LEUPHANA
UNIVERSITÄT LÜNEBURG

Bachelorarbeit

zur Erlangung des Grades Bachelor of Science (B.Sc.)

zum Thema

Inhaltsanalytische Untersuchung zum Thema Nachhaltigkeit mit RapidMiner

Content analytical survey on the subject of sustainability with RapidMiner

Datum der Abgabe: 26.01.2016

Verfasserin: Theresa Schöbel

Hochschule: Leuphana Universität Lüneburg
Studiengang: Umweltwissenschaften B.Sc.
Matrikelnr.:

Betreuer: Prof. Dr. Jan Wilk
Prof. Dr. Guido Barbian

Kurzfassung

Auf Grund der stetig wachsenden Menge an Daten gewinnt die automatische Datenanalyse durch Algorithmen zunehmend an Bedeutung. Im Speziellen trägt die Analyse von Texten ohne manuelles Zutun zu einer erhebliche Erleichterung der Extraktion von relevanten Informationen bei. Sprachliche Informationen können neben der Zuordnung zu Kategorien auf Regeln und Muster untersucht werden. Diese Art der Untersuchung fällt in den Bereich des Text Minings und in der vorliegenden Arbeit geht es darum, eine qualitative Inhaltsanalyse zum Thema Nachhaltigkeit nachzuempfinden. Es soll geprüft werden, in wie weit automatisierte Verfahren in der Lage sind, Ergebnisse einer bereits bestehenden Untersuchung zu erzielen.

In der Durchführung werden mit der OpenSource Software RapidMiner vier Prozesse erstellt, die darauf abzielen, Zeitungsartikel auf ihren Inhalt zu analysieren. Unter anderem werden eine Assoziationsanalyse und eine Klassifikation realisiert, deren Ziel es ist, den Kontext und die Verwendung des Begriffes der Nachhaltigkeit in den Medien zu untersuchen. Die vorliegende Studie will prüfen, ob automatisierten Methoden im Vergleich zu manuellen Verfahren hinreichende Ergebnisse liefern können, sodass die hiesigen Resultate an denen der zu Grunde gelegten Studie von Fischer und Haucke gemessen werden sollen.

Die Ergebnisse zeigen, dass die Realisierung einer Inhaltsanalyse mit RapidMiner möglich ist und zu erheblichen Zeiteinsparungen gegenüber konventionellen Verfahren führt. Sie zeigen jedoch auch, dass sich die Minderung des Aufwandes in der Ergebnisqualität widerspiegelt und somit der alleinige Einsatz von Text Mining Verfahren zur Analyse von spezifischen Kontexten noch nicht ausreichend ist.

Abstract

Due to the continuously growing amount of data, data analysis becomes more and more important. Especially the analysis of text documents without manual support helps significally to extract the relevant information. Besides different possibilities to determine text documents, rules and patterns can be detected or categories can be assigned to language information. This kind of investigations is part of text mining and the present study is about to recreate a qualitative content analysis on the subject of sustainability. The aim is to prove how close automatic procedures can reach the results of a consisting research.

The implementation contains four different processes, made with the open source software RapidMiner, that analyse the content of newspaper articles. An association analysis as well as a classification can be done to explore the context and use of the term of sustainability in public media. The present study finds out whether automatic methods or manual techniques can provide sufficient outcomes by comparing these results with those on the basis of a study by Fischer and Haucke.

The findings show that it is possible to realize a content analysis with RapidMiner and to reduce the amount of time needed with conventional methods – but they also show that the results decline with the effort being diminished. Therefore text mining methods cannot be used on themselves to analyse specific contexts so far.

Inhaltsverzeichnis

Abkürzungsverzeichnis

CRISP DM	Cross Industry Standard Process Model
FN	False Negative
FP	False Positive
KDD	Knowledge Discovery in Databases (Wissensentdeckung in Datenbanken)
OCR	Optical Character Recognition (optische Zeichenerkennung)
TN	True Negative
TP	True Positive

Abbildungsverzeichnis

Tabellenverzeichnis

1. Einleitung

Data Mining und Text Mining Verfahren erfreuen sich immer größerer Beliebtheit. Grund hierfür ist die steigende Anzahl an Daten und Dokumenten in vielerlei Disziplinen. Sei es im Alltag, in den Wissenschaften, in den Medien – überall nimmt die Menge an (unstrukturiertem) Material zu [vgl. Fayyadd 1996: 37], sodass eine effiziente Analyse dieser Datenmengen immer relevanter wird. Beispielsweise sorgt die Anwendung von Data Mining Methoden innerhalb von Knowledge-Discovery in Databases (KDD) zur Lösung jeglicher wirtschaftlicher Optimierungsprobleme für erhebliche Wettbewerbsvorteile. So kann der Verkauf – digital genauso wie im Supermarkt – gesteigert werden, indem durch Auswertung bereits vergangener Verkäufe die Produkte entsprechend platziert werden. Dies ist jedoch nur ein winziges Beispiel, für die vielfältigen Möglichkeiten und Nutzungen der Wissensentdeckung in Datenbanken.

In dieser Studie soll es die Nutzung ähnlich bis identischer Methoden zur Analyse von Texten untersucht werden. Fokus stellt hier das Text Mining, ein Teilgebiet des Data Minings, dar [vgl. Siegmund 2006: 43]. Die steigende Menge an verfassten Schriftstücken, die für sich genommen als unstrukturierte Daten gesehen werden können [vgl. Sharafi 2013: 67], erfordert den Einsatz von automatischen Verfahren, um auch in Bereichen der Bildung und Wissenschaft, effizient Rückschlüsse auf Phänomene der Gesellschaft ziehen zu können. Folglich steht der Inhalt der Texte im Mittelpunkt.

Die Nutzung von Text Mining Verfahren im Bereich der Wissenschaften konzentriert sich zu einem Großteil auf seine Entstehungsbereiche, obwohl auch in anderen Fachbereichen Texte auf ihren Inhalt oder ihre Struktur analysiert werden [vgl. Seidel 2013: 27]. In der sozialwissenschaftlichen Forschung wird zur Analyse von Texten auf die qualitative Inhaltsanalyse zurückgegriffen. Mit der qualitativen Inhaltsanalyse hat Phillip Mayring eine Methode vorgestellt, die Kommunikation auf ihren Inhalt analysieren will, mit dem Anspruch den wissenschaftlichen Kriterien gerecht zu werden. Es handelt sich dabei um ein strukturiertes, regel- und theoriegeleitetes Vorgehen, das quantitative und qualitative Analyseschritte miteinander kombiniert, um das Besondere, Individuelle oder auch Einzelne zu untersuchen [Mayring 2008: 13; vgl. ebd.: 18]. Im Mittelpunkt der qualitativen Inhaltsanalyse steht die Entwicklung eines Kategoriensystems [ebd.: 43]. Ziel ist es Intersubjektivität zu gewähren, sodass jeder die Analyse nicht nur nachvollziehen kann, sondern bei einer eigenen Durchführung auf die gleichen Ergebnisse kommt [ebd.].

