Reihe „Methodeninnovationen in der Kommunikationswissenschaft“

Empirische Methoden gehören zu den basalen Werkzeugen der Wissenschaft, lassen sich jedoch nicht wie Rezepte anwenden. Vielmehr sind im empirischen Forschungsprozess im Hinblick auf Fragestellung und Gegenstand zahlreiche spezifizierte Entscheidungen zu treffen. Mitunter müssen dabei neue, innovative oder nur selten genutzte Wege beschritten werden, etwa in Bezug auf das übergeordnete Untersuchungsdesign, die Stichprobenbildung, Instrumentenentwicklung, Datenerhebung oder die Auswertung der Daten und Befunde. Methodische Herausforderungen ergeben sich auch für Untersuchungsgegenstände z. B. im Rahmen computervermittelter Kommunikation. Mit diesen Aspekten sollen sich die Beiträge der von Wiebke Loosen konzipierten kontinuierlichen Reihe „Methodeninnovationen in der Kommunikationswissenschaft“ auseinandersetzen. Dabei kommen prinzipiell Beiträge aus allen Themenfeldern der Medien- und Kommunikationswissenschaft infrage, die methodisch innovative Vorgehensweisen in eigenen Studien methodologisch reflektieren, andere Studien einer „sekundäranalytischen Methodenreflexion“ unterziehen und deren innovativen bzw. richtunggebenden Charakter herausarbeiten oder die unabhängig von konkreten Einzelstudien theoretisch-methodologisch Methodenentwicklungen nachzeichnen und reflektieren.

Zur Verknüpfung manueller und automatischer Inhaltsanalyse durch maschinelles Lernen

Michael Scharkow

Obwohl die computergestützte Codierung eine Reihe von methodischen und forschungspraktischen Vorteilen bei der Analyse umfangreicher Textmengen bietet, haben sich automatische Verfahren der Textanalyse bislang nicht in der Kommunikationswissenschaft etablieren können. Die wenigen im Fach eingesetzten computergestützten Verfahren, bei denen die automatische Codierung auf Regel- und Wörterbuchmethoden basiert, sind an die klassische manuelle Inhaltsanalyse wenig anschlussfähig. In diesem Beitrag soll das Verfahren der induktiven Textklassifikation vorgestellt werden. Sie bietet durch die Kombination von manuell codierten Texten mit dem Prinzip des maschinellen Lernens zahlreiche Vorteile für die hypothesengeleitete Analyse großer Textmengen. Der Beitrag beschreibt zunächst die Funktionsweise der induktiven Textklassifikation. Die Eignung dieses Verfahrens im kommunikationswissenschaftlichen Forschungsalltag wird anschließend anhand der Codierung von Online-Nachrichten illustriert.

Schlagwörter: Methoden, Textanalyse, computergestützte Verfahren, maschinelles Lernen, Bag-of-Words

1. Einführung

Das rasante Wachstum von digitalen (und digitalisierten) Online-Inhalten ist Chance und Herausforderung für die sozialwissenschaftliche Forschung. King (2011: 719) vergleicht die Verfügbarkeit von riesigen Textmengen im Internet-Zeitalter für unser Fach mit der Situation eines Biologen, der erstmals durch ein Mikroskop blickt. Die besondere methodische Herausforderung bei der Datenerhebung und -auswertung ist die wissen-

schaftliche Nutzbarmachung der schieren Menge von Untersuchungsobjekten zur Beantwortung aktueller Forschungsfragen, die sich über die systematische Analyse (halb-)öffentlicher Kommunikate im Internet beantworten lassen. Obwohl die Kommunikationswissenschaft mit der quantitativen Inhaltsanalyse über ein erprobtes Verfahren verfügt, aus Medieninhalten Inferenzschlüsse über gesellschaftliche Entwicklungen zu ziehen (Krippendorff 2004a), gerät dieses Verfahren bei einer groß angelegten Vermessung von Online-Inhalten an seine praktischen Grenzen. Manuelle Inhaltsanalysen lassen sich mit vertretbarem Aufwand nur für relativ kleine Stichproben bewerkstelligen, klassische computergestützte Verfahren wie die Diktionärcodierung (Stone et al. 1966) gelten als unattraktiv, umständlich sowie weniger valide und werden daher als alternative Verfahren häufig verworfen (Früh 2007; Rössler 2010b).

In diesem Beitrag soll mit der induktiven Textklassifikation ein innovatives und bislang kaum in den Sozialwissenschaften eingesetztes Verfahren der Textanalyse vorgestellt werden. Einführend wird im zweiten Abschnitt des Beitrags das Prinzip der automatischen Inhaltsanalyse beschrieben. Die Funktionsweise der induktiven Textklassifikation und die methodischen Vor- und Nachteile dieses Verfahrens werden im dritten Abschnitt erläutert. Im vierten Abschnitt werden Untersuchungsanlage und Ergebnisse einer Beispielstudie vorgestellt. Da die meisten Studien zur induktiven Textklassifikation bislang nur mit englischsprachigen Inhalten durchgeführt wurden und darin aus kommunikationswissenschaftlicher Sicht eher irrelevante Kategorien untersucht worden sind, wurden im Rahmen dieser Fallstudie deutschsprachige Online-Nachrichten mittels bereits erprobter Codebücher analysiert. Im fünften und letzten Abschnitt geht es um die Frage der Anschlussfähigkeit des vorgestellten Verfahrens an die klassische codebuchbasierte Inhaltsanalyse.

2. Automatisierung von Inhaltsanalysen

2.1 Warum überhaupt automatisch codieren?

Die Verwendung automatischer Codierverfahren wird bislang zumeist im Zusammenhang mit Reliabilitäts- und Effizienzargumenten diskutiert (King & Lowe 2003; Krippendorff 2004a). Tatsächlich lässt sich jedoch zeigen, dass die große Anzahl realisierbarer Codierungen, die durch eine Automatisierung ermöglicht wird, zentral für die Validität der Analyse und die Qualität der darauf basierenden Inferenzschlüsse ist. Um die methodologischen Implikationen der Codierquantität zu verdeutlichen, bietet sich ein Blick auf das Modell der stochastischen Textgenese und -codierung von Benoit et al. (2009) an, das in Abbildung 1 dargestellt ist. Dieses entspricht grundsätzlich dem klassischen Kommunikationsmodell, wie es bei Früh (2007: 43) oder Merten (1995: 16) dargestellt ist: Ziel der Inhaltsanalyse ist es, von codierbaren Texten auf latente, text-externe Merkmale ihres Entstehungs- oder Wirkungskontextes zu schließen. Während bei Früh und Merten der Begriff Inferenz relativ unspezifisch als „interpretativer Schluss" (Früh 2007: 44) behandelt wird, orientieren sich Benoit et al. (2009) explizit an den Prämissen *statistischer* Inferenz und der damit verbundenen *quantitativen* Messung latenter Phänomene, die stets mit Unsicherheit und Ungenauigkeit einhergeht.

Grundsätzlich gilt für jede quantitative Inhaltsanalyse, dass ein statistischer Inferenzschluss von einer Stichprobe an codierten Texten auf eine Grundgesamtheit latenter Kontextmerkmale gezogen wird. Dies gilt selbst für Vollerhebungen, da auch dort Daten durch einen stochastischen Prozess der Textgenese (jede intendierte Aussage π lässt sich auf viele Arten in einen Text τ überführen) und Textcodierung (der Text kann mit vielen

verschiedenen Instrumenten *I* in verschiedenen Codierprozessen *C* codiert werden) generiert werden (Benoit et al. 2009: 498; vgl. auch Behnke 2005).

Aufgrund der Tatsache, dass zumeist eine relativ *kleine* Anzahl Dokumente von *wenigen* Codierern mittels nur *eines* Untersuchungsinstruments codiert wird, sind einerseits inhaltsanalytische Daten mit Messfehlern behaftet, andererseits darauf basierende Inferenzschlüsse mit stichprobenbedingter Unsicherheit verbunden. Obwohl der Inferenzschluss auf Urheber oder Rezipienten einer Mitteilung keineswegs nur von der Qualität und Quantität der Daten δ abhängt, sondern vor allem auch von der Theorie, mit der text-interne und text-externe Merkmale logisch verknüpft werden, muss es ein zentrales Anliegen der Methodenentwicklung sein, die Reichweite, Validität und Reliabilität inhaltsanalytischer Daten zu optimieren. Die Verwendung automatischer Codierverfahren kann dazu auf mehrere Arten beitragen, wie sich an Abbildung 1 zeigen lässt:

1. Durch die Verarbeitung einer größeren Dokumentenmenge erhöhen sich Reichweite und Genauigkeit der Inferenzschlüsse. Je mehr Texte τ letztlich codiert werden, desto kleiner ist der Standardschätzfehler, bezogen auf die eigentlich interessierende Grundgesamtheit an Aussagen π.
2. Die Verwendung automatischer Verfahren ermöglicht es, eine Vielzahl unterschiedlicher Operationalisierungsstrategien *I* umzusetzen (z. B. alternative Codieranweisungen oder Wörterbücher). Durch die Verwendung verschiedener automatischer und manueller Verfahren, d. h. durch eine Triangulation der Messung, lässt sich die Qualität der Analyse und die Validität der Daten steigern.[1]
3. Da bei der Verwendung automatischer Verfahren die Codierkosten nicht mit der Anzahl der Codierungen steigen, lassen sich durch deren gezielte Verwendung Ressourcen sparen, die sich an anderer Stelle besser einsetzen lassen. Wenn etwa „leichte" Variablen automatisch codiert werden, lassen sich „schwere" Variablen ggf. von mehreren Personen oder zu mehreren Zeitpunkten codieren. Diese mehrfache Codierung *C* kann in vielen Fällen die Reliabilität und Validität einer manuellen Inhaltsanalyse erheblich erhöhen.

