universität wien

DIPLOMARBEIT

Titel der Diplomarbeit

Ansätze zur Automatisierung der Inhaltsanalyse

Verfasser
DI Paul Schneeweiß

angestrebter akademischer Grad

Magister der Sozial- und Wirtschaftswissenschaften (Mag. rer. soc. oec.)

Wien, 2012

Studienkennzahl lt. Studienblatt:	A 121
Studienrichtung lt. Studienblatt:	Diplomstudium Soziologie (sozial-/wirtschaftswissenschaftliche Studienrichtung)
Betreuerin / Betreuer:	Ao. Univ.-Prof. Dr. Josef Hörl

Paul Schneeweiß
Holochergasse 32/26
1150 Wien
E-Mail: paul.schneeweiss@gmx.at

„Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe."

Ort, Datum, Unterschrift

Abstract (Deutsch)

Die sozialwissenschaftliche Inhaltsanalyse bietet ein Bündel von erprobten Methoden zur Analyse von Kommunikationsinhalten. In dieser Arbeit werden einleitend die methodischen Vorzüge und Grenzen der Inhaltsanalyse erläutert, sowie ihre unterschiedlichen Typen beschrieben. Die Motivation der Arbeit besteht jedoch darin, die methodische Weiterentwicklung der Inhaltsanalyse, die durch Verfahren aus verschiedenen Disziplinen wie den Informationswissenschaften, der Statistik oder der Computerlinguistik unterstützt wird, aufzuzeigen. Ziel dieser Entwicklungen ist es, textuelle Inhalte automatisiert auszuwerten. In Anbetracht der immensen Mengen an Informationen und Meinungen, die praktisch ohne Zeitverzögerung über das Medium Internet ausgetauscht werden, würde eine automatisierte Inhaltsanalyse von Online-Inhalten ein enormes Potential für die Sozialwissenschaften bieten. Freilich ist eine Automatisierung mit großen Herausforderungen verbunden, die nur durch eine interdisziplinäre Zusammenarbeit und Kombination verschiedener Ansätze überwunden werden können. Kapitel 7 zeigt unterschiedliche Ansätze der automatisierten Inhaltsanalyse, deren Funktionsweise am Beispiel der Sentiment Analyse sowohl theoretisch, als auch im Rahmen empirischer Auswertungen erläutert wird. Ziel ist es dabei die grundsätzliche Funktionsweise der Methoden aufzuzeigen, anstatt komplexe Algorithmen mit Hilfe von Computerprogrammen zu entwickeln, die die Treffsicherheit der automatisierten Inhaltsanalyse optimieren. An Hand der praktischen Beispiele werden in einem abschließenden Kapitel Potentiale und Grenzen der automatisierten Inhaltsanalyse, sowie geeignete Anwendungsmöglichkeiten aufgezeigt.

Abstract (English)

The sociological content analysis provides a set of proven methods for analysis of textual content. In the first chapter of this paper the methodological strengths and limitations of content analysis are described, as well as the different types of content analysis.

The motivation of the work, however, is to point out the further development of content analysis, supported by different disciplines such as information science, statistics and computational linguistics. The aim of these developments is to automatize textual content analysis. Given the vast amounts of information and opinions, that are exchanged with no time delay through the Internet, an automated content analysis of online content would offer enormous potential for the social sciences.

Of course automated text analysis is combined with major challenges which can only be overcome through a combination of different approaches and interdisciplinary cooperation.

Chapter 7 shows different approaches of automated content analysis which functionality is shown by the example of sentiment analysis, theoretically, as well as by empirical analysis.

The goal is to demonstrate the basic functionality of the methods, rather than to develop complex algorithms with the help of computer programs that will optimize the accuracy of automated content analysis.

On the basis of practical examples in a concluding chapter potentials and limitations of automated content analysis, and appropriate applications are presented.

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Empirische Inhaltsanalyse

1.1 Einleitung und Definition

Versucht man die Anfänge der Inhaltsanalyse zu finden, so stoßt man auf die Namen Bernard Berelson und Harold D. Lasswell, die Untersuchungen zu Kriegsberichten während des Zweiten Weltkriegs unternahmen. Sie gelten als die Begründer der Inhaltsanalyse und Bernard Berelson veröffentlichte 1952 das erste Buch zur Inhaltsanalyse mit der mittlerweile klassischen Definition:

„Content analysis is a research technique for the objective, systematic, and quantitative description of the manifest content of communication" (Berelson, 1952)

Diese Definition wird heute noch gerne zitiert, muss jedoch auf Grund des Begriffs „manifest content" insbesondere für die vorliegende Arbeit, die die Online-Inhaltsanalyse als Schwerpunkt hat, erweitert werden. Auch die Möglichkeit der Inhaltsanalyse Bilder, Filme und nicht-gesprochene Kommunikation zu erfassen, erfordert eine Erweiterung Berelsons Definition. Mit den Begriffen „objective" und „systematic" unterstreicht Berelson die Gütekriterien Reliabilität und Validität, auf die in diesem Kapitel näher eingegangen wird. Berelsons Definition umfasst ausschließlich die quantitative Form der Inhaltsanalyse. Lasswell wiederum berücksichtigte bei seiner Analyse von Propagandamaterial aus dem Zweiten Weltkrieg auch qualitative Aspekte und verbesserte die Inhaltsanalyse, indem er mit seinem Team weitere grundlegende Methoden erarbeitete. Erst durch die Propagandaforschung im Zweiten Weltkrieg wurde die Inhaltsanalyse als richtige wissenschaftliche Methode anerkannt. Entscheidend bei dieser Untersuchung war, dass sie sich nicht mehr auf nur inhaltsinterne Merkmale konzentrierte, sondern die Wirkung des Inhalts auf den Adressaten in den Mittelpunkt brachte.

Durch die Verbreitung der elektronischen Datenverarbeitung ab den 50iger Jahren wurde die Verarbeitung von Massendaten ermöglicht, wodurch die Inhaltsanalyse noch attraktiver wurde.

Eine aktuelle Definition von Werner Früh bezieht sich ebenfalls auf die quantitative Inhaltsanalyse: "Die Inhaltsanalyse ist eine empirische Methode zur systematischen, intersubjektiv nachvollziehbaren Beschreibung inhaltlicher und formaler Merkmale von Mitteilungen; (häufig mit dem Ziel einer darauf gestützten interpretativen Inferenz)." (Früh 1998: 25). Früh unterscheidet die quantitative Inhaltsanalyse explizit von der hermeneutischer Textinterpretation und der linguistischen Textanalyse.

Während die Inhaltsanalyse für Berelsons eine reine Forschungstechnik war, ist sie heute zweifellos als eigenständige Methode anerkannt.

Bis heute wurde der methodische Ansatz erweitert und ausdifferenziert, sodass die Inhaltsanalyse in den unterschiedlichsten Wissenschaften, wie der Linguistik, Psychologie, Soziologie, Geisteswissenschaften und Kulturwissenschaften Einzug hielt.

1.2 Typen der Inhaltsanalyse

1.2.1 Frequenzanalyse

Gegenstand der Frequenzanalyse ist die Auszählung von Wörtern, wobei angenommen wird, dass sich aus der Häufigkeit die Wichtigkeit für den Sprecher oder Schreiber ableiten lässt. Dieser Zusammenhang mag in vielen, aber nicht in allen Fällen zutreffen und war daher stets Anstoß zur Kritik unter Sozialwissenschaftlern. Ein Beispiel für die Frequenzanalyse ist die Wertanalyse, bei der Werturteile in eine Tabelle eingetragen und gezählt werden. Der Zusammenhang zwischen den einzelnen Werturteilen kann anschließend analysiert werden. Die Frequenzanalyse war die erste Form der Inhaltsanalyse und ist typischerweise auch mit dem geringsten Aufwand durchzuführen. (Werner Faulstich, 1993)

1.2.2 Valenzanalyse

Die Valenzanalyse stellt eine Weiterentwicklung der Frequenzanalyse dar, indem neben der Aufzählung auch Bewertungen mitaufgenommen werden. So wird beispielsweise angegeben ob im untersuchten Material bestimmte Personen, Themen, etc. in einem positiven, neutralen oder negativen Zusammenhang auftreten. (Werner Faulstich, 1993)

1.2.3 Intensitätsanalyse

Die Intensitätsanalyse ist der Valenzanalyse ähnlich. Sie nimmt jedoch nicht nur die Bewertungen des Analysematerials auf, sondern auch deren Intensität, d.h. wie stark eine Wertung im Text zum Ausdruck kommt. (Werner Faulstich, 1993)

1.2.4 Kontingenzanalyse

Auf Grund der bereits erwähnten Kritikpunkte der Frequenzanalyse wurde das Verfahren weiterentwickelt, indem gezählt wird, wie oft Elemente im Zusammenhang mit anderen sprachlichen Elementen auftreten. Somit kann der Forscher Rückschlüsse auf Assoziationen zwischen zwei Ausdrücken ziehen. Beispielsweise können damit Aussagen danach differenziert werden, ob sie sich auf Männer oder Frauen, oder auf verschiedene Bevölkerungsschichten beziehen. Weiterentwickelte Formen der Kontingenzanalyse sind beispielsweise die Argumentationsanalyse oder die semantische Struktur- und Inhaltsanalyse. (Kromrey, 2006)

1.3 Ablauf der Inhaltsanalyse

1.3.1 Planung

In der Planungsphase der Inhaltsanalyse wird die Forschungsfrage definiert und festgelegt welche Aspekte der Realität für die Forschungsfrage relevant sind. Wie bei anderen empirischen Sozialforschungen ist auch die Relevanz der Forschungsfrage zu begründen. Zu Beginn werden außerdem Hypothesen formuliert, die Vermutungen über Zusammenhänge in der Problemstellung ausdrücken.

Auch die Analyseeinheiten, also die zu untersuchenden Texte oder Medien, sowie gegebenenfalls weitere Teil-Elemente werden in dieser Phase festgelegt. Ist die Grundgesamtheit sehr umfangreich, so wird eine Stichprobe ausgewählt. Die Entscheidung, ob eine Teil- oder Vollerhebung durchgeführt wird, ist oft auch von den zur Verfügung stehenden Ressourcen abhängig. (Rössler, 2005) Wird ein aktuelles Thema an Hand von Zeitungsartikeln analysiert, so ist durchaus eine Vollerhebung möglich. Bei der Analyse von Online-Inhalten wird auf Grund der unüberschaubaren Informationsmenge oft eine Teilerhebung durchgeführt. Wie auch bei anderen empirischen Methoden gibt es unterschiedliche Verfahren der Stichprobenziehung, die sich in die Wahrscheinlichkeitsauswahl, willkürliche und bewusste Auswahl einteilen

lassen. Für Inhaltsanalysen wird häufig die bewusste Auswahl eingesetzt, bei der die Auswahl nach nachvollziehbaren Regeln erfolgt. Bei der willkürlichen Auswahl ist dies nicht der Fall, weshalb von ihr auch abzuraten ist. Die Wahrscheinlichkeitsauswahl ist mit dem größten Aufwand verbunden, da dazu zuerst sehr viele Daten angesammelt werden müssen, aus denen anschließend gezogen wird. Deshalb wird gerne auf Quotenverfahren zurückgegriffen, bei der keine vollständige Liste der Grundgesamtheit erforderlich ist.

1.3.2 Entwicklung

Nachdem in der Planungsphase die Forschungsfrage und die Analyseeinheiten definiert wurden, wird in der Entwicklungsphase das Kategoriensystem, das Herzstück der Inhaltsanalyse, festgelegt.

Es werden strenge Forderungen an das Kategoriensystem gestellt. Die Kategorien werden aus theoretischen Annahmen abgeleitet und werden in ihrer Gesamtheit auch als Kategoriensystem bezeichnet. Für die Kategorien müssen Hypothesen formuliert werden, die auf ihren Wahrheitsgehalt überprüft werden können.

Das Kategoriensystem muss verschiedenen Kriterien entsprechen, um als solches für die Inhaltsanalyse zum Einsatz zu kommen, und es muss alle Aspekte des zu untersuchenden Materials abdecken und gleichzeitig eindeutig sein, d.h. es darf keine Überschneidungen der Ausprägungen geben. Je mehr Kategorien erstellt werden, d.h. je detaillierter die Kategorien definiert werden, desto schwieriger wird die eindeutige Zuordnung. Gleichzeitig sollte der Detailliertheitsgrad für die Forschungsfrage ausreichend sein. Weiters müssen die Ausprägungen einer Kategorie nach genau einer Dimension ausgerichtet sein und die Kategorien voneinander unabhängig sein, um statistische Auswertungen zu ermöglichen. Die Kategorienbildung hat insbesondere auf die Validität der Inhaltsanalyse Auswirkungen, da hier festgelegt wird was gemessen werden soll. (Atteslander, 2003)

Jede Kategorie entspricht einer Variable mit verschiedenen Merkmalsausprägungen, wobei man Variablen folgendermaßen kategorisieren kann (Diekmann, 2009):

- Variablen mit Ausprägungen
    - Diskret: festgelegte Anzahl verschiedener Werte
    - Dichotom: zwei Merkmalsausprägungen
    - Kontinuierlich: jeder beliebige Wert

- Nach Skalenniveau
    - Nominalskala: Ausprägungen ohne Rangordnung
    - Ordinalskala: Festlegung einer Rangordnung
    - Intervallskala: Die Abstände der Ausprägungen sind ident
    - Ratioskala: Es existiert ein absoluter Nullpunkt
- Nach der Position einer Hypothese
    - Unabhängige Variable
    - Abhängige Variable
- Nach Merkmalsebenen

Ausschlaggebend für die Reliabilität ist die exakte Definition von Kodierregeln, die festlegen, nach welchen Kriterien die Inhalte in Zahlen kodiert werden. Dieser Vorgang wird von Kodierern vorgenommen, bei denen es sich um Menschen oder auch Computer handeln kann.
Resultat der Entwicklungsphase ist ein Kodebuch, das die Verschlüsselung der Inhalte enthält und somit dem Kodierer als Werkzeug für die Kodierung dient.

1.3.3 Test

Vor der Kodierung müssen die Kodierer ausreichend geschult werden um eine hohe Reliabilität sicherzustellen. Weiters ist eine Testphase empfehlenswert, in der die Kodierer in einem Probelauf Inhalte kodieren, um bei Bedarf Überarbeitungen im Kategoriensystem vornehmen zu können. Bei der Probekodierung wird typischerweise eine kleinere Stichprobe als bei der Hauptuntersuchung gezogen, das Auswahlverfahren selbst bleibt jedoch gleich.
Sind die Kodierregeln missverständlich, so besteht die Gefahr, dass bei mehreren Kodierungsvorgängen desselben Materials unterschiedliche Resultate folgen. Es wird unterschieden zwischen Identifikationsreliabilität und Kodiererreliabilität. Bei der Kodiererreliabilität wird unterschieden zwischen Interkoderreliabilität, die die Unterschiede zwischen zumindest zwei Kodierern berücksichtigt und der Intrakoderreliabilität, die sich auf Unterschiede mehrerer Kodierungsdurchgänge von einer einzigen Person bezieht. (Atteslander, 2003)

Basierend auf den Erkenntnissen der Testphase wird das Kategoriensystem überarbeitet und im Kodebuch auf mögliche Schwierigkeiten hingewiesen. (Rössler, 2005)

1.3.4 Kodierung

Nachdem in der Testphase die Reliabilität des Messinstruments überprüft wurde und das Kategoriensystem, sowie Kodierbuch gegebenenfalls angepasst wurden, kann mit der eigentlichen Kodierung begonnen werden. Die Kodierung wird entweder von menschlichen Kodierern oder Computern vorgenommen. In beiden Fällen müssen die Kodierer auf die Kodierung durch Einschulung bzw. Programmierung vorbereitet werden. Die Einschulung soll garantieren, dass unterschiedliche Kodierer zum selben Ergebnis gelangen, also die Reliabilität gewährleistet werden kann. Außerdem werden den Kodierern Kodierungsanweisungen gegeben, sodass auch tatsächlich das gemessen wird, was der Forscher beabsichtigt zu messen. Wieviele Kodierer zum Einsatz kommen, ist oft eine Kostenfrage. Auf die Forschungsergebnisse wirkt sich eine Vielfalt an Kodierern positiv aus, da sich dadurch Fehlkodierungen weniger auswirken. Der Umfang der Einschulung ist auch von der Komplexität der Medien abhängig. Gerade die besonderen Eigenschaften von multimedialen und dynamischen Online-Inhalten stellen eine große Herausforderung für die Kodierung dar, wie in Kapitel 1.5 näher erläutert wird.

Nachdem die Einschulung abgeschlossen ist und die Kodierer mit den Kodebüchern vertraut gemacht wurden, wird das Untersuchungsmaterial verteilt bzw. der Zugang bei Online-Analysen ermöglicht. Auch während der Feldphase ist eine ständige Abstimmung unter den Kodierern, sowie Kontakt zum Forscher empfehlenswert. Fehlerquellen bei der Kodierung können beispielsweise Tippfehler, falsche Zuordnungen zu den Kategorien, Fehlinterpretationen, Ablenkungen, unterschiedliche intellektuelle Fähigkeiten der Kodierer oder Selektionseffekte sein. Auf Grund der vielfältigen Fehlerquellen ist eine sorgfältige Planung der Kodierung sehr wichtig. (Rössler, 2005)

1.3.5 Auswertung

Sofern die Daten nicht bereits während der Kodierung digital erfasst wurden, erfolgt dies im ersten Schritt der Auswertungsphase, bei der die Informationen der Kodierbögen in einem Statistikprogramm eingegeben werden. In manchen Fällen

werden Personen speziell zur Datenerfassung eingesetzt. Um Fehleingaben zu vermeiden, erfolgt eine anschließende Fehlerkontrolle, wofür ein Vergleich mit den Originalbögen erfolgt.
Bei vielen Auswertungen werden in der Statistiksoftware Rekodierungen der Variablen vorgenommen oder Variablen in neue Variablen transformiert. Durch Anwendung geeigneter statistischer Verfahren erfolgt die Überprüfung der in der Planungsphase aufgestellten Hypothesen.

1.4 Die computerunterstützte Inhaltsanalyse (CUI)

Bereits ab den ersten praxistauglichen Errungenschaften der Computerwissenschaften in den 50er- und 60er- Jahren, wurde begonnen Computer als Unterstützung bei großen Textmengen einzusetzen. Osgood und Lasswell diskutierten bereits in den 60er Jahren die möglichen Einsätze der computerunterstützten Inhaltsanalyse. Die ersten Programme, wie der General Inquire oder WORDS‘ wurden in diesen Jahren entwickelt. (Cornelia Züll, 2002)
Doch damals hatte man noch mit einigen Hürden zu kämpfen. Die Computer waren noch nicht leistungsfähig genug, um mit wirklich großen Datenmengen umzugehen bzw. zu teuer und unflexibel. Viele Programme waren für den Einsatz auf Großcomputer ausgelegt und waren sehr aufwendig zu bedienen. Auch gab es verglichen mit heute nur eine geringe Anzahl an Texten, die in für den Computer verständlicher digitaler Form vorlagen. Ein weiteres Problem war, dass es nur wenig Software gab, die sich für diesen Zweck eignete. All diese praktischen Hürden sind heute beseitigt und so lässt sich nach einigen Jahren der Stagnation auf diesem Gebiet ab den 90er Jahren ein Aufschwung der computerunterstützten Inhaltsanalyse feststellen. (Cornelia Züll, 2002) Computer sind leistungsstark und preislich erschwinglich und Software zur computerunterstützten Inhaltsanalyse wird von verschiedenen Herstellern angeboten. Die Menge an digitalem Text ist schier unüberschaubar. Tageszeitungen werden im Internet in digitaler Form angeboten und selbst ältere Bücher wurden eingescannt und durch Schrifterkennungssoftware digitalisiert und somit für den Computer auswertbar gemacht. War man früher auf Grund der vorhandenen Ressourcen auf einige Hundert oder Tausend Texte beschränkt, können durch Computerunterstützung heutzutage Zehntausende oder mehr Texte

problemlos analysiert werden. Doch freilich hat der Einsatz von Computern zur Automatisierung von Inhaltsanalysen mit einigen Herausforderungen zu kämpfen, die denen der konventionellen Inhaltsanalyse vor einigen Jahrzehnten ähneln. Man kann zwei Tendenzen erkennen, wie die Inhaltsanalyse durch Computer unterstützt wird. Eine Richtung legt Hauptaugenmerk auf komfortabel unterstützte Schnittstellen, an denen man Sprachkompetenz einschleusen kann. Dieser Ansatz ist eine Kombination maschineller und konventioneller Inhaltsanalyse, so wie es auch der Begriff „Computerunterstützte Inhaltsanalyse“ nahelegt. Eine weitere Richtung versucht den Kodierprozess vollständig zu automatisieren. Zwar gibt es auch Entwicklungen, bei denen versucht wird die Definition des Kategoriensystems zu automatisieren, doch ist die Aussagekraft der Ergebnisse unklar. (Früh, 2007)

Die Automatisierung des Kodierprozesses bietet verschiedene Vorteile: es entfällt die Kodierschulung und personelle Ressourcen können minimiert werden, sowie die Dauer des Kodierprozesses kann um ein Vielfaches reduziert werden. Die Kodierung ist immer nachvollziehbar, sofern nach programmierten Regeln vorgegangen wird, wodurch die Reliabilität der Kodierung immer bei 100% liegt. Dies geschieht jedoch zu Lasten der Validität. Die computerunterstützte Inhaltsanalyse beschränkte sich daher hauptsächlich auf syntaktisches Niveau - die semantische Ebene wird nicht miteinbezogen, was auch den Hauptkritikpunkt darstellt. Daher kam die computerunterstützte Inhaltsanalyse bislang bei einfachen Fragestellungen zum Einsatz, bei denen beispielsweise Worthäufigkeiten oder Häufigkeiten von Wortkombinationen gültige Indikatoren darstellen. Dass gerade in den letzten Jahren an dieser Limitierung der computerunterstützten Inhaltsanalyse durch interdisziplinäre Zusammenarbeiten geforscht wird, soll diese Arbeit aufzeigen.

Verschiedene Autoren empfehlen auf Grund der Validitätsprobleme vor oder nach der Erstellung des Kategoriensystems Expertenbefragungen durchzuführen. Auch wird empfohlen die Ergebnisse stichprobenartig von menschlichen Kodierern überprüfen zu lassen. Es zeigt sich also auch bei diesem Ansatz, der den Prozess der Kodierung automatisieren soll, dass es sich um eine Unterstützungsleistung handelt. Ob hier der Computer den Menschen unterstützt, oder umgekehrt der Computer durch den Menschen unterstützt wird, sei dahingestellt. Es kann jedoch davon ausgegangen werden, dass sich eine vollständige Automatisierung der Inhaltsanalyse nicht so bald

einstellen wird. Wie bereits erwähnt ist man von einer automatischen Generierung von Kategorien durch das Datenmaterial auf Grund von großer Validitätsprobleme noch weit entfernt. Auch soll an dieser Stelle noch einmal betont werden, dass die Analyse von textuellen Inhalten Schwerpunkt dieser Arbeit ist, da bei anderen, visuellen oder audiovisuellen Inhalten weitere Problemfelder hinzukommen. (Rüf, Böcking, & Kummer, 2010)

1.5 Reliabilität und Validität von Online-Inhaltsanalysen

Das Internet bietet mit seiner immensen Größe und vielfältigen Möglichkeiten der Meinungsäußerung großes Potential für sozialwissenschaftliche Untersuchungen. Der dazu verbreitetste Zugang für Untersuchungen textueller Inhalte im Internet ist die der Inhaltsanalyse. Die digitale Textform der Inhalte unterschiedlichster Themenbereiche stellen auf den ersten Blick hervorragende Voraussetzungen für Inhaltsanalysen dar, vergleicht man sie mit der vergleichsweise aufwändigen Analyse von Presseartikeln oder Nachrichtensendungen. Doch die vielfältigen multimedialen Möglichkeiten und die Flüchtigkeit der Inhalte im Internet stellen auch große Herausforderungen an die Gütekriterien wissenschaftlicher Arbeiten dar. Die Besonderheiten von Webinhalten müssen bei der Messmethode berücksichtigt werden. Zu diesen zählt u.a. die Nichtlinearität von Internettexten, womit die Verlinkung von Texten durch Hyperlinks gemeint ist. Herkömmliche Texte und Printmedien weisen diese Eigenschaft nicht auf und sind außerdem nicht durch eine Dynamik wie Internettexte gekennzeichnet.

Diese besonderen Merkmale von Online-Inhalten erfordern eine Reflektion traditioneller Inhaltsanalyseverfahren, um zu reliablen und validen Forschungsergebnissen zu kommen.

Im Folgenden wird erläutert welche Reliabilität und Validität von Online-Inhaltsanalysen zu erwarten ist und wie die speziellen Eigenschaften von Internettexten berücksichtigt werden können.

1.5.1 Reliabilität

Unter Reliabilität wird die Verlässlichkeit wissenschaftlicher Messungen verstanden. Hohe Reliabilität ist dann gegeben, wenn bei Wiederholung einer Messung unter gleichen Bedingungen das gleiche Ergebnis erzielt wird. Die Reliabilität kann verschiedene Stufen des Forschungsprozesses betreffen. Im Gegensatz zur

Inhaltsanalyse von Pressetexten können bei der Online-Inhaltsanalyse bereits auf den ersten Stufen Gefahren für die Reliabilität auftreten. Die besonderen Eigenschaften der Online-Inhaltsanalyse wurden bereits einleitend erläutert. Die Dynamik und Interaktivität des Untersuchungsmaterials stellen besondere Hürden für die Planung des Forschungsprozess dar. Während ein Pressetext als „physisch eindeutig identifizierbare Manifestation der sozialen Wirklichkeit“ betrachtet werden kann, sind Online-Inhalte zunehmend reaktiv. Sie werden bei modernen Internetseiten an den Benutzer individuell zugeschnitten, indem möglichst viele verfügbare Informationen wie aktueller Standort, Sprache, Surf-Verhalten und Internetbrowser-Software des Benutzers herangezogen werden. Diese Personalisierung, sowie die hohe Dynamik und Aktualität von Online-Inhalten sind ohne Zweifel ein großer technologischer Fortschritt, aber gleichzeitig auch eine neue Herausforderung für die Reliabilität der Inhaltsanalyse. Online-Inhalte können zu unvorhersehbaren Zeitpunkten aktualisiert werden oder gänzlich verschwinden. Um diese Probleme zumindest einzuschränken, sollen die Untersuchungseinheiten in einem möglichst engen Zeitfenster gesammelt werden und anschließend in einem Archiv abgelegt werden. Die Archivierungsmethode ist dabei von der Multimedialität und Interaktivität der Inhalte abhängig. Um die Transparenz der Untersuchung zu erhöhen, können auch Screenshots oder Videoaufzeichnungen angefertigt werden. Jedenfalls ist eine umfassende Dokumentation über die Identifikation der Untersuchungseinheiten, technische Randparameter und Archivierungsdatum der Seite anzufertigen. Bei der Online-Inhaltsanalyse ist bereits vor Beginn der Analyse auf eine präzise Definition der Grundgesamtheit und Untersuchungseinheiten zu achten, da auf Grund der Dynamik nachträglich nicht immer alle Inhalte zur Verfügung stehen. (Martin R. Herbers, 2010)

Eine Grundvoraussetzung für die Reliabilität der Online-Inhaltsanalyse ist die exakte Definition der Analyseeinheiten. Hier stellt die Nichtlinearität der Online-Inhalte die größte Herausforderung dar. Texte verweisen durch Hyperlinks auf andere Webangebote, wodurch die genaue Abgrenzung von Texteinheiten erschwert wird. Auch Hyperlinks innerhalb eines Textes müssen speziell behandelt werden. Die Bestimmung der Analyseeinheiten hat unmittelbare Auswirkung auf die Reliabilität, da die Kodierung bei unklaren Abgrenzungen nicht mehr vergleichbar ist. Wie später erläutert wird, hat die Identifikation der Analyseeinheiten auch Auswirkungen auf die Konstruktvalidität der Untersuchung.

Auf der Stufe der Instrumententwicklung hat man bei Online-Inhaltsanalysen besonders mit multimedialen Inhalten umzugehen, die Online-Inhalte nicht nur strukturell mit textuellen Inhalten verknüpfen, sondern sich auch inhaltlich aufeinander beziehen. Die verschiedenen Informationseinheiten richtig zu werten, stellt nach wie vor ein ungelöstes Problem dar. Die Kategorien des Kategorienschemas müssen wechselseitig exklusiv, voneinander unabhängig und eindeutig sein, um reliable und valide Messungen zu ermöglichen. Weitere Anforderungen an das Kategorienschema sind nach Merten Vollständigkeit und theoretische Ableitbarkeit. (Merten, 1995)

Beim Vorgang der Kodierung werden Merkmalsausprägungen zuvor definierten Kodes zugeordnet. Dies wird von Kodierern vorgenommen, die mit dem Kategoriensystem und deren Anwendung vor der Durchführung vertraut gemacht werden müssen. Dieser Schritt ist Voraussetzung für reliable Studienergebnisse und sollte auf Grund der speziellen Eigenschaften von Online-Inhaltsanalysen besonders umfangreich ausfallen. Bei der Kodierung tritt die einleitend erwähnte besondere Eigenschaft des reaktiven Inhalts auf, d.h. Online-Inhalte erscheinen nicht zwangsläufig für alle Kodierer in derselben Form. Dem kann gegengesteuert werden, indem alle Kodierer mit standardisierten technischen Voraussetzungen starten. Dazu zählen vergleichbare Hardwarevoraussetzungen und vor allem Softwareinstallationen- und Einstellungen. Diese sind präzise zu dokumentieren, um reproduzierbare Ergebnisse zu erzeugen.