Der Grund, dass Text Mining Methoden bisweilen gar nicht oder kaum in diesem Bereich genutzt werden, zeigt den aktuellen Forschungsbedarf. Es existieren Arbeiten, die sich mit der konkreten Methodik des Text Minings auseinandersetzen (z.B. L. M. Seidel oder R. Witte und J. Mülle) oder Arbeiten, die sich zur konkreten Anwendung der Text Mining Verfahren für bestimmte Probleme annehmen (wie beispielsweise fürs Customer Relationship

Management) sowie einige wenige, die sich mit einem Vergleich der Methoden und dem Einsatzgebiet außerhalb der Informatik auseinandersetzen (z.B. M. Scharkow).
Eine fachübergreifende Nutzung des Text Minings kann nur dann bewährte Verfahren (wie in den Sozialwissenschaften die qualitative Inhaltsanalyse) ergänzen oder ersetzen, wenn die Ergebnisse den wissenschaftlichen Kriterien entsprechen. Mit anderen Worten: nur wenn valide, reliabel und replizierbare Ergebnisse erzielt werden, kann der Einsatz automatischer Verfahren die Quantität manueller Verfahren mühelos steigern.
Der Vorteil eines größeren Datenumfangs ist eine minimierte statistische Fehlerquote, die wiederum für eine höhere Verlässlichkeit im Bezug auf die Interpretation der Ergebnisse sorgt [vgl. Scharkow 2011: 13]. Aus diesem Grund soll in dieser Arbeit eine qualitative Inhaltsanalyse aus dem Bereich der Nachhaltigkeitskommunikation nachempfunden und somit herausgefunden werden, in wie weit mit der Open Source Software RapidMiner Studio[1] eine qualitative Inhaltsanalyse realisierbar ist.
Der Aufbau gliedert sich in vier Kernteile und beginnt mit einer Betrachtung des KDD. Weiterhin werden Data und Text Mining betrachtet und vor allem voneinander abgegrenzt. Im Anschluss folgen Grundlagen zur Durchfühung der Untersuchung, sodass die Software, die zur Analyse herangezogenen Daten und vorbereitende Analyseschritte vorgestellt werden. Im weiteren Verlauf werden Ansätze und Lösungen besprochen, wie eine qualitative Untersuchung mit RapidMiner realisiert werden kann. Dazu zählen eine Umformung der untersuchten Dokumente, eine Analyse der Stimmung in den Texten, eine Assoziationsanalyse sowie eine Klassifizierung. Vor der praktischen Umsetzung wird der zugehörige theoretische Hintergrund (zum Verständnis) vorangestellt. Daran wird sich eine Gegenüberstellung der Methoden anschließen, woraufhin die Arbeit mit einem Fazit beendet wird.

2. KDD, Data Mining und Text Mining

Bevor mit einer Analyse des Material und einer konkreten Anwendung der Techniken begonnen wird, steht die Klärung der Begrifflichkeiten. Definitionen unterschiedlicher Autoren (wie z.B. J. Hipp, H. Petersohn oder A. Sharafi) zum Thema Knowledge Discovery in Databases belaufen sich unter Anderem auf einen Artikel von Fayyadd, Piatetsky-Shapiro und Padhraic Smyth, die bereits 1996 eine Abgrenzung und Definition von KDD und Data Mining beschrieben haben [vgl. Fayyadd, Piatetsky-Shapiro und Padhraic Smyth 1996: 37].
„In our view, KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process.“ [ebd.: 39] Insgesamt geht es folglich um die Gewinnung von Wissen aus Daten, die normalerweise durch manuelle

[1] Verfügbar unter: https://rapidminer.com/.

Analyse und/oder Interpretation erfolgt, doch auf Grund der steigenden Datenmengen nur unter großem Aufwand erfolgen kann [Sharafi 2013: 51]. Die Anwendung von Methoden und Algorithmen des Data Mining stellen zu diesem Zweck eine computergestützte Lösung dar. Durch die bloße Anwendung von Algorithmen kann jedoch kein Wissen generiert werden, sondern es können lediglich Muster erkannt und extrahiert werden, die wiederum in dem richtigen Kontext und Betrachtung zum eigentlichen Wissen leiten. Auf Grund dessen stellt Data Mining den Kern eines KDD-Prozesses dar, zu dem noch weitere Schritte gehören.

> „The additional steps in the KDD process, such as data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the results of mining, are essential to ensure that useful knowledge is derived from the data.“ [Fayyad 1996: 39]

Einerseits zielt KDD darauf ab ein Werkzeug bereit zu stellen, mit dem so weit wie möglich automatisch Muster erkannt werden können [ebd.: 40], anderseits ist KDD ein iterativer Prozess bei dem währenddessen viele Entscheidungen vom Anwender getroffen werden [ebd.: 42].

Für die einzelnen Schritte eines KDD-Prozesses gibt es mehrere Vorgehensmodelle, die für eine bessere Planung und Umsetzung in Organisationen entwickelt wurden [vgl. Sharafi 2013: 57]. Innerhalb von vier Schritten fasst Sharafi die Gemeinsamkeit der unterschiedlichen Prozessmodelle wie folgt zusammen [ebd.]:

1. Zielfestlegung und Analyse der Domäne
2. Datenvorverarbeitung
3. Methodenanwendung / Analyse (Data Mining)
4. Interpretation der Ergebnisse

Eine Darstellung nach Fayyadd umfasst fünf Schritte auf dem Weg der Wissensexploration und wird in Abbildung 1 veranschaulicht.

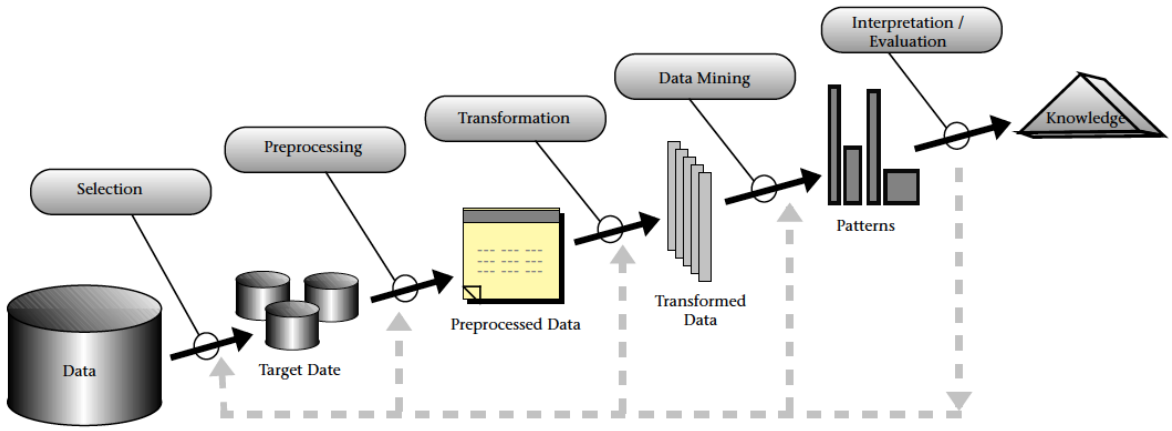

Abbildung 1: Darstellung eines modellhaften Data Mining-Prozessablauf [Fayyadd 1996: 41].

Zuerst erfolgt die Festlegung auf einen konkreten Untersuchungsgegenstand. Ist dieser ausgewählt, können daraufhin das Zielgebiet analysiert und die erforderlichen Daten aus denen Informationen gewonnen werden sollen, ausgewählt werden. Bei der Datenvorverarbeitung geht es dann darum, die Daten auf das Analyseziel vorzubereiten, indem sie auf wesentliche Merkmale reduziert und gegebenenfalls für die Anwendung von Algorithmen in eine nutzbare Darstellungsform transferiert werden. Anschließend kann die eigentliche Analyse mit den auswählten Methoden erfolgen und die erzielten Ergebnisse entsprechend der Ausgangsfrage interpretiert und dargestellt werden.
Ein allgemeines Prozessmodell, dass sich für jegliche Data Mining-Anwendungen nutzen lässt und hier der Vollständigkeit halber angesprochen werden soll, nennt sich CRISP DM, was für Cross Industry Standard Process Model für Data Mining steht. Dies hat sich in der Praxis als Standard durchgesetzt und umfasst im wesentlichen die bereits beschriebenen Schritte [Hipp 2003: 10; vgl. Sharafi 2013: 64f.].