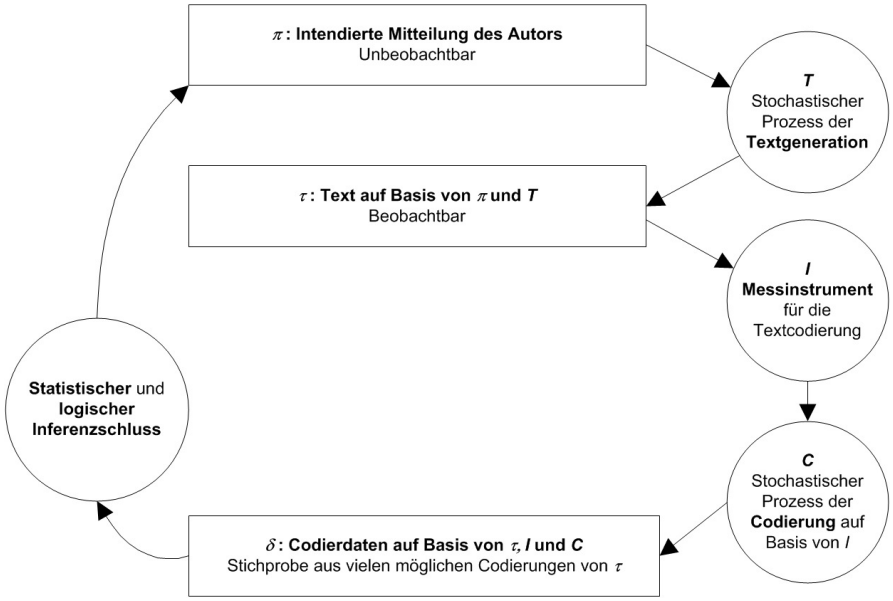
Zusammenfassend lässt sich sagen, dass die Quantität der Codierungen bei der quantitativen Inhaltsanalyse in mehrfacher Hinsicht relevant ist: So sind nicht nur viele Fragestellungen, etwa im Bereich der Online-Kommunikation, nur mit umfangreichen Stichproben sinnvoll zu beantworten.[2] Auch methodologisch ist es wünschenswert, die Zahl der Codierungen (nicht unbedingt der Codiereinheiten!) möglichst umfangreich zu gestalten, weil damit die Qualität der Analyse und letztlich der Inferenzschlüsse erhöht werden kann. Während sich bei begrenzten Ressourcen manuelle Inhaltsanalysen kaum beliebig erweitern lassen, versprechen automatische Codierverfahren beinahe unbegrenzte Skalierbarkeit. Es ist allerdings eine offene Frage, welches computergestützte Verfahren dieses Versprechen auch einlösen kann.

---

1 Die psychologische Testtheorie zeigt, dass sich komplexe Konstrukte mit mehreren Items valider messen lassen als mit einer einzigen Frage (Nunally & Bernstein 1978). Diese Überlegungen wurden bereits von Weber (1983) auf die Inhaltsanalyse übertragen, bislang aber kaum empirisch umgesetzt.

2 So wird in der Rezeptionsforschung schon seit vielen Jahrzehnten die Verknüpfung von Befragung und Inhaltsanalyse angewandt, bei der den Befragten individuell genutzte Inhalte zugeordnet werden (Wolling 2002). Angesichts der vielbeschworenen Fragmentierung des Publikums ist es wahrscheinlich, dass die Menge individuell rezipierter Medieninhalte sich nur noch automatisch adäquat analysieren lässt.

Abbildung 1: Modell der stochastischen Textgenese und -codierung



Vereinfachte und übersetzte Darstellung nach Benoit et al. 2009: 498.

2.2 Von der Diktionärcodierung zur induktiven Textklassifikation

In der Geschichte automatischer Inhaltsanalysen dominieren bislang vor allem zwei Ansätze: Die vollautomatische Co-Occurrence-Analyse und die Diktionärcodierung (Züll & Alexa 2001). Diese lassen sich als Prototypen eines explorativen bzw. deduktiven Vorgehens ansehen (vgl. hierzu ausführlich Scharkow 2010). Da es im vorliegenden Beitrag um die hypothesengeleitete Codierung geht, werden vollautomatische Ansätze, die fast immer auf einer explorativen faktor- oder clusteranalytischen Verfahrenslogik beruhen, an dieser Stelle nicht weiter thematisiert. Zudem konzentriert sich die folgende Darstellung auf statistische, sog. *Bag-of-Words*-Ansätze, bei denen auf rein lexikalischer Ebene codiert wird, d. h. Syntax und Semantik keine Berücksichtigung finden. Obwohl es auch auf dem Gebiet der eher linguistisch motivierten Ansätze in neuerer Zeit erhebliche Weiterentwicklungen gab (vgl. ausführlich Atteveldt 2008), unterscheiden sich diese sowohl in der Verfahrenslogik als auch in der Zielsetzung, nämlich der automatischen Rekonstruktion von Aussagen, erheblich von Bag-of-Words-Verfahren. Diese wortbasierten Ansätze sind untrennbar mit der *thematischen* Inhaltsanalyse nach Roberts (2000: 263) verbunden, deren prominenteste, aber keineswegs einzige Variante die Themenfrequenzanalyse ist (Früh 2007: 147ff.). Bei dieser Form der Analyse wird für jede zuvor definierte Codiereinheit (Beitrag, Satz, Aussage) das Vorkommen eines Themas (bzw. einer Kategorie) codiert, zumeist als dichotome oder ordinale Variable. Im Gegensatz zur semantischen Codierung werden dabei alle Kategorien unabhängig voneinander codiert, d. h. es geht nicht um komplexe, mehrteilige Subjekt-Verb-Objekt-Beziehungen. Verwendet man als Indikatoren für das Vorkommen einer Kategorie ein Schlüsselwort oder eine Wörterliste, spricht man von einer diktionärbasierten Codierung.

Seit den 1960er Jahren ist die computergestützte Diktionärcodierung nicht nur das wichtigste Verfahren der automatischen Textanalyse, vielfach wird sie sogar synonym mit dem Begriff der computergestützten Codierung verwendet (Früh 2007; Rössler 2010b). Da deren Vor- und Nachteile im Vergleich zur manuellen Codierung vielfach diskutiert worden sind, etwa bei Merten (1995: 339ff.) oder Früh (2007: 286ff.), soll an dieser Stelle vor allem auf zwei Probleme der Diktionärcodierung eingegangen werden, die sich durch die Verwendung von Verfahren aus dem maschinellen Lernen beheben lassen. Auch wenn dadurch nicht alle Nachteile von automatischen wortbasierten Analysen verschwinden, werden doch manuelle und computergestützte Verfahren der thematischen Inhaltsanalyse näher zueinandergeführt.

Problem 1: Einzelwortbasierte Kategorien

Bei der klassischen Diktionärcodierung werden Kategorien streng wortorientiert gebildet, d. h. eine Kategorie *c* wird dann vergeben, wenn Wort *w* in der Codiereinheit (zumeist Sätze oder ganze Beiträge) enthalten ist. Dies erfordert eine relativ umfangreiche Texttheorie bei der Operationalisierung, da in den meisten Fällen die Kategorien nicht anhand einzelner Schlüsselwörter, sondern umfangreicher Wortlisten definiert werden. Um ein komplexes Konstrukt in ein Wörterbuch zu überführen, bedarf es einer vollständigen Theorie der themenbezogenen Textgenese. Jedes Wort muss eindeutig zu einer Kategorie zugeordnet werden, wobei gewährleistet sein muss, dass möglichst wenige Fehlklassifikationen entstehen. Die Validität der Codierung steht und fällt mit der Regeldefinition, die nicht nur äußerst aufwendig ist, sondern bei komplexeren Codierungen versagen kann. Das *Wörterbuch Umweltschutz* von Schönbach (1982) mit 245 Indikatorwörtern für die Kategorie *Umweltschutzthema* zeigt, dass schon relativ einfache Kategorien operationale „Verrenkungen" erfordern können. So musste Schönbach beispielsweise die Wörter *Umwelt* und *Abfall* wegen mangelnder Trennschärfe wieder aus der Kategoriendefinition entfernen, um die Anzahl falsch positiv codierter Dokumente zu senken, in denen die Schlüsselbegriffe im falschen Kontext auftauchten.