Maßnahmen zur Reliabilität machen nur dann Sinn, wenn man die erreichte Reliabilität auch überprüfen kann. Bei Online-Inhaltsanalysen kann die Überprüfung der Reliabilität bereits vor der eigentlichen Studie in einer Testphase erfolgen, da hier die Herausforderungen besonders hoch sind. Für die Reliabilitätsstudie wird eine Stichprobe aus dem Analysematerial gezogen. Die Empfehlungen für die Größe der Stichprobe bewegen sich zwischen 50 und 300 Einheiten.

Die Reliabilitätsstudie sollte in zwei Stufen durchgeführt werden: Bestimmung der Identitfikationsreliabilität und Bestimmung der Kodierreliabilität. Die Kodierreliabilität sollte nur für solche Analyseeinheiten erhoben werden, die im ersten Schritt der Reliabilitätsstudie eindeutig identifiziert wurden. Für die Quantifizierung der Reliabilitätswerte stehen verschiedene Koeffizienten zur Verfügung, wobei davon keiner als Standard bezeichnet werden kann. Reliabilitätswerte ab 0,8 oder 0,9 gelten in der Regel als akzeptabel. Reliabilitätswerte werden für jede Variable einzeln berechnet

und ausgegeben. Um Transparenz zu gewährleisten sollte auch an dieser Stelle eine detaillierte Dokumentation der Reliabilitätsstudie angefertigt werden.
Die auf diese Weise zugesicherte Reliabilität hat wesentlichen Einfluss auf die Validität, deren Untersuchung im Folgenden erläutert wird. (Martin R. Herbers, 2010)

1.5.2 Validität

Das Gütekriterium Validität, das auch als Gültigkeit bezeichnet wird, gibt an ob auch tatsächlich das gemessen wurde was beabsichtigt war. Zwischen Reliabilität und Validität besteht insofern ein Zusammenhang, als ohne Reliabilität keine Validität gewährleistet werden kann. Die Validität wird durch das Untersuchungsdesign, das Messinstrument und die einzelnen Forschungsschritte beeinflusst. (Martin R. Herbers, 2010)
Man unterscheidet interne und externe Validität. Die interne Validität besagt, dass die abhängige Variable tatsächlich auf die Veränderung des Stimulus rückführbar ist und nicht auf eine mögliche Störvariable. Sie kann daher durch maximale Kontrolle erhöht werden, wodurch die Untersuchung jedoch schwieriger auf die Realität übertragbar ist und somit die externe Validität reduziert wird.
Weiters kann Validität in die drei Arten Inhaltsvalidität, Kriteriumsvalidität und Konstruktvalidität unterteilt werden. Die Inhaltsvalidität unterstellt einer Messung, dass das untersuchte Phänomen in allen Aspekten erfasst wird. Zur Ermittlung der Inhaltsvalidität werden Experten herangezogen. Die Kriteriumsvalidität (Übereinstimmungs- und Vorhersagevalidität) überprüft das Ergebnis, indem sie externe Kriterien heranzieht und es mit diesen vergleicht. (Ludwig-Mayerhofer, 2004) Damit ein Messinstrument im Sinne der Konstruktvalidität als valide eingestuft werden kann, müssen Zusammenhänge mit anderen vergleichbaren Konstrukten theoretisch und empirisch nachgewiesen werden können. (Martin R. Herbers, 2010)
Zur Sicherung der internen Validität ist eine Standardisierung der Erhebung notwendig. Gerade bei Online-Inhaltsanalysen ist die Ziehung einer Stichprobe erforderlich, da die Grundgesamtheit sehr umfangreich sein kann. Oft werden bei Online-Untersuchungen Stichprobenziehungen vorgenommen, die einfach zugängliche Elemente enthalten, wodurch die Gefahr der Bildung einer Extremgruppe gegeben ist. Eine zweite Ziehung würde somit ein anderes Ergebnis liefern, wodurch die externe und auch die interne Validität negativ beeinflusst werden. Ein Zufallsverfahren zur Stichprobenziehung wäre

auf jeden Fall vorzuziehen, wobei sich dieses auf Grund der schwer überblickbaren Grundgesamtheit und der dynamischen und multimedialen Online-Inhalten oft schwer realisieren lässt.

Auch das Definieren der Analyseeinheiten ist bei Online-Inhaltsanalysen nicht immer ohne weiteres möglich, da die Unterscheidung zwischen eigentlichem Inhalt und Kontext der Website nicht immer eindeutig ist und auch die Behandlung von Hyperlinks definiert werden muss. Eine Herausforderung besteht hier in der reinen Definition der Untersuchungseinheiten und eine weitere in der Automatisierung der Extraktion der Untersuchungseinheiten, falls dies bei der Inhaltsanalyse vorgesehen wird. Diese Schwierigkeiten wirken sich insbesondere auf die Inhaltsvalidität aus, die überprüft, ob die vom Messinstrument erfassten Inhalte diejenigen Inhalte darstellen, die in ihrer Vollständigkeit gemessen werden sollen. Für die Bewertung der Inhaltsvalidität muss die Analyseeinheit festgelegt werden. Wird die Analyseeinheit auf Wortebene festgelegt, sind beispielsweise validere Ergebnisse möglich als auf Bild-Ebene. Die Genauigkeit der Messung steigt also mit sinkender Komplexität der Analyseeinheit.

Die multimedialen Eigenschaften von Online-Inhalten erschweren auch die Kategorienbildung und können sich folglich auch auf die Konstruktvalidität auswirken. Während für eindeutige Zeichen ein Kategorienschema relativ problemlos definiert werden kann, gestaltet sich dies bei interaktiven und dynamischen Inhalten schwieriger.

Der Prozess der Kodierung wirkt sich ebenfalls auf die Validität der Forschungsergebnisse aus. Invalide Ergebnisse können entstehen, wenn die Kodierer nicht ausreichend eingeschult wurden bzw. ihre Auffassungen von der des Forschers abweichen. (Martin R. Herbers, 2010)

1.6 Anwendungen und Einschätzung der Inhaltsanalyse

Nachdem die Grundzüge der Inhaltsanalyse in diesem Kapitel erläutert wurden und dabei bereits einzelne Vorteile und Anwendungsgebiete genannt wurden, sollen diese abschließend noch einmal zusammengefasst werden und die Inhaltsanalyse als empirische Methode bewertet werden.

Die Inhaltsanalyse ermöglicht die Analyse von Aussagen von Personen, die nicht mehr erreichbar oder bereits verstorben sind, was vielfältige Anwendungen ermöglicht. Der

Forscher ist nicht auf die Kooperation von Versuchspersonen angewiesen. Dies sind wesentliche Vorteile gegenüber der Befragung, die diesbezüglich wesentlich mehr Aufwand erfordert und auch stärker vom Faktor Zeit abhängig ist (Brosius, Koschel, & Haas, 2008). Als non-reaktives Verfahren tritt man bei der Inhaltsanalyse nicht in Kontakt mit den Versuchspersonen und beeinflusst somit das Untersuchungsergebnis durch die Untersuchung nicht. Hat man das Untersuchungsmaterial einmal identifiziert und ist die Entwicklungsphase abgeschlossen, so lässt sich die Untersuchung beliebig oft reproduzieren. Somit können beispielsweise die Messinstrumente oder das Kategorienschema nachträglich angepasst werden, ohne das Untersuchungsmaterial zu verändern. Aus all diesen Vorteilen und der unkomplizierten Handhabung der Methode folgen auch vergleichsweise geringe Kosten für die Durchführung und das Personal. Herausforderungen der Inhaltsanalyse liegen ohne Zweifel in der Vorbereitungsphase. So werden hohe Ansprüche an das Verständnis und die Ziele des Forschers gestellt, sowie an das Kodierungssystem und die Kodiererschulungen. Fehler im Kodierungssystem schlagen sich insbesondere in der Validität nieder, während unklare Anweisungen und Missverständnisse bei der Kodiererschulung negative Auswirkungen auf die Reliabilität haben. Die Grenzen der Inhaltsanalyse sind durch das vorhandene Untersuchungsmaterial gegeben. Liegen für bestimmte Themen schlichtweg keine Informationen vor, so muss der Forscher auf andere Methoden zurückgreifen. Wie auch in der Umfrageforschung, wird bei der Inhaltsanalyse typischerweise mit Stichproben gearbeitet, wodurch repräsentative Aussagen ermöglicht werden sollen. Gerade bei der Inhaltsanalyse ist die Definition der Grundgesamtheit nicht immer eindeutig durchführbar, insbesondere wenn verschiedene Medien miteinbezogen werden. Speziell an der quantitativen Inhaltsanalyse wurde kritisiert, dass sich allein aus der Häufigkeit von Textelementen nicht immer Rückschlüsse auf die Absichten des Autors ziehen lassen. So führt Kracauer an, dass Auslassungen von Textelementen ebenso von Bedeutung sein können und daher der Text durch eine gekonnte Art des Lesens erfasst werden müsse. Somit ist Kracauer ein starker Verfechter der qualitativen Inhaltsanalyse, die er über die quantitative stellte. (Kracauer, 1952) Doch auch an der qualitativen Form der Inhaltsanalyse wurde Kritik geübt. Hier lassen sich die Ansprüche der Objektivität und Systematik kaum erfüllen, da der Forscher immer mit einer gewissen

Erwartungshaltung an einen Text herangeht und das Forschungsergebnis bei unterschiedlichen Forschern stark variiert.

In den Anfängen der Inhaltsanalyse kam die Inhaltsanalyse als „Frühwarnsystem“ zum Einsatz, indem die deutsche Kriegspropaganda im Zweiten Weltkrieg analysiert wurde. Beliebte Anwendungsgebiete sind heutzutage die Untersuchung von Wahlprogrammen oder politischen Texten, Tagebucheinträge in der Psychologie oder die Untersuchung von Schüleraufsätzen für die Sprachwissenschaften. Die Kriminalistik setzt Inhaltsanalysen zum Mustervergleich anonymer Texte ein. Auch in der Konsumentenforschung und bei der Untersuchung von Werbetexten wird die Inhaltsanalyse eingesetzt. Die Aufzählung der Einsatzgebiete der Inhaltsanalyse könnte noch weiter fortgesetzt werden und es ist anzunehmen, dass durch die zunehmende Vielfalt an textuellen und leicht zugänglichen Informationen im Internet noch weitere Anwendungen folgen.

2 Automatisierte Inhaltsanalyse

2.1 Einleitung

An automatischen Textanalyseverfahren wurde schon vor der Verbreitung des Internets gearbeitet, z.B. in der Nachrichtenfaktorenforschung, der Kampagnenberichterstattung, Analysen von Wahlprogrammen oder bei der Auswertung offener Befragungsitems. Dennoch führen automatische Textanalyseverfahren in der Sozialforschung ein Schattendasein und auf den ersten Blick scheint es, als würde die Entwicklung seit Jahren stagnieren. Tatsächlich wurde daraus jedoch ein sehr interdisziplinäres Forschungsfeld – vor allem durch die Zusammenarbeit der Informatik, der künstlichen Intelligenz, der Computerlinguistik und der Statistik konnten einige innovative und vielversprechende Methoden entwickelt werden. Scharkov bemerkt, dass dies in den Sozialwissenschaften nur teilweise wahrgenommen wurde. (Scharkow, 2010)

Die Tatsache, dass verschiedene Wissenschaften an ähnlichen Methoden und Anwendungen forschen, führt dazu, dass für ähnliche Betätigungsfelder unterschiedliche Bezeichnungen existieren. So ist in den Informationswissenschaften etwa der Begriff Text Mining verbreitet, womit die Sammlung und Analyse von unstrukturierten Texten bezeichnet wird. Eine präzisere Definition lautet: „Mit dem

Terminus Text Mining werden computergestützte Verfahren für die semantische Analyse von Texten bezeichnet, welche die automatische bzw. semi-automatische Strukturierung von Texten, insbesondere sehr großen Mengen von Texten, unterstützen.“ (Heyer, Quasthoff, & Wittig, 2008)

Die Anwendungsgebiete liegen etwa in der Markt- und Meinungsforschung, der Spam-Filterung von E-Mails, der automatischen E-Mail-Beantwortung von Dienstleitungsunternehmen, der Plagiatserkennung von wissenschaftlichen Arbeiten, der Internetsuche, der automatisierten Textzusammenfassung und vielen weiteren.

Es lässt sich also erahnen, dass es sich um ein sehr weites Forschungsfeld handelt, das um den Rahmen dieser Arbeit nicht zu sprengen, auf die Anwendungsmöglichkeiten für Inhaltsanalysen beschränkt wurde. Dementsprechend konzentriert sich auch die Erläuterung der Analyseverfahren auf dieses Anwendungsgebiet.

2.2 Verfahren automatischer Textanalyse

In der Literatur findet man unterschiedliche Ansätze zur Klassifizierung der Verfahren der automatischen Textanalyse. Die Unterschiede haben mit den verschiedenen Ursprüngen der Verfahren und der Vielzahl an Wissenschaften zu tun, die sich mit diesem Forschungsfeld beschäftigen. Es lässt sich eine grobe Einteilung in unüberwachte und überwachte Verfahren machen, wobei sich durch unüberwachte Verfahren der gesamte Forschungsablauf automatisieren lässt und sie weniger komplex sind als überwachte Verfahren. Überwachte Verfahren sind wesentlich aufwändiger und sind auf Grund ihres Potentials Schwerpunkt der Methodenentwicklung.

2.2.1 Unüberwachte Verfahren

Eines der einfachsten Verfahren dieser Kategorie ist die Berechnung von Text-, Satz- und Wortstatistiken. Die einfache Zählung von Wörtern oder Ermittlung von Satzlängen ist mit relativ wenig Aufwand durchführbar und wird auch von üblichen Textverarbeitungsprogrammen angeboten. Dennoch lassen sich durch Textstatistiken interessante Schlüsse auf die Texte ziehen. Einsatzgebiete sind beispielsweise die Autorschaftsforschung oder die Ermittlung von Lesbarkeitsmaßen.

Bei einem weiteren Verfahren, der Co-Occurence-Analyse, wird davon ausgegangen, dass semantisch ähnliche Wörter auch im Text nahe beieinanderliegen und deren Auftreten ermittelt und für Clusteranalysen verwendet werden kann. Co-Occurence-

Analysen sind explorative Verfahren, d.h. es wird kein Vorwissen über die Textinhalte benötigt, weshalb sie meist dafür verwendet werden, um sich einen Überblick über die Texte zu verschaffen. Zu explorativen Verfahren zählt auch Dokumenten-Clusterung, bei dem versucht wird Texte semantisch zu gruppieren. Dafür stehen verschiedene Cluster-Algorithmen zur Verfügung, wie z.B. k-Means. Eine Eigenschaft von Cluster-Verfahren ist, dass der Forscher keinen Einfluss auf die Kategorienbildung hat. Die Cluster werden durch den jeweiligen Algorithmus im Zuge der Durchführung bestimmt und müssen anschließend vom Forscher inhaltlich interpretiert werden. Anwendung findet die automatische Kategorisierung von Texten beispielsweise bei Online-Suchmaschinen, die Webseiten in sinnvolle Kategorien einteilen. (Welker & Wünsch, 2010)

Am Beispiel des Cluster-Verfahren wird auch die Bezeichnung „Unüberwachte Verfahren“ nachvollziehbar: Der Forscher kann die Kategorien nicht selbst bestimmen, sondern sie ergeben sich durch die Daten und den angewendeten Algorithmus selbst.

Bei den hier vorgestellten Verfahren ist eine perfekte Reliabilität gegeben, da die Algorithmen bei Analyse der gleichen Texte auch immer exakt dieselben Ergebnisse liefern. Fraglicher ist jedoch die Validität, die je nach Anwendung sehr gering sein kann.

2.2.2 Überwachte Verfahren

Überwachte Verfahren haben eine große Bedeutung für die Inhaltsanalyse, sind jedoch auch aufwändiger durchzuführen. Man unterscheidet zwischen deduktiven und induktiven Verfahren.

2.2.2.1 Deduktiv

Bei deduktiven Verfahren werden vom Forscher selbst Regeln und Anweisungen aufgestellt, nach denen ein Algorithmus die Klassifizierung bzw. Kodierung der Texte vornimmt. Dementsprechend groß ist für die Anwendung deduktiver Verfahren auch die Vorarbeit des Forschers. Mitunter ergeben sich sehr komplexe Kodierregeln, die erst mit viel Aufwand gefunden und dem eingesetzten Algorithmus beigebracht werden müssen. Deduktive Verfahren wurden bislang häufiger als induktive Verfahren eingesetzt, da es sich bei letzteren um ein relativ junges Forschungsfeld handelt.

Die wohl ältesten Verfahren dieser Kategorie sind die relativ einfachen diktionärbasierten Verfahren. Es werden vom Forscher Kategorien vorgegeben, denen einzelne Wörter oder Wortstämme zugeordnet werden. Anschließend durchsucht eine Software Texte nach diesen Wörtern und nimmt die Kodierung vor. Zwar ist die Arbeitsweise dieses Verfahrens trivial, doch kann die dafür notwendige Vorarbeit sehr umfangreich sein. Moderne Software ermöglicht auch die Definition von komplexeren Regeln, sogenannten regulären Ausdrücken, oder logischen Verknüpfungen von Termen. Da das Ergebnis deterministisch ist, liegt die Reliabilität bei 100%. Die Validität leidet durch das simple Vorgehen jedoch, da z.B. Rechtschreibfehler und Homonyme nicht erkannt werden. Bei komplexen Fragestellungen ist auch fraglich, ob die Aufstellung eines umfassenden Diktionärs nicht zeitaufwendiger als die manuelle Kodierung ist. (Welker & Wünsch, 2010)

Weitere deduktive Verfahren sind regelbasierte Verfahren. In diesem Fall sind es Regeln, die vom Forscher vorgegeben werden und der Informationsextraktion dienen. Texte können dadurch in eine Graphen- oder Baumstruktur zerlegt werden, die sich dann automatisch analysieren lassen. Regelbasierte Verfahren lassen sich gut anwenden, wenn die vorliegenden Sätze eine ähnliche Struktur haben, wie es beispielsweise bei Nachrichten-Tickermeldungen der Fall ist. Durch relativ einfache Regeln lassen sich diese Meldungen z.B. in Subjekt-Objekt-Prädikat-Beziehungen zerlegen, die sich dann getrennt analysieren lassen. Die Auflösung der Sätze in derartige Strukturen wird auch als Parsing bezeichnet. Methoden aus dem Bereich des Natural Language Processings sind sehr moderne Parser, die syntaktisch-semantische Analysen ermöglichen.

2.2.2.2 Induktiv

Das Prinzip der induktiven Verfahren ist, dass in einer Trainingsphase einem Algorithmus richtig klassifizierte Texte zur Verfügung gestellt werden, dieser auf Grund dessen selbst Regeln aufstellt und anschließend auch unbekannte Texte richtig klassifiziert. Deshalb spricht man hier auch von „lernenden Algorithmen“. Eine manuelle Vorgabe von Regeln seitens des Forschers ist nicht erforderlich. Stattdessen müssen den Algorithmen richtig kodierte Texte bereitgestellt werden – der manuelle Aufwand ist jedoch deutlich geringer als das Aufstellen eines Regelwerks oder Diktionärs. Induktive Verfahren sind ein noch relativ junges Forschungsgebiet, das vor allem durch die Zusammenarbeit der Statistik und der Informatik vorangetrieben wird.

Dieses induktive Vorgehen ist auch bei der manuellen Inhaltsanalyse zu beobachten, denn auch dort werden den Kodierern richtig kodierte Beispieltexte vorgelegt, die sie zur Kodierung neuer Texte verwenden. Der Vorteil der induktiven automatischen Klassifikation gegenüber deduktiven Verfahren ist, dass das Aufstellen komplexer umfangreicher Regeln oder eines Diktionärs völlig entfällt. Zur automatischen Klassifikation werden Verfahren aus dem Bereich des maschinellen Lernens eingesetzt, einem Unterbereich der Künstlichen Intelligenz. Am häufigsten werden die Algorithmen Naive Bayes und Support-Vektor-Maschinen für die Anwendung der automatischen Textklassifikation eingesetzt. Genau diese Verfahren haben großes Potential für die empirische Sozialforschung, wo sie jedoch noch kaum wahrgenommen werden. Im Medium Internet werden hingegen Texte wie Nachrichten oder Filmbewertungen mit derartigen Verfahren bereits automatisch klassifiziert. Diese offensichtliche Distanz zwischen den Methoden der Computerwissenschaften und der empirischen Sozialforschung ist sicherlich auch auf die hohe Anforderung an die Güte seitens der empirischen Sozialforschung zurückzuführen. Schließlich erfüllen die erwähnten Klassifizierungen im Internet hauptsächlich den Zweck die Suche von Texten zu erleichtern, neue Informationen aus Texten zu extrahieren oder auch Texte automatisiert zusammenzufassen, doch handelt es sich dabei um keine Untersuchungen, die wissenschaftlichen Kriterien entsprechen müssen.

Die Reliabilität des Verfahrens hängt von der Reliabilität der manuellen Kodierung des Trainingsmaterials ab. Die Validität kann überprüft werden, indem die automatisiert vorgenommene Kodierung mit einer manuellen Kodierung verglichen wird. Die manuell vorgenommene Kodierung wird also als Standard herangezogen. In der Regel sind mehrere Durchführungen der automatisierten Kodierung erforderlich, um schließlich den optimalen Algorithmus und eine geeignete Konfiguration der Parameter zu finden.

2.3 Ablauf der automatisierten Textanalyse

Der prinzipielle Ablauf der automatisierten Textanalyse hat sich seit der ersten Anwendung nicht verändert, jedoch haben sich die in den einzelnen Schritten verwendeten Verfahren zum Teil erheblich verbessert.

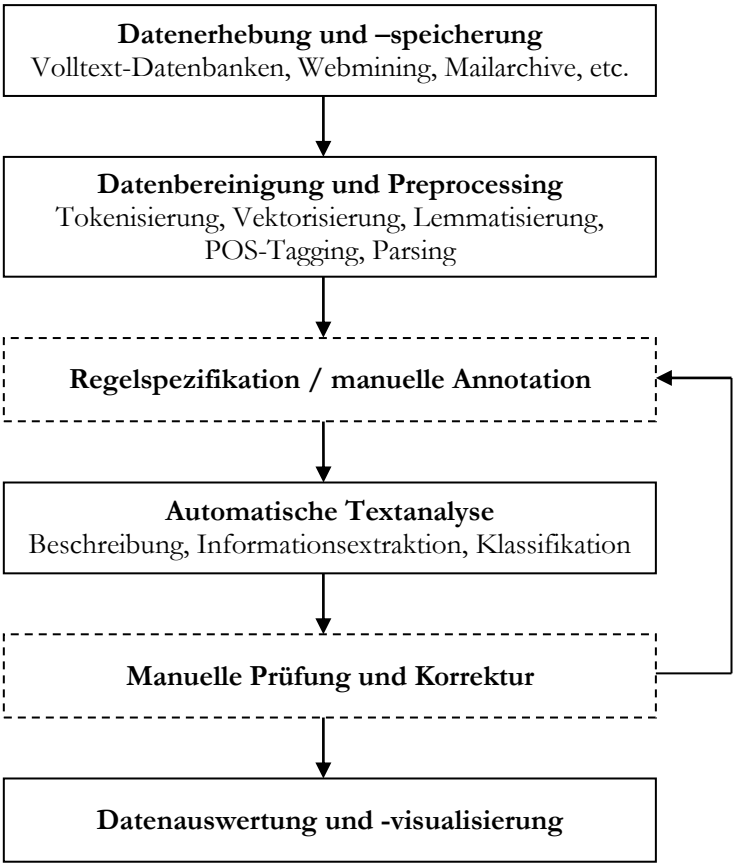


Abbildung 1: Ablauf der automatisierten Inhaltsanalyse (Scharkow, 2010)

Der Ablauf beginnt mit der Extraktion und gegebenenfalls Zwischenspeicherung der zu analysierenden Texte aus verschiedenen Quellen. Anschließend werden die Texte in eine einheitliche Struktur gebracht und verschiedene Methoden angewendet, um die Texte zu vereinfachen, ohne dabei die Bedeutung der Texte zu verändern. Dieser Vorbearbeitungsschritt hat unmittelbare Auswirkung auf die Güte der automatisierten Inhaltsanalyse und ist auch Gegenstand aktueller Forschung. Die aufbereiteten Texte werden anschließend mit Hilfe von Regeln oder Methoden des maschinellen Lernens

analysiert und automatisch kategorisiert. Nachdem die Textanalyse abgeschlossen ist, kann eine manuelle Prüfung erfolgen, und daraus abgeleitete Erkenntnisse in die Regelspezifikation miteinfließen, sodass sich ein iterativer Prozess ergeben kann. Schließlich können die kategorisierten Texte ausgewertet und visualisiert werden.

2.3.1 Datenerhebung und –speicherung

Während zum Zeitpunkt der ersten Anwendungen in den 1950er/1960er Jahren der Bestand an digitalisierten Texten noch vergleichsweise gering war, sind heute unüberschaubare Mengen an Texten sowohl online als auch offline in digitalisierter Form verfügbar. Die Herausforderung besteht heute viel mehr darin, die für die Untersuchung relevanten Informationen zu sammeln und zu verwalten. Auch dazu wird spezielle Software angeboten, die Schnittstellen zu den verschiedenen Textquellen mitliefern. Datenquellen können Mailarchive, Webseiten, Volltext-Datenbanken, Internetforen, Zeitungsarchive, eingescannte und digitalisierte Bücher uvm. sein. Nachdem die Texte extrahiert und eventuell in einer zentralen Form gespeichert wurden, erfolgt die Bereinigung und Vorbereitung des Textmaterials.

2.3.2 Datenbereinigung und Preprocessing

2.3.2.1 Datenbereinigung

Da die extrahierten Texte aus verschiedenen Quellen stammen können und in unterschiedlichen Formaten und Strukturen vorliegen, müssen die Texte in einem ersten Schritt vereinheitlicht werden. Dieses Format kann etwa ein standardisiertes Format wie ASCII, Unicode-Text, HTML oder XML sein oder die Texte können auch in eine Datenbank importiert werden. Bei dieser Standardisierung werden auch nicht relevante oder nicht-textuelle Inhalte entfernt, sodass man schließlich eine einheitliche, bereinigte Datenbasis zur Verfügung hat.

Die weiteren Preprocessing-Schritte sind stark vom angewendeten Textanalyse-Verfahren abhängig. Üblicherweise werden aus dem Text die sogenannten Features extrahiert, also Wörter bzw. Wortgruppen definierter Länge. Durch Wortgruppen, die aus zwei- oder mehr Wörtern bestehen, kann erfahrungsgemäß die semantische Vielfalt der Sprache gut erfasst werden und so mit Hilfe von statistischen Verfahren gut analysiert werden.

2.3.2.2 Tokenisierung

Die Segmentierung der Texte in einzelne Einheiten, wie Wörter, Sätze oder Absätze wird Tokenisierung genannt. Die Zerlegung in einzelne Wörter kann beispielsweise durch das White-Space-Verfahren erfolgen, bei dem Leerzeichen und Interpunktionszeichen als Trennzeichen interpretiert werden. Fehlerquellen dieses einfachen Verfahrens sind, dass z.B. Abkürzungen, durch Bindestrich getrennte Wörter, durch Leerzeichen getrennte Eigennamen (z.B. New York) oder auch Dezimalzahlen fehlerhaft zerlegt werden. (Kai-Uwe Carstensen, 2010) Komplexere Verfahren versuchen diese Fehler zu beseitigen. Zu denen zählen beispielsweise Methoden die reguläre Ausdrücke verwenden, um genaue Regeln zu definieren, wie Texte in Token zerlegt werden sollen.

2.3.2.3 Stemming und Lemmatisierung

Eine weitere Vorbearbeitung der Texte ist die Zurückführung der Wörter auf einen gemeinsamen Wortstamm, was in der Fachsprache auch als Stemming bezeichnet wird. Dieser Wortstamm muss nicht zwingend in der jeweiligen Sprache existieren, was Stemming von Lemmatisierung unterscheidet. Die Aufbereitung des Textes durch diese Verfahren ist für viele Textanalyse-Verfahren erforderlich, um zu besseren Ergebnissen zu kommen, beispielsweise für Verfahren des maschinellen Lernens. Aber auch der Speicherbedarf des Vokabulars kann dadurch reduziert werden.