Wie bereits erläutert, fallen unter Data Mining die Verwendung von Algorithmen zur Mustererkennung. Ans KDD angrenzende Fachbereiche, wie das maschinelle Lernen, das Wissensmanagement oder die Statistik sowie Wahrscheinlichkeitsrechnung, bilden ein Diversa aus dem sich die Methoden des Data Mining entwickelt haben [Fayyadd 1996: 39]. Aus dieser Übersicht wird deutlich, dass die zur Analyse verwendeten Daten mit einer gewissen Struktur, den Methoden deutlich leichter unterliegen, als ohne. Somit kann zwischen strukturierten und unstrukturierten Daten unterschieden werden [vgl. Sharafi 2013: 67], woraus sich auch der wesentliche Unterschied zwischen Data und Text Mining ableiten lässt. Beim Text Mining werden Texte, die zu unstrukturierten Daten zählen [vgl. ebd.], verarbeitet, wobei bei genauerer Betrachtung deutlich wird, dass auch Texte eine gewisse Struktur, wie die Semantik oder Grammatik, aufweisen [vgl. ebd.]. Aus diesem Grund ist es möglich, Texte mit den Verfahren des Data Minings zu untersuchen. Insgesamt kann Text Mining analog zum Data Mining definiert werden, indem es als „Menge von Techniken zum Entdecken und automatischen Extrahieren von neuen, zuvor unbekannten Informationen aus Texten“ zusammengefasst wird [Siegmund 2006: 42f.]. Es stellt ein eigenständiges, spezielles Gebiet des Data Minings dar und sein Ursprung wird mitunter in der Inhaltsanalyse gesehen [Leong/Ewing/Pitt 2004:188 zitiert nach Sharafi 2013: 80].

3. Material und Hinführung

Nach kurzer Einführung ins Themengebiet, werden im kommenden Abschnitt konkrete Grundlagen folgen. Dabei sollen die genutzte Software, das Material und erste allgemeine Vorbereitungsschritte erläutert werden.

3. 1. Arbeiten mit RapidMiner & Datenselektion

Zur Anwendung der Data bzw. Text Mining Techniken für eine qualitative Inhaltsanalyse wird die Software RapidMiner Studio herangezogen. Dabei handelt es sich um eine Open Source Lösung für Data Mining Anwendungen mit der Prozesse erstellt werden, indem einzelne Operatoren per drag and drop zu einem Prozess kombiniert werden. Je nach Arbeitsziel können eine Reihe von Erweiterungen herangezogen werden, wie beispielsweise ein Package zur Text Verarbeitung. In dieser Arbeit wird mit der RapidMiner Studio Version 6.5 gearbeitet sowie allen verfügbaren Erweiterungen[2].

Innerhalb von insgesamt 14.907 Zeitungsartikeln, die sich aus drei Jahrgängen (2001, 2007, 2013) sechs großer Medien (Frankfurter Allgemeine Zeitung, Süddeutsche Zeitung, Die Zeit, Die Welt, TAZ, Der Spiegel) zusammensetzten, soll der Begriff „nachhaltig*“ betrachtet und seine Verwendung analysiert werden. Die spezielle Schreibweise drückt aus, dass hinter „nachhaltig“ noch beliebige Zeichen in die Suche mit einbezogen werden, wodurch sowohl der Begriff *nachhaltig* als auch *Nachhaltigkeit* und weitere gefunden werden. Somit ist das Zeichen „*“ ein Platzhalter bzw. eine sogenannte Wildcard [vgl. Schicker 2014: 111].

Eine solche Auseinandersetzung hat mit den identischen Daten bereits stattgefunden (vgl. Abschnitt 3. 2.), sodass die mittels Text Mining erzielten Ergebnisse mit denen der bestehenden Studie von Fischer und Haucke verglichen werden können. Davon ausgehend wird die Möglichkeit geschaffen eine Bilanz zu ziehen, wie effektiv der Einsatz von Algorithmen gegenüber einer manuellen Auswertung für eine solche Aufgabe sein kann. Darüber entsteht die Möglichkeit die Qualität der automatischen Analyse zu prüfen und den Einsatz von Text Mining für eine Inhaltsanalyse zu bewerten.

3. 2. Hintergründe zur Grundlagenstudie

In der herangezogenen Grundlagenstudie „‘To sustain, or not to sustain…?’ An empirical analysis of the usage of ‘sustainability’ in German newspapers“ von Daniel Fischer und Franziska Haucke wurde in zwei Schritten die Verwendung und Bedeutung des Begriffes *nachhaltig** in deutschen Printmedien analysiert. Zu Beginn wurde eine Trendanalyse in den genannten Medien über die Jahre 1995 - 2014 durchgeführt, um einen Überblick über die Frequenz der Nutzung des Begriffes zu bekommen [Fischer/Haucke 2015: 1]. Die Jahrgänge 2001, 2007 und 2013, die auch in der vorliegenden Arbeit untersucht werden sollen, wurden daraufhin in der Tiefe analysiert. Anlass für eine solche Untersuchung gibt die Rolle der Medien, die zur Bildung der öffentlichen Meinung und als Verbindung zwischen Wissenschaft und Gesellschaft erheblich zur Definition des Terms der Nachhaltigkeit beiträgt [ebd.: 2]. Da das Konzept der Nachhaltigkeit seit seiner Einführung (1987 hat die

[2] Hiervon ausgenommen sind folgende Erweiterungen: Octave Extension, Rapid Development und RapidProM.

Brundtland-Kommission die weitreichendste Definition nachhaltiger Entwicklung veröffentlicht [vgl. Michelsen/Adomßent 2014: 12].), je nach Kontext seiner Nutzung unterschiedliche Bedeutungen erhält, erschließt die Studie durch ein Darlegen des Ist-Zustandes, ob der Begriff im Journalismus geschärft oder gar weniger genutzt werden sollte [Fischer/Haucke 2015: 13]. Obwohl die Studie selbst aus zwei Teiluntersuchungen besteht, wird im folgenden nur auf die Tiefenanalyse eingegangen, da diese den Teil ausmacht, der mittels Text Mining nachempfunden werden soll.

Die Auswahl der herangezogenen Zeitungsverlage wurde so getroffen, dass sowohl jegliche politische Orientierung abgedeckt wird, als auch eine große Reichweite der Medien gewährleistet ist [ebd.: 4]. Innerhalb der Jahre, zu denen die Frequenz der Verwendung von *nachhaltig** untersucht wurde, wurden drei Jahre ausgewählt, um die Bedeutung des Begriffs im jeweiligen Kontext zu analysieren. Ausschlaggebend für die Jahre 2001, 2007 und 2013 war erstens, dass sich die semantische Bedeutung eines Begriffes nicht innerhalb von ein oder zwei Jahren ändert und zweitens, die Jahre mit internationalen Konferenzen gemieden werden sollten [vgl. ebd.: 5], um möglichst allgemeingültige Resultate zu erzielen. In Jahren, in denen Folgekonferenzen zur Rio-Konferenz[3] 1992 gehalten wurden (2002, 2012) [Michelsen/Adomßent 2014: 19ff.; Fischer/Haucke 2015: 5], ist davon auszugehen, dass das Thema Nachhaltigkeit eine erhöhte Aufmerksamkeit in den Medien erfährt.

Des Weiteren wurden lediglich Artikel untersucht, die mehr als 300 Wörter beinhalten, damit „eine präzisere qualitative Analyse der jeweiligen Nutzung“ [Fischer/Haucke 2015: 5] möglich wird.

Die Tiefenanalyse erfolgte mit einem Kategoriensystem. Dafür wurden aus bestehenden Studien und Theorien Codes entwickelt (deduktive Kategorienbildung) [ebd.: 6], die die unterschiedliche Nutzung des Nachhaltigkeitsbegriffes abbilden sollen. Insgesamt sind auf diese Weise 10 verschiedene Bedeutungen ausgearbeitet worden (s. Tabelle 1).