Problem 2: Streng deterministische Codierung

Das Problem der deduktiven wortbasierten Inhaltsanalyse wird zusätzlich verschärft durch die zumeist streng deterministische Codierlogik der Software. Eine Kategorie wird mit 1 codiert, sobald eine bestimmte Anzahl von Wörtern aus der Liste in der Untersuchungseinheit gefunden wurde, ggf. auch schon dann, wenn ein beliebiges Schlüsselwort vorkommt. Obwohl es plausibel ist, dass unterschiedliche Schlüsselbegriffe ein unterschiedliches Gewicht bei der Entscheidung für oder gegen eine Kategorie haben, werden in den meisten Fällen ungewichtete Wortlisten verwendet, weil die Gewichtung eine noch elaboriertere Texttheorie erfordern würde und praktisch äußerst aufwendig wäre. Dies ist auch der Grund dafür, dass bei der klassischen Diktionärcodierung selten Negativ-Schlüsselwörter definiert werden, die explizit gegen die Vergabe einer Kategorie sprächen. Angesichts der Unschärfe und Ambiguität von Sprache scheint eine probabilistische Codierung bei der wortbasierten Textanalyse deutlich sinnvoller als eine deterministische.

Die streng deduktive Operationalisierung und die deterministische Codierung sind zwei Ursachen dafür, dass diktionärbasierte Verfahren oft nicht als sinnvolle Alternative oder Ergänzung zu manuellen Analysen betrachtet werden. In den meisten Fällen wird es leichter sein, eine gut funktionierende Codieranweisung für Menschen zu schreiben als eine valide Wörterliste bzw. komplexe Codierregeln. Diktionäre sind hochgradig

themen- und sprachspezifisch, und vielfach liegen im Gegensatz zu Codebüchern in der Regel keine vorgefertigten Wortlisten vor, auf die man zurückgreifen könnte. Die Entwicklung eines Diktionärs für eine Kategorie beansprucht daher nicht selten so viel Zeit wie die manuelle Codierung dieser Kategorie (Rössler 2010b: 191), womit das Effizienzargument automatischer Verfahren entfällt.

Die überwachte Textanalyse, die ursprünglich im Bereich des statistischen maschinellen Lernens beheimatet ist (Sebastiani 2002), verbindet die Stärken der manuellen Inhaltsanalyse (Flexibilität bei der Operationalisierung, Validität, Anschlussfähigkeit) und der automatischen Diktionärcodierung (Reproduzierbarkeit, Effizienz) miteinander. Wie bei den Wörterbuchverfahren eignet sich der Ansatz nur für die thematische Textanalyse, und wie zuvor erfolgt die Codierung nur auf Wortebene, ohne dass syntaktische Informationen des Textes verwendet werden. Der zentrale Unterschied zwischen den Ansätzen liegt darin, dass im Gegensatz zu Wörterbuchverfahren bei der induktiven Textklassifikation die Wortliste der relevanten Indikatoren nicht vom Forscher a priori und streng deduktiv entwickelt wird, sondern induktiv anhand zuvor geleisteter manueller Codierungen. Durch dieses Vorgehen schließt die induktive Textklassifikation unmittelbar an die Praxis der konventionellen manuellen Inhaltsanalyse an und setzt diese mit einer probabilistischen automatischen Codierung neuer Texte fort. Bereits vorhandene manuelle Codierentscheidungen dienen dazu, Regeln für die Codierung neuer Dokumente zu generieren. Dieser Prozess bezeichnet das maschinelle Lernen. Die genaue Funktionsweise einer induktiven Textklassifikation und mögliche Trainingsstrategien werden im folgenden Abschnitt ausführlich erläutert.

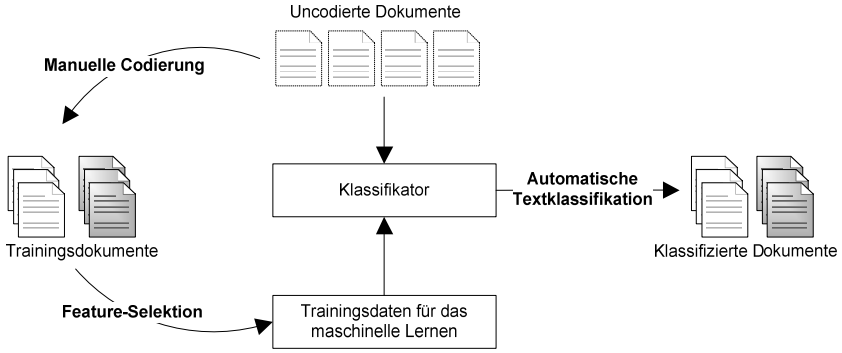
3. Inhaltsanalyse und maschinelles Lernen

3.1 Funktionsweise induktiver Textklassifikation

Die induktive Textklassifikation wendet Algorithmen des überwachten maschinellen Lernens (*supervised learning*) auf die Analyse von digitalen Texten an. Das Verfahren besteht dabei aus der induktiven Trainingsphase, in der das Computerprogramm aus manuell vorcodierten Beispieldokumenten ein statistisches Klassifikationsmodell entwickelt, und der eigentlichen Klassifikationsphase, bei der neue Dokumente automatisch einer Kategorie zugeordnet werden (vgl. Abbildung 2). Der Trainings- und Klassifikationsprozess schließt damit unmittelbar an die klassische manuelle Codierung an. Überwachte Verfahren der Textklassifikation machen sich auf diese Weise die Tatsache zunutze, dass die beispielhafte Codierung von ausgewählten Texten eigentlich immer deutlich weniger aufwendig ist als die Formulierung von komplexen Regelsätzen oder umfassenden Diktionären (Sebastiani 2002).

Forschungspraktisch bedeutet dieses Vorgehen, dass ein Algorithmus mit wenigen Texten und den Ergebnissen der manuellen (korrekten) Codierung zuerst trainiert wird und daraus mittels statistischer Verfahren ein „probabilistisches Diktionär" (Pennings & Keman 2002: 70) entsteht, das dann für alle nachfolgenden automatischen Klassifikationen genutzt wird. Dieses probabilistische Diktionär unterscheidet sich nicht nur in der Genese von deduktiv entwickelten Wörterbüchern, sondern auch dadurch, dass jedes Wort mit einem speziellen Gewicht in die Klassifikation eingeht. Während bei manuell erstellten Diktionären die Wörter als Indikatoren für eine Kategorie gleich gewichtet sind (z. B. kann sowohl das Wort „Fußball" als auch das Wort „Schiedsrichter" das Vorliegen einer Sportmeldung indizieren), wird bei der induktiven Textklassifikation für jedes Wort eine empirische Wahrscheinlichkeit ermittelt, wie stark es zur Unterscheidung der Klassifikationskategorien beiträgt. Die Software muss bei der Textklas-

Abbildung 2: Funktionsweise überwachter Textklassifikation



Eigene Darstellung nach Evans et al. 2007: 1011.

sifikation zwei unterschiedliche Aufgaben lösen, die die Klassifikationsqualität entscheidend beeinflussen: Feature-Selektion und statistische Klassifikation.

Feature-Selektion

Bei der Feature-Selektion geht es darum, aus unstrukturiertem Textmaterial möglichst relevante Elemente (*Features*) zu extrahieren, die dann in eine numerische Repräsentation des Textes umgewandelt werden. Die eigentlichen Klassifikationsalgorithmen basieren ausschließlich auf der statistischen Modellierung von Zahlencodes und sind daher unabhängig von der konkreten Beschaffenheit des Codiermaterials. Bei Textanalysen sind die Features zumeist Einzelwörter und Wortgruppen (N-Gramme); es sind jedoch auch kleinere (Zeichen) oder größere Einheiten (Sätze) denkbar. Induktive Klassifikationsverfahren lassen sich auch auf Audio, Bild- oder Videomaterial anwenden, wobei sich natürlich die extrahierten Features (Töne, Farben, Schnitte, etc.) unterscheiden (Bartlett et al. 1999; Radhakrishan et al. 2004; Scaringella et al. 2006).