Es gibt eine Reihe von Stemming-Verfahren, die sich durch ihre Komplexität unterscheiden, und abhängig von der Sprache des Textes unterschiedlich gute Ergebnisse liefern. Auf die genaue Funktionsweise der Algorithmen von Stemming-Verfahren wird in dieser Arbeit nicht eingegangen. Ein verbreitetes Verfahren ist zum Beispiel der Porter-Stemmer-Algorithmus, der eine Reihe von Verkürzungsregeln auf ein Wort anwendet, bis es eine minimale Anzahl von Silben aufweist. (Porter, 1980) Es sei erwähnt, dass die Rückführung der Wörter auf einen gemeinsamen Wortstamm auf Basis von linguistischem Wissen erfolgen kann, oder durch rein statistische Verfahren. Letztere versuchen, vereinfacht ausgedrückt, einen gemeinsamen Wortstamm für ähnliche Wörter zu finden. Die meisten Stemming-Verfahren wurden ursprünglich für die englische Sprache entwickelt und anschließend für die deutsche Sprache portiert. Neben der Sprachabhängigkeit kann auch eine Abhängigkeit vom Anwendungsbereich

miteinbezogen werden. Eine Schwierigkeit bei der Durchführung des Stemmings ist ein gutes Gleichgewicht zwischen den beiden Fehlerquellen „Overstemming“ und „Understemming“ zu finden. Overstemming würde Wörter mit unterschiedlicher Bedeutung auf denselben Wortstamm zurückführen (z.B. kommunismus, kommunikation, kommunizieren → kommun). In diesem Fall würde der Algorithmus eine zu lange Zeichenkette der Wörter abschneiden. Understemming ist die Gefahr, eine zu kurze Zeichenkette abzuschneiden. Gleichbedeutende Wörter werden in unterschiedliche Wortstämme zerlegt (z.B. kommunismus → kommun, kommunikation → kommunika, kommunizieren→kommuniz). (Frake, 1992) Neben der Korrektheit der Resultate kann auch die Laufzeit des eingesetzten Algorithmus eine Rolle spielen. Gerade sehr komplexe Verfahren können bei großen Textmengen zu unpraktikablen Laufzeiten führen.

2.3.2.4 POS-Tagging

Die Anforderung der Untersuchung kann es auch erfordern, die Wortformen der einzelnen Wörter zu bestimmen. Dieses Verfahren wird als Part-of-speech Tagging bezeichnet. Auch dazu gibt es eine Reihe von Algorithmen unterschiedlicher Komplexität, die im Rahmen dieser Art jedoch nicht genauer erläutert werden. Zur korrekten Zuordnung der Wortform, muss sowohl die Definition des Wortes, als auch der Kontext des Wortes betrachtet werden. (Scharkow, 2010) Für die deutsche Sprache wird oft das Stuttgart-Tübingen-Tagset (STTS) als Basis eingesetzt, welches sich als Standard zur Benennung von Wortarten in Form von Kürzeln für die deutsche Sprache durchgesetzt hat. (Stuttgart, 1996)

2.3.2.5 Phrase Recognition

Durch Phrase Recognition können Wörter zur Wortgruppen oder Phrasen zusammengefasst werden. Auch das Auffinden von Personen, Firmennamen oder Ortschaften in Texten fällt in diese Kategorie. Ob Phrase Recognition erforderlich ist, ist weitgehend vom Anwendungsfall abhängig. (Scharkow, 2010)

2.3.2.6 Parsing

Unter Parsing wird die grammatikalische Analyse von Texten verstanden. Damit wird der Satzbau analysiert und für jedes Wort dessen Stellung im Satz eruiert (z.B. Subjekt, Prädikat, Objekt). (Scharkow, 2010)

2.3.2.7 Weitere Vorbearbeitungen

In der Praxis können noch weitere Vorbearbeitungen des Textes durchgeführt werden, um das Ergebnis der automatischen Textanalyse zu verbessern. Dazu zählen beispielsweise die Auflösung von Synonymen und Anaphora, das Erkennen von Negationen oder die Entfernung von häufig vorkommenden Wörtern, die keine wesentliche semantische Aussagekraft für die Untersuchung haben (z.B. Bindewörter, Artikel, etc.).

Als Resultat des Preprocessings hat sich die Struktur einer Dokument-Term-Matrix als nützlich erwiesen. In dieser Darstellungsform entsprechen die Zeilen Dokumenten und die Spalten enthalten die extrahierten Features und die Zellen Informationen zum Vorkommen des Features im jeweiligen Dokument. Dadurch wird die Analyse mit Statistikprogrammen wie SPSS oder R erheblich erleichtert.
Die Vielzahl an Vorbearbeitungsschritten ist einerseits notwendig um die Komplexität und die Feature-Anzahl zu reduzieren, wodurch viele Textanalyseverfahren erst ermöglicht werden. Andererseits ist sie nicht ganz unumstritten, da der Text sowohl syntaktisch, als auch semantisch dadurch verändert wird und die Validität darunter leiden kann.

3 Diktionärbasierte Verfahren

3.1 Klassische diktionärbasierte Verfahren

Diktionärbasierte Verfahren der Textanalyse lassen sich den deduktiven Verfahren zuordnen und stellen die klassische Methode zur automatischen Textanalyse dar. Das Grundprinzip der diktionärbasierten Verfahren ist relativ simpel blieb bislang auch unverändert.
Der Forscher entwirft das Kategoriensystem und ordnet den einzelnen Kategorien Wörter oder Wortstämme zu, die einen Indikator für die entsprechende Kategorie

darstellen. Diese Auflistung an Wörtern mit zugeordneter Kategorie wird auch als Diktionär bzw. Lexikon bezeichnet. Eine Analysesoftware kann auf Basis des Lexikons Texteinheiten nach den Wörtern durchsuchen und somit automatisch kategorisieren. (Cornelia Züll, 2002)

Ein ähnlicher Ansatz ist die Freitextrecherche, bei der Suchanfragen beispielsweise an Online-Suchmaschinen geschickt werden und gewünschte Texte zurückliefert. Die Suchbegriffe können dabei auch durch logische Operatoren miteinander verknüpft werden.

Diktionärbasierte Verfahren werden auch als „bag of words"-Ansätze bezeichnet. (García Flore, Gillar, Ferret, & de Chalenda) Texte werden demnach zunächst als ungeordnete Aneinanderreihung von Wörtern verstanden - Wörter eines Textes sind voneinander unabhängig und gleichwertig. Wie später erläutert wird, gibt es auch Weiterentwicklungen der diktionärbasierten Verfahren, die sich vom einfach „bag of words"-Ansatz abheben.

Auch bei diktionärbasierten Verfahren lassen sich selbstverständlich komplexere Suchregeln und Verknüpfungen von Wörtern definieren um die Flexibilität und Trefferquote zu erhöhen.

Die Reliabilität ist bei diktionärbasierten Verfahren vollständig gegeben, da das Ergebnis durch das Lexikon und den Suchprozess eindeutig determiniert ist. Dafür leidet die Validität unter dem sehr einfachen Vorgehen und es kann nicht immer gewährleistet werden, dass die Methode tatsächlich das misst, was der Forscher beabsichtigt zu messen. Dennoch gibt es Anwendungsfälle, für die diktionärbasierte Verfahren durchaus interessant sein können. Eine solche Anwendung wäre beispielsweise eine Medienresonanzanalyse, bei der Texte nach Vorkommen von Markennamen oder Eigennamen durchsucht und kategorisiert werden. (Welker & Wünsch, 2010) Problematischer ist die Anwendung der Methode bei Analysen nach thematischen Fragestellungen in Texten. Rechtschreibfehler, Metaphern oder Negationen sind nur einige der Probleme, die bei der Analyse auftreten. Ein vollständiges und richtiges Lexikon zu erstellen ist oft mit sehr viel Aufwand verbunden, der sich nur lohnt, wenn er geringer als bei einer manuellen Analyse ist. Ist ein geeignetes Lexikon einmal erstellt, so können damit allerdings auf sehr effiziente Weise Texte kategorisiert werden. Die Kodierung kann dann innerhalb von Sekunden

oder wenigen Minuten erfolgen und auch eine Stichprobenziehung ist in vielen Fällen dann nicht erforderlich. Existiert ein geeignetes Lexikon, so kann dieses ebenso einfach auf neue Texte des Themenbereichs angewendet werden bzw. kann die Kodierung problemlos und beliebig oft wiederholt werden.

3.2 Diktionär-Typen

Wie der Name bereits vermuten lässt, ist das Diktionär zentraler Bestandteil des Verfahrens. Von ihm hängt zu einem großen Teil die Güte ab und auch der Aufwand einer Untersuchung mittels diktionärbasiertem Verfahren besteht zu einem Großteil aus der Erstellung und Anpassung des Diktionärs. Es können folgende vier Arten von Diktionären unterschieden werden (Cornelia Züll, 2002):

- Nutzerdefinierte Diktionäre, die für einen bestimmten Textteil entwickelt werden. Das Diktionär wird vom Forscher selbst erstellt und ist folglich speziell für den ausgewählten Text optimiert.
- Themenspezifische Diktionäre, die für bestimmte Fragestellungen erstellt und optimiert werden. Diese Diktionäre sind in der Regel allgemein zugänglich und sind mit verschiedenen Computerprogrammen kompatibel. Ein Beispiel für diesen Diktionär-Typ ist etwa das Dresdner Angstwörterbuch. (Berth, 2000)
- Programminterne Diktionäre, die nur in Kombination mit einem bestimmten Computerprogramm verwendet werden können. Auch diese Diktionäre werden für eine spezielle Fragestellung entwickelt.
- Allgemeingültige Diktionäre, die nicht an spezielle Themen oder Fragestellungen gebunden sind, sondern universell einsetzbar sind. Freilich ist von themenspezifischen Diktionären eine höhere Güte zu erwarten, doch sind diese nicht für alle Domänen verfügbar. Bekannte Diktionäre dieses Typs sind etwa jene des General Inquirers. (Cornelia Züll, 2002)

3.3 Weiterentwicklungen der diktionärbasierten Verfahren

Die zahlreichen Vorteile und Potentiale zur Kosteneinsparung waren auch Antrieb für viel Forschungsarbeit und Versuche, die Fehlerquellen der Methode zu reduzieren. Es

setzte sich ein Ansatz durch, der sich in seinem grundsätzlichen Ablauf in den meisten Anwendungen wiederfindet.

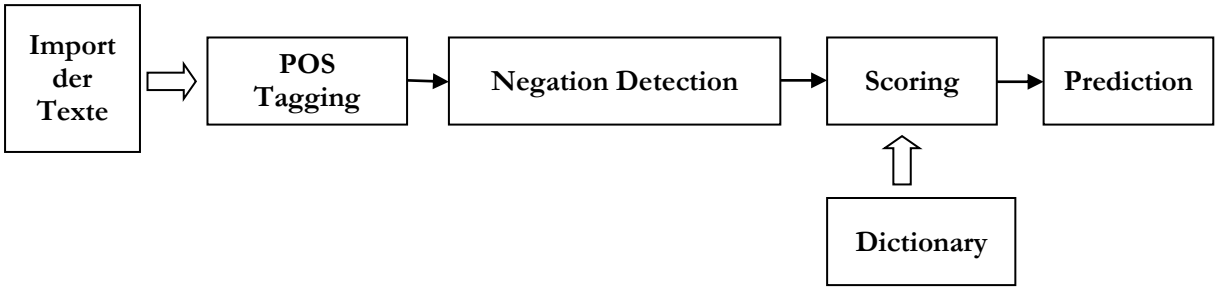


Abbildung 2: Ablauf diktionärbasierte Inhaltsanalyse (Bruno Ohana, 2011)

Texte werden in einem ersten Schritt aus einer oder mehreren Quellen extrahiert und in eine einheitliche Struktur gebracht, bevor mit dem weiteren Prozess fortgefahren wird. Mit Hilfe eines POS Tagging-Verfahrens werden anschließend die Wortarten der einzelnen Wörter bestimmt. Um die Güte minimal zu verbessern, können danach Methoden zur Erkennung von Negationen angewendet werden, die das Ergebnis der Kategorisierung verfälschen würden. Erst nach diesen Vorbearbeitungsschritten beginnt die Kategorisierung der Texteinheiten, indem mit Hilfe eines Diktionärs, welches zu den kategorisierten Wörtern auch die entsprechenden Wortarten enthält, die Texte automatisch kodiert (Scoring). Die einzelnen Bearbeitungsschritte werden im Folgenden genauer erläutert.

3.3.1 POS (Part of Speech) Tagging

POS Tagging ist die wohl wichtigste Voraussetzung für die Weiterentwicklung der klassischen diktionärbasierten Verfahren. Texte werden dadurch nicht mehr als bloße Aneinanderreihung von gleichwertigen Wörtern aufgefasst, sondern in ihre grundlegenden Bestandteile zerlegt (Parts of Speech). Damit sind Worttypen wie Verben, Adjektive, Adverbien, Nomen, Artikel, Pronomen etc. gemeint, die durch den Prozess des POS Taggings identifiziert werden. (Group, Stanford Log-linear Part-Of-Speech Tagger, 2011)

Die Wortart wird mit einem Schrägstrich getrennt direkt an das Wort angehängt. Soll beispielsweise der Satz *„Peter mag saftige Äpfel“* getaggt werden, so würde das korrekte Ergebnis folgendermaßen aussehen: *„Peter/NE mag/VVFIN saftige/ADJA*

Äpfel/NN". Wie in dem Beispiel ersichtlich, existieren für die verschiedenen Wortarten Abkürzungen (NE=Eigenname, VVFIN=finites Vollverb, ADJA=Adjektiv, NN=sonstige Nomen, etc.)

Ursprünglich wurden die Wortarten manuell gekennzeichnet. Durch Fortschritte in der Computerlinguistik wurden jedoch mittlerweile verschiedene Methoden entwickelt, die freilich nicht immer fehlerfrei arbeiten, in der Regel jedoch sehr gute Ergebnisse liefern. Gute POS Tagger liefern einen Accuracy-Wert von über 90%. (Kevin Gimpel, 20114) Voraussetzung für das POS Tagging ist eine Tokenisierung des Textes, d.h. die Identifizierung der einzelnen Wörter (siehe Kapitel 2.3.2.2). Die wohl größten Herausforderungen beim POS Tagging ist einerseits die Tatsache, dass der Typ eines Worts von seinem Kontext abhängig ist - ein Wörterbuch, das alle Wörter einer Sprache mit zugehörigem Worttyp enthält, ist daher nicht ausreichend. Andererseits sind unbekannte Wörter in Texten eine weitere Fehlerquelle für das POS Tagging oder auch seltene Satzkonstruktionen und selbstverständlich grammatische Fehler in Sätzen. Üblicherweise ziehen die Methoden des POS Taggings Statistiken heran und schätzen die Wahrscheinlichkeit, mit der eine Wortart auf eine andere Wortart folgt.

Die Verfahren des POS Taggings können grundsätzlich wie folgt kategorisiert werden:

- Regelbasierte Verfahren
  Es werden Regeln definiert, an Hand derer die Wortarten von Texten erkannt werden können. Diese werden am Text abgearbeitet. Regelbasierte Verfahren sind abhängig vom Kontext und der Sprache, was sie wenig flexibel macht.
- Statistische Ansätze
  Durch Ansätze des Maschinellen Lernens und der Verwendung von bereits getaggten Trainingsdatensätzen können die Wortarten neuer Texte geschätzt werden.

Auf die detaillierte Funktionsweise der einzelnen Methoden wird in dieser Arbeit nicht eingegangen, da es sich um ein eigenes Forschungsfeld handelt, zu dem bereits viel Literatur verfügbar ist. Sollen POS Tagging Verfahren angewendet werden, so müssen diese nicht unbedingt selbst entwickelt werden, da schon eine Reiche vorgefertigter Anwendungen existieren. Für die deutsche Sprache existieren beispielsweise ein POS Tagger der Universität Saarbrücken namens TnT (Trigrams'n'Tags), der auf

statistischen Methoden basiert, FreeTragger der Universität Stuttgart oder Morphy der Universität Morphy. Im englischen Sprachraum ist der POS Tagger der Universität Stanford stark verbreitet, der auf statistischen Methoden basiert. (Group, Stanford Log-linear Part-Of-Speech Tagger, 2003)

3.3.2 Negation Detection

Eine der Fehlerquellen, die große Auswirkungen auf die Validität des Verfahrens haben kann, ist die falsche Interpretation bzw. Nicht-Interpretation von Negationen in Texten. Eine Negation im Zusammenhang mit Adjektiven bedeutet eine Umkehrung der Polarität der Stimmung, die jedoch nur schwer erkannt werden kann. Tritt eine Negation in einem Text auf, so kann sie sich entweder auf einzelne Wörter eines Satzes oder auch auf den gesamten Satz beziehen. Darum ist ein erster Schritt der Negation Detection die korrekte Einschätzung des Geltungsbereichs (scope) der Negation. Dazu werden häufig linguistische Regeln herangezogen, doch gibt es in diesem Forschungsbereich nur vergleichsweise wenig Fortschritte. (Dadvar, Hauff, & de Jong, 2011)

Die Komplexität der Negation Detection kann schon daran erahnt werden, dass Negationen nicht nur durch an Adjektive vorangestellte Wörter wie „nicht“, „kein“, „zu“ ausdrücken, sondern auch durch Präfixe (un-sauber, un-schön, dis-funktional, etc.) oder Suffixe (wert-los, zusammenhangs-los, etc.). Auch implizite Negationen ohne Negationswörter sind denkbar („Die Qualität des Produktes war unter meinen Erwartungen“). Während man das Auftreten von Negationen in Form von Präfixen oder Suffixen durch explizite Aufnahme der negierten Wörter in das Lexikon lösen kann, ist das Auffinden von impliziten Negationen weitaus komplexer. Negation Detection konzentriert sich daher vorwiegend auf die explizite Angabe von Negationen durch Negationswörter und das richtige Erkennen deren Wirkungsbereiche. Der Wirkungsbereich wird durch die sogenannte „window size“ angegeben, die die Anzahl an Wörtern vom Auftreten des Negationswortes bis zum letzten Wort, auf das sich die Negation auswirkt, angibt. (Dadvar, Hauff, & de Jong, 2011) Folgendes Beispiel soll die „window size“ verdeutlichen:

„Die Qualität der Fotos war nicht zufriedenstellend.“

In diesem Fall wäre eine window size von 1 ausreichend, da das Adjektiv „zufriedenstellend“ unmittelbar auf das Negationswort folgt. Bei folgendem Beispielsatz ist dies nicht der Fall.

„Ich mag die Qualität der Fotos der Kamera nicht.“

Zwischen dem Wort „mag“, welches die Stimmung des Satzes wiedergibt, und dem Negationswort „nicht“ befinden sich 6 Wörter. Um die Negation richtig zu erkennen wäre also eine höhere window size erforderlich. Die optimale Wahl der windows size ist experimentell zu ermitteln und auch von der Sprache abhängig.
Eine weitere Verbesserung der Negation Detection kann durch Einbeziehung des Worttyps erreicht werden, unter der Voraussetzung dass POS Tagging eingesetzt wird. D.h. alle Wörter des Textes müssen in einem vorherigen Schritt durch ihren Worttyp gekennzeichnet werden.
Negation Detection ist eine sehr komplexe Aufgabe. Verschiedene Forschungen zeigen jedoch, dass die hier vorgestellten Ansätze die Accuracy der Sentiment Analyse nur minimal erhöhen. (Dadvar, Hauff, & de Jong, 2011) Häufiger als simple explizite Negationen werden auf indirekte Weise Stimmungen und Negationen in Texten wiedergegeben, sowie Sarkasmus oder Metaphern verwendet. Verlässliche Methoden diese Phänomene richtig zu interpretieren gibt es jedoch bislang noch nicht.

3.3.3 Scoring

Lexika oder Diktionäre enthalten die für die Forschungsfrage erforderlichen Kategorien, denen Indikatoren zugeordnet werden können. Dabei kann es sich um einzelne Wörter, Wortteile, oder auch ganze Phrasen handeln. Ein Computerprogramm durchsucht im Prozess des Scorings den oder die Texte nach den Indikatoren und ordnet sie dementsprechend den Kategorien zu.
Erfreulicherweise existieren für verbreitete Anwendungen bereits Lexika, die den immensen Vorteil haben, dass der bei diktionärbasierten Verfahren sehr aufwändige Schritt der Erstellung des Lexikons völlig entfällt und sich das Verfahren unmittelbar anwenden lässt.

Für Sentiment Analysen in der englischen Sprache ist SentiWordNet[1] als lexikalische Ressource sehr beliebt und verbreitet, da es öffentlich zugänglich ist und für Forschungszwecke frei verwendet werden kann. (Stefano Baccianella, 2010) Detaillierte Lexika wie SentiWordNet enthalten nicht bloß eine einfache Auflistung an Wörtern und deren Zuordnung zu Kategorien, sondern auch diverse Zusatzinformationen, die die Trefferquote der Sentiment Analysen verbessern können. (Esuli & Sebastiani)

Wörter in SentiWordNet werden durch das in Abbildung 3: SentiWordNet Polarity ersichtliche Dreieck beschrieben. Die SO-Achse gibt an, zu welchem Grad das Wort eine (subjektive) Meinung wiedergibt oder ob es sich um einen objektiven Begriff handelt. Diese Achse kategorisiert ein Wort also in die beiden Kategorien „subjektiv“ und „objektiv“. Die horizontale Achse gibt die positiv/negativ Polarität an. Der SO-Dimension wurde in der Vergangenheit nur wenig Beachtung geschenkt, obwohl sie für semantische Analysen von großer Bedeutung ist. Im SentiWordNet Lexikon werden jedem Wort 3 numerische Werte zugewiesen: Objektiv, Positiv und Negativ. Dabei werden den Kategorien nicht nur binäre Werte (ja oder nein) zugewiesen, sondern ein kontinuierlicher Zahlenwert, der von 0 bis 1 reicht. Bemerkenswert an diesem Ansatz ist, dass ein Wort sowohl ein Wert für eine positive, als auch für eine negative Stimmung zugeordnet werden kann.

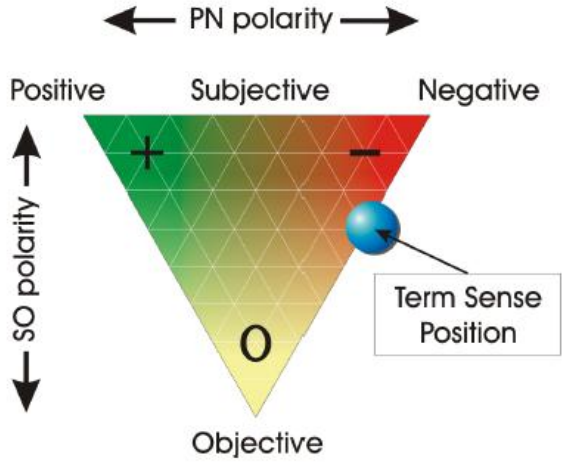


Abbildung 3: SentiWordNet Polarity (Esuli & Sebastiani)

[1] http://sentiwordnet.isti.cnr.it/

SentiWordNet liefert auch Informationen über die Wortart mit. Abbildung 4: SentiWordNet Adjektive verdeutlicht die Informationen, die SentiWordNet liefert an Hand eines Beispiels. Zum englischen Wort „estimable“ erden drei verschiedene Bedeutungen gefunden mit jeweils unterschiedlichen Kategorisierungen.

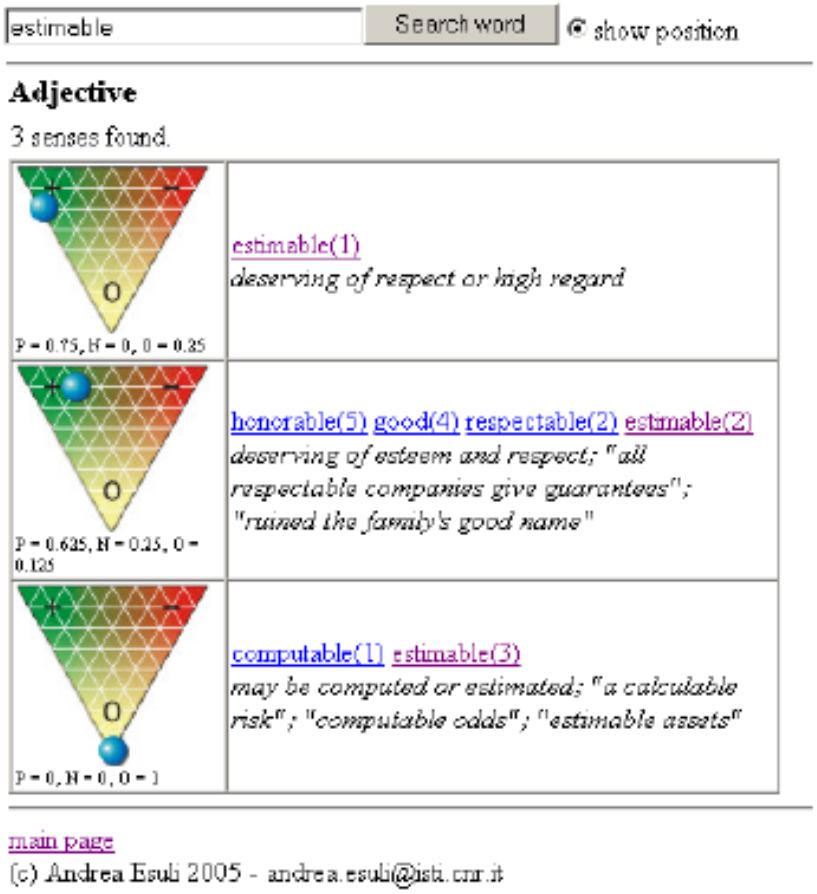


Abbildung 4: SentiWordNet Adjektive (Esuli & Sebastiani)

Wie bereits erwähnt ist SentiWordNet das verbreitetste frei verfügbare Lexikon für semantische Analysen. Es gibt jedoch noch zahlreiche weitere vorgefertigte Diktionäre, die auch für andere Sprachen verfügbar sind.

Die im vorgegangenen Schritt des POS Taggings (Kapitel 2.3.2.4) vorgenommene Kategorisierung der Wörter in Wortarten und die entsprechenden Zusatzinformationen im Diktionär ermöglichen eine weitaus bessere Klassifizierung als es bei klassischen lexikalischen Ansätzen möglich war. Dieser Prozess der Bewertung von Texten und Texteinheiten auf Basis des Lexikons wird auch als „Scoring“ bezeichnet.

Es sei darauf hingewiesen, dass die Kategorisierung mit Hilfe von diktionärbasierten Verfahren nicht nur für semantische Analysen, sondern für beliebige Kategorisierungen anwendbar ist.

4 Co-occurence-Verfahren

Ein weiteres Verfahren der computerunterstützten Inhaltsanalyse ist das Co-occurence-Verfahren, welches das gemeinsame Auftreten von Wörtern in Texten analysiert. In der Regel gehen dabei nur die häufigsten Wörter in die Auswertung ein. Damit unterscheidet es sich grundlegend von diktionärbasierten Verfahren, wo für die Analyse ein „Wörterbuch“ erstellt wird bzw. zur Verfügung stehen muss. Folglich müssen vor der Analyse auch keine Kategorien definiert werden, da sich durch das explorative Vorgehen des Co-occurence-Ansatzes mögliche Kategorisierungen nach der Auswertung der gefundenen Assoziationsmuster ergeben. Auch Hypothesen werden nicht wie beim diktionärbasierten Ansatz vor der Analyse erstellt.

4.1 Cooccurrence

Die Co-occurence von Wörtern kann auf unterschiedliche Weiße bestimmt werden. Positional Co-occurence ist der traditionelle Ansatz. Demzufolge wird das gemeinsame Auftreten von Wörtern als solches gewertet, sofern die Wörter innerhalb einer definierten Distanz in einem Text oder Textteil vorkommen. Die Distanz wird in den meisten Fällen durch die Anzahl an Wörtern gemessen, kann stattdessen aber auch an Hand von Wortgruppen, Sätzen, Absätzen, etc. bestimmt werden. Kritikpunkt an diesem Ansatz ist das Fehlen einer theoretischen Grundlage der bloßen Auszählung von Wörtern oder Textteilen. (Evert, 2004)

Von der Positional Co-occurence unterscheidet sich die Relational Co-occurence. Hier werden linguistische Informationen hinzugenommen, um das gemeinsame Auftreten von Wörtern zu bestimmen. Ein Beispiel eines solchen Falles wäre ein Adjektiv, das ein Nomen näher beschreibt oder die Kombination eines Verbs mit einem Nomen. Um diese Beziehungen herstellen zu können, müssen die Texte in einem Vorbearbeitungsschritt von Algorithmen interpretiert und Wortarten bestimmt werden. Vom theoretischen Standpunkt liefert die Einbeziehung linguistischer Informationen

verlässlichere Ergebnisse, sofern die Bestimmung der Wortarten fehlerfrei erfolgt. (Evert, 2004)

4.2 Assoziationsmaße

Nachdem gemeinsame Vorkommen von Wörtern gefunden wurden, ist es Ziel des Verfahrens an Hand einer Kennzahl die Signifikanz bzw. Stärke des Zusammenhangs zu bewerten. Diese Assoziationsmaße deuten dann auf einen semantischen Zusammenhang der Wörter hin, wie z.B. Ober- und Unterbegriffe, typische Eigenschaften von bestimmten Nomen oder Tätigkeiten von Personennamen. (Riggert, 2009)

Es gibt eine Reihe verschiedener Assoziationsmaße, die andere Ergebnissen liefern können. Die meisten davon bauen auf einer Kontingenztabelle auf, die die tatsächlich beobachteten Häufigkeiten enthält.

