, Marc, Johannes Daxenberger, Oliver Quiring und Iryna Gurevych (2018). „Developing Automated Measures to Predict Incivility in Public Online Discussions on the Facebook Sites of Established News Media“. In: *68th Annual Conference of the International Communication Association (ICA)*. Prague, Czech Republic.

Abbildungsverzeichnis

3.1	Forschungsprozess für die manuelle Inhaltsanalyse	23
3.2	Reliabilitätsmaß (Holsti, 1969)	27
3.3	Workflow für die automatische Textklassifikation	31
3.4	Forschungsprozess automatische Inhaltsanalyse, IST-Zustand	41
4.1	Warum-Frage	53
4.2	Beispiel LIME	58
4.3	Beispiel Shapley	59
4.4	Beispiel rationale	60
4.5	Beispiel Integrated Gradients	61
4.6	Beispiel VisBERT	62
4.7	Beispiel concepts	63
4.8	Irreführende Erklärung eines Neuronalen Netzes	65
4.9	Beispiel Naive Bayes	66
5.1	Forschungsprozess automatische Inhaltsanalyse, möglicher SOLL-Zustand	74

Tabellenverzeichnis

3.1	Maße für die Klassifikatorevaluation	33
4.1	Methoden für Erklärbarkeit und Interpretierbarkeit	56

Colophon

This thesis was typeset with $\text{\LaTeX}$ 2 ϵ . It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download the *Clean Thesis* style at <http://cleanthesis.der-ric.de/>.