[3] Als Rio-Konferenz wird die Konferenz der Vereinten Nationen über Umwelt und Entwicklung bezeichnet, die das Ziel hat, internationale Übereinkommen und Vereinbarungen zu treffen, um dem Handlungsbedarfs zu nachhaltiger Entwicklung gerecht zu werden [vgl. Michelsen/Adomßent 2014: 15].

Tabelle 1: Codes für die jeweilige Bedeutung, die *nachhaltig** je nach Kontext einnehmen kann. [Fischer/ Haucke 2015: 6]

Code	**Description: Sustainab* is used...**
Names	... as one part of a fixed term or name (e.g. a policy, an organization etc.)
Everyday language	... in the meaning of durable/permanent or if something was really intense or powerful.
Ecological	... with respect to the conservation of natural resources
Socio-cultural	... to address issues of (distributional) justice and conditions that allow human beings to meet their needs
Economical	... to refer to development that ensures that the economic system can continue to function in the future
Integrated	... to relate to the idea of sustainable development as it has emerged in the context of the UN: as the integrative consideration of ecological, economic and socio-cultural dimensions of development.
Critique	... to question the (usage of the) term as such or to criticize its vagueness
Deliberative	... to emphasize that the idea is not a fix concept but needs to be deliberatively specified
Responsibility	... to point out the consequences of our actions today on others living today and on future generations
Unclear	... in a way that cannot be related to any of the aforementioned meanings

Je Abschnitt mit *nachhaltig** ist händisch ein Code zugewiesen worden, sodass am Ende eine Gegenüberstellung stattfinden konnte, wie sich die Nutzung insgesamt darstellt und auch über die Jahre verändert hat. Fischer und Haucke sind zu dem Schluss gekommen, dass eine deutliche Zunahme des Terms zu verzeichnen ist und der Begriff im Lauf der Zeit weniger in einem alltagssprachlichen Sinne und mehr im Sinne des wissenschaftlichen Verständnisses genutzt wird [Fischer/Haucke 2015: 13].

3. 3. Allgemeines Preprocessing

Für eine qualitativen Inhaltsanalyse mit RapidMiner, steht zu Beginn die Aufbereitung der jeweiligen Dokumente. Damit die Texte mit Algorithmen untersucht werden können, ist das sogenannte Preprocessing bzw. die Datenvorverarbeitung einer der entschiedensten Schritte im Text Mining (vgl. Abschnitt 2). Die Bedeutung diesen Schrittes spiegelt sich auch in dem Arbeitsaufwand dafür wider, sodass Schätzungen 50% der Arbeitszeit dafür veranschlagen [vgl. Sharafi 2013: 60].

Um verlässliche und aussagekräftige Ergebnisse zu erzielen muss gewährleistet sein, dass alle Daten innerhalb eines Prozesses, gleichermaßen verarbeitet werden können. Im vorliegenden Fall wird ausschließlich mit .pdf-Files (Portable Document Format) gearbeitet. Dies portable Dokumentenformat von Adobe ermöglicht einen lückenlosen, plattform-

unabhängigen Austausch von elektronischen Dokumenten [Bütefisch/Petermann 2014: 128], sodass jeder Anwender die Datei originalgetreu betrachten und ausdrucken kann [ebd.]. Die gespeicherten Informationen in einer .pdf hängen trotz allem von der Software ab, mit der sie erzeugt wurden und so kann es sein, dass ein Text als Bild gespeichert ist. Aus diesem Grund wurde für alle Dokumente eine optische Zeichenerkennung durchgeführt (OCR), um vorhandene Bilder mit Text in maschinen-lesbaren Text zu konvertieren.
Des Weiteren umfasst das Preprocessing die Aufbereitung der Texte. Im Allgemeinen besteht dieser Schritt aus folgenden möglichen Teilschritten:

- Tokenisierung
- Stoppwörter entfernen / Filtern
- Stemming

Je nach Fragestellung muss der Text in Wörter, Sätze oder ganze Paragraphen zerteilt werden, was als Tokenisierung bezeichnet wird [vgl. Sharafi 2013: 85]. Somit wird der Text in Terme bzw. Tokens transformiert. Diese können wiederum selektiert werden, indem beispielsweise wenig aussagekräftige Terme von der Analyse ausgeschlossen werden. Stoppwörter, wie Artikel oder Präpositionen – also Wörter die überdurchschnittlich oft in Texten vorkommen – zählen zu solchen Termen mit wenig Aussagekraft [vgl. Ferber 2003: 21]. Um die Ergebnisse noch weiter zu präzisieren, können Wörter auf ihren Wortstamm zurückgeführt werden, wodurch „die Reduzierung der morphologischen Varianz von Begriffen erreicht werden [soll].“[Sharafi 2013: 88].
Im Zuge einer Assoziationsanalyse ist ein solches Stemming beispielsweise von Vorteil, da Worte trotz unterschiedlicher Schreibweise als gleiche Worte gezählt werden (vgl. Abschnitt 4. 2). Somit bleibt für das Preprocessing festzuhalten, dass die Art der Aufbereitung abhängig vom Ziel der Textanalyse ist.

4. Prozesse einer Inhaltsanalyse mit RapidMiner

Im folgenden wird die praktische Herangehensweise einer Inhaltsanalyse mit RapidMiner dargelegt. Die einzelnen Prozesse werden mit ihrem Ziel, dem zugehörigem theoretischen Hintergrund sowie ihren Ergebnissen vorgestellt. Des Weiteren werden die Ergebnisse im Hinblick auf die Grundlagenstudie diskutiert.

4. 1. Erzeugen von Passus mit *nachhaltig**

Da der Kontext des Gebrauches von *nachhaltig** bei der qualitativen Inhaltsanalyse im Fokus steht, wird zunächst ein Prozess gestaltet, der die Dokumente auf diese Kontexte reduziert. Die Texte werden mittels *Process Documents* strukturiert (s. Abb. 2), indem je ein Segment

bzw. Token von einem bis zum nächsten Punktzeichen erstellt wird. Auf diese Weise wird jedes Dokument in seine Sätze unterteilt, woraufhin diese (in einem Subprozess des *Process Documents* Operators) auf *nachhaltig** gefiltert werden. An dieser Stelle kann entschieden werden, ob beispielsweise lediglich die Sätze, in denen *nachhaltig** direkt enthalten ist, dargestellt werden sollen oder, ob zur Darstellung des Kontextes auch Sätze davor und danach mit einbezogen werden sollen. Da RapidMiner jedoch die Tokens als Ergebnis (hier alle ganzen Sätze) jeden für sich und alphabetisch sortiert ausgibt, birgt die Erweiterung auf Sätze vor und nach einem Satz mit *nachhaltig** keinen Vorteil, da der Bezug der Sätze zueinander verloren geht. Schlussendlich kann also nur die Filterung auf alle Sätze, die *nachhaltig** enthalten, sinnvoll sein. Um demnach ganze Paragraphen zum Codieren zu erhalten, mag ein Rückgriff auf andere Software zur computerunterstützten Inhaltsanalyse Erfolg versprechender sein [vgl. Mayring 2008: 101].

Übrig bleiben im Ergebnis insgesamt 19.302 Sätze, die *nachhaltig** enthalten, was in der Summe mit den Codierungen, die in der Grundlagenstudie vorgenommen wurden näherungsweise übereinstimmt [vgl. Fischer/Haucke 2015: 7]. Diese werden in einer Excel Tabelle gespeichert, wobei für jeden Satz eine Zeile angelegt wird (s. Anhang A, nachhaltig_Saetze.xlsx). Eine manuelle Codierung kann durch diesen Schritt bzw. Prozess beschleunigt werden, da die zu codierenden Stellen nicht einzeln definiert werden müssen. Außerdem können, von der Excel Tabelle ausgehend, weitere inhaltsanalytische Schritte vorgenommen werden, wie zum Beispiel das Anlegen einer Spalte für eine Codierung.