	Wort 1=v	Wort 1≠v	
Wort 2=u	O_{11}	O_{12}	$\boldsymbol{R_1}$
Wort 2≠u	O_{21}	O_{22}	$\boldsymbol{R_2}$
	$\boldsymbol{C_1}$	$\boldsymbol{C_2}$	**N**

Abbildung 5: beobachtete Häufigkeiten (Evert, 2004)

Dem gegenübergestellt wird eine Matrix, die die erwarteten Häufigkeiten enthält.

	Wort 1=v	Wort 1≠v
Wort 2=u	$E_{11=}\frac{R_1 C_1}{N}$	$E_{11=}\frac{R_1 C_2}{N}$
Wort 2≠u	$E_{21=}\frac{R_2 C_1}{N}$	$E_{22=}\frac{R_2 C_2}{N}$

Abbildung 6: erwartete Häufigkeiten (Evert, 2004)

Basierend auf der Kontingenztabelle mit tatsächlich beobachteter und erwarteter Häufigkeit können statistische Tests entwickelt werden. Auf die detaillierte Berechnung der Kennzahlen wird in dieser Arbeit nicht näher eingegangen. Die am häufigsten

angewendeten Assoziationsmaße sind Mutual Information, Likellihood-Ratio-Test, Log-Likelihood Poisson-Verteilung oder das Tanimoto-Maß.

4.3 Clusteranalyse

Durch Assoziationsmaße lässt sich auswerten, welche Wortpaare signifikant oft gemeinsam in Texten auftreten. Ausgehend von häufigen Wortpaaren können anschließend Clusterverfahren angewendet werden, deren Aufgabe es ist, Gruppen zu bilden, die innerhalb eine hohe Homogenität aufweisen und gegenüber anderen Gruppen eine hohe Heterogenität. Kennzeichnend für Clusterverfahren ist die Eigenschaft, dass die Gruppen nicht selbst definiert werden, sondern durch das Clusterverfahren selbst erstellt werden. Daher zählen sie zu den unüberwachten Verfahren und unterscheiden sich beispielsweise von Klassifikationsverfahren. (Bortz, 1999) Somit können etwa Substantive gruppiert werden, die häufig mit den gleichen Verben auftreten und sich daher vermutlich semantische Eigenschaften teilen. Die Bedeutung eines Clusters muss vom Forscher interpretiert werden. Clusterverfahren können in hierarchische und partitionierende Verfahren eingeteilt werden. Häufig angewendet wird der k-means Algorithmus.

4.4 Beurteilung des Verfahrens

Ein Vorteil des Co-occurence Verfahrens gegenüber des diktionären Ansatzes ist die Tatsache, dass kein Diktionär vor der Analyse erstellt werden muss. Daraus ergibt sich auch der Vorteil, dass eine weitgehende Unabhängigkeit von Sprachen gegeben ist. Die Kategorien werden durch das explorative Vorgehen selbst aus dem Text extrahiert. Der Aufwand ist dadurch als geringer einzuschätzen. Landmann & Züll können diesen Vorteil jedoch nicht bestätigen (Juliane Landmann, 2004), da eine umfassende Vorbearbeitung des Textes notwendig ist. Weiters wird kritisiert, dass es bei den angebotenen Computerprogrammen unzureichende Entscheidungshilfen zur Ergebnisinterpretation gibt. Das Co-occurence Verfahren wird demnach in erster Linie als Ergänzung zu diktionärbasierten Ansätzen empfohlen, indem die ermittelten Cluster als Grundlage zur Kategorienbildung dienen.

5 Überwachtes Lernen

Die wohl neusten Methoden zur potentiellen Automatisierung der Inhaltsanalyse, die in dieser Arbeit vorgestellt werden sollen, sind sogenannte Verfahren des überwachten Lernens. Der Oberbegriff dieser Kategorie ist das Maschinelle Lernen, dessen Prinzip es ist, dass ein künstliches System aus vorgegebenen Beispielen Gesetzmäßigen erkennt und diese anschließend auch auf unbekannte Daten anwenden kann. Durch Methoden des maschinellen Lernens sind Systeme also in der Lage, an Hand vorgegebener Daten zu verallgemeinern und zu lernen. Kennzeichnend für die hier betrachtete Methode des überwachten Lernens ist, dass der lernende Algorithmus für die Trainingsphase vorgegebene Beispiele benötigt, um anschließend auch unbekannte Daten automatisch klassifizieren zu können. Demgegenüber steht das unüberwachte Lernen, bei dem keine Beispieldaten erforderlich sind. Der Algorithmus versucht in diesem Fall selbst die Daten zu erklären, indem etwa Cluster gebildet werden, die Daten mit großer Ähnlichkeit zusammenfassen.
Um eine Inhaltsanalyse zu automatisieren, scheint die Methode des überwachten Lernens besonders attraktiv, da das Kategorienschema vor der Datenanalyse bereits feststeht.

5.1 Datenaufbereitung

Die Güte der Verfahren des überwachten Lernens ist zu einem großen Teil von der Aufbereitung der Texte abhängig. Grundsätzlich haben alle Vorbearbeitungsschritte das Ziel, die Texte so weit wie möglich zu vereinfachen, ohne dabei deren Bedeutung zu verändern. Teilweise ist diese Gratwanderung durchaus möglich, da natürliche Sprachen redundante Teile beinhalten. In den meisten Fällen entsteht durch das Reduzieren der Texte jedoch geringe semantische Veränderung bzw. nicht ausreichendes Vereinfachen der Texte verschlechtert die Güte des Verfahrens, sodass die Datenaufbereitung einen wichtigen Stellenwert bei der Entwicklung einnehmen sollte.
Die einzelnen Vorbearbeitungsschritte, die sich in der Praxis bewährt haben, sind im Rahmen des Kapitels 2.3.2 beschrieben. In einem Beispiel in Kapitel 7.3.3 wird überwachtes Lernen verwendet, um Kinokritiken automatisiert in positive und negative Kritiken zu klassifizieren. Dort werden auch verschiedene Möglichkeiten aufgezeigt, um Texte vor der Anwendung des Verfahrens zu vereinfachen. Es sei jedoch angemerkt,

dass kein standardisiertes Vorgehen existiert, wie Texte optimal aufbereitet werden. Dies hängt insbesondere vom Anwendungsfall ab und ist Gegenstand aktueller Forschung.

5.2 Verfahren des überwachten Lernens

Um überwachtes Lernen zu implementieren, existieren verschiedene Algorithmen, die auf statistischen und mathematischen Prinzipien beruhen. Alle Methoden des überwachten Lernens sind prinzipiell zur Kategorisierung von Texten verwendbar, doch haben sich für dieses Anwendungsgebiet einige Methoden besonders bewährt. (Carsten Felde, 2006) Die auf Grund ihrer guten Klassifikationsresultate am häufigsten eingesetzten Methoden sind Support Vektor Maschine und Naive Bayes. Deren detaillierte Funktionsweise ist nicht Schwerpunkt dieser Arbeit, weshalb sie nur kurz vorgestellt werden.

5.2.1 Support Vektor Maschine

Dieser Ansatz basiert auf der statischen Lerntheorie und zeichnet sich dadurch aus, sehr gut aus vorgegebenen Beispieldatensätzen Gesetzmäßigkeiten extrahieren zu können und dadurch zu generalisieren. Das Prinzip der Support Vektor Maschine ist, für die zu klassifizierenden Daten eine Hyperebene zu finden, sodass die vorgegebenen Kategorien optimal getrennt werden, d.h. ihr Abstand maximal ist. Die Methode der Support Vektor Maschine wird daher auch als „large margin classification“ bezeichnet.
Neben der bereits erwähnten hohen Generalisierungsfähigkeit ist dieses Verfahren auch in der Lage Klassifikationen vergleichsweise schnell durchzuführen, was bei sehr großen Textmengen ein Vorteil sein kann. Weiters ist die Analyse einer Vielzahl an Dimensionen problemlos möglich. Demgegenüber stehen die Nachteile, dass die Trainigsphase vergleichsweise lange dauert und bei Hinzunahme neuer Trainingsdaten der gesamte Trainingsprozess neu ausgeführt werden muss. Neben der Klassifikation von Textdokumenten werden Support Vektor Maschinen beispielsweise zur Bild- und Schrifterkennung oder in der Bioinformatik eingesetzt. (Eitrich, 2003)

5.2.2 Naive Bayes

Die Naive Bayes Methode basiert auf der Wahrscheinlichkeit, mit der ein Objekt zu einer Klasse gehört und somit auf dem Bayes-Theorem, welches die Berechnung

bedingter Wahrscheinlichkeiten beschreibt. Eine Grundannahme des Naive Bayes-Klassifikators ist die Unabhängigkeit zwischen den Attributwerten. Tatsächlich ist diese Annahme selten erfüllt, doch sie reduziert das Klassifizierungsproblem erheblich und da als Resultat die Klasse mit der höchsten Wahrscheinlichkeit gewählt wird, erzielt die Methode in der Praxis gute bis sehr gute Klassifizierungsergebnisse. (Hüftle, 2006)
Das Verfahren ist in der Textanalyse weit verbreitet, wo es durch seine hohe Trainings- und Klassifizierungsgeschwindigkeit gerne für zeitkritische Anwendungen eingesetzt wird. Im Gegensatz zur Support Vektor Maschine kann diese Klassifikationsmethode auch inkrementell lernen, d.h. bei Hinzunahme neuer Beispieldaten muss nicht die gesamte Trainingsphase wiederholt werden, sondern nur das Training für die neuen Daten. Ein Nachteil der Naive Bayes-Klassifikationen ist, dass sie bei hochdimensionalen Klassifikationsproblemen nicht mehr effizient sind. (Hüftle, 2006)

5.3 Beurteilung des Verfahrens

Verfahren des überwachten Lernens sind in der Lage auf Basis vorklassifizierter Beispieldatensätze Regeln abzuleiten und anschließend unbekannte Daten zu klassifizieren. Die Bereitstellung vorklassifizierter Daten für die Trainingsphase kann mit einem beträchtlichen Aufwand verbunden sein, worin auch der wohl größte Nachteil dieses Ansatzes liegt. Die Güte der Klassifikation übersteigt jedoch in der Regel jene des diktionärbasierten Ansatzes. Um eine gute Klassifikationsleistung zu erzielen, ist eine Vorbearbeitung der Texte unbedingt erforderlich. Nur durch die Vereinfachung der zu analysierenden Texte sind die Verfahren des maschinellen Lernens auch in der Lage in der Trainingsphase Regeln abzuleiten und zu generalisieren.

6 Opinion Mining

6.1 Gegenstand des Opinion Minings

Opinion Mining, oder auch Sentiment Analysis genannt, ist eine Unterdisziplin von Text Mining, die sich mit der automatisierten Extraktion von Meinungen und Stimmungen aus unstrukturierten Texten beschäftigt.
Ziel ist es, Meinungsäußerungen vorher definierten Kategorien (z.B. positiv, neutral, negativ) zuzuordnen. Erweitert kann die Zuordnung in Kategorien um die Angabe der Intensität der jeweiligen Stimmung. Opinion Mining ist vor allem in der

Marktforschung verbreitet, wenn es darum geht Foren, Blogs und soziale Netzwerke im Internet nach Meinungen zu Produkten zu durchsuchen. Im sogenannten Web 2.0, dem „Mitmach-Web“, spielt Dynamik und die Meinungsäußerung von Internetnutzern eine große Rolle. Meinungen zu aktuellen Themen oder Produkte von Unternehmen werden ununterbrochen ausgetauscht und so sammelt sich eine Vielzahl wertvoller Information an, die jedoch auf Grund ihrer unstrukturierten Form für Marktforschungen bislang nur schwer zugänglich war. (Lee, 2004)

Mit Hilfe von Opinion Mining kann sich die Marketingabteilung eines Unternehmens ein Bild davon machen, wie ein Produkt von den Benutzern angenommen wird, indem Internetforen automatisiert durchsucht werden. Je nach Komplexität der Verfahren werden als Ergebnis die Meinung verschiedener Eigenschaften und Kriterien des Produkts, sowie demographische Informationen der Autoren zurückgeliefert. So mögen etwa Käufer einer höheren Altersgruppe die Geräumigkeit eines neuen Automodells schätzen, während jüngere das Design des Modells bemängeln. Besonders interessant ist Opinion Mining auch für Wahlkämpfe in der Politik, wo Medien automatisiert nach Meinungen zu Wahlkampfauftritten durchsucht werden können.

Opinion Mining hat insbesondere auf Grund der großen Mengen und der hohen Aktualität von Texten im Internet sehr großes Potential. Besonders attraktiv ist die Verwendung von Texten aus Internetforen auch auf Grund der kostenfreien Verfügbarkeit und da die Meinungen in natürlichen Kommunikationssituationen ausgetauscht werden, sind die Inhalte sehr realitätsgetreu.

Im Bereich des Opinion Minings gibt es derzeit drei verschiedene Forschungsschwerpunkte (Kaiser, 2009):

- Stimmungsklassifikation: Hier werden Dokumente als Datenbasis verwendet und entsprechend ihrer Stimmung in Bezug auf bestimmte Objekte klassifiziert. Beispielsweise können auf diese Weise Kommentare von Kinofilmen oder Produktreviews analysiert werden.
- Eigenschaftsbasiertes Opinion Mining: Dieser Ansatz analysiert Texte auf Satzebene und versucht die Stimmung bezüglich Objekteigenschaften zu extrahieren. Das eigenschaftsbasierte Opinion Mining geht also einen Schritt weiter als die Stimmungsklassifikation und ist auch entsprechend komplexer und aufwändiger durchzuführen.

- Vergleichsbasiertes Opinion Mining: Das vergleichsbasierte Opinion Mining versucht Texte zu finden, in denen Objekte miteinander verglichen werden (z.B.: „Kamera A schießt schärfere Bilder als Kamera B“) um das bevorzugte Objekt zu identifizieren.

6.2 Herausforderungen von Opinion Mining

Das große Interesse an Opinion Mining in den letzten Jahren hängt vor allem mit den vielversprechenden Anwendungsmöglichkeiten zusammen. Dabei darf man die Herausforderungen und Schwierigkeiten nicht vergessen, die mit der Analyse von unstrukturierten Texten verbunden sind. Schließlich geht es etwa nicht bloß um die Ermittlung von Häufigkeiten bestimmter Worte in Texten, sondern um die semantische Analyse zur Ermittlung von positiven oder negativen Stimmungen. Die Ausgangssituation stellt sich für den Computer ähnlich dar, als würde ein Mensch einen Text in fremder Sprache vor sich haben und müsste ihn kategorisieren. Wie in Kapitel 2 erläutert wird, existieren verschiedene Verfahren, um dem Computer die „fremde Sprache“ hinsichtlich der geforderten Kategorisierung in positive und negative Stimmungen beizubringen.

Doch die Schwierigkeiten beginnen bereits vor der eigentlichen semantischen Textanalyse. In der praktischen Anwendung müssen zuerst Texte in zu untersuchende Einheiten strukturiert werden. Je nach Anwendung können die Einheiten einzelne Wörter, Sätze, Absätze, Kommentare, etc. sein. Es gilt also eine genaue Vorgangsweise zu finden, wie diese Texteinheiten erkannt werden können. Möchte man Texte etwa in einzelne Sätze strukturieren, so ist sicherlich der Punkt ein gutes Indiz für ein Satzende. Doch auch hier gibt es zahlreiche Ausnahmen, wie Abkürzungen oder Aufzählungen, die ebenfalls mit einem Punkt enden, jedoch kein Satzende bedeuten. Durch die Tatsache, dass Sätze mit einem Großbuchstaben beginnen, kann man die Trefferquote erhöhen. Bei der Strukturierung in Kommentare in Internetforen müssen wiederum andere Indizien zur Trennung der Textblöcke herangezogen werden, die von den Internetseiten abhängen können. Im besten Fall bieten Betreiber von Internetforen oder sozialen Netzwerken bereits Schnittstellen an, die die einzelnen Benutzerkommentare zurückliefern.

Nachdem die Texte in die gewünschten Einheiten strukturiert wurden, beginnt die semantische Analyse, die durch unterschiedliche Verfahren bewerkstelligt werden kann. Eine Herausforderung ist hier die Tatsache, dass Wörter in einer Situation eine positive und in einer anderen eine negative Stimmung wiedergeben. Betrachtet man das Wort „lang“ im Kontext der Lebensdauer eines Produktes, so handelt es sich eindeutig um eine positive Beschreibung. Bezieht sich das Wort „lang“ jedoch auf die Dauer eines Kinofilms oder die Wartezeit in einem Restaurant, so würde es eine negative Stimmung widerspiegeln. Daraus lässt sich einerseits schließen, dass die bloße Suche nach vorher kategorisierten Begriffen in Texten problematisch ist und andererseits, dass die Treffsicherheit von Opinion Mining erhöht werden kann, wenn sie auf ein Themengebiet konzentriert wird. (Pang & Lee, 2008)

Auch auf die Annahme, dass ähnliche Textbausteine auch ähnliche Meinungen widerspiegeln, kann nicht vertraut werden, denkt man etwa an die Verneinung von Aussagen. Obwohl die beiden Sätze „Der Kinofilm hat mir gut gefallen.“ und „Der Kinofilm hat mir nicht gut gefallen.“ syntaktisch sehr ähnlich sind, ist ihre Stimmung genau entgegengesetzt. (Barber, 2010)

Ein weiterer Grund, der die automatisierte Extraktion von Stimmungen aus Texten erschwert, ist die gleichzeitige Verwendung von positiven und negativen Ausdrücken in einzelnen Sätzen, betrachtet man etwa den Satz „Der Film war langweilig, obwohl der Schauspieler eine hervorragende Leistung ablieferte“. Selbst für Menschen ist es nicht immer einfach die Meinung aus Texten richtig abzuschätzen. Der Satz „Der Film war genauso gut wie sein letzter“ würde zusätzliches Hintergrundwissen erfordern. (Barber, 2010) Weitere Schwierigkeiten treten mit ironischen oder sarkastischen Aussagen auf. (Pang & Lee, 2008)

Gerade in Internetforen ist ein Schreibstil üblich, der sich durch Abkürzungen und einen eigenen Dialekt auszeichnet, was die Textanalyse weiters erschwert.

Es wird schnell deutlich, dass Opinion Mining-Verfahren auf Grund dieser großen Herausforderungen nie völlig fehlerfreie Ergebnisse liefern können. Ziel ist es vielmehr eine Trefferquote zu erreichen, die für eine bestimmte Anwendung nützliche Ergebnisse liefert.

6.3 Anwendungen von Opinion Mining

Die in dieser Arbeit bereits erwähnte Anwendung für Review-basierte Webseiten ist wohl das klassische Beispiel für Opinion Mining. Damit ist die Extraktion von Stimmungen aus Meinungen zu Produkten von Unternehmen oder Filmen gemeint. Aber man kann denselben Ansatz beispielsweise auch dafür verwenden, um Stimmungen gegenüber Politikern oder tagesaktuelle Themen zu analysieren. Für Webseiten, die neben textuellen Reviews auch eine quantitative Bewertung von Objekten ermöglichen, können Opinion Mining Verfahren dazu verwendet werden, Korrekturen vorzunehmen bzw. die Bewertungsqualität zu verbessern, wenn etwa die aus dem textuellen Review extrahierte Stimmung sehr stark von der quantitativ abgegebenen Bewertung abweicht. (Pang & Lee, 2008)

Eine weitere Anwendung ist die Verbesserung von „recommendation systems“, worunter im Internet eingesetzte Mechanismen verstanden werden, die dem Benutzer passende Produkte vorschlagen. Opinion Mining kann hier eingesetzt werden, um dem Benutzer bevorzugt positiv bewertete Objekte vorzuschlagen.

Für die Betriebswirtschaft und das Marketing eines Unternehmens bietet Opinion Mining eine Möglichkeit Verbesserungspotentiale für Produkte auszuloten. Um herauszufinden, warum sich Produkte nicht wie erwartet verkaufen, könnten selbstverständlich auch Umfragen verwendet werden. Doch wird es sich als schwierig erweisen, genau die Personen zu finden, die sich gegen den Kauf eines Produktes entschieden haben. Auch das manuelle Durchforsten von Diskussionsforen ist eine Möglichkeit, um sich einen Überblick über die Kritikpunkte eines Produktes zu verschaffen. Opinion Mining kann diesen Vorgang automatisiert übernehmen, und anschließend automatisch Verbesserungspotentiale für Produkte zurückliefern. Doch nicht nur für Unternehmen, auch für Regierungen lässt sich dieser Ansatz verwenden, um Unzufriedenheit rechtzeitig zu erkennen und gegenlenken zu können. Für die Sozialwissenschaften wiederum lassen sich daraus interessante Forschungsfragen ableiten, z.B. um herauszufinden, welche politische Themen oder Personen von welchen Personengruppen eher positiv oder negativ bewertet werden. (Pang & Lee, 2008)

Auch im wissenschaftlichen Bereich kann Opinion Mining zur Analyse von Zitierungen eingesetzt werden, indem aus dem Text extrahiert wird, ob ein Autor oder eine Quelle in positiven oder negativen Zusammenhang zitiert wird.

7 Empirische Untersuchung

7.1 Ziel der empirischen Untersuchung

Im empirischen Teil der Arbeit sollen die Ansätze, die in den vorherigen Kapiteln theoretisch beschrieben wurden, in praktischen Anwendungsfällen umgesetzt werden. Ziel der Untersuchung ist eine Einführung in die Implementierung der Methoden durch Verwendung verbreiteter und benutzerfreundlicher Computerprogramme. Im Vordergrund der Demonstration stehen die Vermittlung eines Einblicks in die praktische Umsetzung und die aus den Beispielen abgeleiteten Grenzen aber auch Potentiale der Verfahren. Ziel der empirischen Untersuchung und der praktischen Umsetzung der Verfahren ist es ausdrücklich nicht, zu möglichst verlässlichen Ergebnissen durch Verwendung der automatisierten Inhaltsanalyse zu kommen. Zu diesem Forschungsbereich existiert eine Reihe wissenschaftlicher Arbeiten, die unter Zuhilfenahme komplexer und kombinierter Methoden versuchen die Treffsicherheit zu erhöhen. Doch damit würde nicht nur der Rahmen dieser Diplomarbeit gesprengt werden, sondern auch trotz des interdisziplinären Anspruchs dieser Arbeit, das Wissenschaftsgebiet der Soziologie verlassen werden, da sich die sogenannten Text Mining Verfahren großteils Methoden der Informatik, der Statistik oder der Computerlinguistik bedienen.

Es wurden zwei unterschiedliche Untersuchungen durchgeführt bzw. Datensets herangezogen, die wiederum jeweils mit Hilfe von zwei unterschiedlichen Verfahren der automatisierten Textanalyse analysiert wurden. Die Datengrundlage des ersten Beispiels beinhaltet textuelle Bewertungen von Kinofilmen. Dieser Datensatz ist bereits in positiv und negativ Texte klassifiziert. Dieses Beispiel ist für Sentiment Analysen sehr beliebt und wurde auch in anderen wissenschaftlichen Arbeiten als Test verschiedener Verfahren herangezogen. Die Texte liegen in englischer Sprache vor, was auch bei der Durchführung der Untersuchung durch Anpassung der verwendeten Methoden berücksichtigt wurde.

Im zweiten Beispiel werden Texte aus Internetforen herangezogen, die ebenfalls in die Kategorien positiv und negativ klassifiziert werden sollen. Zur Überprüfung bzw. für das Training der Verfahren wurden die Texte durch manuelle Kodierer klassifiziert.

Auf beide Beispieldatensätze werden die in dieser Arbeit bereits theoretisch beschriebenen Ansätze angewandt. Dabei handelt es sich einerseits um den klassischen diktionärbasierten Ansatz und andererseits um die Methode des maschinellen Lernens.
Es sei noch einmal darauf hingewiesen, dass es sich hierbei um keine Gegenüberstellung der beiden Verfahren hinsichtlich Güte handelt. Für beide Verfahren existieren verschiedene Möglichkeiten der Verbesserung der Trefferquote, sowie auch die Möglichkeit der Kombination der beiden Verfahren. Vielmehr soll ein Einblick in die unterschiedlichen Funktionsweisen gewährleistet und Potentiale, sowie Grenzen der Verfahren aufgezeigt werden.

7.2 Kennzahlen zur Messung der Güte

Bevor mit der Vorstellung des diktionärbasierten Ansatzes und der Methode des maschinellen Lernens begonnen wird, werden Methoden vorgestellt, mit denen die Güte der Methoden gemessen werden kann. Die Aufstellung einer Confusion Matrix wird auch bei empirischen Untersuchungen angewandt, im Speziellen bei Experimenten, indem eine Kreuztabelle mit vorhergesagten und tatsächlichen Werten erstellt wird. Die Receiver Operating Characteristics ist ein grafisches Verfahren, das ebenfalls auf Berechnungen der Confusion Matrix basiert.

7.2.1 Confusion Matrix

Die Confusion Matrix ist ein Bewertungsmaßstab, der die aktuellen Klassifikationsergebnisse mit den vorhergesagten vergleicht und bewertet. Ein häufig zitiertes Beispiel einer Confusion Matrix ist die Bewertung eines Labortests, der feststellen soll, ob eine Person an einer bestimmten Krankheit erkrankt ist oder nicht. Wie auch der in dieser Arbeit verwendete Klassifikator soll der medizinische Labortest Fälle in zwei Gruppen einteilen, nämlich in „krank“ und „gesund“:

		Predicted	
		Negative	Positive
Actual	Negative	a	b
	Positive	c	d

Abbildung 7: Confusion Matrix

Die Confusion Matrix enthält nun die Anzahl vier unterschiedlicher Fälle:

- Richtig negativ (a): Die Anzahl richtiger Vorhersagen, wenn der tatsächliche Wert negativ ist
- Falsch positiv (b): Die Anzahl falscher Vorhersagen, wenn der tatsächliche Wert positiv ist
- Falsch negativ (c): Die Anzahl falscher Vorhersagen, wenn der tatsächliche Wert negativ ist
- Richtig positiv (d): Die Anzahl richtiger Vorhersagen, wenn der tatsächliche Wert positiv ist. (Claude Sammut, Geoffrey I. Webb, 2011)

Ausgehend von diesen 4 Werten können Kennzahlen berechnet werden, die die Güte der Klassifikation repräsentieren

Accuracy beschreibt das Verhältnis der korrekten Vorhersagen zur Gesamtanzahl der Fälle. Kommen mehr negative als positive Ergebnisse bei einem Experiment vor, so liefern diese Kennzahlen keine genauen Ergebnisse mehr. Accuracy unterscheidet nicht zwischen verschiedenen Fehlertypen, weshalb die Aussagekraft zu hinterfragen ist. Dennoch wird die Accuracy sehr häufig als Genauigkeitsmaß eingesetzt.

$$AC = \frac{a + d}{a + b + c + d}$$

Recall gibt den Anteil der positiven Fälle an, die korrekt identifiziert sind und wird auch als True Positive Rate (TP) bezeichnet.

$$TP = \frac{d}{c + d}$$

Die *False Positive Rate (FP)* gibt den Anteil der negativen Fälle an, die nicht korrekt identifiziert wurden.

$$FP = \frac{b}{a+b}$$

Die *True Negative Rate (TN)* gibt den Anteil der negativen Fälle an, die korrekt identifiziert wurden.

$$TN = \frac{a}{a+b}$$

Die *False Negative Rate (FN)* gibt den Anteil positiver Fälle an, die fälschlicherweise als positiv identifiziert wurden.

$$FN = \frac{c}{c+d}$$

Precision ist die Anzahl positiver Fälle, die richtig identifiziert wurden.

$$P = \frac{d}{b+d}$$

7.2.2 Receiver Operating Characteristic (ROC)

Im ROC-Graphen wird die True Positive Rate (FP) der False Positive Rate (FP) gegenübergestellt. Der ROC-Graph kann für Zwei-Klassen-Klassifikationsprobleme eingesetzt werden. Die Fläche unter der Kurve (area under curve) kann als Qualitätsmaßstab verwendet werden. Je größer die Fläche, desto besser der Klassifikator. Ist die ROC-Kurve der Diagonalen sehr nahe, so ist dies ein Indiz für eine zufällige Trefferquote. In diesem Fall wäre die True Positive Rate gleich der False Positive Rate. Bei korrekter Vorhersage, steigt die ROC-Kurve zunächst senkrecht (Fehlerquote=0%). Danach steigt die False Positive Rate an. Abbildung 8: ROC Curve skizziert die ROC Kurve schematisch, in Kapitel 7.3.3.8 wird sie zur Analyse der Klassifikationsgüte verwendet.