Statistische Klassifikation

In den letzten Jahrzehnten wurde eine Vielzahl unterschiedlicher Klassifikationsverfahren für die Textanalyse eingesetzt (Manning & Schütze 1999; Sebastiani 2002), wobei bislang die als *Naive Bayes* und *Support Vector Machines* bezeichneten Verfahren die besten Ergebnisse liefern (Joachims 2002; Hillard et al. 2008). Da von beiden Verfahren der naive bayesianische Klassifikator weniger komplex ist, soll der Algorithmus nachfolgend kurz erläutert werden. *Naive Bayes*-Klassifikatoren zeichnen sich durch ihre Effizienz (Geschwindigkeit bei Training und Klassifikation) und Effektivität (hohe Genauigkeit) aus. Ihr Funktionsprinzip ist so simpel, dass ihre Funktionsweise auch für Forscher ohne statistisches Spezialwissen nachvollziehbar ist. Eine ausführliche Darstellung des Verfahrens findet sich bei Manning, Raghavan und Schütze (2008, Kap. 13). Für den generalisierbaren Fall eines binären Klassifikators muss für jedes Dokument d die Wahrscheinlichkeit berechnet werden, dass es zur Klasse c (oder zur Alternativklasse $\bar{c}$) gehört. Dieses Problem kann auf die Ebene einzelner Features (z. B. Wörter w) heruntergebrochen werden. Die Frage lautet dann: Wie wahrscheinlich gehört Dokument d zur Klasse c, wenn Wort w darin vorkommt? Dies ist im Prinzip auch die Frage, die

sich jeder Forscher bei der Konstruktion eines Diktionärs stellt. Mit Hilfe des Bayestheorems lässt sich diese Frage für jede Wort-Dokument-Kombination beantworten:

$$P(c \mid w) \;\alpha\; P(c)\, P(w \mid c) \qquad (1)$$

Hierbei bezeichnet $P(c|w)$ die bedingte Wahrscheinlichkeit für Klasse c gegeben Wort w. $P(c)$ ist die Priorwahrscheinlichkeit der Klasse, zumeist hergeleitet aus der relativen Häufigkeit der Klasse in den Trainingsdaten Nc/N. $P(w|c)$ ist die Wahrscheinlichkeit, dass das Wort w in Texten der Klasse c vorkommt. Auch letztere lässt sich aus den relativen Häufigkeiten der Trainingsdaten bestimmen. Verknüpft man nun die Wahrscheinlichkeiten aller n_{d} Wörter im Dokument, ergibt sich:

$$P(c \mid d) \;\alpha\; P(c) \prod_{k=1}^{n_d} P(w_k \mid c) \qquad (2)$$

Die Wahrscheinlichkeit, dass ein Dokument d zur Klasse c gehört, lässt sich aus der Priorwahrscheinlichkeit $P(c)$ und dem Produkt der bedingten Wahrscheinlichkeiten, dass jedes enthaltene Wort w in Dokumenten der Klasse vorkommt, bestimmen. Wie in Formel (2) zu erkennen ist, wird diese bayesianische Klassifikation deshalb als naiv bezeichnet, weil sie bei der Berechung des Produkts der Wahrscheinlichkeiten von der statistischen Unabhängigkeit der Wörter ausgeht. Obwohl diese Annahme bei natürlichsprachlichen Texten nicht zutrifft, sind naive bayesianische Klassifikatoren erstaunlich leistungsfähig. Die Unabhängigkeit der Wörter- bzw. Feature-Wahrscheinlichkeiten ist auch der Grund für die erstaunliche Effizienz des Verfahrens. Beim Training mit neuen Dokumenten müssen jeweils nur die einzelnen Wahrscheinlichkeiten für die vorkommenden Wörter $P(w|c)$ aktualisiert werden.

3.2 Trainingsstrategien für das maschinelle Lernen

Da die manuelle Codierung des Trainingsmaterials fast immer den Löwenanteil der verfügbaren Ressourcen bindet, ist es das Ziel der überwachten Klassifikation, mit möglichst wenig Trainingsdaten eine hohe Reliabilität und Validität der automatischen Codierung zu erzielen. Um dieses Ziel zu erreichen, sind mehrere Trainingsstrategien denkbar. Einfach und zuverlässig ist dabei das *blockweise* Lernen, bei dem alle Trainingsdaten dem Klassifikator in einem Zug vorgelegt werden: Der Klassifikator entwickelt das statistische Modell und ist dann einsatzbereit.

Liegen die Trainingsdaten nicht bereits komplett vor, bietet sich ein *inkrementeller* Trainingsprozess an, bei dem der Klassifikator schrittweise mit dem Trainingsmaterial und den Ergebnissen der manuellen Codierungen konfrontiert wird. Nach einer bestimmten Anzahl an Trainingseinheiten wird dann ein Evaluationszwischenschritt durchgeführt um zu überprüfen, ob bereits eine akzeptable Klassifikationsqualität erreicht worden ist. Auf diese Weise werden nicht mehr manuelle Codierungen durchgeführt als nötig. Ein zweiter Vorteil inkrementellen Trainings liegt in der Möglichkeit, bei kontinuierlichen Analysen auf Themen- oder Kategorienveränderungen reagieren zu können, indem neues Trainingsmaterial manuell codiert und damit das Klassifikationsmodell aktualisiert wird.

Der inkrementelle Trainingsprozess ist – bezogen auf den lernenden Klassifikator – passiv: Der Algorithmus bekommt sein Lernmaterial zur Verfügung gestellt, ohne selbst darauf Einfluss zu nehmen. Dies ist vergleichsweise ineffektiv und ineffizient, weil nicht

in jedem Dokument gleich viele Informationen zur Verbesserung der Klassifikationsleistung enthalten sind. Trainiert man einen Klassifikator mit zufällig oder bewusst ausgewählten Dokumenten, ist die Wahrscheinlichkeit groß, dass die meisten von ihnen bereits vom Klassifikator valide eingeordnet werden können und daher keinen großen Einfluss auf die statistischen Parameter des Klassifikationsmodells haben. Auch bei der manuellen Codiererschulung werden deshalb selten Dokumente ausführlich besprochen, deren Kategorisierung offensichtlich ist. Vielmehr wenden sich die Codierer vor allem bei problematischen Dokumenten an den Untersuchungsleiter, um die korrekte Kategorisierung zu erfahren. Dieses Prinzip des *aktiven Lernens* ist bei induktiven Klassifikationsverfahren von großem Nutzen, weil es den Bedarf an vorcodierten Beispielen erheblich senken kann (Lewis & Gale 1994). Praktisch übernimmt hierbei der Klassifikator selbst die Aufgabe, optimales Trainingsmaterial zu suchen, das zuerst manuell codiert werden soll. Ausgewählt werden hierfür solche Dokumente, bei denen der erwartete Zugewinn an klassifikationsrelevanten Informationen besonders hoch ist. Beim *Uncertainty Sampling* macht man sich die Tatsache zunutze, dass für alle Klassifikationsentscheidungen ein quantifizierbares Maß an Unsicherheit angegeben wird. Aus einer Menge an Dokumenten mit unbekannter Kategorisierung wählt der Algorithmus demnach diejenigen aus, bei denen die Unsicherheit bezüglich der Klassenzugehörigkeit am größten ist.

3.3 Vor- und Nachteile des Verfahrens für die Inhaltsanalyse

Welche Vorteile sprechen nun für den Einsatz induktiver Klassifikationsverfahren bei der Analyse von Medieninhalten, welche Nachteile dagegen? Die hier vorgestellten Verfahren sind zunächst einmal themen- und sprachunabhängig, da sie lediglich auf den Textdaten und den Kategorien der manuellen Codierung basieren. Für die Verwendung des Klassifikationsalgorithmus ist es unerheblich, in welcher Sprache die Dokumente vorliegen, solange die Trainingsdaten konsistent sind. Dies ist ein entscheidender Unterschied zu deduktiven Verfahren, die zwangsläufig sprachspezifisch angepasst werden müssen, weil sich sowohl das Vokabular als auch die Syntax der analysierten Texte je nach Sprache unterscheiden. Da der Klassifikator naiv ist, d. h. lediglich versucht, eine möglichst optimale Regel für die Entscheidung zwischen mehreren möglichen Kategorien anhand des Trainingsmaterials zu finden, kann das Verfahren potenziell für jede geschlossene, manuell codierbare Variable angewandt werden. Ob sich die Klassifikation zuverlässig und valide automatisch umsetzen lässt, ist letztlich immer eine empirische Frage.

Ein weiterer Vorteil der überwachten induktiven Textklassifikation ist in der unmittelbaren Anschlussfähigkeit an die gängige manuelle Codierpraxis zu sehen. Jede manuell codierte Variable eines Codebuchs kann für das Training und den Test eines Klassifikators genutzt werden, sofern das Textmaterial digital vorliegt. So wird als Nebenprodukt der manuellen Codierung ein Klassifikator für künftige automatische Codierungen parallel entwickelt. Die überwachte Textklassifikation kann daher eher als Ergänzung und Erweiterung der klassischen manuellen Praxis der Inhaltsanalyse verstanden werden, während deduktiv-automatische Verfahren zumeist eine grundsätzliche Abkehr von der manuellen Analyse verlangen.

Schließlich versprechen induktive Verfahren eine größtmögliche Reproduzierbarkeit und Anschlussfähigkeit inhaltsanalytischer Forschungsprojekte: Das „Wissen" des Klassifikators befindet sich in einer einzelnen Datei, die zwischen Forschern ausgetauscht, getestet und weiterentwickelt werden kann. So können die Ergebnisse vorangegangener Analysen entweder direkt als Datei oder zumindest indirekt (durch Training

des Klassifikators mit den Original-Dokumenten und den entsprechenden Codesheets oder -dateien) weiterverwendet werden. Das Abspeichern des Klassifikatorwissens ermöglicht einen kumulativen Prozess inhaltsanalytischer Forschung.