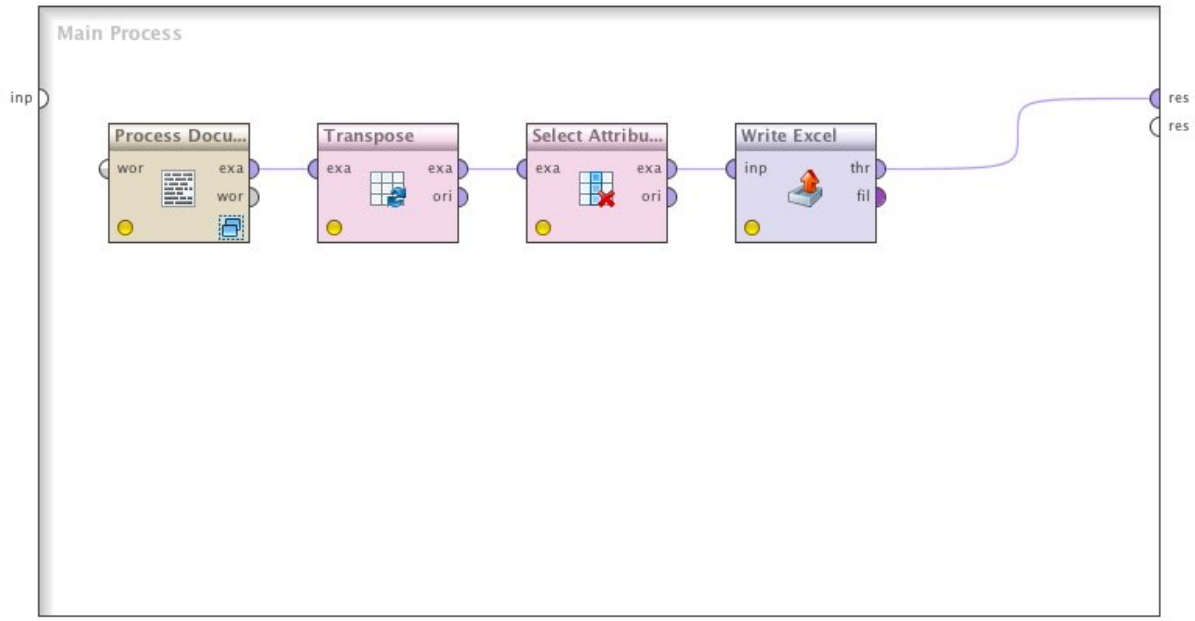

Abbildung 2: Darstellung des Prozesses zur Selektion der Sätze mit *nachhaltig** in RapidMiner.

4. 2. FP-Growth

Eine Methode des Data Minings, die sich zur Analyse von Texten heranziehen lässt, ist die Assoziationsanalyse. Sie zielt darauf ab gewisse Regelhaftigkeiten innerhalb eines großen Datensatzes ausfindig zu machen, und hat infolgedessen Assoziationsregeln zum Ergebnis [Petersohn 2005: 101].

Um herauszufinden, wie der Begriff der Nachhaltigkeit in den Medien genutzt wird, kann untersucht werden, ob sich Regeln für eine gemeinsame Verwendung von Worten finden lassen. Tritt der Term *nachhaltig** beispielsweise signifikant häufig mit einem anderen, aussagekräftigen Term auf, so kann je nach Bedeutung eines solchen Terms auf die Art der Verwendung geschlossen werden. Formal lässt sich eine Assoziationsanalyse beschreiben, indem zu erst eine Menge $I = \{i_1, i_2, \ldots i_n\}$ mit allen vorkommenden Objekten (*items*) zugrunde gelegt wird. Eine Menge T beschreibt eine Kombination der Objekte und wird Transaktion genannt. Je nach Anzahl der beherbergten Objekte wird von einem *k-itemset* gesprochen, wodurch k der Repräsentant der Größe eines *itemsets* ist. Folglich ist T eine Teilmenge von I ($T \subseteq I$). Alle Transaktionen sind in einer Menge D, einem „set of transactions“, zusammengefasst und für eine Menge X wird untersucht, ob es Teil von T ist. [ebd.: 102f.]

Die folgende Assoziationsregel beschreibt die logische Implikation zweier Items X und Y.

$$\{X \Rightarrow Y \,|\, X \in I, Y \in I \text{ und } X \cap Y = \varnothing\}$$

(4.2 - 1 [Hildebrandt / Boztuğ 2007: 223])

Im vorliegenden Fall ist die Menge an Objekten die Menge an Tokens aus allen Zeitungsartikeln. Somit ist ein Zeitungsartikel eine Transaktion, weil dort eine Teilmenge an Tokens zusammengeschlossen ist. Das *set of transactions* gestaltet sich entsprechend aus allen untersuchten Zeitungsartikeln. Diese Grundgesamtheit soll darauf untersucht werden, ob im Zusammenhang mit dem Begriff *nachhaltig** Gemeinsamkeiten gefunden werden können. Damit davon ausgehend aussagekräftige Regeln erstellt werden, gibt es Kriterien der Güte zur Einschränkung. Wird beispielsweise ein Zusammenhang im Sinne von „Wenn – Dann“ gefunden, beschreibt die Konfidenz die Stärke des Zusammenhalts. In Abhängigkeit von der Anzahl der Transaktionen, die diese Regel befolgen [Hildebrandt/Boztuğ 2007: 224], drückt sie aus, dass eine zufällig gezogene Transaktion mit der Präsmisse X auch Y enthält [Hipp 2003: 16]. Grundlage für diese bedingte Wahrscheinlichkeit bildet der Support, der die statistische Signifikanz der Regel beschreibt [Hildebrandt/Boztuğ 2007: 224]. Im Prinzip wird durch den Support ausgedrückt, wie hoch das relative Vorkommen einer spezifischen Kombination von *items* innerhalb einer Transaktionsmenge ist [ebd.].

Zur Durchführung einer Assoziationsanalyse stehen eine Reihe von Algorithmen zur Verfügung [vgl. Peterson 2005: 103]. Für die nachstehende Untersuchung soll der sogenannte FP-Growth herangezogen werden, dessen Vorteil im Verarbeiten großer Datenmengen liegt [vgl. RapidMiner 6.5: FP-Growth Operator Beschreibung]. Indem eine komprimierte Kopie der Daten in Form einer – als FP-Tree bezeichneten – Baumstruktur erstellt wird, wird für jede Transaktion die Menge häufiger *items* berechnet [ebd.].

Konkret gestaltet sich die Umsetzung der Assoziationsanalyse in RapidMiner aus den in Abschnitt 3.3 erläuterten Text Mining Schritten. Zu Beginn werden Tokens aus den Worten der Zeitungsartikel erzeugt. Diese werden auf ihren Wortstamm zurückgeführt, von Groß- und Kleinschreibung befreit und um Stoppwörter reduziert. Daraufhin wird eine Dokument-Term-Matrix erzeugt, in der binär gekennzeichnet ist, welcher Token in welchem Dokument vorkommt. Davon ausgehend kann der FP-Growth angewandt werden, woraufhin aus den gebildeten häufigen *itemsets* Assoziationsregeln aufgestellt werden (s. Abb. 3). Zu beachten ist, dass an dieser Stelle keinerlei Einschränkungen vorgenommen wurden, um die Zeitungsseiten auf die Absätze mit *nachhaltig** einzugrenzen. Stattdessen sind als regulierende Kriterien das Wort „nachhaltig“ als Bestand eines häufigen *itemsets* festgelegt, sowie ein Schwellenwert für den Support und die Konfidenz zur Regelerzeugung vorgegeben worden.