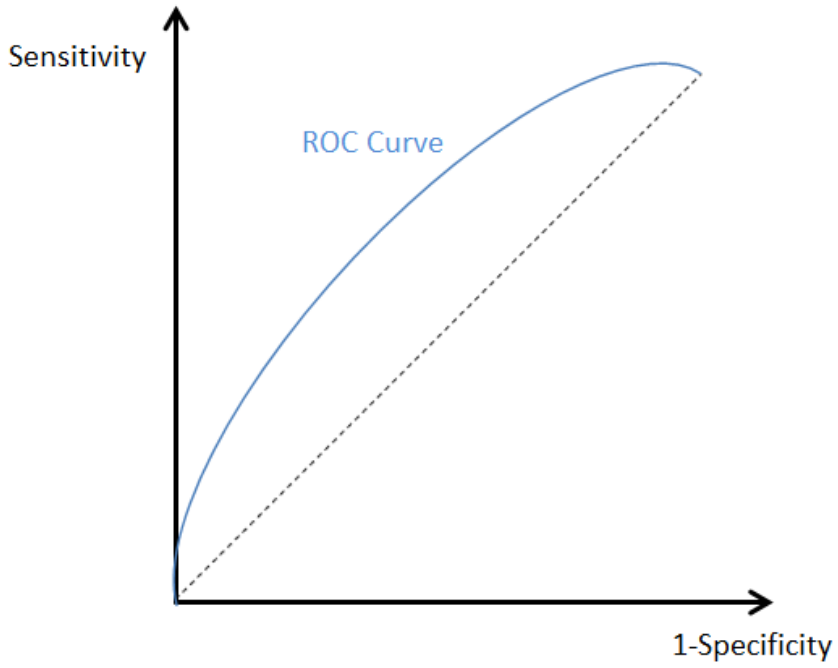


Abbildung 8: ROC Curve

Ein häufig zitiertes Bewertungsschema für die Interpretation der AUC ist weiter unten angeführt. Dieses kann freilich nur als Richtwert gelten. Die tatsächliche Bewertung der AUC ist vom jeweiligen Anwendungsfall abhängig. (Fawcett, 2003)

- .90-1 = excellent (A)
- .80-.90 = good (B)
- .70-.80 = fair (C)
- .60-.70 = poor (D)
- .50-.60 = fail (F)

7.3 Analyse von Movie Reviews

7.3.1 Beschreibung der Datenquelle

Datenquelle der Analyse ist eine im Internet frei verfügbare Auflistung von englischsprachigen Filmkritiken. (Lee, 2004) Die Reviews stammen von der

Filmdatenbank imdb[2] (International Movie Database). Dort können Benutzer Kritiken zu Kinofilmen verfassen und gleichzeitig eine quantitative Bewertung auf einer mehrstufigen Skala abgeben. Diese Bewertungsinformation wurde selbstverständlich aus den Reviews entfernt. Die Datenquelle umfasst 1000 positive und 1000 negative Kritiken.

Um die Reviews in positive und negative Texte zu klassifizieren, wurde folgendermaßen vorgegangen. Es wurde die quantitative Bewertung verwendet, die die Benutzer selbst angegeben haben. Voraussetzung zur Verwertung der Information ist die Angabe der Obergrenze des Ratings (z.B. 8/10 oder vier von fünf Sternen, etc.). Für verschiedene Rating-Systeme wurde dann folgende Auflösung definiert:

- Bei 5-Sterne-Systemen
    - Positiv: 3,5 und mehr Sterne
    - Negativ: 2 und weniger Sterne
- Bei 4-Sterne Systemen
    - Positiv: 3 und mehr Sterne
    - Negativ 1,5 und weniger Sterne
- Buchstabensystem
    - Positiv: B und mehr
    - Negativ: C- und weniger

Bevor die Datensätze ausgewertet werden, ist es empfehlenswert, sich durch stichprobenartiges Lesen der Texte einen Überblick zu verschaffen.

Folgendes Zitat stammt aus einer typischen Filmkritik, die einleitend mit der Beschreibung von Filmsequenzen beginnt.

"the movie opens with blackness , and only distant , alien-like underwater sounds . then it comes , the first ominous bars of composer john williams' now infamous score . dah-dum . from there , director steven spielberg wastes no time , taking us into the water on a midnight swim with a beautiful girl that turns deadly" (cv003_11664.txt)

[2] http://www.imdb.com

Da die hier kurz beschriebene Handlung des Films in keinem Zusammenhang zu der Bewertung bzw. der Stimmung der Filmkritik steht, aber dennoch Wörter enthält, die eine positive oder negative Stimmung vermittelt, ist diese wahrscheinlich eine mögliche Fehlerquelle für die automatisierte Inhaltsanalyse.

Typischerweise ist die eigentliche Bewertung des Films am Ende der Filmkritik zu finden, wie es auch hier der Fall ist:

"at least the make-up was realistic . with some emotional moments in the poorly written script , ghosts of mississippi lacked in heart , when its predecessor , a time to kill , brought tears to everyone's eyes . don't get me wrong , the movie wasn't all that bad , but if you've seen grisham's masterpiece , then don't expect this one to be an excellent film ." (cv498_8832.txt)

An diesem Beispiel lässt sich zeigen, dass eine Einteilung in die beiden Kategorien "positiv" und "negativ" auch bei manueller Kodierung nicht immer einfach möglich ist. Manche Texte enthalten eine differenziertere Bewertung, bei der man mit 2 Kategorien einer Stimmungsdimension an die Grenzen stößt.

Neben diesen Auszügen aus Filmkritiken existiert jedoch auch eine Reihe von Texten, die Stimmung relativ eindeutig vermitteln, indem klar positiv und negativ besetzte Wörter verwendet werden, wie dieses Zitat stellvertretend zeigt:

„even that aspect of the film fails , throwing in a convenient , ridiculous and unsatisfying wrap to things . it's been a while since i walked away from a movie theater in an angry mood . what makes it all the more remarkable is that i rarely remember a comedy making me so angry for wasting my time at it's ineptitude ." (cv164_23451.txt)

Der Datensatz „Movie Review Data" ist eine sehr beliebte und verbreitete Quelle für Tests von Sentiment Analysen. In vielen wissenschaftlichen Untersuchungen wird er zum Test von Methoden verwendet, da eine lange Liste von Vergleichsmaterial vorhanden ist.

Für diese Arbeit wurde die Datenquelle gewählt, da sich die Funktionsweise der verschiedenen Verfahren der Sentiment Analyse damit sehr gut demonstrieren lassen.
Movie Review Data kann frei von der Webseite Movie Review Data[3] heruntergeladen werden.

7.3.2 Diktionärbasiertes Verfahren

7.3.2.1 Computerunterstützte Inhaltsanalyse mittels Lexicoder

Lexicoder ist eine für nichtkommerzielle und akademische Zwecke frei verfügbare Software zur automatisierten Inhaltsanalyse mittels diktionärbasiertem Verfahren.
Der Einsatzbereich des Programms ist sehr begrenzt, diktionärbasierte Inhaltsanalysen wie sie für diese Arbeit entwickelt werden, lassen sich mit Lexicoder jedoch sehr gut durchführen.
Die Arbeit mit Lexicoder beginnt, indem das Diktionär und die auszuwertenden Dokumente eingefügt werden.

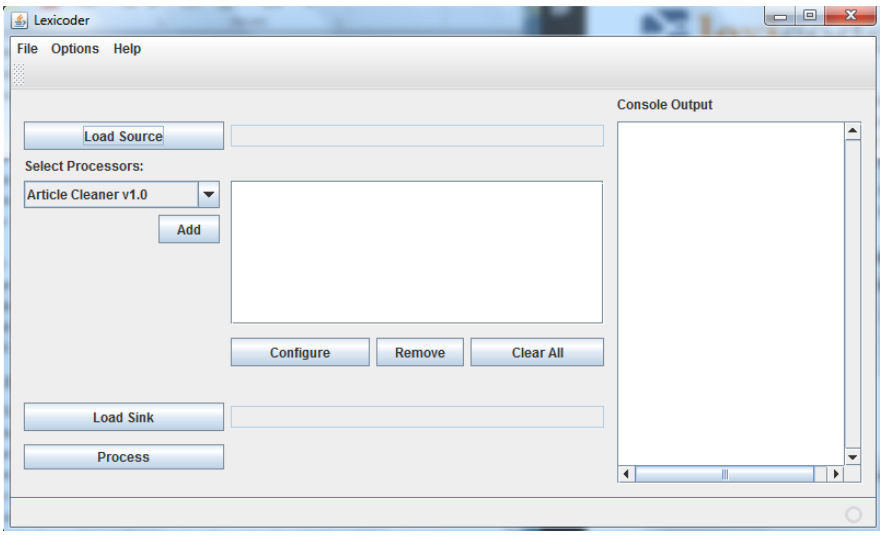


Abbildung 9: Lexicoder Oberfläche

Die Oberfläche von Lexicoder ist dem Funktionsumfang entsprechend relativ schlicht gehalten. Im Folgenden werden die für diese Diplomarbeit erforderlichen Komponenten erläutert und beschrieben wie mittels Lexicoder eine automatisierte Inhaltsanalyse durchgeführt wird.

[3] http://www.cs.cornell.edu/people/pabo/movie-review-data/

7.3.2.2 Installation

Lexicoder kann von der Seite Snsoroka[4] für akademische Zwecke kostenfrei bezogen werden. Dazu muss ein Anfrageformular ausgefüllt werden – das Programm wird innerhalb einiger Tage per Mail zugesandt. Für das Programm ist keine Installation erforderlich, es kann einfach gestartet werden und läuft auf allen gängigen Betriebssystemen, die Java unterstützen.

Die Durchführung einer automatisierten Inhaltsanalyse mittels Lexicoder kann grundsätzlich in vier Schritte eingeteilt werden:

- Load Source
- Select Processors
- Load Sink
- Process

Diese werden im Folgenden an Hand der in Kapitel 7.3.1 vorgestellten Filmkritiken beschrieben.

7.3.2.3 Load Source

Lexicoder erwartet sich die auszuwertenden Texte in einem bestimmten Format. So müssen die Texteinheiten in einer einzigen Textdatei mit zwei durch Tabulator getrennten Spalten enthalten sein. Die erste Spalte muss mit „ID“ beschriftet sein und die zweite Spalte mit „Body“. Jedem Text muss daher eine eindeutige Nummer zugeordnet werden. Eine Möglichkeit die Texte in die entsprechende Struktur zu überführen ist es, sie in die Zeilen eines Tabellenkalkulationsprogramms wie Excel zu kopieren und mit einer fortlaufenden Nummer zu versehen. Die Tabelle kann anschließend als „tab-delimited“ Textdatei abgespeichert und mit Lexicoder verwendet werden. Umfasst die Untersuchung eine sehr große Anzahl an Texten, so empfiehlt es sich die Texte mit Hilfe von spezieller Software automatisiert in das entsprechende Format zu bringen. Für diese Arbeit wurde das Hilfstool TXTCollector[5] verwendet, welches mehrere einzelne Textdateien in eine einzige Datei anfügt. Anschließend wurde das Resultat in ein Excel-Dokument kopiert, wo eine ID-Spalte mit aufsteigender Nummer hinzugefügt wurde.

[4] http://www.snsoroka.com/lexiKoder
[5] http://bluefive.pair.com/txtcollector.htm

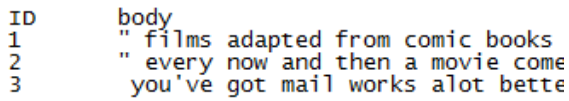


Abbildung 10: Lexicoder Format

7.3.2.4 Select Processors

Lexicoder bietet verschiedene Prozesse, mit denen die Texte vorbearbeitet oder analysiert werden können. Es sollen jene Prozesse beschrieben werden, die auch für diese Arbeit von Relevanz sind.

Der Prozess „Article Cleaner“ dient dazu, Punktationen in Texten zu entfernen. Dazu zählen Satzzeichen wie Punkte, Kommata, Fragezeichen, Klammern, Doppelpunkte etc. Diese können die Identifikation von Wörtern in Texten verhindern und dadurch die Auswertung verfälschen. Im Grunde entspricht die Funktion einem einfachen „Suchen und Ersetzen“. Werden Punktationen oder erwünschte Ersetzungen von diesem Prozess nicht abgedeckt, so können diese problemlos mit jedem herkömmlichen Textverarbeitungsprogramm vorgenommen werden.

Für die Durchführung der Analyse der Filmkritiken wurde der Vorbearbeitungsprozess „Article Cleaner“ aktiviert, und anschließend der „Dictionary Counter“.

7.3.2.5 Dictionary Counter

Das Diktionär besteht grundsätzlich aus Kategorien und Pattern, wobei Pattern den Kategorien untergeordnet sind und Indikatoren für die zugeordnete Kategorie darstellen. Auch hier erwartet sich Lexicoder eine bestimmte Struktur, in der das Diktionär vorliegen muss. Es handelt sich um das XML Format und kann am besten durch folgendes Beispiel verdeutlicht werden:

```
<?xml version= “1.0” encoding=UTF-8” standalone= “no”?>
<dictionary style= “Lexicoder” name= “Test Dictionary”>
<cnode name= “Animals”>
<pnode name= “fox” />
<pnode name= “cow” />
<pnode name= “dog” />
</cnode>
```

```
<cnode name= "Colours">
<pnode name="black" />
<pnode name="brown" />
<pnode name= "red" />
</cnode>
</dictionary>
```

Um Wortlisten in dieses Format zu überführen, empfiehlt sich ein Tabellenkalkulationsprogramm wie Excel. Die zusätzlichen Textelemente können damit einfach in die erste Spalte eingefügt werden, die zweite Spalte enthält die eigentliche Wortliste und die dritte Spalte die Zeichen „/>“, die das Ende eines Patterns darstellen. Anschließend kann der gesamte Textblock wieder in eine Textdatei kopiert werden und von Lexicoder verarbeitet werden.

Die Funktionsweise des „Dictionary Counters“ ist schnell beschrieben: Für jeden Text wird die Anzahl an Wörtern je Kategorie ermittelt. Doch gibt es Unterschiede wie diese Anzahl ermittelt wird. Lexicoder geht wie folgt vor:

Der Dictionary Counter arbeitet die Kategorien sequenziell ab. Wird ein Wort der ersten Kategorie im Text gefunden, so würde es in der zweiten Kategorie ignoriert werden. Existiert ein Wort also in mehreren Kategorien, so wird nur das Wort der ersten Kategorie verwendet.

In Lexicoder kann eingestellt werden, ob der Dictionary Counter „case-sensitive“ arbeiten soll, d.h. ob zwischen Groß- und Kleinschreibung unterschieden werden soll. Für die meisten Fälle ist es anzuraten nicht zwischen Groß- und Kleinschreibung zu unterscheiden, es sei denn es ist für die Analyse relevant.

Eine Besonderheit von Lexicoder, die bei der Textanalyse bedacht werden muss, ist dass Wörter aus dem Diktionär auch dann in Texten gefunden werden, wenn sie Teil einer Zeichenfolge sind.

Ist beispielsweise im Diktionär das Pattern

```
<pnode name= "gut" />
```

enthalten, so werden die Wörter “gute”, “guter” oder “gutes” gefunden. In diesem Fall kann das durchaus erwünscht sein, da die aufgelisteten Wörter von dem Wort aus dem Diktionär abgeleitet sind und ebenfalls eine positive Stimmung wiedergeben. Doch auf die gleiche Weise wird etwa auch das Wort „ungut“ gefunden, was auf Grund der gegensätzlichen Bedeutung des Wortes unerwünscht ist. Diese Problematik kann jedoch einfach gelöst werden, indem ein führendes Leerzeichen an den Eintrag im Diktionär angefügt wird.

```
<pnode name= " gut" />
```

Entsprechend werden ausschließlich Wörter gefunden, die mit „gut“ beginnen.
Für die Durchführung dieser Arbeit wurde ein Diktionär für Sentiment Analysen verwendet, das auf der Lexicoder-Seite verfügbar ist. Es enthält 4567 Patterns für die beiden Kategorien „positiv“ und „negativ“.
Ein Auszug des Diktionärs der positiven Kategorie ist hier angeführt:

```
<pnode name=" GOLDEN"></pnode>
<pnode name=" GOOD "></pnode>
<pnode name=" GOODIE"></pnode>
<pnode name=" GOODNESS"></pnode>
```

Wie zu erkennen ist, wurden wie erläutert auch Leerzeichen angefügt bzw. angehängt, wenn das Wort nicht als Teil eines Wortes gefunden werden soll.

Um Negationen zu erkennen, wird auch ein eigenes Diktionär mitgeliefert, das die negierten Patterns enthält, d.h. mit dem Wort „NOT“ davor.

```
<pnode name=" NOT GOLDEN"></pnode>
<pnode name=" NOT GOOD "></pnode>
<pnode name=" NOT GOODIE"></pnode>
<pnode name=" NOT GOODNESS"></pnode>
```

Entsprechend ergeben sich in diesem Diktionär die Kategorien “Neg_Positiv” für negierte positive Wörter und “Neg_Negativ” für negierte, negative Wörter, die durch die doppelte Verneinung einer positiven Stimmung entsprechen. Ein Beispiel für die letzte Kategorie wäre etwa „NOT BAD“.

Um dieses anzuwenden, muss ein zweiter Durchlauf der Analyse durchgeführt werden und anschließend die als negiert erkannten Wörter in die Berechnung miteinbezogen werden. Dazu wird folgende Formel verwendet.

Anzahl der positiven Wörter

```
Normal(Positiv)-Negiert(Neg_Positiv)+Negiert(Neg_Negativ)
```

Anzahl negativer Wörter

```
Normal(Negativ)-Negiert(Neg_Negativ)+Negiert(Neg_Positiv)
```

Wobei Normal das Ergebnis der Inhaltsanalyse bei Verwendung des normalen, nicht-negierten Diktionärs und Negiert das Ergebnis bei Verwendung des negierten Diktionärs bedeutet. Die Verwendung des negierten Diktionärs wird in dieser Arbeit in einem separaten Durchgang herangezogen.

7.3.2.6 Load Sink

Mittels der Load Sink Option wird die Zieldatei angegeben, in die das Ergebnis der Analyse gespeichert werden soll. Die Datei kann mittels einer herkömmlichen Textbearbeitungssoftware geöffnet werden

7.3.2.7 Process

Durch einen Klick auf den Process-Button wird schließlich der gesamte Prozess mit den vorgenommenen Einstellungen gestartet und das Ergebnis der Textanalyse in die Zieldatei geschrieben. Bis alle Filmkritiken vollständig analysiert sind, benötigt Lexicoder einige Minuten. Das Ergebnis wird im nächsten Kapitel erläutert.

7.3.2.8 Resultat

Es fanden zwei Durchläufe der Analyse statt. Der erste Durchlauf verwendete das normale, nicht negierte Diktionär. Im zweiten Durchlauf wurde das negierte Diktionär verwendet. Dementsprechend werden zwei Resultate ausgewiesen.

1 .Durchgang

In der ersten Auswertung wurde das nicht negierte Diktionär verwendet und auf 1000 positive und 1000 negative Filmkritiken angewandt. Es wurden anschließend die richtig vorhergesagten Stimmungen gezählt und in die Confusion Matrix eingetragen.

		Predicted	
		Negative	Positive
Actual	Negative	681	319
	Positive	361	639

Abbildung 11: Confusion Matrix 1 (diktionärbasierter Ansatz)

$$Accuracy = \frac{681 + 639}{681 + 319 + 361 + 639} = 66{,}0\%$$

Die Accuracy ist eine Maßzahl dafür, wie richtig die Texte klassifiziert wurden. Es wird die Anzahl der richtig qualifizierten Texte durch die Gesamtanzahl der Texte dividiert. Wenn man bedenkt, dass eine zufällige Vorhersage ein Ergebnis von 50% liefern würde, so ist eine Accuracy von 66% ein Wert, der für viele Anwendungen zu gering sein wird. Dennoch ist er deutlich über dem Zufallswert, was zumindest das Prinzip der Methode verdeutlicht. Es sei an dieser Stelle noch einmal betont, dass im Rahmen dieser Arbeit nicht versucht wurde die Trefferquote der Methode zu optimieren. Dazu existiert eine Reihe von erfolgreichen, aber zum Teil wesentlich komplexeren Ansätzen.

2 .Durchgang: negiertes Diktionär

Zusätzlich wurde die Analyse unter Verwendung eines negierten Diktionärs wiederholt. Alle Pattern enthalten in diesem ein vorangestelltes „NOT „. Anschließend wurde

ausgehend vom Ergebnis des ersten Durchlaufs mit Hilfe der in Kapitel 7.2 angeführten Formeln die korrigierten Werte ermittelt.
Ein Beispiel soll die Funktionsweise des negierten Diktionärs verdeutlichen. Das negierte Diktionär enthält unter der Kategorie „negativ“ den Pattern:

```
<pnode name=" NOT WORTH "></pnode>
```

Im folgenden Zitat einer Filmkritik wird der Pattern richtig erkannt und als negativer Eintrag erkannt:

“the bachelor' is a painfully clumsy mess strung together with a few brief moments of surprising poignancy . awaiting these moments by enduring the rest of the film is certainly ***not worth*** *your time or money”*

Die im ersten Durchgang ermittelte Anzahl an positiven und negativen Wörtern muss daher korrigiert werden, indem 1 der Anzahl an positiven Wörtern subtrahiert und zur Anzahl an negativen Wörtern addiert wird.

		Predicted	
		Negative	Positive
Actual	Negative	687	313
	Positive	361	639

Abbildung 12: Confusion Matrix 2 (diktionärbasierter Ansatz)

$$Accuracy = \frac{687 + 639}{681 + 313 + 361 + 639} = 66{,}3\%$$

An der Confusion Matrix lässt sich ablesen, dass bei Berücksichtigung von Negationen bei den Texten mit negativer Stimmung 6 Texte korrigiert und als positiv klassifiziert wurden. Bei den Texten der positiven Klasse gab es keine Änderung.

Die Accuracy konnte dadurch von 66,0% auf 66,3% erhöht werden, was einer minimalen Verbesserung entspricht. In anderen wissenschaftlichen Arbeiten wurden ähnliche Beobachtungen gemacht (Dadvar, Hauff, & de Jong, 2011).

7.3.2.9 Analyse der Fehlkategorisierungen

Um die Ursache für die relativ hohe Fehlerquote zu finden, werden stichprobenmäßig falsch klassifizierte Filmkritiken ausgewählt und näher untersucht. Bei der Analyse der Fehlkategorisierungen lassen sich eindeutig Gemeinsamkeiten beobachten, die stellvertretend an Hand einiger Beispiele verdeutlicht werden sollen.

Bezüge auf andere Filme

*„because i absolutely **hated** the previous ryan/hanks teaming , sleepless in seattle"*

Einige Filmkritiken beinhalten Bezüge auf andere Filme desselben Regisseurs oder Schauspieler und bewerten diese gleichzeitig. In diesem Fall wird dieser negativ bewertet - das Wort „ hate" ist im Diktionär vorhanden, wodurch dieser Satz als „negative Stimmung" klassifiziert wird. Der eigentliche Film wurde vom Autor der Filmkritik jedoch positiv bewertet. Diese Fehler stellen eine große Herausforderung dar - für den diktionärbasierten Ansatz ist es so gut wie unmöglich zu erkennen, worauf sich ein Satz des Autors tatsächlich bezieht.

Filmzusammenfassungen und Slang-Ausdrücke, Tippfehler

*"**kicking** freshman **ass** and trying to impress the muff . also hanging around these **losers** is babs , future universal studios employee and serious **bitch** . now let's just take a peak next door , at the delta house . over here , anything goes : you wanna throw **shit** out the window ? okay . you wanna **crush** a bunch of beer cans"*

Ein sehr häufiges Problem bei der Analyse von Filmkritiken ist die Tatsache, dass Filmkritiken in den meisten Fällen mit einer kurzen Zusammenfassung des Films beginnen. In diesem Beispiel kommt noch eine vulgäre Sprache des Autors hinzu, wodurch in einem kurzen Absatz 6 Wörter gezählt werden, die im Diktionär als

„negativ“ klassifiziert sind. Da in der Untersuchung die Stimmung bzgl. des Films gemessen werden soll, und nicht die Stimmung, die der Inhalt des Films transportiert, kommt es in diesem Fall zu einer eindeutigen Fehlklassifikation. Die eigentliche Bewertung des Films wird vom Autor erst im letzten Satz der Kritik wiedergegeben:

„ this is the fucnniest movie int he history of the world . do yourself a favor and go see it ."

Die tatsächliche Bewertung des Films ist also positiv. In diesem Fall wurde ausgerechnet beim positiven Wort „funniest“ ein Tippfehler eingebaut, sodass selbst dieses nicht korrekt klassifiziert werden kann. Tipp- und Rechtschreibfehler konnten bei der stichprobenmäßigen Untersuchung der Fehlkategorisierungen jedoch nicht als häufiges Problem ausgemacht werden.

Negationen

Im 2. Durchlauf der Untersuchung wurde das negierte Diktionär verwendet, um eine minimale Verbesserung zu erzielen. Dass nicht alle Negationen richtig erkannt wurden, liegt zum Teil an der Verwendung von Füllwörtern zwischen dem Negationswort „not“ und dem stimmungstragenden Wort. Einige Beispiele aus den Filmkritiken verdeutlichen dies:

“it's certainly not a bad”, “not much ambition”, “it's not a fatal misstep”

Das Diktionär enthält allerdings nur Pattern ohne Füllwort, wie etwa “not bad”. Ein Ansatz diese Fehlerquelle zu minimieren, ist die Verwendung einer „Window size“, die die Wirkungsreichweite des Negationswortes durch die Anzahl an Wörtern angibt. (siehe Kapitel 3.3.2). Das weiters Zitat aus einer Filmkritik *„isn't very funny“* beinhaltet das Negationswort „not“ als Abkürzung. In diesem Fall wurde die Negation jedoch auch auf Grund des Füllwortes „very“ nicht erkannt. Ein Lösungsansatz wäre auch abgekürzte Versionen von Negationswörtern in das Diktionär aufzunehmen.

Doch stellen Negationen (Dadvar, Hauff, & de Jong, 2011) eine vergleichsweise geringe Fehlerquelle dar, sodass durch die Verwendung einer „Window size“ und Erweiterung des Diktionärs nur minimale Verbesserungen zu erwarten sind.

Problematische Formulierungen

Folgendes Beispiel zeigt, dass negative Stimmungen auch vermittelt werden können, ohne dabei ein einziges negativ besetztes Wort zu verwenden:

“so why aren't we seeing storyboards of those segments (especially the parody moments from the matrix) ? i wonder what's funnier : the fact that any moment in this film required storyboards or that di$ney thought fans of this film would want to see them ? deleted scenes ? sure . audio commentary ? you betcha . but storyboards ? come on .”

Für den diktionärbasierten Ansatz kann hier keine negative Stimmung erkannt werden. Vergleichbar ist diese Fehlerquelle mit der Verwendung von ironischen Formulierungen oder Sarkasmus.

Ein Großteil der Fehlkategorisierungen beruht darauf, dass sich Stimmungen nicht auf den eigentlichen Film beziehen. Beinahe jede Filmkritik beinhaltet typischerweise eine kurze Zusammenfassung des Inhalts, die wiederum positive oder negativ beladene Wörter enthalten kann. Überwiegt eine Inhaltsangabe anteilsmäßig in der Filmkritik, so ergeben sich zwangsläufig Fehlkategorisierungen bzw. ist die Kategorisierung dann vom Zufall abhängig, je nachdem ob positive oder negativ beladene Wörter in der Zusammenfassung verwendet wurden. Bezüge auf andere Filme oder Themen sind ebenfalls eine Fehlerquelle, doch treten diese meist nur in einzelnen Sätzen auf und fallen daher nicht im selben Ausmaß in das Gewicht. Die erstgenannte Fehlerquelle war bei dieser Untersuchung mit Abstand die Größte, und gleichzeitig auch die wohl am schwersten zu beseitigen.

Für die weiter genannten Fehlerquellen der Negationen, Tippfehler und Rechtschreibfehler existieren bereits Ansätze diese zu erkennen. Doch treten sie

vergleichsweise selten auf, sodass hier nur eine minimale Verbesserung erwartet werden kann.
Neben komplexen Möglichkeiten die hier genannten Fehlerquellen zu erkennen, besteht selbstverständlich auch die Möglichkeit das Diktionär zu verbessern, beispielsweise indem es für einen speziellen Kontext, wie in diesem Fall Filmkritiken, optimiert wird.

7.3.3 Überwachtes Lernen

7.3.3.1 RapidMiner

Zur Durchführung der automatisierten Inhaltsanalyse mittels überwachten Lernens im Rahmen der Diplomarbeit wurde ein Programm gesucht, das einerseits frei verfügbar und andererseits leicht erlernbar ist. Es wurde auch bewusst ein Ansatz gesucht, der die Ergebnisse nicht vollautomatisiert berechnet. Der gesamte Ablauf und sämtliche Aufbereitungsschritte sollen manuell konfigurierbar sein, um die Funktionsweise der Methode demonstrieren zu können.
Nach intensiver Recherche wurde RapidMiner aus dem Open Source Bereich als geeignete Software ausgewählt. RapidMiner wurde seit dem Jahr 2001 vom Lehrstuhl für künstliche Intelligenz der Technischen Universität Dortmund entwickelt und wurde ursprünglich unter dem Namen YALE (Yet Another Learning Environment) geführt. Die Anwendungsfelder liegen in den Bereichen Data Mining, Text Mining, maschinelles Lernen und weiteren Formen der Wissensentdeckung.
RapidMiner wurde zur Durchführung des praktischen Teils der Diplomarbeit herangezogen, da keinerlei Programmierkenntnisse erforderlich sind um Text Mining Prozesse durchzuführen. Die gesamte Automatisierung der Inhaltsanalyse kann durch Zusammensetzen grafischer Elemente durchgeführt werden.
In diesem Kapitel soll das Programm RapidMiner vorgestellt, erforderliche Voraussetzungen beschrieben werden, so wie die Implementierung des Prozesses zur automatischen Durchführung von Inhaltsanalysen genau erläutert werden.