Ein erster Nachteil induktiver Klassifikationsverfahren war bislang der hohe technischen Aufwand beim Einsatz der Klassifikationssoftware. Mit der Veröffentlichung freier, leicht bedienbarer Software (Jurka et al. 2011) ist dies mittlerweile kein Hindernis mehr. Die *technische* Möglichkeit, ein manuelles Kategoriensystem für die Automatisierung nutzbar zu machen, bedeutet jedoch keinesfalls, dass dies auch *inhaltlich* zuverlässig und valide gelingt, d. h. dass sich eine Variable für die induktive Klassifikation eignet. Der Computer verhält sich wie ein eifriger, aber unselbstständiger Codierer, der zwar schnell lernt, aber kein eigenes Wissen zur Anwendung bringen kann. Dies unterscheidet ihn von menschlichen Codierern. Insbesondere Kategorien, bei deren Codierung semantische Informationen jenseits des einzelnen Wortes verarbeitet werden müssen, werden sich nicht zuverlässig automatisieren lassen. Je mehr sich eine Codierung von der konkreten Textebene entfernt und auf die gelenkte Interpretation eines sprachlich und inhaltlich vorgebildeten Codierers verlässt, wie etwa bei der Codierung eines Nachrichtenfaktors wie „Überraschung" oder bei Bewertungsvariablen, desto unwahrscheinlicher ist es, dass die induktive Klassifikation gelingt. Diese Einschränkung gilt freilich für alle automatischen Verfahren. Da die überwachte Textklassifikation immer auf manuellen Codierungen beruht, kann die Reliabilität und Validität nie höher sein als die der Trainingsdaten. Zum Messfehler in der manuellen Codierung kommt noch der klassifikatorspezifische Messfehler hinzu, so dass man immer mit einem Qualitätsverlust in der automatischen Codierung rechnen muss.

Schließlich muss noch betont werden, dass sich die hier vorgestellten Klassifikationsverfahren nur für geschlossene, nominalskalierte Variablen eignen. Obwohl es auch Entwicklungen auf dem Gebiet ordinaler Klassifikationsalgorithmen gibt (Pang & Lee 2008), wird in der Forschungspraxis zumeist mit multinomialen bzw. mehreren binären Variablen gearbeitet. Die Ranginformationen bei mehrfach gestuften Kategorien gehen dabei verloren. Gänzlich ungeeignet sind überwachte Klassifikationsverfahren für (halb-)offene Codierungen, bei denen a priori unbekannte Informationen (Akteure, Orte etc.) aus dem Text extrahiert werden müssen. Auch ist es nicht ohne Weiteres möglich, Themendominanz statt bloßem Vorkommen zu codieren, weil dies implizit eine Segmentierung des Beitrags in thematische Abschnitte verlangt, die anschließend nach Häufigkeit oder Reihenfolge des Auftretens gewichtet werden. Da die Klassifikation immer auf allen Wörtern der Untersuchungseinheit basiert, kann hier nur An- und Abwesenheit einer Kategorie codiert werden. Ähnliches gilt bei der Codierung von Bewertungen auf Beitragsebene, die vom Codierer (a) die Identifikation einer oder mehrerer Aussagen, (b) deren Codierung und ggf. (c) eine Aggregation der Bewertungen erfordert. Dieser Nachteil lässt sich jedoch ggf. durch eine alternative Operationalisierungsstrategie umgehen, die am Ende des Beitrags noch diskutiert werden wird. Zuvor wird die Anwendung der überwachten Textklassifikation nachfolgend anhand einer Inhaltsanalyse von Online-Nachrichten illustriert.

4. Anwendungsbeispiel: Analyse von Online-Nachrichten

4.1 Untersuchungsanlage und Datengrundlage

Um die Anwendungsmöglichkeiten des maschinellen Lernens für die Inhaltsanalyse zu evaluieren, sind vor allem zwei zentrale Fragen zu beantworten:

1. Wie hoch ist die Klassifikationsqualität, d. h. wie hoch ist vor allem die Coder-Klassifikator-Übereinstimmung?
2. Wie effizient ist das maschinelle Lernen, d. h. wie viel Training ist für eine ausreichende Codierqualität notwendig?

Beide Fragen sind gleichermaßen wichtig für eine mögliche Entscheidung, maschinelles Lernen in der kommunikationswissenschaftlichen Forschungspraxis einzusetzen. Im Rahmen einer umfangreichen Inhaltsanalyse wurden die Angebote von zwölf großen deutschen Nachrichtenwebsites (Spiegel, Focus, Zeit, FAZ, SZ, FR, Welt, Bild, Tagesschau, Heute, Tagesspiegel, WAZ) analysiert. Die Grundgesamtheit besteht aus allen Artikeln, die in den RSS-Feeds der oben genannten Online-Angebote zwischen dem 1.6.2008 und dem 31.5.2009 verlinkt worden sind. Diese rund 208.000 Beiträge wurden auf Tagesbasis vollständig automatisch mit einem eigens programmierten Tool aus den RSS-Feeds extrahiert, heruntergeladen, bereinigt und in einer MySQL-Datenbank archiviert. Nach der Datenerhebung wurde per einfacher Zufallsauswahl eine Stichprobe von 1000 Artikeln gezogen, von denen nach Ausschluss problematischer Dokumente (Beiträge mit weniger als einem Satz Fließtext) für die eigentliche Analyse $N = 933$ Beiträge übrig blieben.[3]

Alle Beiträge der Stichprobe wurden zunächst manuell von sieben Personen codiert. Die Codierung erfolgte anhand bereits publizierter Codebücher. Konkret wurden sowohl klassische Themenvariablen (Politik, Sport, Kriminalität) als auch Nachrichtenfaktoren wie „Kontroverse" und „Prominenz" verwendet. Die Themenvariablen sind dabei dichotom, die beiden Nachrichtenfaktoren dreistufig ordinal skaliert, wobei der Klassifikator die Ordinalität der Variablen nicht berücksichtigt, sondern deren Ausprägungen lediglich als unterschiedliche Kategorien behandelt. Um die externe Validität der Studie zu gewährleisten, wurden alle Definitionen und Codieranweisungen den Arbeiten von Bruns und Marcinkowski (1997), Fretwurst (2008) und GÖFAK Medienforschung (2010) entnommen. Nach zwei Codiererschulungen wurden die Dokumente per Zufall auf die Codierer verteilt, wobei rund ein Drittel aller Beiträge zwecks Reliabilitätsbestimmung mindestens zwei Personen zugeteilt wurde. Der Intercoder-Reliabilitätstest erfolgte also während der Feldarbeit. Für die überwachte Klassifikation wurde der Klassifikator OSBF-Lua von Assis (2006) ausgewählt, der im Kern ein *Naive-Bayes-Classifier* ist und der auf orthogonalisierten Bigrammen statt Einzelwörtern basiert, was die Klassifikationsqualität deutlich erhöht (Siefkes et al. 2004). OSBF-Lua ist als Open-Source-Software frei verfügbar und kann daher auch flexibel angepasst werden. Die Schnittstelle zum Klassifikator besteht nur aus den beiden Befehlen *train* und *classify*, mit denen Dokumente entsprechend weiterverarbeitet werden.

4.2 Qualität der überwachten Klassifikation

Die Qualität der Klassifikation lässt sich bei maschinellem Lernen durch *Train-Test*-Verfahren prüfen, bei denen der Klassifikator mit einer Auswahl an vorcodierten Dokumenten trainiert wird und damit anschließend selbst eine neue Dokumentenstichprobe klassifiziert. Die Ergebnisse der Klassifikation können im Anschluss mit den manuellen Codierungen verglichen werden. Die Ergebnisse eines solchen *Train-Test*-Durchgangs bilden die Grundlage für weitere statistische Analysen. Konkret besteht ein einzelner Evaluationsdurchgang für eine Variable V aus dem Codebuch aus folgenden Schritten:

3 Die ausgeschlossenen Beiträge bestanden zumeist aus einem einzelnen Link zu einem Audio- oder Video-Beitrag, etwa aus der Reihe kicker.tv bei Spiegel Online.

1. Für alle 933 Dokumente der Stichprobe wird der manuell vergebene Code für *V* bestimmt. Existieren aufgrund einer Mehrfachcodierung mehrere Codes, wird aus der Menge der vorhandenen Codes per einfacher Zufallsauswahl ein Wert gezogen. Dies hat bei übereinstimmender Codierung keine Konsequenzen, bei Nichtübereinstimmung entspricht die Wahrscheinlichkeit, einen bestimmten Code *x* zu erhalten, der relativen Häufigkeit dieses Wertes in den Mehrfachcodierungen $P(x)$. Nach dieser Auswahl liegt eine Liste mit je einem als richtig definierten „kanonischen" Code pro Dokument vor.
2. Die Dokumentenstichprobe wird für die *10-fold Cross Validation* (Manning & Schütze 1999: 210ff.) zufällig in zehn Partitionen gleicher Größe aufgeteilt. Auch die Zusammensetzung dieser Partitionen variiert zwischen den einzelnen Evaluationsdurchgängen.
3. Eine Partition wird als Test-Set zurückgelegt, mit den Dokumenten und Codes der anderen neun Partitionen wird der Klassifikator trainiert. Anschließend klassifiziert der Algorithmus die Testdokumente. Die Ergebnisse werden im Anschluss an die Klassifikation mit den kanonischen manuellen Codes verglichen. Dieser Vorgang wird für alle Partitionen wiederholt, so dass am Ende für jedes Dokument ein Paar mit manuellen und automatisch vergebenen Codes existiert. Diese Daten werden für die Berechnung der Reliabilität der Klassifikation verwendet.