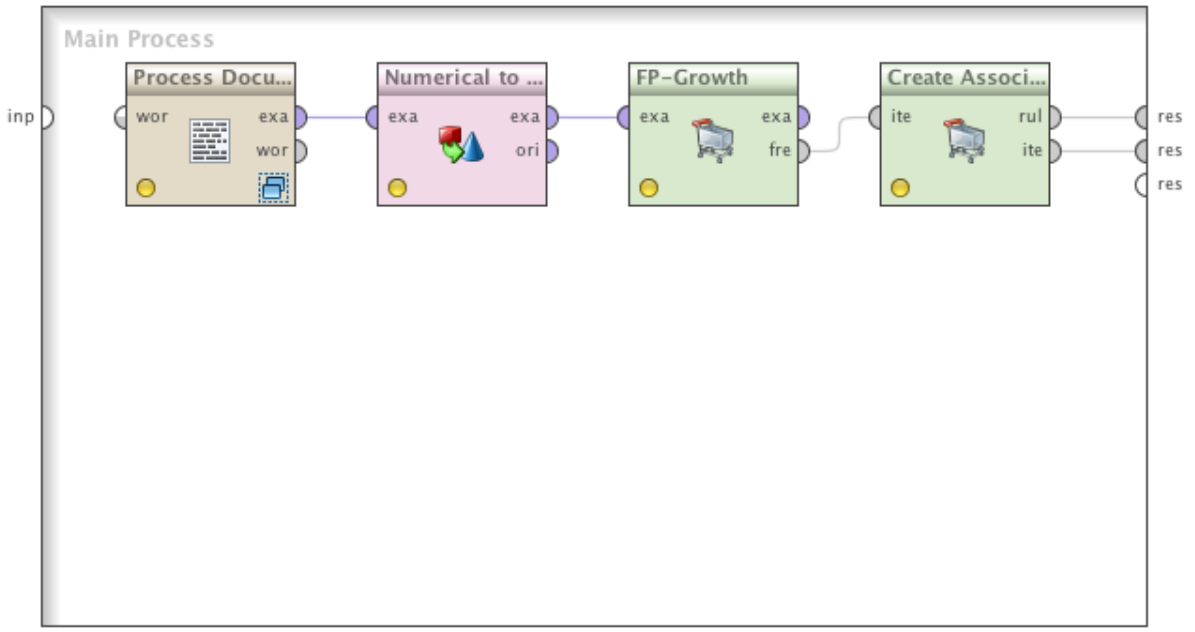

Abbildung 3: Prozess der Assoziationsanalyse mit FP-Growth in RapidMiner.

Je höher der minimale Support gewählt wird, desto präziser wird der generelle Kontext abgebildet, in dem *nachhaltig** genutzt wird. Andererseits steigt das Maß der Differenzierung, je niedriger der Schwellwert angelegt wird. Um sowohl aussagekräftige

Worte in Kombination mit dem gesuchten Term zu finden, als auch die Menge der Ergebnisse im Rahmen zu halten, wurde ein minimaler Support von 0.6 zu Grunde gelegt. Mit diesen Rahmenbedingungen konnten 35 häufige Worte und Wortkombinationen gefunden werden, die in Abbildung 4 abgebildet sind.

Die Liste umfasst sehr allgemeine Begriffe, sodass die daraus gebildeten Regeln keinerlei Rückschlüsse auf eine inhaltliche Verwendung des Termes zulassen. Beispielsweise folgt mit einer Wahrscheinlichkeit von über 90 % auf das Wort „nachhaltig“ das Wort „nich“ und auch mit abnehmender Konfidenz (auf bis zu 0.7 Vorgabe), werden die Konklusionen nicht aussagekräftiger. In umgekehrter Weise, sodass „nachhaltig“ die Konklusion darstellt, erzeugen die gebildeten Regeln kein gehaltvolleres Bild. Mit bis zu maximal 75 %iger Wahrscheinlichkeit ist auf die Worte „zeitung“, „jahr“, „nich“ das Wort „nachhaltig“ die Folge (s. Abb. 4).

No. of Sets: 35
Total Max. Size: 3
Min. Size: 1
Max. Size: 3
Contains Item:
Update View

Size	Support	Item 1	Item 2	Item 3
1	0.971	nich		
1	0.851	jahr		
1	0.809	zeitung		
1	0.766	nachhaltig		
1	0.759	deutsch		
1	0.734	lang		
1	0.714	gross		
1	0.703	neu		
1	0.695	rech		
2	0.831	nich	jahr	
2	0.784	nich	zeitung	
2	0.745	nich	nachhaltig	
2	0.737	nich	deutsch	
2	0.714	nich	lang	
2	0.701	nich	gross	
2	0.687	nich	neu	
2	0.682	nich	rech	
2	0.680	jahr	zeitung	
2	0.652	jahr	nachhaltig	
2	0.654	jahr	deutsch	
2	0.636	jahr	lang	
2	0.637	jahr	gross	
2	0.623	jahr	neu	
2	0.604	jahr	rech	
2	0.607	zeitung	nachhaltig	
2	0.628	zeitung	deutsch	
2	0.610	zeitung	lang	
2	0.616	deutsch	lang	
3	0.663	nich	jahr	zeitung
3	0.637	nich	jahr	nachhaltig
3	0.638	nich	jahr	deutsch
3	0.622	nich	jahr	lang
3	0.627	nich	jahr	gross
3	0.611	nich	jahr	neu
3	0.608	nich	zeitung	deutsch

Abbildung 4: Häufige Wortkombinationen, die mittels des FP-Growth gefunden werden konnten.

Ein plausibler Grund dafür, dass Worte, wie „zeitung“ oder „jahr“ als häufige Worte, ausgemacht wurden, ist die Art der untersuchten Dokumente. Bei allen handelt es sich um Zeitungsartikel und es ist anzunehmen, dass beispielsweise in Kopf- oder Fußzeile jeden Textes die genannten Worte vorkommen. Allein in drei der herangezogenen Medien ist der Term „zeitung“ im Namen enthalten und der Hauptanteil aller Codierungen der Grundlagenstudie konnte für Artikel der „Frankfurter Allgemeine Zeitung“ gemacht werden [vgl. Fischer/Haucke 2015: 4].
Die Menge an gefundenen Regeln könnte mit abnehmender minimalen Konfidenz erhöht werden, jedoch ist dies nicht sinnvoll, da in gleichem Maße die Wahrscheinlichkeit für rein zufällig gefundene Verbindungen steigt. Somit ist zu erkennen, dass eine Assoziationsanalyse für diese Daten keinen inhaltlichen Gewinn im Bezug auf Nachhaltigkeit hervorbringt.

4. 3. Sentiment - Analysis

Eine weitere Möglichkeit den Datensatz zu untersuchen besteht darin, die Stimmung in den Artikeln zu analysieren. Da der in den Medien erzeugte Beiklang eine wesentliche Rolle für das Verständnis und die Definition des Nachhaltigkeitsbegriffes spielt [ebd.: 3], liegt es nahe diesen zu untersuchen. Einen Überblick zum Beiklang – ob ein Beitrag eher positiv oder negativ betont ist – liefert eine Stimmungsanalyse. Dabei werden positive und negative Worte eines Artikels gezählt und gegeneinander abgewogen. Die Grundlage dazu bildet ein von der Uni Leipzig entwickelter Wortschatz[4], der für beide (positive und negative) Stimmungen charakteristische Wörter beherbergt.
Die Umsetzung in RapidMiner findet in Anlehnung an einen von F. Wortmann online vorgestellten Prozess statt (s. Abb. 5)[vgl. Wortmann 2013].