7.3.3.2 Installation

RapidMiner kann von der Webseite RapidMiner[6] heruntergeladen und installiert werden. Es läuft unter der Affero GNU Public License und kann somit z.B. für diese wissenschaftliche Arbeit frei verwendet werden.

Für die Anforderung dieser Arbeit muss die Erweiterung „Text Processing“ installiert werden, was nach der Installation im Menü „Help\Manage Extension“ bewerkstelligt werden kann.

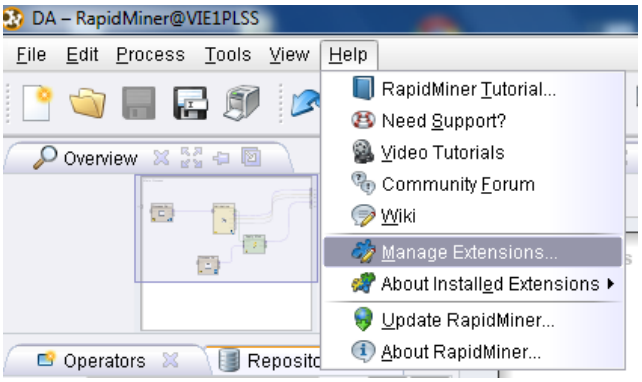


Abbildung 13: RapidMiner Erweiterungen

Die „Text Processing“ – Erweiterung bietet zusätzliche Funktionen zur Textverarbeitung, die zur Durchführung der automatisierten Inhaltsanalyse benötigt werden.

7.3.3.3 Prozessmodellierung

Bevor mit der Modellierung des Prozesses begonnen werden kann, muss ein neues Projekt angelegt werden. Dazu wird der Menüpunkt File/New gewählt. Es erscheint die Hauptoberfläche von RapidMiner mit einem noch leeren Arbeitsbereich. Auf diesen können Operatoren aus der linken Bildschirmhälfte gezogen und aneinandergereiht werden.

[6] http://sourceforge.net/projects/rapidminer/

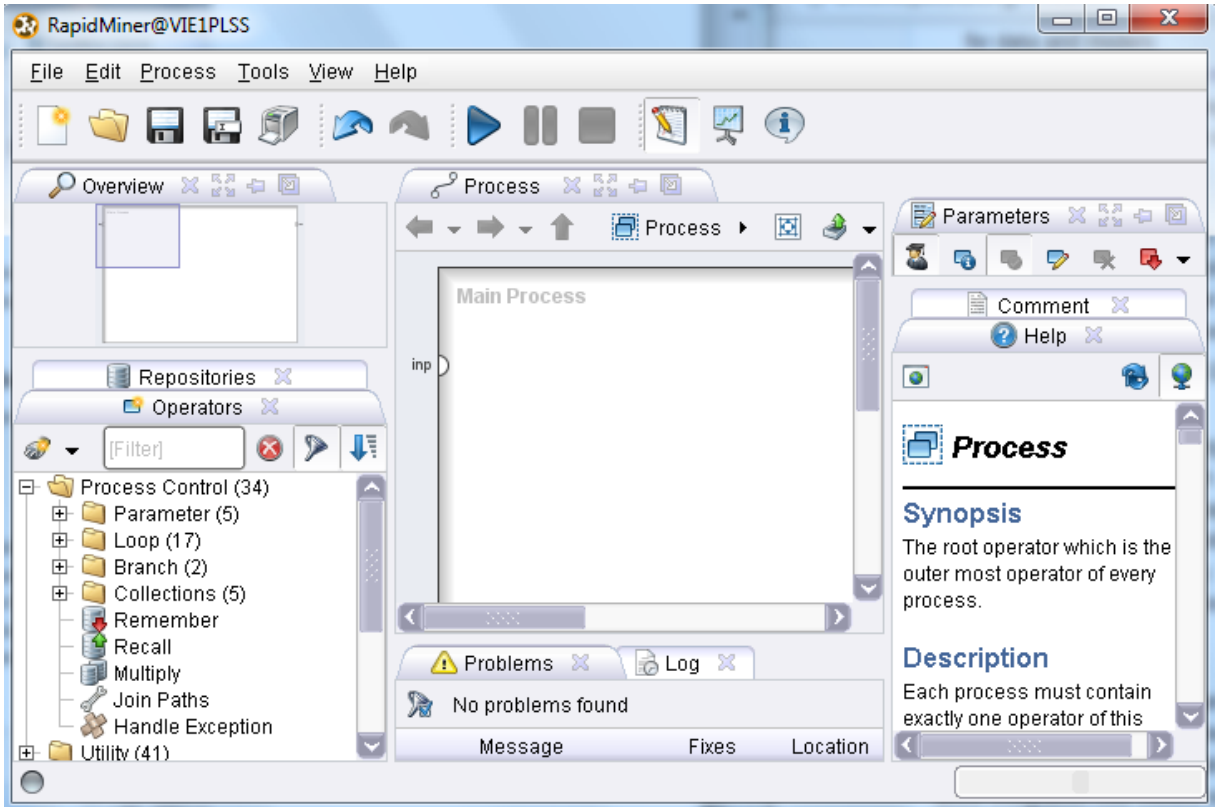


Abbildung 14: RapidMiner Oberfläche

Zu den Operatoren werden automatisch Erklärungen angezeigt und Einstellungsmöglichkeiten beschrieben. Dadurch eignet sich das Programm auch sehr gut für unerfahrene Anwender.

Nachdem der Prozess fertig modelliert wurde, kann er mit der Schaltfläche „Run Process“ gestartet werden.

7.3.3.4 Datenquelle

Wie später noch genauer erläutert wird, basiert die angewandte Methode auf maschinellem Lernen. D.h. dem Prozess werden Beispielfälle von positiven und negativen Texten zur Verfügung gestellt, inklusive der Information welche der Fälle eine positive oder negative Stimmung beinhalten. Dadurch kann ein Modell abgeleitet werden, das auch neue Texte in positive und negative Texte klassifiziert.

Die Beispieltexte werden in zwei Verzeichnisse „neg“ und „pos“ eingefügt.

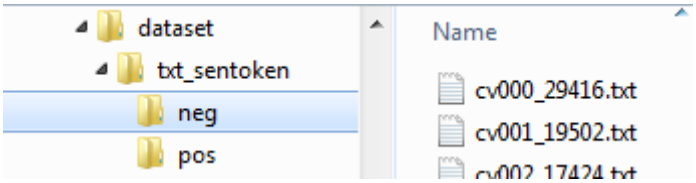


Abbildung 15: RapidMiner Filmkritiken

Die Texte in den Verzeichnissen können später im Text Mining Prozess ausgelesen werden und als Trainingsdaten verwendet werden.

7.3.3.5 Design des Text Mining Prozesses

Im Design-Fenster von RapidMiner erfolgt die Implementierung des Textverarbeitungsprozesses. Hier werden mit verschiedenen Bausteinen Text-Dokumente eingelesen, aufbereitet, ein maschineller Lernalgorithmus angewandt und ein Modell entwickelt und schließlich die Güte des Modells getestet. Das Bild zeigt den Hauptprozess der Implementierung. Die einzelnen Elemente des Prozesses sollen im Folgenden näher beschrieben werden.

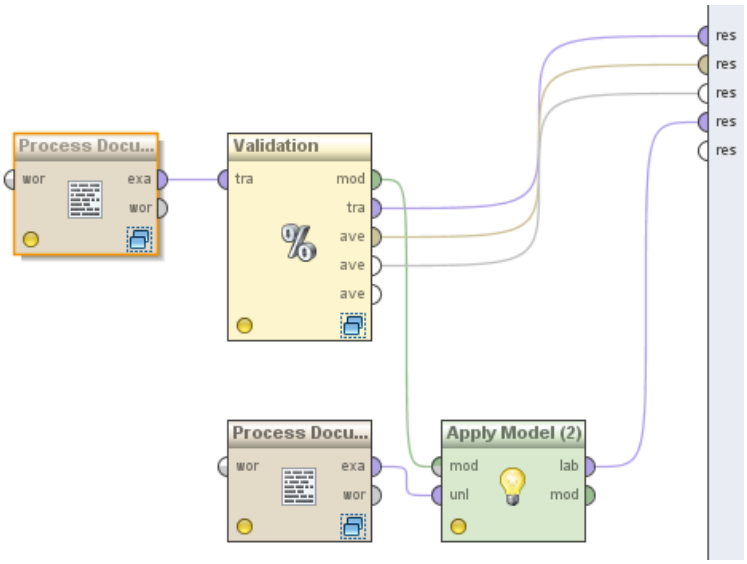


Abbildung 16: Design des Text Mining-Prozesses

7.3.3.6 Process Document

Der erste Schritt des Prozesses dient dem Einlesen von Text-Dokumenten. Weiters werden verschiedene Aufbereitungsschritte durchgeführt, die die Texte vereinfachen, um ein möglichst generisches Klassifikationsmodell erstellen zu können.

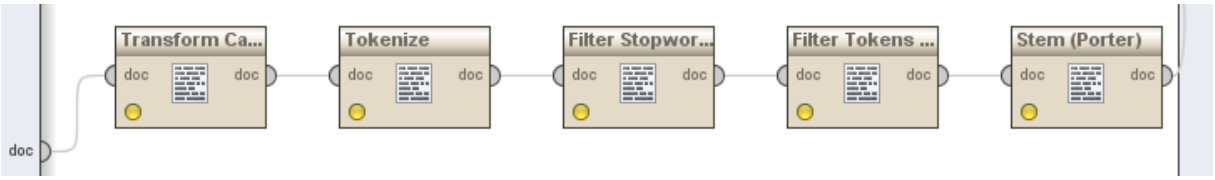


Abbildung 17: Process Document

Transform Cases

Hier werden alle Texte in Kleinbuchstaben umgewandelt. Groß- und kleingeschriebene Wörter werden dadurch im Modell gleichwertig betrachtet. Das Modell wird dadurch verbessert, ohne dass darunter die Bedeutung der Texte nennenswert leidet.

Tokenize

Anschließend erfolgt die Tokenisierung der Texte. Für Menschen ist die Unterscheidung der Wörter offensichtlich. Für das Computerprogramm müssen jedoch Regeln definiert werden. Für diese Arbeit wurde die Methode „non letters“ gewählt, d.h. nach jedem Zeichen, das kein Buchstabe ist, wird ein neues Wort angenommen. Im häufigsten Fall handelt es sich dabei um das Leerzeichen. Ein Token könnte auch eine andere Texteinheit wie etwa einen Satz oder einen Absatz bezeichnen. In unserem Fall wurden Wörter als Token verwendet.

Filter Stopwords

Texte bestehen zu einem großen Teil aus Wörter, die wenig oder keinen Einfluss auf die Bedeutung des Textes haben. Diese Füllwörter würden die Qualität des Modells verschlechtern und sollten daher vor der Erstellung des Modells ausgeschlossen werden. Es gibt für verschiedene Sprachen bewährte Stoplists, die frei zur Verfügung stehen. Die hier verwendete Stoplist enthält u.a. Wörter wie and, are, because, at, etc.

Filter Tokens

Zur Verbesserung der Datenqualität können hier sehr kurze oder sehr lange Zeichenfolge ausgeschlossen werden. Dabei könnte es sich beispielsweise um Tippfehler oder Importfehler von Textdaten handeln. In diesem Prozess werden nur Wörter mit mindestens 2 und maximal 50 Zeichen herangezogen.

Stem (Porter)

Stemming ist zum Aufbau des Klassifikationsmodells eine sehr wichtige Vorbearbeitung. Wörter werden durch diesen Algorithmus auf ihren Wortstamm verkürzt, indem sie auf eine Minimalanzahl von Silben verkürzt werden. Die auf den Wortstamm reduzierten Wörter müssen dabei nicht dem linguistisch korrekten Wortstamm entsprechen. Vielmehr ist es Ziel des Algorithmus, semantisch verwandte Wörter auf eine gemeinsame Zeichenkette zurückzuführen. Dieser Idealfall kann nicht immer fehlerfrei erreicht werden. Ein guter Stemming-Algorithmus findet einen Mittelweg zwischen Overstemming (zu viel wird abgeschnitten) und Understemming (zu wenig wird abgeschnitten).

7.3.3.7 Validation

Im Validation-Baustein sind einerseits der Algorithmus zur Erstellung eines Modells und andererseits ein Mechanismus zur Überprüfung bzw. Validierung des Modells enthalten. Die Abbildung zeigt die Trennung dieser Schritte in die Bereiche „Training“ und „Testing“

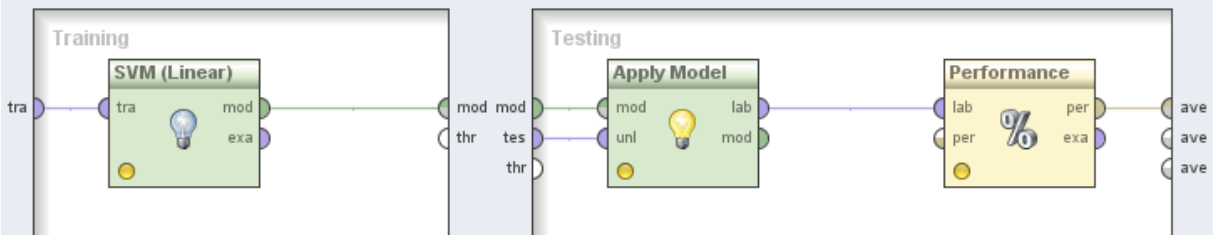


Abbildung 18: Validation

Die in den Prozess importierten Texte werden in Trainings-Datensätze und Test-Datensätze aufgeteilt. Die Trainingsdatensätze dienen der Erstellung des Modells.

Hierfür wird auch die Information, ob es sich um einen Text mit positiver oder negativer Stimmung handelt, berücksichtigt. Das Modell kann durch diese Beispieltexte lernen, positive von negativen Stimmungen in Texten zu unterscheiden und anschließend auch neue Texte in die Kategorien „positiv“ und „negativ“ klassifizieren. Um die Qualität der Klassifizierung zu testen, werden die Test-Datensätze herangezogen. Das Ergebnis des Validation-Bausteins ist ein Performance-Vektor, der verschiedene Kennzahlen zur Qualität der Klassifizierung enthält.
Die Auswahl der Test- und Trainingsdatensätze kann auf verschiedene Wege erfolgen. Für diese Arbeit wurde die Einstellung „Stratified sampling“ verwendet. Diese Methode wählt zufällig einzelne Texte aus und teilt sie in Trainings- und Testdatensätze ein, wobei darauf geachtet wird, dass die Verteilung in den beiden Stichproben der der Grundgesamtheit entspricht.

SVM (Linear)

Als Klassifikationsmethode wurde die Support Vektor Maschine ausgewählt, die im Bereich Text Mining häufig zum Einsatz kommt. Die Support Vektor Maschine erwartet Trainingsdaten, für die die zugehörigen Klassen bereits bekannt sind. Diese werden vom Validations-Baustein im vorherigen Schritt zufällig ausgewählt. Wie bereits in Kapitel 5.2.1 beschrieben, besteht die grundlegende Funktionsweise darin, dass jeder Datensatz durch einen Vektor repräsentiert wird. Die Support Vektor Maschine versucht nun eine Hyperebene in diesem Raum zu finden, die die Trainingsdaten in Klassen einteilt. Ziel ist es, den Abstand der Vektoren, die der Hyperebene am nächsten liegen, zu maximieren. Die detaillierte Vorgehensweise und mathematische Erklärung der Funktionsweise ist weitaus komplexer, doch ist die genaue Funktionsweise der Methode nicht Schwerpunkt dieser Arbeit. Weiters existiert eine Reihe von Parametern, die die Funktionsweise und den Lernvorgang der Support Vektor Maschine beeinflussen. Für die vorliegende Arbeit wurde mit den Standardeinstellungen gearbeitet.

Apply Model

Der nächste Baustein “Apply Model” befindet sich im Bereich “Testing”. Das auf Basis der Trainingsdatensätze erstellte Modell wird nun auf die Testdatensätze angewandt.

Resultat sind durch das Modell klassifizierte Datensätze bzw. in konkret diesem Fall Texte, die mit den Labels „positiv“ oder „negativ“ klassifiziert wurden.

Performance

Das Performance-Element berechnet verschiedene Performance-Kennzahlen, die die Güte des Modells beschreiben. Dazu wird die Klassifikationsleistung des Modells ermittelt, indem die vorgenommene Klassifikation mit der tatsächlichen verglichen wird. Der Baustein erwartet sich dementsprechend als Eingabeparameter Testdatensätze mit den Attributen „vorhergesagtes label“ und „tatsächliches label“.
Für binomiale Klassifikationsprobleme werden folgende Kennzahlen ermittelt:

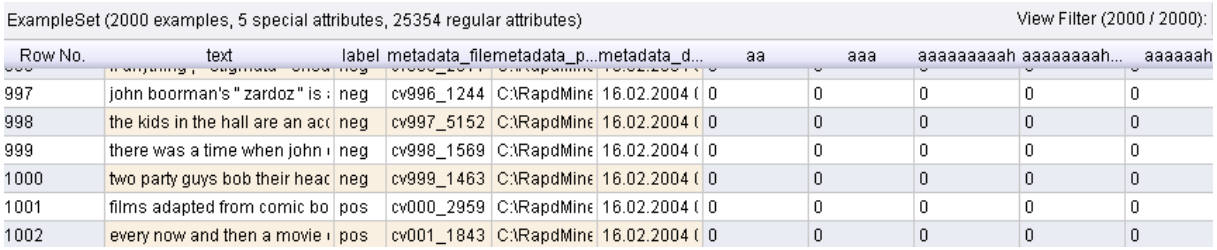
- Accuracy
- Precision
- Recall
- AUC (optimistic)
- AUC (neutral)
- AUC (pessimistic)

Diese Kennzahlen sind auch das wichtigste Resultat des Praxisteils. Sie geben beispielsweise an, mit welcher Wahrscheinlichkeit Texte richtig qualifiziert werden.

7.3.3.8 Resultat

Example Set

Nachdem der Prozess abgeschlossen ist, bekommt der Benutzer verschiedene Resultate präsentiert. Das Example Set zeigt die aufbereitete Datenbasis für das Modell. Darin enthalten sind die ursprünglichen Texte und das tatsächliche Label „positiv/negativ“. Die Abbildung zeigt in den Spalten alle aus den Texten extrahierte Token, die wie in Kapitel 7.3.3 beschrieben, aufbereitet wurden. Die Zellen beinhalten die Anzahl an Vorkommen der Token in den Texten. Somit wurde für jeden Text ein Vektor generiert, der durch das Modell mathematisch weiterverarbeitet werden kann.

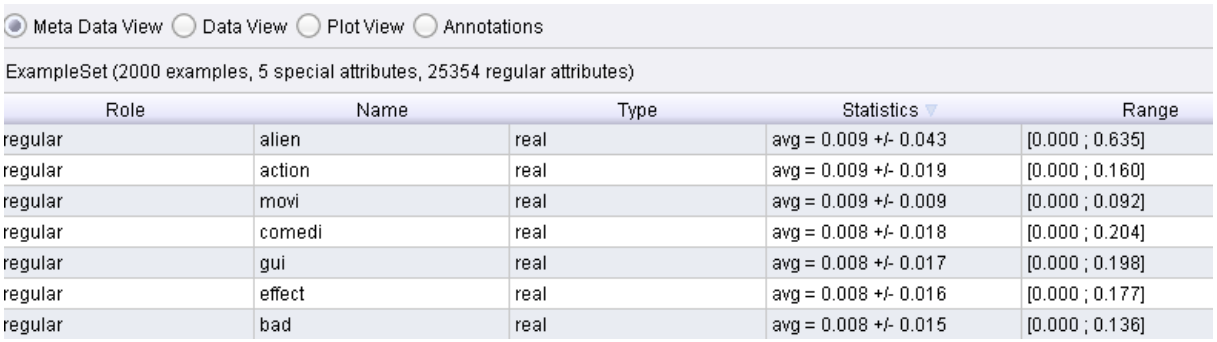


ExampleSet (2000 examples, 5 special attributes, 25354 regular attributes) View Filter (2000 / 2000):

Row No.	text	label	metadata_file	metadata_p...	metadata_d...	aa	aaa	aaaaaaaaah	aaaaaaaah...	aaaaaah
997	john boorman's " zardoz " is :	neg	cv996_1244	C:\RapdMine	16.02.2004 (	0	0	0	0	0
998	the kids in the hall are an acc	neg	cv997_5152	C:\RapdMine	16.02.2004 (	0	0	0	0	0
999	there was a time when john ı	neg	cv998_1569	C:\RapdMine	16.02.2004 (	0	0	0	0	0
1000	two party guys bob their heac	neg	cv999_1463	C:\RapdMine	16.02.2004 (	0	0	0	0	0
1001	films adapted from comic bo	pos	cv000_2959	C:\RapdMine	16.02.2004 (	0	0	0	0	0
1002	every now and then a movie ı	pos	cv001_1843	C:\RapdMine	16.02.2004 (	0	0	0	0	0

Abbildung 19: Example Set

In einer weiteren Ansicht des Example Sets wird zu jedem Token dessen Häufigkeit angezeigt. In der Abbildung ist auch ersichtlich, dass die durch das Stemming-Verfahren gekürzten Wortstämme in keinem Lexikon zu finden sind. Sie werden speziell für die Anforderungen von Text Mining Verfahren erstellt.



Meta Data View Data View Plot View Annotations

ExampleSet (2000 examples, 5 special attributes, 25354 regular attributes)

Role	Name	Type	Statistics	Range
regular	alien	real	avg = 0.009 +/- 0.043	[0.000 ; 0.635]
regular	action	real	avg = 0.009 +/- 0.019	[0.000 ; 0.160]
regular	movi	real	avg = 0.009 +/- 0.009	[0.000 ; 0.092]
regular	comedi	real	avg = 0.008 +/- 0.018	[0.000 ; 0.204]
regular	gui	real	avg = 0.008 +/- 0.017	[0.000 ; 0.198]
regular	effect	real	avg = 0.008 +/- 0.016	[0.000 ; 0.177]
regular	bad	real	avg = 0.008 +/- 0.015	[0.000 ; 0.136]

Abbildung 20: Example Set Meta Data

Performance Vector

Das wohl wichtigste Resultat des entworfenen Prozesses ist der Performance Vector. Er enthält verschiedene Kennzahlen, die die Güte des Modells beschreiben. Somit gibt der Performance Vector darüber Auskunft, ob das Modell für einen bestimmten Anwendungszweck brauchbar ist oder nicht.

```
Accuracy: 78.05% +/- 0.75% (mikro: 78.05%)
```

		Predicted	
		Negative	Positive
Actual	Negative	795	234
	Positive	205	766

Abbildung 21: Confusion Matrix Filmkritiken (maschinelles Lernen)

```
precision: 78.89% +/- 0.86% (mikro: 78.89%)
recall: 76.60% +/- 0.60% (mikro: 76.60%)
AUC (optimistic): 0.865 +/- 0.015 (mikro: 0.865)
AUC: 0.865 +/- 0.015 (mikro: 0.865)
AUC (pessimistic): 0.865 +/- 0.015 (mikro: 0.865)
```

Die Berechnung des Performance Vektors bzw. der Confusion Matrix und der daraus abgeleiteten Kennzahlen wird automatisch ermittelt, indem die Resultate der 3 Analyse-Durchgänge herangezogen werden. Die häufig herangezogene Kennzahl Accuracy beträgt 78,05%, d.h. dieser Anteil an Texten wurde korrekt klassifiziert. Die Bewertung dieser Kennzahl ist freilich von den Anforderungen der Untersuchung abhängig. In Anbetracht der Tatsache, dass der Prozess mit relativ einfachen Methoden erstellt wurde und keine Feinjustierung der Parameter vorgenommen wurde, ist das Ergebnis positiv. Vergleichbare wissenschaftliche Untersuchungen, bei denen Sentiment Analysen mit Hilfe der Methode des maschinellen Lernens durchgeführt wurden, kamen auf Ergebnisse in derselben Größenordnung. (Pang & Lee, 2008)

ROC (Receiver Operating Characteristic)

Die ROC ist ein grafisches Verfahren zur Beurteilung von Klassifikationsproblemen. Wie in Kapitel 7.2.2 erläutert, werden zur Darstellung der Receiver Operating Characteristic die fp (false positive)-Rate und die tp (true positive)-Rate herangezogen. Die tp-Rate gibt gemessen an der Gesamtanzahl der positiven Fälle die korrekt als positiv klassifizierten Fälle an. Die fp-Rate gibt den Anteil der Fälle an, die fälschlicherweise als negativ klassifiziert wurden.

Im ROC Diagramm (Abbildung 22: Filmkritiken ROC Graph) ist die tp-rate auf der Y-Achse und die fp-Rate auf der X-Achse aufgetragen. Die Fläche unter der Kurve (AUC-area unter curve) ist eine Maßzahl für die Qualität der Klassifikationsmethode. Gemäß dem in Kapitel 7.2.2 angeführten Bewertungsschema entspricht der in dieser Untersuchung erreichte Wert von 0,865 einem guten Ergebnis.

Abbildung 22: Filmkritiken ROC Graph

7.3.3.9 Analyse der Fehlkategorisierungen

Nach der Durchführung der Text Analyse mittels des diktionärbasierten Ansatzes konnten die Fehlkategorisierungen sehr einfach nachvollzogen werden, indem die in den Texten gezählten Wörter der positiven und negativen Kategorie überprüft wurden. Die Regeln der Klassifizierung beim diktionären Ansatz sind somit völlig transparent und das Ergebnis nachvollziehbar. Dies ist beim Ansatz des maschinellen Lernens nicht der Fall. Die Lösungsfindung des hier eingesetzten Algorithmus, der Support Vektor Maschine, ist nicht ohne weiteres nachvollziehbar, wenn nicht die Berechnungsregeln des Lernverfahrens überprüft werden.

In der Auswertung werden allerdings Werte mitgeliefert, die die Wahrscheinlichkeiten für eine Zuordnung in die positive oder negative Kategorie wiedergeben.
Eine Möglichkeit Probleme bei der Klassifizierung aufzuspüren, ist solche Fälle, die mit einer besonders hohen Wahrscheinlichkeit fehlklassifiziert wurden, zu betrachten.
Eine Analyse einer Stichprobe an fehlkategorisierten Texten führt zu einem ähnlichen Ergebnis wie bei der Durchführung mittels diktionärbasierten Ansatz.

Beinahe alle fehlkategorisierten Texte beinhalten eine sehr ausführliche Zusammenfassung des Filminhalts, dessen vermittelte Stimmung in keinem Zusammenhang mit der eigentlichen Filmbewertung steht. Dadurch werden auch manche Filmkritiken mit eindeutig negativer Bewertung falsch klassifiziert, wie etwa folgendes Beispiel, das mit einer ausführlichen Filmzusammenfassung einleitet und mit einer kurzen Filmbewertung abschließt:

“right now , it's tops on my list of the worst of 2001”

Wie auch vom diktionärbasierten Verfahren, kann diese Fehlerquelle durch maschinelles Lernen nicht ausgeschalten werden.
Einige fehlqualifizierte Texte sind auch durch eine manuelle Analyse nicht eindeutig einer positiven oder negativen Kategorie zuordenbar, wie z.B. die abschließende Bewertung dieser Filmkritik:

„ …this movie is easier to follow than the last one . which means the plot isn't quite as crowded . this movie isn't anything special though ….”

Oder dieser:

" ravenous " starts wonderfully and continues to shock and scare until it gets to the finale , where it looses focus and gets too primitive and rather boring . however these little failures don't diminish the impression”

Wie auch das diktionärbasierte Verfahren, basiert auch die hier verwendete Implementierung des maschinellen Lernens auf der Analyse einzelner Wörter. Der Vorteil des Lernverfahrens besteht darin, dass der Algorithmus nach der Lernphase optimal an den jeweiligen Themenbereich angepasst ist.
Dennoch können Negationen oder Bezüge auf andere Filme oder Themen nicht erkannt werden und führen daher zu Fehlklassifikationen.

7.3.3.10 Ergebnis der Untersuchung

Es konnte gezeigt werden, dass mittels der Software RapidMiner mit relativ wenig Aufwand ein Prozess entwickelt werden kann, der aus vorklassifizierten Texten (Trainingsdaten) lernt und anschließend in der Lage ist auch unbekannte Texte zu klassifizieren. In Anbetracht der vergleichsweise einfachen Vorgehensweise und Verwendung von Standardeinstellungen der Komponenten ist die erreichte Güte von ca. 78% (Accuracy) durchaus zufriedenstellend. Vergleichbare wissenschaftliche Untersuchungen (vgl. (Vaithyanathan, 2002)), die denselben Datensatz verwendet haben, kommen auf ähnliche Ergebnisse.
Selbstverständlich muss angemerkt werden, dass bei diesen Untersuchungen die gesamte Aufgabe der Extraktion der Texte ausgeklammert wurde, da bereits in einzelne Dateien strukturierte Texteinheiten verwendet wurden. Dies kann je nach Anwendungsfall ebenfalls Teil einer automatisierten Inhaltsanalyse sein. Die automatisierte Inhaltsanalyse von Filmkritiken kann durchaus als anspruchsvolle Aufgabe bezeichnet werden, da Filmkritiken häufig mit Verweisen auf andere Filme, einer bilderreichen Sprache und auch ironischen Elementen versehen sind. Andererseits erleichtert die Tatsache, dass sich alle Filmkritiken eindeutig auf einen Film beziehen die Untersuchung - im Gegensatz zu Diskussionsforen, wo sich naturgemäß Texte auf vorangegangene Meinungen beziehen können.