Durch den unvermeidlichen Messfehler bei der manuellen Codierung und die zufällige Zusammensetzung von Test- und Trainingsmaterial variieren die Ergebnisse der einzelnen Evaluationsdurchgänge. Um zuverlässige Inferenzschlüsse über die Klassifikationsqualität ziehen zu können und gleichzeitig den Rechenaufwand für die Replikationen zu minimieren, wurde jede Kreuzvalidierung vier Mal wiederholt.

Die Ergebnisse des Intercoder-Reliabilitätstests und der Klassifikatorevaluation sind in Tabelle 1 dargestellt. In der ersten Spalte sind die Häufigkeiten der Positiv-Kategorie dargestellt. So beinhalten 19 Prozent der Beiträge in der Stichprobe das Thema Bundespolitik. Die Intercoder-Reliabilität CR_{m} und α_{m} entspricht durchgehend den Werten vergleichbarer Studien, wobei die Reliabilität der Nachrichtenfaktorvariablen erwartungsgemäß niedriger ausfällt als die der Themenvariablen (vgl. Raupp & Vogelgesang 2009; Eilders et al. 2010). Insgesamt kann man die manuelle Codierung als gelungen bezeichnen, so dass das Trainings- und Testmaterial für das maschinelle Lernen zwar keineswegs optimal, aber von typischer Qualität ist.

Wie gut funktioniert die überwachte Textklassifikation? Betrachtet man zunächst die einfache Prozentübereinstimmung nach Holsti, fällt das Urteil überaus positiv aus. Im Durchschnitt ist die automatische Codierung durch den Klassifikator nur knapp sechs Prozent weniger zuverlässig als die manuelle Codierung. Diesen Qualitätsverlust könnte man bei einem Verfahren, dass nach der Trainingsphase vollautomatisch tausende Dokumente klassifiziert, billigend in Kauf nehmen, zumal die prozentuale Coder-Klassifikator-Übereinstimmung meist über den üblichen Daumenregeln liegt (Lauf 2001; Rössler 2010a). Dies ist aber nur die halbe Wahrheit: Betrachtet man den zufallskorrigierten Koeffizienten von Krippendorff (2004b), für dessen Anwendung nicht zuletzt die Tatsache spricht, dass er unabhängig von der Verteilung der Variablen zu interpretieren ist, fällt das Ergebnis deutlich negativer aus. Die automatische Klassifikation der Nachrichtenfaktoren genügt nicht den Qualitätsstandards der Lehrbücher. Die zufallskorrigierte Coder-Klassifikator-Reliabilität ist rund 20 Prozentpunkte niedriger als die Reliabilität nach Holsti. Einen derartigen Verlust an Codierqualität kann man sicher nur in den seltensten Fällen rechtfertigen.

Tabelle 1: Vergleich von Intercoder- und Klassifikationsreliabilität

		CR Holsti			*Krippendorffs α*		
	P	CR_m	CR_a	ΔCR	α_m	α_a	$\Delta\alpha$
Bundesdeutsche Politik	19	.90	.86	-.04	.69	.55	-.14
Internationale Politik	19	.93	.89	-.04	.76	.61	-.15
Sport	15	.99	.96	-.03	.98	.84	-.14
Kriminalität	16	.92	.86	-.06	.67	.36	-.31
Kontroverse	41	.69	.62	-.07	.49	.30	-.19
Prominenz	50	.71	.60	-.11	.72	.45	-.27

Die Subskripte a und m bezeichnen die automatische bzw. manuelle Codierung, P die relative Häufigkeit positiv codierter Dokumente.

Aus den Ergebnissen der Evaluation lassen sich noch zwei weitere interessante Befunde ablesen: Erstens besteht ein starker linearer Zusammenhang ($r > .8$) zwischen der manuellen und überwachten Codierqualität. Zwar kann letztere nie höher sein als erstere, jedoch zeigt sich, dass einfachere Variablen sich besser automatisieren lassen. Zweitens ist bei der Inspektion der Klassifikationsmatrizen eine systematische Fehlklassifikation der Dokumente zu erkennen – der Anteil falsch negativ codierter Beiträge (*Recall*) ist zumeist deutlich höher als der Anteil falsch positiver Codierungen (*Precision*). Beispielsweise hat der Klassifikationsalgorithmus nur drei Prozent der Beiträge abweichend von den Codierern als internationale Politik klassifiziert, aber acht Prozent der von den Codierern entsprechend zugeordneten Dokumente übersehen.

Wie ist nun die Klassifikationsqualität des maschinellen Lernens einzuschätzen? Diese Frage kann nur mit Blick auf die manuelle Codierung beantwortet werden. Bezogen auf die einfache Prozentübereinstimmung ist der Qualitätsverlust relativ gering. Bei Sportmeldungen kann man davon ausgehen, dass der Computer über 95 Prozent aller Dokumente korrekt klassifiziert. Ist man in nachfolgenden Analysen darauf angewiesen, dass alle Variablenausprägungen gleich zuverlässig codiert werden, zeigt ein Blick auf die zufallskorrigierten Reliabilitäten, dass die überwachte Klassifikation nur in den seltensten Fällen die Lehrbuchanforderungen von Krippendorff (2004b) erfüllt. Insbesondere die automatische Codierung der Nachrichtenfaktoren oder des Themas Kriminalität sind als nicht gelungen zu bezeichnen. Ein möglicher Grund hierfür liegt im unzureichenden Trainingsmaterial: Da der Computer keinerlei Kontextwissen darüber besitzt, welche Personen prominent oder welche Handlungen kriminell sind, muss erst eine große Menge an unterschiedlichem Trainingsmaterial vorgelegt werden, um ein zuverlässiges statistisches Modell der Klassifikation zu generieren. Hier zeigt sich deutlich, wie sehr man sich beim Einsatz menschlicher Codierer auf deren Vorwissen bezüglich der Kategorien des Codebuchs verlässt. Diese Möglichkeit hat man beim maschinellen

Lernen nicht, weshalb gerade semantische oder pragmatische (Bewertungs-)Variablen besonders schwer automatisierbar sind.[4]

4.3 Effizienz des maschinellen Lernens

Typischerweise lässt sich ein Lernprozess durch eine wiederholte Evaluation der Leistungen analysieren, dies gilt auch für das maschinelle Lernen. Um den Trainingsverlauf bei verschiedenen Lernstrategien zu modellieren, bietet sich eine Simulation dieses Prozesses an. Die experimentelle Evaluation des Lernprozesses verläuft in folgenden Schritten, die für jede Variable der Inhaltsanalyse zwanzigmal wiederholt werden:

1. Aus der Dokumentenstichprobe werden 233 zufällig ausgewählte Dokumente als Test-Set zurückgehalten, die anderen 700 Artikel werden wie zuvor zum Training verwendet. Der Klassifikator wird zu Beginn mit einem zufällig daraus ausgewählten Initial-Set von 50 Dokumenten trainiert.
2. Dem Klassifikator werden schrittweise 50 weitere Trainingsdokumente vorgelegt, wobei je nach Experimentalbedingung (a) alle Dokumente zufällig aus dem Trainings-Set ausgewählt werden (passives Lernen) oder (b) der Klassifikator aus den noch nicht verwendeten Dokumenten 50 auswählen kann, deren Codierung dann bekannt gemacht wird (aktives Lernen).
3. Nach jedem Trainingsschritt werden alle Dokumente des Test-Sets klassifiziert und die Evaluationsergebnisse gespeichert. Diese Klassifikationsentscheidungen fließen aber nicht in das Training des Klassifikators ein, so dass dieser die Testdokumente immer wieder als neu behandelt.