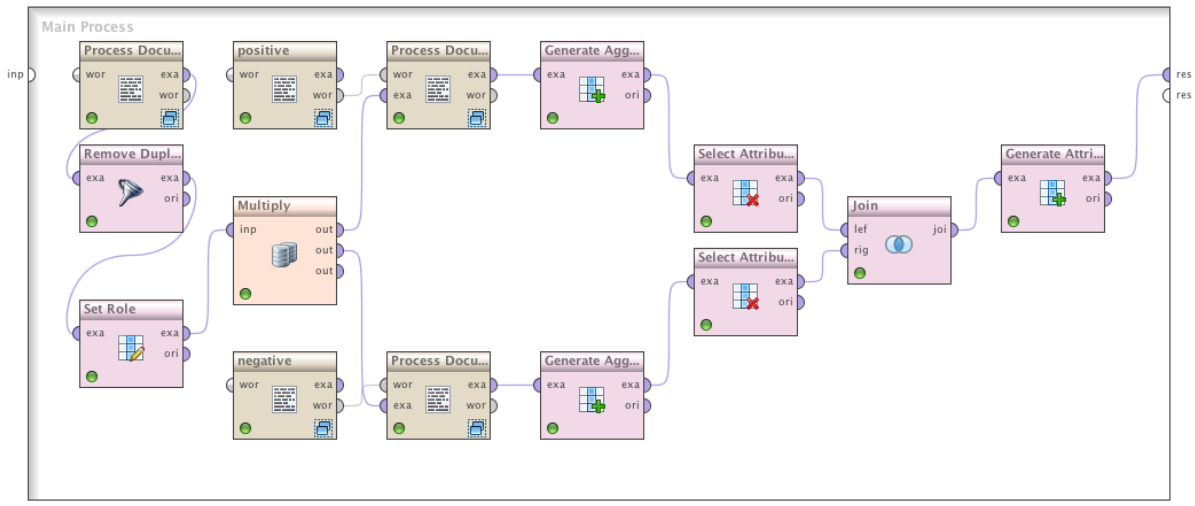

Abbildung 5: Prozess zur Stimmungsanalyse.

[4] Verfügbar unter: http://wortschatz.informatik.uni-leipzig.de/download/

Da es sich bei den Dokumenten der Zeitungsartikel um ganze Zeitungsseiten handelt ist es notwenig Eingrenzungen vorzunehmen. Lediglich die Artikel, in denen *nachhaltig** vorkommt, sollen untersucht werden, sodass die Datenvorverarbeitung eine Tokenisierung in Sätze mit diesem Term vorsieht. Bei einer durchschnittlichen Satzlänge von ca. 20 Wörtern [Stark 1992: 162] und dem Hintergrundwissen, dass in der Grundlagenstudie Artikel mit mindestens 300 Wörtern Länge gewählt wurden [Fischer/Haucke 2015: 5], lassen sich die Bereiche mit *nachhaltig** problemlos eingrenzen. Unter der Annahme der Term stünde jeweils in einer mittleren Position des Artikels, werden Bereiche von durchschnittlich ca. 300 Wörtern erreicht, indem sieben Sätze vor und nach *nachhaltig** sowie der entscheidende Satz an sich mit einbezogen werden.
Weitere Schritte des Prozesses beinhalten die Tokenisierung der sowohl positiven und negativen Wortliste, als auch der im Vorfeld generierten Sätze mit und um *nachhaltig**. Hierbei werden erneut Stoppwörter herausgefiltert, Groß- und Kleinschreibungen aufgehoben und die Terme auf ihren Wortstamm zurückgeführt. Darauf folgend wird eine Dokumenten-Term-Matrix erstellt, in der dokumentiert wird, ob in einem Dokument ein positiver oder negativer Term vorkommt. Mit Hilfe des *Generate Aggregation* Operators wird die jeweilige Matrix um eine Spalte erweitert, in der die Summe positiver oder negativer Wörter je Dokument berechnet wird.
Die Auswahl der betrachteten Items bzw. Attribute wird mittels *Select Attributes* auf jeweils die Summe der positiven bzw. negativen Wörter reduziert, woraufhin die Verwendung einer Join-Funktion folgt. Diese Funktion stammt aus dem Bereich der Datenbanken bzw. der Programmiersprache um auf Datenbanken zuzugreifen [Schicker 2014: 4] und führt eine Verbindung zweier Tabellen aus. Hier ist der Typ „inner“-Join gewählt worden, sodass ein Attribut zur Verbindung übereinstimmen muss [ebd.: 46]. Für die jeweils positiven bzw. negativen Wörter soll über die Attribute *text* und *metadata_path* eine Verbindung geschaffen werden, weil diese in beiden Prozess-Zweigen identisch sind. Im Ergebnis entsteht eine Bilanzierungsmöglichkeit, die im letzten Operator des Prozesses mathematisch vorgenommen wird (*Generate Attribute*). Aus der absoluten Häufigkeit der positiven bzw. negativen Wörter wird die Differenz berechnet, aus der hervorgeht, welche Wörter überwiegen. Dies ermöglicht Aussagen zur Stimmung der Texte bzw. Dokumente.
Aus allen untersuchten Dokumenten verbleiben 13.997 in der Ergebnismenge. Gründe für die (akzeptable) Reduktion um 6,1 % im Vergleich zur Grundgesamtheit sind folgende:

- Fehler bei der strukturellen Textverarbeitung
- Reduzierung von Duplikaten
- Keine explizite Nutzung des konkreten Terms

Anhand eines Beispiels soll der „Fehler bei der strukturellen Verarbeitung“ erläutert werden. Dokument 2013_1003.pdf ist ein aus mehreren Spalten Text bestehender Zeitungsartikel.

Anstatt dass der Text spaltenweise verarbeitet wird, erfolgt die Interpretation der Sätze über die gesamte Bildbreite von links nach rechts. Somit entstehen Wortkonstruktionen, die sich aus mehreren Sätzen zusammensetzen und ohne jeglichen Sinnzusammenhang sind. Im vorliegen Beispiel befindet sich der gesuchte Term *nachhaltig** an einem Zeilenumbruch, sodass Beginn und Ende des Begriffes aufgeteilt und mit Worten anderer Sätze vermischt ist. Daraus ergibt sich bei der Termsuche in der Verarbeitung eine leere Ergebnismenge, die bei der Reduzierung von Duplikaten – also Transaktionen, die die gleiche Zeichenkette beinhalten – herausgefiltert werden. Um diesen Fehler so weit wie möglich abzufangen, sind reguläre Ausdrücke zum Einsatz gekommen. Dabei handelt es sich um allgemeine Beschreibungen für eine Menge von Zeichenfolgen [Rapid-I-Wiki 2010: Regular expressions].

Weiterhin befinden sich zwischen den Zeitungsartikeln welche, die sich zwar inhaltlich mit Themen der Nachhaltigkeit auseinandersetzen mögen, solange jedoch der Term *nachhaltig** nicht explizit genutzt wird, kann dieser zur weiteren Untersuchung auch nicht gefunden werden. Folglich landen solche Artikel ebenfalls nicht in der Ergebnismenge.

In den verbleibenden Dokumenten kommen im Mittel 19 positive und 7 - 8 negative Worte vor, sodass eine durchschnittliche Stimmung von 11 positiv überwiegenden Worten entsteht. Diese geringen Worthäufigkeiten sind auf die relativ kleinen untersuchten Abschnitte von maximal 300 Worten zurückzuführen, wobei diese Anzahl durch das Preprocessing noch weiter verringert wurde.