7.4 Analyse von Foreneinträgen

7.4.1 Beschreibung der Datenquelle

In Kapitel 7.3 wurden Filmkritiken analysiert, die in englischer Sprache vorlagen. Bei der zweiten Untersuchung werden Foreneinträge in deutscher Sprache nach ihrer Stimmung klassifiziert. Damit unterscheiden sich die Texte nicht nur durch ihre

Sprache, sondern auch in weiteren Faktoren, was in diesem Kapitel weiter erläutert wird.

Als Quelle für Foreneinträge wurde die Online-Community dol2day[7] (democracy online today) herangezogen. Dol2Day ist eine Politik-Community, in der sich Benutzer zu selbstgewählten, meist politischen Themen, austauschen können. Darüber hinaus können Benutzer virtuellen Parteien beitreten und solche gründen, sowie in regelmäßigen Abständen einen Internetkanzler bzw. eine Internetregierung wählen. Es gibt noch viele weitere Funktionen und Community-Elemente, wodurch dol2day auch als Demokratie-Simulation bezeichnet werden kann. Erwähnenswert ist weiters, dass die Benutzer der Plattform auch „Real Life"-Treffen organisieren.

Diskutiert wird bei dol2day zu den unterschiedlichsten Themen aus politischen, wirtschaftlichen, kulturellen und weiteren verwandten Bereichen. Am Anfang der Diskussion steht eine Frage oder auch eine Aussage, die dann von den Benutzern kommentiert werden kann.

Ein Beispiel aus dol2day ist z.B. die Diskussionsfrage *„Hartz-IV-Satz soll 2012 um 10 Euro steigen - Was sagst Du dazu?"*

Daraufhin folgen verschiedene Kommentare, wie etwa:

„Soll meinen Segen haben. Auch in Zeiten knapper Kassen kommt man nicht drum herum die Sätze hin und wieder der Preisentwicklung anzupassen."

Dieser Kommentar kann beispielsweise eindeutig als positiv klassifiziert werden. Bei der weiteren manuellen Vor-Analyse der Diskussionsbeiträge stellte sich die gesamte Untersuchung jedoch bald als problematisch heraus. Auch wenn die ursprünglich gestellte Diskussionsfrage die Benutzer auffordert eine Antwort darauf zu geben, entwickeln die dazugehörigen Kommentare bald ein „Eigenleben".

[7] http://www.dol2day.com/

Der folgende Kommentar bewertet nicht die Diskussionsfrage, sondern enthält einen Gegenvorschlag, der eine automatisierte und selbst manuelle Analyse der Stimmung beinahe unmöglich macht.

„Man sollte die Kopplung an die Inflation und das Einkommen aufheben und die Sätze somit langsam und vorsichtig absenken. Irgendwann muss der ständige Ausbau des Sozialstaates auch mal gestoppt werden, denn er nimmt den Menschen die Würde für sich selbst zu sorgen.

Neben diesem Kommentartyp, besteht die grundsätzliche Problematik, dass sich nach dem Entstehen einer Diskussion Kommentare nicht mehr auf die ursprüngliche Diskussionsfrage beziehen sondern aufeinander. In Foren wird dies auch häufig durch das Zeichen „@ (at)“ gefolgt vom Forenteilnehmernamen gekennzeichnet, auf den sich das Kommentar bezieht. Häufig wird die Aussage, auf die Bezug genommen wird, noch einmal zitiert und anschließend kommentiert, wie folgendes Beispiel zeigt:

„@ LLange

Zitat: Sozialleistungen erfüllen einen Zweck. Sie sollen Menschen ein Leben in Würde ermöglichen.
Das ist der entscheidende Punkt.
Und deswegen wird es langfristig auf ein bedingungsloses Grundeinkommen hinauslaufen - nicht zuletzt aufgrund der weiterhin rasant fortschreitenden Automatisierung und Rationalisierung“

Die automatisierte Inhaltsanalyse würde in diesem Fall auf zumindest zwei Probleme stoßen. Zum einen ist ein Teil des Kommentars eines vorherigen Benutzers als Zitat im Text enthalten. Die Inhaltsanalyse würde somit auch diesen Teil des Textes auswerten, obwohl er nicht Teil der Meinung und Stimmung des Verfassers des Kommentars ist. Somit ist die Auswertung des Zitates nicht erwünscht. Ein Ansatz wäre daher zu versuchen Zitate aus Texten zu entfernen, was auf automatisierte Weise nur mit aufwendigen Regeln möglich ist.

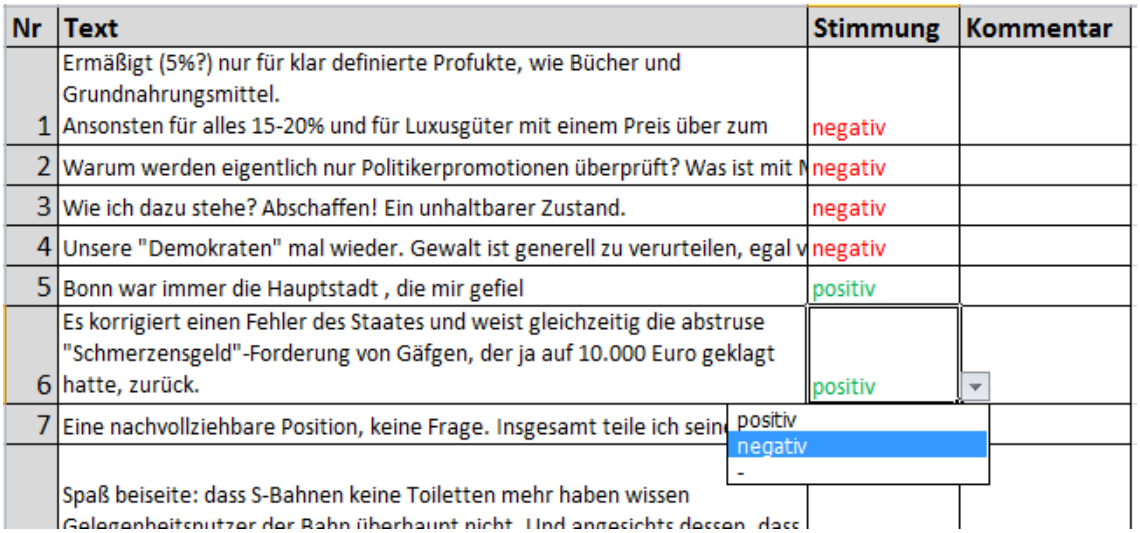
Das schwerwiegendere Problem ist nun aber, dass sich auch der Kommentar nicht auf die ursprüngliche Diskussionsfrage bezieht, sondern auf die Aussage des Zitats. Somit würde die automatisierte Inhaltsanalyse nicht die Stimmung hinsichtlich der Diskussionsfrage auswerten. Auch manuelle Auswertungen sind in solchen Fällen mit Mehraufwand verbunden. Für automatisierte Inhaltsanalysen müsste ein Weg gefunden werden, Zitate und Referenzen auf andere Kommentare zu erkennen. Dann könnte als Lösungsansatz der Problematik die Stimmung des referenzierten Kommentars ermittelt werden, welches sich auf die ursprüngliche Diskussionsfrage bezieht. Anschließend kann die Stimmung des eigentlichen Kommentars ausgewertet werden. Eine positive Stimmung würde als Zustimmung, eine negative Stimmung als Ablehnung des referenzierten Kommentars und dessen Stimmung interpretiert werden. Ein Kommentar kann jedoch auch mehrere Zitate und Referenzen auf andere Kommentare enthalten, was die Komplexität der Auswertung weiters erhöht.

Die ausgeführten Schwierigkeiten bei Auftreten von Diskussionen waren im Rahmen der Diplomarbeit nicht lösbar. Aus diesem Grund wurden für die Analyse ausschließlich Kommentare herangezogen, die sich eindeutig auf die Diskussionsfrage beziehen – die Repräsentativität der Untersuchung ist damit allerdings nicht mehr gewährleistet.

7.4.2 Manuelle Kategorisierung der Foreneinträge

Um die Güte der automatisierten Inhaltsanalyse bewerten zu können, wurden die Foreneinträge zuvor manuell analysiert und in positive und negative Kategorien eingeteilt. Es wurden insgesamt 800 Foreneinträge extrahiert, für deren manuelle Bewertung 5 Kodierer engagiert wurden. Jede der Testpersonen bekam die Aufgabe 160 Texte in die Kategorien „positive Stimmung“ und „negative Stimmung“ einzuordnen. Bei der Extraktion der Foreneinträge wurden Zitate und Foreneinträge, die sich auf vorherige Kommentare beziehen, entfernt.

Zur manuellen Kodierung wurde eine einfache Tabelle im Tabellenkalkulationsprogramm Excel erstellt, in dem die Kodierer mittels Auswahlfeldern die Texte positiven und negativen Kategorien zuordnen können (Abbildung 23: manuelle Kodierung). Eine dritte Antwortkategorie wurde für den Ausnahmefall vorgesehen, wenn ein Text keiner Stimmungskategorie zugeordnet werden kann. Die zu analysierenden Texte wurden in dieser Form per E-Mail an die Kodierer versendet.



Nr	Text	Stimmung	Kommentar
1	Ermäßigt (5%?) nur für klar definierte Profukte, wie Bücher und Grundnahrungsmittel. Ansonsten für alles 15-20% und für Luxusgüter mit einem Preis über zum	negativ	
2	Warum werden eigentlich nur Politikerpromotionen überprüft? Was ist mit N	negativ	
3	Wie ich dazu stehe? Abschaffen! Ein unhaltbarer Zustand.	negativ	
4	Unsere "Demokraten" mal wieder. Gewalt ist generell zu verurteilen, egal v	negativ	
5	Bonn war immer die Hauptstadt , die mir gefiel	positiv	
6	Es korrigiert einen Fehler des Staates und weist gleichzeitig die abstruse "Schmerzensgeld"-Forderung von Gäfgen, der ja auf 10.000 Euro geklagt hatte, zurück.	positiv	
7	Eine nachvollziehbare Position, keine Frage. Insgesamt teile ich sein		
	Spaß beiseite: dass S-Bahnen keine Toiletten mehr haben wissen Gelegenheitsnutzer der Bahn überhaupt nicht. Und angesichts dessen, dass		

positiv
negativ
-

Abbildung 23: manuelle Kodierung

Vor der Durchführung der manuellen Kategorisierungen durch die 5 Kodierer erfolgte eine Einschulung in den Kodierprozess. Es wurden folgende Definitionen und Regeln für die Kodierung getroffen.

Kodierleitfaden

Jedem Kodierer werden 160 zufällig ausgewählte Texte aus politischen Onlineforen zugeteilt, die die Meinung von Internetnutzern zu unterschiedlichen Themen enthalten. Im Kodierprozess soll die Stimmung der Texte durch den Kodierer eingeschätzt werden und der Kategorie „positive Stimmung“ oder „negative Stimmung“ zugeordnet werden. Kann in einem Text weder eine positive, noch eine negative Stimmung erkannt werden, so kann der Text als unbekannt („-„) gekennzeichnet werden. Die Kodierung soll gemäß folgender Definitionen und Regeln durchgeführt werden.

Kategorie: positive Stimmung

Definition

Texte, in denen der Autor beabsichtigt eine positive Stimmung zu vermitteln, sollen der Kategorie „positive Stimmung“ zugeordnet werden.

Ankerbeispiele

„Das wäre endlich mal wieder ein großer zivilisatorischer Fortschritt. Unbedingt umsetzten!“

„Unbedingt! Es geht auch, wie verschiedene durchgerechnete Modelle aus kirchlichen -, gewerkschaftlichen -und parteikreisen darlegten.“

„Das einzig Wahre - es wäre endlich mal ein radikaler Umbruch im Wirtschaftssystem.“

Kodierregel

Bei der Kodierung der Texte ist darauf zu achten, dass die vermittelte Grundstimmung des gesamten Textes bewertet wird, anstatt dass Teilbereiche betrachtet werden. Die Intensität der Stimmung spielt dabei keine Rolle, sodass auch minimal positive Tendenzen der Kategorie zugeordnet werden.

Kategorie: negative Stimmung

Definition

Texte, in denen der Autor beabsichtigt eine negative Stimmung zu vermitteln, sollen der Kategorie „negativ Stimmung“ zugeordnet werden.

Ankerbeispiele

„Eine so völlig bescheuerte Idee kann wohl nur von Leuten kommen, die sich das Internet von der Sekretärin ausdrucken und in die Vorlagenmappe legen lassen.“

„Schwachsinn war klar das sowas wieder mal nur von der EU kommt.“

„Wie kommt man denn nur auf solche Ideen?!“

Kodierregel

Bei der Kodierung der Texte ist darauf zu achten, dass die vermittelte Grundstimmung des gesamten Textes bewertet wird, anstatt dass Teilbereiche betrachtet werden. Die Intensität der Stimmung spielt dabei keine Rolle, sodass auch minimal negative Tendenzen der Kategorie zugeordnet werden.

Resultat

Nach ca. 2 Wochen wurden die 800 kodierten Texte von den 5 Kodierern per E-Mail retourniert. Bei der Kodierung traten keine Schwierigkeiten oder Missverständnisse auf. Die Häufigkeit in den Kategorien ergab sich wie folgt:

- Positive Stimmung: 345
- Negative Stimmung: 447
- Unbekannt (-): 8

Dieses durch manuelle Kodierung erzielte Ergebnis wird somit als tatsächliche Kategorisierung angenommen und der durch automatisierte Inhaltsanalyseverfahren vorhergesagtem Ergebnis gegenübergestellt. Die 8 nicht zugeordneten Texte enthielten keine auswertbare Stimmung und wurden für die automatisierten Verfahren ausgeschlossen, da bei diesen nur 2 Kategorien vorgesehen wurden. Dies ist selbstverständlich, wie bereits erwähnt, keine empirische repräsentative Untersuchung. Sie dient vielmehr der Überprüfung und Demonstration der diktionärbasierten Verfahrens und des Ansatzes mittels überwachten Lernens.

7.4.3 Diktionärbasiertes Verfahren

7.4.3.1 Import des Diktionärs

Für den praktischen Test des diktionärbasierten Ansatzes wurde das für Forschungszwecke frei verfügbare Diktionär SentiWS herangezogen. Darin sind 1650 negative und 1818 positive Wörter enthalten, wobei auch eine Gewichtung inkludiert ist, die für diese Arbeit jedoch nicht verwendet wurde. Auch die Zusatzinformation der Wortart wurde in dieser Untersuchung ignoriert.
Damit das Lexikon in Lexicoder importiert werden kann, wurde es in das entsprechende Format gebracht und in einer Textdatei abgespeichert.

```
<cnode name="POSITIVE">
...
<pnode name=" günstigsten "></pnode>
<pnode name=" günstigster "></pnode>
<pnode name=" günstigstes "></pnode>
```

```
<pnode name=" gut "></pnode>
<pnode name=" gute "></pnode>
<pnode name=" Güte "></pnode>
<pnode name=" gutem "></pnode>
...


<cnode name="NEGATIVE">
...
<pnode name=" grottenschlecht "></pnode>
<pnode name=" grottenübel "></pnode>
<pnode name=" gruselig "></pnode>
<pnode name=" gruselige "></pnode>
<pnode name=" gruseligem "></pnode>
<pnode name=" gruseligen "></pnode>
<pnode name=" gruseliger "></pnode>
...
```

Das Diktionär kann in diesem Format in Lexicoder importiert werden (Load Source).

7.4.3.2 Import der Texteinheiten

Auch die zu analysierenden Texte müssen in das geforderte Format gebracht werden, um in Lexicoder importiert werden zu können. Die genaue Vorgehensweise wurde bereits in Kapitel 7.3.2.3 erläutert.

7.4.3.3 Resultat

Die Ergebnisse wurden mit Hilfe eines Tabellenkalkulationsprogramms ausgewertet und in die Confusion Matrix eingetragen. Basierend auf dieser wurde der Accuracy-Wert berechnet.

		Predicted	
		Negative	Positive
Actual	Negative	281	166
	Positive	139	206

Abbildung 24: Confusion Matrix Forum (diktionärbasierter Ansatz)

$$Accuracy = \frac{281 + 206}{281 + 166 + 139 + 206} = 61{,}5\%$$

Dividiert man die richtig klassifizierten Texte durch die Gesamtanzahl der Texte so erhält man die Kennzahl Accuracy. Sie beträgt in dieser Untersuchung 61,5%. Dieser Wert wäre für die meisten Anwendungen mit Sicherheit nicht zufriedenstellend, bedenkt man dass er nicht weit vom Zufallswert von 50% entfernt ist.

Ein Grund für den niedrigen Accuracy Wert der Klassifikation ist in dem verwendeten Diktionär zu suchen. Dieses enthält auch Wörter mit einer minimalen positiven oder negativen Stimmung, die in dieser Analyse gleichgewichtet wurden. Die Klassifikationsleistung kann durch Berücksichtigung der gewichteten Stimmung verbessert werden oder durch Überarbeitung des Diktionärs. Die Überarbeitung des Diktionärs kann erfolgen, indem an Hand einer Stichprobe Indikatoren für positive oder negative Stimmungen ausgewählt werden. Dadurch wird das Diktionär für die vorliegende Domäne des politischen Forums optimiert. Eine konkrete Anwendung dieses Vorgehens ist hier ausgeführt: (Kaczmirek, Baier, & Züll, 2010).

7.4.4 Maschinelles Lernen

7.4.4.1 Implementierung des Prozesses

Der Aufbau des Prozesses wurde in Kapitel 7.3.3.5 bereits ausführlich beschrieben. Für diese Untersuchung mussten geringe Anpassungen vorgenommen werden, die sich durch den Wechsel von der englischen zur deutschen Sprache ergeben. Konkret handelt es sich um folgende 2 Punkte im Vorbearbeitungsschritt

Filter Stopwords

Im Schritt „Filter Stopwords“ werden unbedeutende Wörter aus den Texten einfach entfernt. Diese Füllwörter werden in einer Textdatei gespeichert, auf die das Programm RapidMiner zugreifen kann. Ein Auszug aus der deutschen Stoplist:

…
sonst
über
um
und
uns
unse
unsem
…

Stem (Porter)

Im Stemming-Schritt werden Wörter durch einen speziellen Algorithmus auf einen Stamm vereinfacht, sodass inhaltlich ähnliche oder gleichbedeutende Wörter auf denselben Stamm reduziert werden können. Es existieren für jede Sprache eigene Algorithmen, die diese Vereinfachung vornehmen. Entsprechend wird für die Untersuchung der deutschen Texte die deutsche Variante des Stemming Algorithmus verwendet.

7.4.4.2 Resultat

Performance Vector

Der Performance Vektor von RapidMiner beinhaltet die Confusion Matrix und die zur Einschätzung der Güte wichtigsten Kennzahlen.

```
Accuracy: 75.6%
```

		Predicted	
		Negative	Positive
Actual	Negative	309	138
	Positive	55	290

Abbildung 25: Confusion Matrix Forum (maschinelles Lernen)

Der Accuracy Wert von 75,6% ist etwas geringer als wie bei der Analyse der Filmbewertungen in englischer Sprache. Dennoch ist der Wert durchaus zufriedenstellend, betrachtet man den relativ simplen Analyseprozess. Eine Ursache für den im Vergleich zur ersten Untersuchung mittels maschinellen Lernens niedrigere Wert ist die niedrigere Textanzahl. Dadurch ist die in der Trainingsphase herangezogenen Textanzahl deutlich geringer, was zu einer niedrigeren Accuracy führt. Verstärkend kommt die Tatsache hinzu, dass nicht nur die Textanzahl, sondern auch die Wortanzahl innerhalb der Texte deutlich geringer ist.

7.5 Resultat der empirischen Untersuchung

Im empirischen Teil der Arbeit wurde versucht mit Hilfe von zwei unterschiedlichen Ansätzen der automatisierten Inhaltsanalyse Texte aus zwei unterschiedlichen Typen von Texten zu analysieren. Bereits mehrmals wurde erwähnt, dass es nicht Zielsetzung des praktischen Teils der Arbeit ist, die Methoden zu optimieren und optimale Klassifizierungsergebnisse zu erreichen. Dennoch kann man einerseits deutliche Unterschiede der Funktionsweise des diktionärbasierten Verfahrens und des Verfahrens des maschinellen Lernens erkennen und andererseits Eigenschaften der automatisierten Inhaltsanalyse im Allgemeinen ableiten.

Die empirische Untersuchung der Filmkritiken beinhaltete die Analyse von 1000 positiven und 1000 negativen Bewertungen von Kinofilmen. Die Bewertungen wurden von Internetnutzern verfasst und beinhalteten dementsprechend viele Dialektausdrücke verschiedener Ursprünge.

Die Texte wurden zuerst mittels der klassischen Methode der automatisierten Inhaltsanalyse ausgewertet, nämlich mittels eines diktionärbasierten Verfahrens. Dazu wurde ein frei verfügbares Diktionär verwendet und das für Forschungszwecke ebenfalls frei verfügbare Programm Lexicoder. Bevor das Diktionär und die Filmkritiken mit Hilfe des Programms ausgewertet werden können, mussten sie in eine spezielle Struktur gebracht werden, worin der größte Aufwand dieser Untersuchung bestand. Im ersten Durchgang wurde eine Accuracy von 66% erreicht, d.h. dieser Anteil der Texte wurde richtig in die Kategorien „positiv“ und „negativ“ eingeordnet. Im zweiten Durchgang wurden Negationen berücksichtigt, wodurch der Accuracy-Wert um 0,3 Prozentpunkte gesteigert werden konnte. Dass die Verbesserung derart minimal ausfiel mag überraschend erscheinen, deckt sich jedoch mit anderen Forschungsergebnissen. (Dadvar, Hauff, & de Jong, 2011) Die größte Anzahl an Fehlkategorisierungen entstand durch die Tatsache, dass Filmkritiken in vielen Fällen mit einer Filmzusammenfassung beginnen, die die Auswertung verfälscht. Dennoch befindet sich der erreichte Wert von 66,3% deutlich über dem Zufallswert von 50%, wodurch die prinzipielle Funktionsweise des diktionärbasierten Verfahrens gut dargestellt werden konnte.

Das gleiche Rohmaterial wurde anschließend mit Hilfe einer Methode aus dem Bereich des maschinellen Lernens analysiert. Das Verfahren unterscheidet sich grundlegend von dem des diktionärbasierten Ansatzes, hat aber dasselbe Ziel: nämlich die Kategorisierung von Texten in solche mit positiver und negativer Stimmung. Obwohl das Verfahren relativ einfach umgesetzt wurde und Standardeinstellungen verwendet wurden, konnte ein Accuracy-Wert von 78,05% erreicht werden. Zwar soll die Untersuchung keineswegs ein Performance-Vergleich der beiden Verfahren sein, doch zeigt sich auch in anderen wissenschaftlichen Untersuchungen, dass Verfahren des maschinellen Lernens eine vergleichsweise hohe Trefferquote aufweisen. Der vor allem für kleiner angelegte Untersuchungen schwerwiegende Nachteil dieser Methode liegt jedoch darin, dass eine nicht unerhebliche Menge an vorklassifizierten Texten für die Lernphase erforderlich ist. Das bedeutet, dass bevor mit der automatisierten Inhaltsanalyse begonnen werden kann, Texte manuell klassifiziert werden müssen. Ist die Trainingsphase einmal abgeschlossen, so können jedoch beliebig viele Texte automatisiert analysiert werden.

In einer zweiten Untersuchung wurden deutsche Foreneinträge politischer Webseiten als Untersuchungsgegenstand gewählt. Dadurch unterscheiden sich die Texte nicht nur durch ihre Sprache von der ersten Untersuchung, sondern auch thematisch.

Bereits bei der Auswahl der Texte zeigte sich die größte Hürde für die automatisierte Inhaltsanalyse von Foreneinträgen im Allgemeinen. Diese liegt an der Kommunikationsform des Forums an sich, nämlich dass sich Personen untereinander austauschen, diskutieren und auf einander Bezug nehmen. Für die automatisierte Inhaltsanalyse ist dies insofern problematisch, als nicht mehr die Meinung zum Ursprungsthema analysiert wird, sondern möglicherweise eine Meinung, die sich auf einen anderen Kommentar bezieht. Um in Internetforen Diskussionen nachvollziehbar zu machen, wird der Textteil, auf den sich ein Kommentar bezieht, üblicherweise zitiert. Ein Computerprogramm kann diese Zitate nur schwer von dem Textteil, der die eigentliche Meinung eines Kommentators enthält, trennen.

Die hier aufgezeigten Hürden konnten im Rahmen der Diplomarbeit nicht überwunden werden, weshalb eine manuelle Korrektur vorgenommen wurde, bevor mit der Analyse der Texte fortgefahren wurde.

Für den diktionärbasierten Ansatz wurde ein frei verfügbares Diktionär in deutscher Sprache verwendet. Die Accuracy betrug 61,5%, was keinem zufriedenstellenden Ergebnis entspricht. Eine Erklärung für diesen niedrigen Wert ist im gewählten Diktionär zu suchen, welches Wörter mit sehr unterschiedlichen Gewichtungen enthält, die bei dieser Analyse nicht berücksichtigt wurden. Auch wenn bereits vorgefertigte Lexika frei zur Verfügung stehen, empfiehlt es sich diese an die konkrete Anwendung anzupassen. Indem das Lexikon an die „Domäne“ der Untersuchung angepasst wird, kann die Trefferquote deutlich erhöht werden.

Der maschinelle Ansatz lieferte eine Trefferquote von 75,6%, was ebenfalls unter dem Wert der ersten Untersuchung liegt. Ein Grund für den niedrigeren Wert ist eindeutig die geringe Anzahl an Texten und auch die geringe Wortanzahl der Texte. Dadurch hat das System nur wenige Texte in der Trainingsphase zur Verfügung und kann in der Testphase die Texte nicht ausreichend gut klassifizieren.

An Hand der beiden unterschiedlichen Untersuchungen konnten die Methoden diktionärbasierter Ansatz und maschinelles Lernen demonstriert werden. Beide Ansätze unterscheiden sich grundlegend voneinander, wobei von der Methode des maschinellen

Lernens eine höhere Trefferquote erwartet werden kann. In beiden Fällen ist eine Anpassung der Methode an die Forschungsfrage zu empfehlen. Für den diktionärbasierten Ansatz wird empfohlen das Diktionär an die Domäne der Untersuchung anzupassen und für das maschinelle Lernen ist eine beachtliche Anzahl an manuell vorklassifizierten Texten erforderlich. Somit eignen sich beide Verfahren für die Analyse größerer Textmengen, da sich nur dann der Anfangsaufwand lohnt. Dieser ist beim diktionärbasierten Ansatz etwas geringer, da die Anpassung des Diktionärs weniger Aufwand als die manuelle Klassifikation von Texten darstellt.

8 Potentielle Anwendungen der automatisierten Inhaltsanalyse

In dieser Arbeit wurden Ansätze demonstriert, deren Motivation es ist, unstrukturierte Texte automatisiert zu analysieren. Im empirischen Teil der Arbeit wurde an Hand konkreter Beispiele eine automatisierte Stimmungsanalyse von Texten durchgeführt. Doch mit diesen und ähnlichen Methoden lassen sich selbstverständlich auch viele weitere Analysen automatisiert durchführen, die über die einfache Zuordnung in Texte positiver und negativer Stimmung hinausgehen. Zwar ist die Güte der automatisierten Verfahren derzeit für viele Anwendungen noch nicht ausreichend, doch geht man davon aus, dass die Entwicklung ähnlich rasant weitergeht wie in den vergangenen Jahren, so ergeben sich bald viele spannende Anwendungsgebiete. Das Potential und mögliche Anwendungen, die zum Teil schon derzeit Realität sind, werden im Folgenden vorgestellt.

Großes Potential von Verfahren der automatisierten Inhaltsanalyse verspricht sich etwa die Marktforschung. Die sehr offene Weise, in der Internetnutzer in Foren, Blogs oder sozialen Netzwerken ihre Meinung über Unternehmen und deren Produkte kundtun, ist für die Marktforschung ausgesprochen wertvoll. Schließlich werden dort genau die Fragen beantwortet, die bisher nur mit umfangreichen Studien analysiert werden konnten. Personen beschreiben ungeschont wie sie über ein Unternehmen oder Produkt denken, warum sie ein Produkt kaufen oder nicht kaufen und was ihnen an anderen Produkten besser gefällt. Während in dieser Arbeit die bloße Kategorisierung in positive und negative Stimmungen demonstriert wurde, werden auch Methoden angeboten, die

beispielsweise Themen identifizieren, die besonders häufig im Zusammenhang mit einem Produkt- oder Unternehmensnamen genannt werden.