Die Ergebnisse dieser simulierten Trainingsprozesse sind in Abbildung 3 dargestellt. Die X-Achse stellt die Anzahl an Trainingsdokumenten, die Y-Achse die Klassifikationsreliabilität nach Krippendorff (2004b) dar.[5]

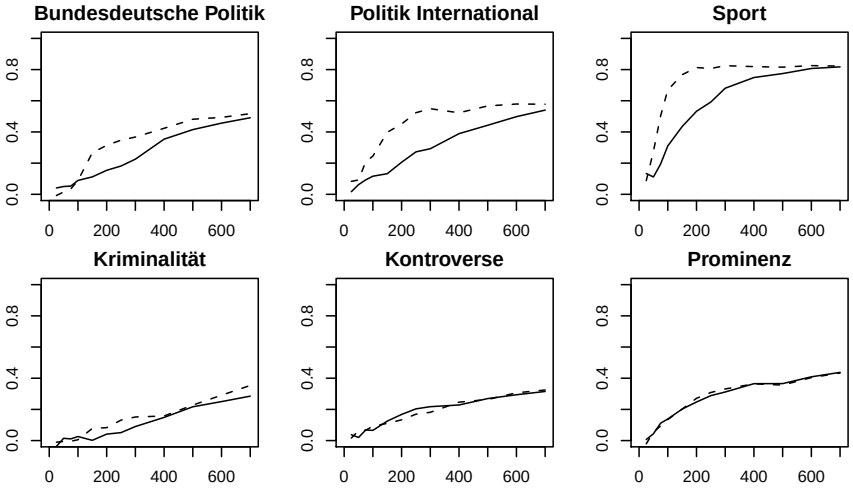
Auf den ersten Blick kann man erkennen, dass die Lernkurven je nach Kategorie sehr unterschiedlich ausfallen. Der Lernzuwachs bei den Politik- und Sportvariablen ist zu Beginn relativ stark, lässt dann aber nach. Die Nachrichtenfaktorvariablen Kriminalität und Kontroverse lernt der Klassifikator vergleichsweise langsam und eher linear. Hier scheint die Codierung einiger hundert weiterer Trainingsdokumente notwendig und sinnvoll. Insgesamt lässt sich festhalten, dass bei den Themenvariablen mindestens 400 Dokumente im passiven Trainingsmodus notwendig sind, um eine akzeptable Klassifikationsqualität zu gewährleisten. Danach sind zum Teil schon Sättigungseffekte zu erkennen.

Der Vergleich der Kurven von passivem und aktivem Lernen zeigt, das letztere Lernstrategie sich in jedem Fall für das inkrementelle Training anbietet. Der Effizienzgewinn bei Internationaler Politik und Sport ist so hoch, dass sich mit der Hälfte des Trainingsmaterials eine vergleichbare Klassifikationsqualität erreichen lässt. Zudem führt aktives Lernen in keinem Fall zu einer Verlangsamung des Trainingsprozesses. Im schlimmsten Fall, wie bei den Nachrichtenfaktoren, ist der Lernzuwachs genauso groß oder gering wie bei einer passiven Trainingsstrategie. Die Analyse des Lernverlaufs bei einer inkrementellen Trainingsstrategie kann wichtige Hinweise zur Optimierung der Feldarbeit liefern, weil relativ früh klar ist, wie sich die Klassifikationsqualität entwickeln wird.

4 Die zuverlässige Operationalisierung von semantischen oder pragmatischen Kategorien ist jedoch auch beim Einsatz menschlicher Codierer schwierig (vgl. Merten 1995).

5 Für die Analyse ist die einfache paarweise Übereinstimmung nach Holsti ungeeignet, da dieser Wert bereits nach dem ersten Training fast seine maximale Ausprägung erreichte, auch wenn das tatsächliche Lernen kaum stattgefunden hatte.

Abbildung 3: Verlauf der Klassifikationsreliabilität (Krippendorffs α) bei aktivem (gestrichelte Linie) und passivem Lernen (durchgezogene Linie)



Neben der generellen *Schwierigkeit* einer Variablen, die sich ggf. nicht ändern lässt, sind dafür vor allem zwei Faktoren maßgeblich: die Qualität und die Quantität des Trainingsmaterials. Die Qualität lässt sich leicht über die Intercoder-Reliabilität der manuellen Codierung bestimmen. Ist diese unzureichend, wird der Klassifikator ggf. mit widersprüchlichen Trainingsdaten konfrontiert, was die Entwicklung eines Klassifikationsmodells für die Kategorien erschwert. Gerade bei relativ schief verteilten Variablen ist jedoch die Quantität der Trainingsdaten für die Entwicklung der Klassifikationsqualität von größerer Bedeutung als deren Qualität (Sheng et al. 2008). Da gerade bei ungleichmäßig verteilten Variablen eine Vielzahl von Trainingsdokumenten nötig ist, um überhaupt eine Handvoll an Beispielen pro Kategorie zu erhalten, lohnt es sich, ggf. durch systematisch geschichtete Auswahl, den Lernprozess zu beschleunigen.

5. Diskussion und Ausblick

Die induktive Textklassifikation, wie sie in diesem Beitrag vorgestellt wurde, ist sicher kein Allheilmittel für jedes inhaltsanalytische Problem und macht andere automatische oder gar manuelle Ansätze nicht obsolet. Der besondere Reiz des maschinellen Lernens liegt zweifellos in der möglichen Kombination mit klassischen manuellen Inhaltsanalysen. Sobald Texte und manuelle Codierungen in digitaler Form vorliegen, lässt sich ein Klassifikator trainieren und testen, ohne dass weitere manuelle Arbeitsschritte nötig sind. Die empirische Frage, ob eine automatische Weitercodierung neuer Dokumente sinnvoll und erfolgsversprechend ist, lässt sich zudem anhand der oben vorgestellten Evaluationsverfahren prüfen, wobei die Qualitätskriterien direkt mit denen der manuellen Inhaltsanalyse vergleichbar sind. Auch in den Fällen, in denen die automatische Klassifikation keine ausreichende Qualität erreicht, kann die Nutzung eines lernenden Algorithmus wertvolle Hinweise für die Verbesserung des Codebuchs liefern. Weiterhin hat das Verfahren den Vorteil, dass umfangreiche manuelle Bereinigungsschritte, wie sie vielfach bei anderen automatischen Analysen notwendig sind (Hotho et al. 2005), ent-

fallen können, was eine weitestgehende Automatisierung des gesamten Forschungsprozesses ermöglicht. Im vorliegenden Beitrag wurde das Verfahren anhand einer klassischen Themenfrequenz- bzw. Nachrichtenfaktorstudie illustriert. Es zeigte sich, dass einige Variablen relativ schnell gelernt und zuverlässig codiert werden konnten, andere deutlich mehr Trainingsmaterial benötigten bzw. gar nicht zuverlässig automatisierbar waren. Eine Möglichkeit, die automatische Klassifikation zu verbessern, ist der Einsatz mehrerer unabhängiger Klassifikatoren (*Ensemble Classification*, Hillard et al. 2008) bzw. die Verknüpfung verschiedener deduktiver und induktiver Codierstrategien.

Für welche Forschungsfragen bietet sich das Verfahren nun überhaupt an? Kurz gesagt: für alle Inhaltsanalysen mit klar (formal) definierten Untersuchungseinheiten, in denen davon auszugehen ist, dass die für die Codierung erforderlichen Informationen auf der lexikalischen Ebene liegen. Dies bedeutet m. E. jedoch nicht, dass damit die vielen komplexen Kategorien der Frame- oder Bewertungsanalyse von vornherein unmöglich zu automatisieren sind. Man muss sich jedoch darüber im Klaren sein, wie viele Teilschritte man bislang von seinen Codierern verlangt, wenn diese etwa einen holistischen Frame oder eine komplexe Bewertung codieren müssen. Gerade wenn es um klassische Bewertungsvariablen oder Frame-Analysen geht, liegt die eigentliche Herausforderung nicht in der Codierung vorliegender Aussagen, sondern in deren Extraktion aus dem Text (vgl. Gerhards et al. 2007). Dies kann das vorgestellte Verfahren nicht leisten, wohl aber kann eine negative Bewertung von einer positiven unterschieden werden (Pang & Lee 2008). Um nun trotzdem eine Variable wie „Bewertung von Angela Merkel" automatisch codieren zu können, muss die Codieraufgabe in zwei Aufgaben geteilt werden: die Identifikation einer Texteinheit (z. B. einer Aussage), in der Angela Merkel vorkommt, und die Klassifikation dieser Texteinheit. Das erstgenannte Problem der semantischen, d. h. nicht formalen, *Unitization* ist bislang m. W. nicht automatisch zu lösen. Hier bleibt in vielen Fällen nur die manuelle Codierung. Trotzdem sollte man sich gerade bei komplexen Codieranweisungen stets fragen, inwieweit sich diese in weniger komplexe Teile zerlegen lassen, um den (menschlichen oder softwarebasierten) Codierern die Arbeit zu erleichtern und so die Reliabilität der Codierung zu steigern, wie dies zuletzt Matthes und Kohring (2008) für die Frame-Analyse vorgeschlagen haben. Hier lässt sich ggf. gezielt auf die Vorteile der induktiven Klassifikation zurückgreifen.

Für die Kommunikationswissenschaft ist die Feature-Selektion, d. h. die Frage, welche Stimulusmerkmale überhaupt in Zahlencodes transformiert werden sollen, angesichts der Multimedialität von Online-Inhalten (Rössler 2010a) ein hochrelevantes Forschungsthema. Auch wenn die Codierung von Texten in absehbarer Zeit das Standardverfahren der Inhaltsanalyse bleiben wird, lässt sich durch die Berücksichtigung von zusätzlichen Informationen – etwa in Form audiovisuellen Zusatzmaterials, Metadaten zum Entstehungs- oder Verwendungskontext einer Mitteilung oder gestalterischen Merkmalen – die Quantität und Qualität manuell und automatisch codierbarer Informationen erheblich steigern. Beispielsweise nutzen lernende Spamfilter-Programme nicht nur den Text einer E-Mail zur Identifikation unerwünschter Mitteilungen, sondern auch den Absender, das Sendedatum oder das verwendete Mailprogramm. Ebenso ist es vorstellbar, dass sich die Klassifikation einer Nachricht mit der Variablen „Prominenz" verbessern ließe, wenn mittels Bilderkennung begleitende Fotos mitklassifiziert würden. Die entsprechenden Gesichtserkennungsalgorithmen werden bereits praktisch von Firmen wie Google, Apple oder Facebook genutzt, haben bislang aber kaum Eingang in die Kommunikationswissenschaft gefunden. Hier sind Methodenexperimente gefragt, in denen systematisch die für die Klassifikation relevanten Merkmale von Medieninhalten untersucht werden.