Bisher wurde von Unternehmen bei Marktforschungsinstituten eine Studie in Auftrag gegeben um Meinungen und Verbesserungspotentiale von Produkten einzuholen. Diese Vorgehensweise würde sich mit automatisierter Textanalyse insofern verändern, als relevante Meinungen aktiv aus verschiedenen Quellen gefiltert und dem Unternehmen präsentiert werden. Unternehmen können so besonders schnell reagieren und Gegenmaßnahmen einleiten. Kombiniert man dies mit Methoden der Netzwerkanalyse, so lassen sich die einflussreichsten Meinungsträger in Internetplattformen identifizieren und diese genauer analysieren.

Um die Stimmung von Begriffen auf Internetplattformen auszuwerten, werden bereits verschiedene Produkte angeboten. Ein einfaches Beispiel ist Tweetfeel[8], welches Einträge der Plattform Twitter[9] analysiert.

Twitter ist eine Internetplattform, auf der Personen in kurzen Textnachrichten ihre Meinung äußern können. Dadurch ist Twitter prädestiniert für die Analyse von Stimmungen. Dementsprechend existiert eine Reihe von Produkten, die diese kurzgefassten Meinungen automatisch analysieren und klassifizieren. Unternehmen können sich dadurch rasch einen Überblick verschaffen, wie gut ihr Unternehmen oder Produkte bei Kunden ankommen.

Versuchsweise wurde das Programm Tweetfeel getestet, indem nach der Stimmung der Stadt Wien (Suchbegriff „Vienna“) gesucht wurde.

[8] http://www.Tweetfeel.com/
[9] https://twitter.com/

Abbildung 26: Tweetfeel

Als Ergebnis erhält man eine Klassifizierung in positive und negative Kommentare, die auch direkt aufgelistet werden. Dadurch hat man auch die Möglichkeit nach den Ursachen von positiven oder negativen Ergebnissen zu suchen.

In diesem Fall wurden von 76 Einträgen über Wien 51 als positiv und 25 als negativ beurteilt. Tatsächlich wurden die Kommentare überwiegend richtig klassifiziert, wie etwa „I dont like vienna. I want to go back to Prague.“ als negativ oder „morning to you and yes vienna is a beautiful city...did you try the "Sachertorte" ? x” als positiv. Doch es zeigen sich auch hier grundsätzliche Probleme bei der automatisierten Inhaltsanalyse. Der Eintrag „but tbh, i don't really like vienna sausage.. “ wird als negativ klassifiziert, bezieht sich jedoch nicht auf die Stadt Wien selbst, sondern auf Wiener Würstchen, weshalb der Eintrag für die Bewertung irrelevant wäre. Auch das Kommentar "vienna is the greatest song ever" bezieht sich auf ein Lied und nicht auf die Stadt Wien und dürfte nicht zur Auswertung herangezogen werden. Aus dem kommerziellen Bereich gibt es selbstverständlich viele weitere Angebote, die den manuellen Aufwand von Textanalysen minimieren.

Rapid-I bietet beispielsweise das Produkt „RapidSentilyzer“ an, welches im Internet automatisch nach Meinungen zu Unternehmen oder Produkten sucht und diese klassifiziert. Diese Produkte werden auch unter den Begriffen „Competitive bzw. Customer intelligence“ geführt. Sie werden hauptsächlich von Unternehmen eingesetzt,

um Meinungen von Kunden über eigene Produkte automatisiert zu analysieren und dadurch Marketing- und PR-Kampagnen zu evaluieren. Außerdem können Mitbewerber und deren Schwächen und Stärken beobachtet werden oder Vorhersagen über die Entwicklung von Kundenbedürfnissen getroffen werden. Die Analysen können in bestimmten Intervallen automatisiert durchgeführt werden, sodass sich die Entwicklung der Stimmung im Zeitkontext betrachten lässt. Als Quellen für Meinungen werden Nachrichtenseiten, Foren und Blogeinträge herangezogen. RapidSentilyzer basiert auf dem in dieser Arbeit eingesetzten RapidMiner, ist jedoch vollständig vorkonfiguriert, um dem Benutzer möglichst schnell und auf einfache Weise das Endresultat, nämlich die Stimmung über Produkte oder Unternehmen, zu präsentieren. Das Produkt ist also für diesen konkreten Anwendungsfall optimiert, was vor der kostenpflichtigen Anschaffung bedacht werden sollte. Wird es im vorgesehenen Anwendungsbereich eingesetzt, bietet das Produkt eindeutige Vorteile. So werden nur relevante Quellen an das System angeschlossen und nach Neuigkeiten und Meinungen durchsucht. Weiters entfällt der gesamte Trainingsprozess, da das System bereits anhand von Beispieltexten aus dem relevanten Themengebiet trainiert wurde. Zusätzlich werden die Ergebnisse grafisch aufbereitet, sodass der Anwender die Stimmung über sein Unternehmen oder des Mitbewerbers übersichtlich in aggregierter Form analysieren kann.

Neben Rapid Sentilyzer existieren selbstverständlich noch zahlreiche weitere Werkzeuge am Markt, die auf Text Mining und auch auf Sentiment Analysis spezialisiert sind. SAS Sentiment Analysis ist ein professionelle Softwarelösung zur Stimmungserkennung, SPSS Text Mining for Clementine ist eine Erweiterung von SPSS, die ebenfalls Text Mining Methoden beinhaltet und Mathematica bietet ebenfalls Werkzeuge zur Mustererkennung in unstrukturierten Texten an, um nur einige verbreitete kommerzielle Text Mining Programme zu nennen. Aus dem Open Source Bereich sind die verbreitete Software GATE und tm: Text Mining Package für die Statistiksoftware R zu nennen.

Gern zitiert wird eine erfolgreiche Anwendung der hier beschriebenen Verfahren der Deutschen Post AG. Diese konnte negative Meinungen, die ihren Ursprung in einem Blogeintrag zum E-Postbrief haben, frühzeitig erkennen. Es ging dabei um ein neues Service der Deutschen Post AG, dem E-Postbrief, der von dem Blogger Richard Gutjahr als sehr unsicher eingestuft wurde. Rasch verbreitete sich seine Einschätzung in

diversen Internetforen und sozialen Netzwerken und wurde dort heftigst diskutiert. Durch die automatisierte Analyse verschiedener Internetquellen konnten die Negativschlagzeilen sehr früh erkannt werden und das Unternehmen reagieren, indem es die Internetnutzer direkt kontaktierte und aufklärte. Die schnelle Reaktionszeit und die direkte Kommunikation wurde von den Kunden sehr positiv aufgenommen und der Schaden somit behoben bzw. begrenzt (Peter Gentsch A.-M. Z., 2011).

Es zeigte sich, dass das unmittelbare Reagieren in Internetplattformen ein wichtiger Faktor im Krisenmanagement ist, da sich Negativschlagzeilen dort sehr rasch verbreiten. Automatisierte Inhaltsanalysen haben hier den Vorteil, dass sie Texte sehr schnell auswerten können. Da in diesem Fall die Reaktionszeit die größte Rolle spielt, kann auf eine hohe Trefferquote der automatisierten Auswertung eher verzichtet werden.

Frühwarnsysteme dieser Art haben auch für Regierungen einen großen Nutzen. Jüngste politische Bewegungen organisieren sich meist über Internetplattformen, wo sich innerhalb kürzester Zeit Personen gruppieren und Offline-Veranstaltungen organisieren können. Im Jahr 2009 organisierten sich Studierende der Universität Wien über Internetplattformen wie Facebook oder Twitter, besetzten das Auditorium Maximum der Universität Wien und tauschten sich online über die aktuelle Situation und das weitere Vorgehen aus. Im Jahr 2011 mobilisierten sich nach der Atom-Katastrophe in Fukushima Atomkraftgegner in Deutschland und auch über Deutschlands Grenzen hinaus über das Medium Internet und veranstaltete Proteste gegen laufende Atomprogramme. Gleichzeitig tauchte auch erstmals der Begriff Protest 2.0 auf, der die neue Stufe von Protestorganisationen verdeutlicht. Doch auch bei den Revolutionen in Ägypten und Tunesien im Jahr 2011 spielten das Internet und soziale Netzwerke eine zentrale Rolle (Fuchs, 2006). Im Grunde wird jeder Unmut über politische Zustände im Internet kundgetan.

Wie Unternehmen, sind auch Regierungen dazu gezwungen richtig und vor allem schnell auf negative Meinungsäußerungen zu reagieren. Methoden der automatisierten Inhaltsanalyse sind dafür als Frühwarnsysteme gut geeignet.

Die automatisierte Inhaltsanalyse eignet sich auch zur Überwachung von Wahlkampf-Kampagnen, wie sie beispielsweise von der Partei Christlich Demokratische Union Deutschlands (CDU) eingesetzt wurde. Durch die rasche Analyse von Meinungen auf

diversen Internetplattformen kann sich die Partei schnell ein Bild darüber machen, was, wann, wo, in welchem Umfang und in welcher Weise über Parteien, Politiker und Bundestagswahlen diskutiert wird (Peter Gentsch M. H., 2009). Auf besonders intensiv diskutierte Themen kann so rasch reagiert werden und Reaktionen auf Wahlkampagnen können unmittelbar analysiert werden. Selbstverständlich zielen Früherkennungssysteme in erster Linie darauf ab, auf bestimmte Themen aufmerksam zu werden. Für detaillierte, gezielte und repräsentative Analysen können Untersuchungen mit Methoden der klassischen Meinungsforschung angeschlossen werden.

Die automatisierte Inhaltsanalyse kann aber auch als Unterstützung für andere empirische Untersuchungen wie Befragungen dienen: nämlich zur Auswertung offener Fragen. Diese werden gerne eingesetzt wenn wenig Vorwissen über den Untersuchungsgegenstand vorhanden sind und Begründungen für Meinungen oder Einstellungen ermittelt werden sollen. Offene Fragen sind einfach zu erstellen, jedoch umso schwieriger auszuwerten. Es handelt sich um qualitative Daten, die von Kodierern gelesen und interpretiert werden müssen. Genau hier kann die automatisierte Inhaltsanalyse helfen, um große Mengen offener Fragen auszuwerten und zum Beispiel Kategorien zuzuordnen oder aber auch mittels Clusterverfahren zuvor unbekannten Gruppen zuzuordnen und so Strukturen in den freien Antwortkategorien zu erkennen. Eine Anwendung der automatisierten Inhaltsanalyse zur Auswertung offener Fragen mittels eines diktionärbasierten Ansatzes wird etwa hier vorgestellt: (Kaczmirek, Baier, & Züll, 2010). Basierend auf einer Teilmenge der offenen Fragen wird ein Diktionär erstellt, der dem Vokabular und der Domäne des Fragebogens angepasst ist und so eine zufriedenstellende Trefferquote erreicht.

Weiters finden Methoden der automatisierten Inhaltsanalyse in der Vorhersage von Aktienpreisen Anwendung. Ausgewählte Nachrichtenplattformen werden nach Meinungen und Neuigkeiten von Unternehmen durchsucht und automatisch in positive und negative Texte klassifiziert. Daraus können Vorhersagen über die zukünftige Preisentwicklung abgeleitet werden (Tang, Yang, & Zhou, 2009). Freilich ist dieses Verfahren als Ergänzung und nicht als Ersatz klassischer Methoden zu sehen. Gerade bei der automatisierten Analyse von Nachrichtenartikel muss angemerkt werden, dass sich hier andere Anforderungen offenbaren als es bei Foren, Blogs oder sozialen

Netzwerken. Schließlich beinhalten News viel weniger klare Positionen, sondern erheben den Anspruch eine neutrale Stellung wiederzugeben (Alexandra Balahur, 2009).

Die hier angeführten unterschiedlichen Anwendungsgebiete werden durch Eigenschaft der automatisierten Inhaltsanalyse ermöglicht, große Textmengen in sehr kurzer Zeit analysieren zu können. Die Güte von deutlich unter 100% Accuracy ist für diese Anwendungen zweitrangig. Somit bieten sich die automatisierten Verfahren als Ergänzung klassischer Methoden an, nämlich als Frühwarnsysteme oder wenn große Textmengen vorstrukturiert werden sollen. Dies überrascht nicht, da die traditionelle Inhaltsanalyse zu Beginn auch als Frühwarnsystem Anwendung fand, nämlich bei der Analyse deutscher Kriegspropaganda im Zweiten Weltkrieg. Mit dieser Grundlage kann mit klassischen Methoden der empirischen Sozialforschung, wie Befragungen oder qualitative Inhaltsanalysen fortgesetzt werden, um spezifische Forschungsfragen beantworten zu können. Ob in Zukunft auch diese Aufgabe automatisiert bewältigbar werden können, hängt von den Weiterentwicklungen der Methoden ab und ist derzeit nur schwer abzuschätzen.

9 Zusammenfassung

Als Ziel der Arbeit wurde angestrebt, die methodische Weiterentwicklung der Inhaltsanalyse aufzuzeigen, die durch die Zusammenarbeit verschiedener Wissenschaftsdisziplinen vorangetrieben wird. Durch den enormen Zuwachs an frei verfügbarem Meinungsaustausch zu vielfältigen Themen auf Internetplattform und deren Aktualität ergibt ohne Zweifel sich ein großes Potential für sozialwissenschaftliche Analysen. Doch bringt die Analyse von Online-Inhalten gegenüber statischen Texten auch neue Herausforderungen mit sich. Schließlich wird hier nicht mehr manifester Inhalt untersucht, wie es Berelsons ursprüngliche Definition der Inhaltsanalyse beschreibt, sondern dynamischer Inhalt, der darüber hinaus durch Verwendung von Hyperlinks nichtlinear ist. Das Problem der Flüchtigkeit des Inhalts kann durch die Sicherung der Internettexte zu einem bestimmten Zeitpunkt behoben werden, doch muss eingestanden werden, dass weitere Eigenschaften von Online-Inhalten, wie Interaktivität, Multimedialität oder Personalisierung von Internetseiten die Inhaltsanalyse vor Probleme stellt, die derzeit noch nicht für alle Untersuchungen in

zufriedenstellendem Maße behoben werden können. Durch die Verfügbarkeit der immensen Textmengen wurde in den letzten Jahren an Methoden geforscht, die in der Lage sind, semantische Analysen an Texten automatisiert durchzuführen. Der klassische Ansatz Kategorisierungen von Texten zu automatisieren ist die Verwendung eines Diktionärs, welches Indikatoren für die jeweiligen Kategorien enthält, und durch ein Computerprogramm auf die zu analysierenden Texte angewendet wird. Dieser Ansatz wurde etwa durch die Hinzunahme von Informationen über die jeweiligen Wortarten oder durch Verfahren zur Erkennung von Negationen verbessert. Dennoch ist die Güte der diktionärbasierten Verfahren für viele Anwendungen nicht zufriedenstellend.

Ein moderner Ansatz aus dem Bereich der künstlichen Intelligenz verspricht für die meisten Anwendungen eine bessere Klassifikationsleistung als diktionärbasierte Ansätze. Dieser Methode werden in einer ersten Phase vorklassifizierte Texte zur Verfügung gestellt, aus der automatisch Regeln abgeleitet werden. Aus diesem Grund wird dieser Ansatz auch als überwachtes Lernen bezeichnet. Nach der Lernphase können Texte automatisch klassifiziert werden. Bei diesen Verfahren des überwachten Lernens hat die Vorbearbeitung der Texte einen hohen Stellenwert. Erst durch eine syntaktische Vereinfachung der Texte ist die Methode in der Lage in der Trainingsphase ausreichend zu generalisieren und anschließend gute Klassifikationsergebnisse zu erzielen.

Die Funktionsweise des diktionärbasierten Ansatzes und der Methode des überwachten Lernens wird in Kapitel 7 an Hand von praktischen Beispielen von Sentiment Analysen erläutert. Beide Ansätze erfordern einen gewissen Anfangsaufwand. Im einen Fall zur Erstellung eines Diktionärs und im anderen Fall zur manuellen Klassifizierung von Texten, die für die Trainingsphase des überwachten Lernens benötigt werden. Anschließend können Texte praktisch ohne Aufwand automatisch klassifiziert werden. Die Güte der Verfahren mag für viele empirische Untersuchungen noch zu gering sein. Die meisten wissenschaftlichen Arbeiten kommen bei der Evaluierung von Text Mining Verfahren einen Accuracy-Wert von ungefähr 80%.

Dennoch haben automatisierte Verfahren Vorteile, die sie für Anwendungen prädestinieren, bei denen ihre Nachteile eine untergeordnete Rolle spielen. Zu den Vorteilen zählt die Geschwindigkeit, mit der sehr große Textmengen ausgewertet werden können. Entsprechend sind Frühwarnsysteme eine Anwendungsmöglichkeit, die

Internetplattformen nach Meinungen von Internetnutzern durchsuchen, analysieren und Auskunft über die aktuelle Stimmung zu Unternehmen, Produkten oder gesellschaftlichen Themen zurückliefern. Anschließend kann mit klassischen empirischen Methoden fortgefahren werden, um spezifische Forschungsfragen zu untersuchen. Die Auswertung offener Fragen bei Befragungen ist ein weiteres Beispiel, wo automatisierte Inhaltsanalyseverfahren bereits eingesetzt werden. (Kaczmirek, Baier, & Züll, 2010) In Kapitel 8 werden weitere Anwendungen aufgezählt und näher beschrieben.

Somit konnten in dieser Arbeit sowohl Grenzen als auch Potentiale von Methoden zur Automatisierung der Inhaltsanalyse aufgezeigt werden.

Ob in Zukunft auch empirische Untersuchungen, die eine hohe Validität erfordern, automatisiert durchgeführt werden können, hängt von der Weiterentwicklung der Methoden und der interdisziplinären Zusammenarbeit auf diesem Gebiet ab. Betrachtet man die Herausforderungen, die auch im empirischen Teil (siehe Kapitel 7) dieser Arbeit sichtbar wurden, ist eine mit manuellen Inhaltsanalysen vergleichbare Validität wohl kaum erreichbar. Als Ergänzung oder zu Vorstrukturierung manueller Analysen werden automatisierte Verfahren jedoch schon heute erfolgreich eingesetzt.

Dass Methoden der automatisierten Inhaltsanalyse in immer mehr soziologischen Untersuchungen eingesetzt werden (Cornelia Züll, 2002) verdeutlicht auch, dass mit zunehmender Verlässlichkeit der Resultate, Forscher immer mehr Vertrauen in durch Computerprogramme automatisierte Analysen gewinnen. Andererseits erkennt man am Fehlen standardisierter Methoden und einer einheitlichen Terminologie die Aktualität und den Forschungsbedarf auf diesem Feld.

10 Literaturverzeichnis

García Flore, J., Gillar, L., Ferret, O., & de Chalenda, G. (kein Datum). *Bag-of-senses versus bag-of-words: comparing semantic and lexical.* Fontenay aux Rose: DS/DICST.

Alexandra Balahur, R. S. (2009). *Rethinking Sentiment Analysis in the News: from Theory to Practice and back.* European Comission - Joint Research Center.

Atteslander, P. (2003). *Methoden der empirischen Sozialforschung.* Berlin: Walter de Gruyter GmbH.

Barber, I. (9. 7 2010). *SearchBusinessAnalytics.com*. Abgerufen am 2. 5 2011 von http://searchbusinessanalytics.techtarget.com/definition/opinion-mining-sentiment-mining

Berelson, B. (1952). *Content analysis in communication research.* Free Press.

Berth, H. (3. 7 2000). *Technische Universität Dresden*. Abgerufen am 30. 12 2011 von http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2002/02_02.pdf

Bortz, J. (1999). *Statistik für Sozialwissenschaftler.* Berlin: Springer.

Brosius, H.-B., Koschel, F., & Haas, A. (2008). *Methoden der empirischen Kommunikationsforschung: Eine Einführung.* Wiesbaden: VS Verlag für Sozialwissenschaften.

Bruno Ohana, B. T. (1. 6 2011). *Sentiment Classification with RapidMiner*. Abgerufen am 3. 11 2011 von http://kmandcomputing.blogspot.com/

Carsten Felde, H. B. (2006). *Evaluation von Algorithmen zur Textklassifikation.* Freiberg: TECHNICAL UNIVERSITY BERGAKADEMIE FREIBERG.

Claude Sammut, Geoffrey I. Webb. (2011). *Encyclopedia of Machine Learning.* New York: Springer.

Cornelia Züll, J. L. (2002). *Computerunterstützte Inhaltsanalyse: Literaturbericht zu neueren Anwendungen.* Mannheim: ZUMA.

Dadvar, M., Hauff, C., & de Jong, F. (2011). *Scope of Negation Detection in Sentiment Analysis.* University of Twente.

Diekmann, A. (2009). *Empirische Sozialforschung.* Hamburg: Rowohlt.

Eitrich, T. (2003). *Support-Vektor-Maschinen und ihre Anwendung auf Datensätze aus der Forschung.* Forschungszentrum Jülich in der Helmholtz-Gemeinschaft.

Esuli, A., & Sebastiani, F. (kein Datum). *SENTIWORDNET: A Publicly Available Lexical Resource.* Pisa: Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche.

Evert, S. (2004). *The Statistics of Word Cooccurrences* . Universität Stuttgart.

Fawcett, T. (2003). *ROC Graphs: Notes and Practical Considerations for Data Mining Researches.* Pao Alto: HP Laboratories Palo Alto.

Frake, W. (1992). *Stemming Algorithm.* Baeza-Yate: Frakes.

Früh, W. (2007). *Inhaltsanalyse.* Konstanz: UVK Verlagesgesellschaft mbH.

Fuchs, C. (2006). *The Self-Organization of Cyberprotest.* ICT&S Center.

Group, T. S. (2003). *Stanford Log-linear Part-Of-Speech Tagger*. Abgerufen am 29. 12 2011 von http://nlp.stanford.edu/software/tagger.shtml

Group, T. S. (14. 9 2011). *Stanford Log-linear Part-Of-Speech Tagger*. Abgerufen am 3. 11 2011 von http://nlp.stanford.edu/software/tagger.shtml

Heyer, G., Quasthoff, U., & Wittig, T. (2008). *Text Mining: Wissensrohstoff Text.* Witten: W3L.

Hüftle, M. (2006). *Methoden zur Klassifikation.* Hannover.

Juliane Landmann, C. Z. (2004). *Computerunterstützte Inhaltsanalyse ohne Diktionär?* ZUMA Nachrichten.

Kaczmirek, L., Baier, C., & Züll, C. (2010). Wie empfinden Teilnehmer die Fragen in Online-Befragungen? Entwicklung eines Diktionärs für die automatische Codierung freier Antworten. In M. Welker, & C. Wünsch, *Die Online-Inhaltsanalyse.* Köln: Halem.

Kaiser, C. (2009). Opinion Mining im Web 2.0 – Konzept und Fallbeispiel. In M. Knoll, & A. Meier, *Web & Data Mining* (S. 90-99). dpunkt.

Kai-Uwe Carstensen, C. E. (2010). *Computerlinguistik und Sprachtechnologie. Eine Einführung.* Heidelberg: Spektrum Akademischer Verlag.

Kevin Gimpel, N. S. (20114). *Part-of-Speech Tagging for Twitter: Annotation, Features and Experiments.* Portland: Association for Computational Linguistics.

Kracauer, S. (1952). *The Challenge of Qualitative Content Analysis.* Public Opinion Quarterly 16.

Kromrey, H. (2006). *Empirische Sozialforschung.* Stuttgart: Lucius & Lucius Verlagsgesellschaft mbH.

Lee, B. P. (2004). *A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts.* Proceedings of the ACL.

Ludwig-Mayerhofer, W. (20. 12 2004). *ILMES - Internet-Lexikon der Methoden der empirischen Sozialforschung*. Abgerufen am 1. 7 2011 von http://www.lrz.de/~wlm/ilm_v5.htm

Maral Dadvar, C. H. (2011). *Scope of Negation Detection in Sentiment Analysis.* Amsterdam: Dutch research school for Information and Knowledge Systems (SIKS).

Martin R. Herbers, A. F. (2010). Reliabilität und Validität bei Online-Inhaltsanalyse. In C. W. Martin Welker, *Die Online-Inhaltsanalyse* (S. 240). Köln: Herbert von Halem Verla.

Merten, K. (1995). *Inhaltsanalyse: Einführung in Theorie, Methode und Praxis.* VS Verlag für Sozialwissenschaften.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, Vol. 2, No 1-2*.

Peter Gentsch, A.-M. Z. (11. 1 2011). *Wie messe ich den Erfolg meines Social Media Engagement?* Abgerufen am 28. 12 2011 von http://www.slideshare.net/UNISGresearch/nextcc-fokusgruppe-sm-measurement-prof-peter-gentsch

Peter Gentsch, M. H. (28. 9 2009). *Mehr Erfolg im Marketing mit dem Social Web Radar*. Abgerufen am 28. 12 2011 von http://www.wobook.com/WBxC2E40qw0b-18-a/BIG_Mehr-Erfolg-im-Marketing-mit-dem-Social-Web-Radar_demexco2009_46S/Page-18.html

Porter, M. (1980). *An algorithm for suffix stripping.* Program.

Riggert, W. (2009). *ECM - Enterprise Content Management: Konzepte und Techniken rund um Dokumente* . Wiesbaden: Vieweg+Teubner Verlag.

Rössler, P. (2005). *Inhaltsanalyse.* Stuttgart: UTB.

Rüf, F., Böcking, S., & Kummer, S. (2010). Automatisierte Inhaltsanalysen im Internet: Möglichkeiten und Grenzen am Beispiel des SINDBAD-Knowledge-Generators. In M. Welker, & C. Wünsch, *Die Online-Inhaltsanalyse* (S. 313-339). Köln: Herbert von Halem Verlag.

Scharkow, M. (2010). Lesen und lesen lassen - Zum State of the Art automatischer Textanalyse. In M. Welker, & C. Wünsch, *Die Online-Inhaltsanalyse* (S. 340-364). Köln: Herbert von Halem.

Stefano Baccianella, A. E. (2010). *SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.* LREC'10.

Stuttgart, I. f. (23. 7 1996). *Die Wortformen der geschlossenen Wortarten im Stuttgart-Tübingen Tagset (STTS).* Abgerufen am 28. 12 2011 von Stuttgart-Tübingen Tagset (STTS): http://www.sfs.uni-tuebingen.de/Elwis/stts/Wortlisten/WortFormen.html

Tang, X., Yang, C., & Zhou, J. (2009). *Stock Price Forecasting by Combining News Mining and Time Series Analysis.* IEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.

Vaithyanathan, B. P. (2002). *Thumbs up? {Sentiment} Classification using Machine Learning Techniques.* Proceedings of the 2002 Conference on Empirical Methods in Natural.

Welker, M., & Wünsch, C. (2010). *Die Online-Inhaltsanalyse.* Köln: Herbert von Halem Verlag.

Werner Faulstich, H.-W. L. (1993). *Arbeitstechniken für Studenten der Literaturwissenschaft.* Tübingen: Narr.

LEBENSLAUF – PAUL SCHNEEWEISS

Persönliche Daten

Name:	Paul Schneeweiß
E-Mail:	paul.schneeweiss@gmx.at
Geburtsdaten:	3. November 1982 in Krems a.d. Donau

Ausbildung

2004 – 2012	Diplomstudium Soziologie (rechts-, sozial- und wirtschaftswissenschaftliche Studienrichtung) Universität Wien
2007 – 2009	Diplomstudium Wirtschaftsinformatik Technische Universität Wien
2007 – 2007	Auslandssemester in Spanien Universidad de Alicante
2004 – 2007	Bakkalaureatsstudium Wirtschaftsinformatik Universität Wien
1999 – 2003	Höhere Technische Lehranstalt St. Pölten Ausbildungszweig Elektronik

Berufliche Erfahrungen

Seit 08/2009	Business Intelligence Consulting Firma Altran Österreich
04/2009-05/2010	Business Intelligence Entwicklung Firma ERESNET
02/2006-01/2009	Studienmitarbeit, Datenbankentwicklung Freiberufler bei TNS Info Research Austria
08/2008-09/2008	Softwareentwicklung Auslandspraktikum in Brasilien an der Universität INATEL
03/2008-04/2008	Datenqualitätsverbesserung Freiberufler bei Firma Fujitsu Siemens
07/2002-07-2002	Softwareentwicklung Praktikum bei Firma Siemens
08/2001-08/2001	Informatik Instandhaltung Praktikum im Krankenhaus Krems

Sprachkenntnisse	Deutsch (Muttersprache) Englisch (Fließend) Spanisch (Gut) Portugiesisch (Grundkenntnisse)
IT-Kenntnisse	
	Business Intelligence ETL, DWH Architekturen und Modellierungstechniken, MDX, MS SSIS, SSRS, SSAS, Informatica PowerCenter, Pentaho Data Integration - Kettle, SAP Business Objects Web Intelligence, Universe Designer, Crystal Reports, Xcelsius, InfoView
	Datenbanken MS SQL Server, MS Access, Oracle, MySQL
Statistik-Kenntnisse	SPSS, SPSS Clementine, Rapid Miner, Data Mining, Text Mining
Sonstige Kenntnisse	Erfahrung in der empirischen Sozialforschung Betriebswirtschaftliche Kenntnisse Auslandserfahrung
Zertifizierungen	Oracle Database: SQL Certified Expert MCITP: MS SQL Server 2008, Business Intelligence Developer MCITP: MS SQL Server 2008, Database Developer MCTS: MS SQL Server 2008, Implementation and Maintenance SAP Certified Application Associate - BusinessObjects Web Intelligence XI 3.x ITIL V3 Foundation EBC*L European Business Competence Licence (Level B)