Um dem stetig wachsenden Material an digitalen Medieninhalten gerecht zu werden, ist eine Automatisierung nicht nur der Codierung, sondern möglichst vieler Schritte im Forschungsprozess der Inhaltsanalyse notwendig – von der Datenerhebung bis zur statistischen Analyse. Das maschinelle Lernen bietet hier Lösungsmöglichkeiten für viele methodische Problemstellungen, auch und gerade dann, wenn man weiterhin auf menschlichen Sachverstand nicht verzichten kann und will.

Literatur

Assis, F. (2006): OSBF-Lua – A Text Classification Module for Lua: The Importance of the Training Method. *Proceedings of the 15th TREC*. Gaithersburg.

Atteveldt, W. v. (2008): *Semantic Network Analysis: Techniques for Extracting, Representing, and Querying Media Content*. Charleston: BookSurge Publishers.

Bartlett, M.; Hager, J.; Ekman, P. & Sejnowski, T. (1999): Measuring Facial Expressions by Computer Image Analysis. *Psychophysiology*, *36*(2), 253-263.

Behnke, J. (2005): Lassen sich Signifikanztests auf Vollerhebungen anwenden? Einige Anmerkungen. *Politische Vierteljahresschrift*, *46*(1), O1-O15.

Benoit, K.; Laver, M. & Mikhaylov, S. (2009): Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions. *American Journal of Political Science*, *53*(2), 495-513.

Bruns, T. & Marcinkowski, F. (1997): *Politische Information im Fernsehen*. Opladen: Leske + Budrich.

Eilders, C.; Geißler, S.; Hallermayer, M.; Noghero, M. & Schnurr, J.-M. (2010): Zivilgesellschaftliche Konstruktionen politischer Realität. Eine vergleichende Analyse zu Themen und Nachrichtenfaktoren in politischen Weblogs und professionellem Journalismus. *Medien und Kommunikationswissenschaft*, *58*(1), 46-62.

Evans, M.; McIntosh, W.; Lin, J. & Cates, C. (2007): Recounting the Courts? Applying Automated Content Analysis to Enhance Empirical Legal Research. *Journal of Empirical Legal Studies*, *4*(4), 1007-1039.

Fretwurst, B. (2008): *Nachrichten im Interesse der Zuschauer. Eine konzeptionelle und empirische Neubestimmung der Nachrichtenwerttheorie*. Konstanz: UVK.

Früh, W. (2007): *Inhaltsanalyse. Theorie und Praxis*. Konstanz: UVK.

Gerhards, J.; Offerhaus, A. & Roose, J. (2007): Die öffentliche Zuschreibung von Verantwortung. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *59*(1), 105-124.

GÖFAK Medienforschung (2010): *Fernsehanalyse zum Bundestagswahlkampf 2009. Methodenbericht GLES1401 der German Longitudinal Election Study.* www.gesis.org/fileadmin/upload/dienstleistung/forschungsdatenzentren/gles/SecureDownload/frageboegen/GLES1401-Pre1.0 %20-%20Methodenbericht.pdf [12.09.2011].

Hillard, D.; Purpura, S. & Wilkerson, J. (2008): Computer-Assisted Topic Classification for Mixed-Methods Social Science Research. *Journal of Information Technology & Politics*, *4*(4), 31-46.

Hotho, A.; Nürnberger, A. & Paaß, G. (2005): A Brief Survey of Text Mining. *LDV Forum*, *20*(1), 19-62.

Joachims, T. (2002): *Learning to Classify Text Using Support Vector Machines*. Boston: Kluwer Academic Publishers.

Jurka, T.; Collingwood, L.; Boydstun, A.; Grossman, E. & van Atteveld, W. (2011): RTextTools: A Supervised Learning Package for Text Classification. http://install.rtexttools.com/files/RTextTools_GettingStarted.pdf [12.09.2011].

King, G. & Lowe, W. (2003): An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design. *International Organization*, *57*(3), 617-642.

King, G. (2011): Ensuring the Data-Rich Future of the Social Sciences. *Science*, *331*(6018), 719-721.

Krippendorff, K. (2004a): *Content Analysis: An Introduction to its Methodology*, 2. ed. London: Sage.

Krippendorff, K. (2004b): Reliability in Content Analysis. Some Common Misconceptions and Recommendations. *Human Communication Research*, *30*(3), 411-433.

Lauf, E. (2001): „.96 nach Holsti". Zur Reliabilität von Inhaltsanalysen und deren Darstellung in kommunikationswissenschaftlichen Fachzeitschriften. *Publizistik*, *46*(1), 57-68.
Lewis, D. & Gale, W. (1994): A Sequential Algorithm for Training Text Classifiers. *Proceedings of the 17th ACM SIGIR* (pp. 3-12). Dublin.
Manning, C. D. & Schütze, H. (1999): *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.
Manning, C. D.; Raghavan, P. & Schütze, H. (2008): *Introduction to Information Retrieval*. New York: Cambridge University Press.
Matthes, J. & Kohring, M. (2008): The Content Analysis of Media Frames: Toward Improving Reliability and Validity. *Journal of Communication*, *58*(2), 258-279.
Merten, K. (1995): *Inhaltsanalyse: Einführung in die Theorie, Methode und Praxis*. Opladen: Westdeutscher Verlag.
Nunally, J. & Bernstein, I. (1978): *Psychometric Theory*. New York: McGraw-Hill.
Pang, B. & Lee, L. (2008): Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2), 1-135.
Pennings, P. & Keman, H. (2002): Towards a New Methodology of Estimating Party Policy Positions. *Quality and Quantity*, *36*(1), 55-79.
Radhakrishan, R.; Xiong, Z.; Divakaran, A. & Ishikawa, Y. (2004): Generation of Sports Highlights Using a Combination of Supervised & Unsupervised Learning in Audio Domain. *Proceedings of the International Conference on Pacific Rim Conference on Multimedia*, vol. 2 (pp. 935-939). Singapore.
Raupp, J. & Vogelgesang, J. (2009): *Medienresonanzanalyse: Eine Einführung in Theorie und Praxis*. Wiesbaden: VS Verlag.
Roberts, C. W. (2000). A Conceptual Framework for Quantitative Text Analysis. *Quality and Quantity*, *34*(3), 259-274.
Rössler, P. (2010a): Das Medium ist nicht die Botschaft. In: M. Welker & C. Wünsch (Hrsg.), *Die Online-Inhaltsanalyse* (S. 31-43). Köln: Halem.
Rössler, P. (2010b): *Inhaltsanalyse*. Konstanz: UVK.
Scaringella, N.; Zoia, G. & Mlynek, D. (2006): Automatic Genre Classification of Music Content: A Survey. *Signal Processing Magazine, IEEE*, *23*(2), 133-141.
Scharkow, M. (2010): Lesen und lesen lassen. Zum State of the Art automatischer Textanalyse. In: M. Welker & C. Wünsch (Hrsg.), *Die Online-Inhaltsanalyse* (S. 340-364). Köln: Halem.
Schönbach, K. (1982): The Issues of the Seventies. *Publizistik*, *27*(1-2), 129-140.
Sebastiani, F. (2002): Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, *34*(1), 1-47.
Sheng, V.; Provost, F. & Ipeirotis, P. (2008): Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. *Proceedings of the 14th ACM SIGKDD* (pp. 614–622). Las Vegas.
Siefkes, C.; Assis, F.; Chhabra, S. & Yerazunis, W. S. (2004): Combining Winnow and Orthogonal Sparse Bigrams for Incremental Spam Filtering. *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases* (pp. 410-421). Pisa.
Stone, P.; Dunphy, D.; Smith, M. & Ogilvie, D. (1966): *The General Inquirer: A Computer Approach to Content Analysis*. Cambridge: The MIT Press.
Weber, R. P. (1983): Measurement Models for Content Analysis. *Quality and Quantity*, *17*(2), 127-149.
Wolling, J. (2002): Methodenkombination in der Medienwirkungsforschung. Der Entscheidungsprozess bei der Verknüpfung von Umfrage- und Inhaltsanalysedaten. *ZUMA-Nachrichten*, 50, 54-85.
Züll, C. & Alexa, M. (2001): Automatisches Codieren von Textdaten. Ein Überblick über neue Entwicklungen. In: W. Wirth & E. Lauf (Hrsg.), *Inhaltsanalyse – Perspektiven, Probleme, Potenziale* (pp. 303-317). Köln: Halem.