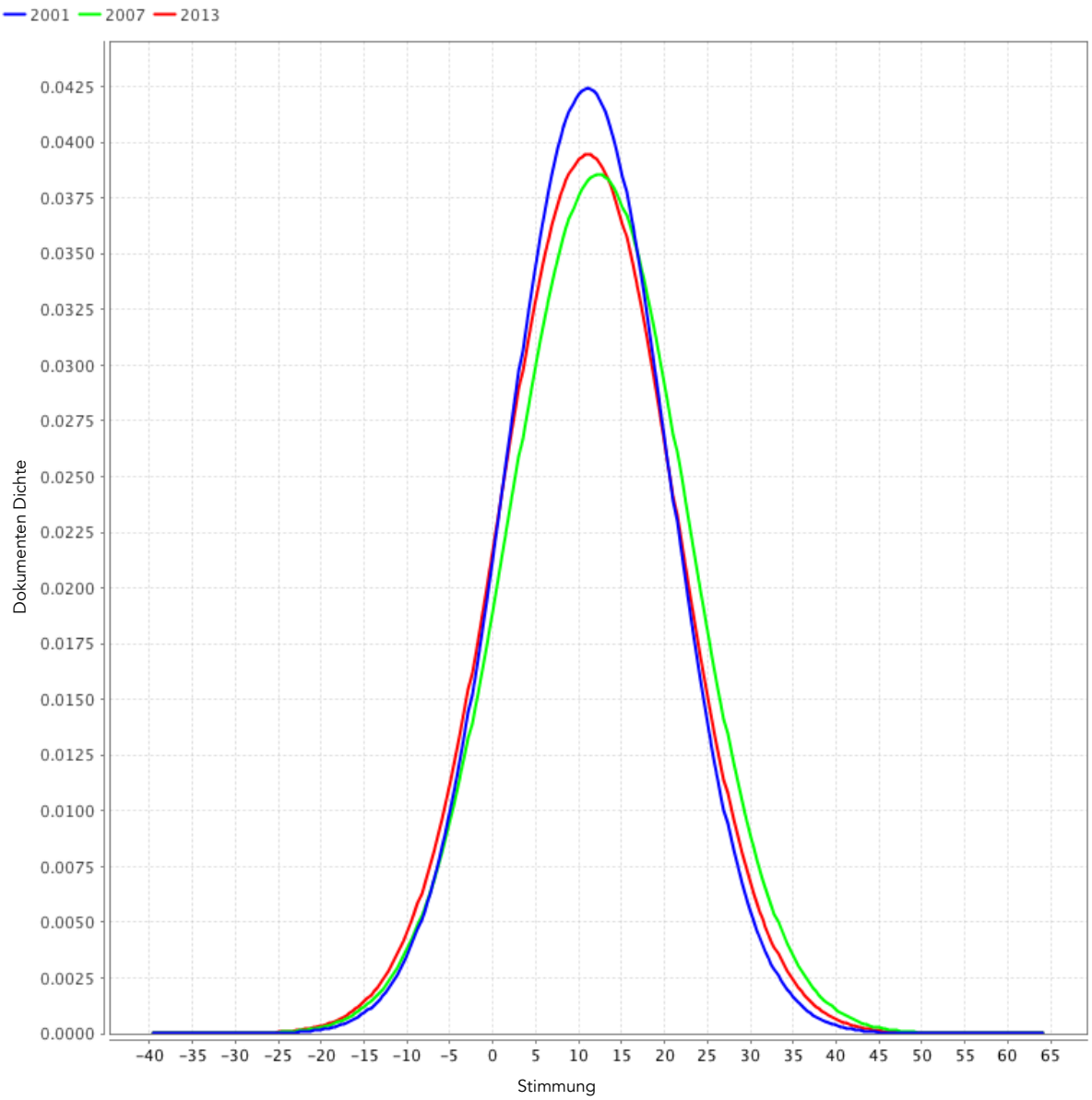

Abbildung 6: Verteilung der Stimmung zu Artikeln mit *nachhaltig** je Jahr. Density repräsentiert die Anzahl der Dokumente.

Wie in Abbildung 6 zu erkennen ist, hat sich die generelle Stimmung über die Jahre nicht grundlegend verändert, jedoch fällt in 2007 der Ton geringfügig positiver aus, als in 2001 und 2013.

Die vorgestellte Methode liefert einen schnellen Überblick hinsichtlich des angeschlagenen Tones in Abschnitten von Zeitungsartikeln mit *nachhaltig**, wobei die Qualität der Ergebnisse ganz klar von einer präzisen Textlesbarkeit abhängt. Sobald Worte (wie es im Bezug auf *nachhaltig** aufgefallen ist) nicht in einem Stück erkannt werden, sind Fehler nicht auszuschließen. Weiterhin ist eine Stimmungsanalyse keine Inhaltsanalyse im Sinne der Grundlagenstudie, sodass mögliche Herangehensweisen dazu im Folgenden erörtert werden.

4. 4. Klassifikation und Clustering

Zum Nachempfinden der eigentlichen Inhaltsanalyse eignen sich zwei Methoden des Text Minings aus dem Bereich der Kategorisierung. Hierbei handelt es sich um die Klassifizierung und das Clustering. Bei einer Klassifizierung soll auf Grundlage einer Menge korrekt zugewiesener Dokumente ein Modell erzeugt werden, das weitere Dokumente (ohne manuelle Betrachtung) einer der bestehenden Kategorien zuweist [Hipp 2003: 13]. Beim Clustering hingegen, sollen ohne Vorgabe von Kategorien unabhängige Teilmengen der Dokumente gebildet werden [ebd.: 14]. Die Objekte einer Teilmenge sollen dabei größtmöglich homogen sein, während die Teilmengen zueinander möglichst heterogen sein sollen [ebd.].
Anders ausgedrückt besteht der Unterschied beider Verfahren im Maß der Überwachung des Prozesses. „Algorithmen, die aus Daten unter Benutzung von Klasseninformationen lernen, werden auch überwachte Lernalgorithmen genannt. Die Clusteralgorithmen […] sind im Gegensatz dazu unüberwachte Lernalgorithmen“ [Runkler 2010: 124]. Während bei einer Klassifizierung eine Inhaltsanalyse folglich quantifiziert werden kann, wird bei einem Clustering herausgestellt, in wie weit sich Algorithmen zum Durchführen einer Inhaltsanalyse bewähren können. Ähnlich wie bei der Assoziationsanalyse, stehen zur Durchführung jeweils verschiedene Algorithmen zur Verfügung. Der sogenannte Naive Bayes ist ein Klassifikationsverfahren, das auf Wahrscheinlichkeiten basiert und sich bei einer großen Anzahl an Attributen (bzw. Termen) eignet [McGuigan 2013: 207]. Indem die relative Häufigkeit des Attributes hinsichtlich der relativen Häufigkeit der Klasse herausgestellt wird [ebd.], wird die wahrscheinlichste Klasse durch Multiplikation der Wahrscheinlichkeitswerte der Attribute berechnet [ebd.: 208].
Beim Clustering ist der sogenannte *k-means* einer der bekanntesten Algorithmen [Sharafi 2013: 70]. Um die größtmögliche Heterogenität einer Gruppierung von Dokumenten zu einer anderen zu erreichen, werden die Distanzen der Dokumente zueinander berechnet. Als Zentrum einer Gruppe zählt bei dieser Methode der Durchschnitt, sodass diese jeweils die größtmögliche Distanz zueinander aufweisen sollen. Das *k* im Namen gibt dabei die Anzahl der zu bildenden Gruppen an, die im Vorfeld bekannt sein muss [ebd.].

Um überprüfen zu können, wie genau und zuverlässig ein Klassifikationsmodell arbeitet, können Evaluationsmaße aus dem Information Retrieval herangezogen werden [Evert et al. 2010: 155]. Zur Erklärung wird eine binäre Textklassifikation vorangestellt, sodass vier mögliche Ausgänge einer Klassifikation entstehen (s. Tabelle 2). Sollen Beispielsweise Texte mit Farbworten in die Kategorien „rot“ und „schwarz“ sortiert werden, kann „rot“ in Klasse „rot“ (TP) oder in Klasse „schwarz“ (FN) gelangen, genauso wie „schwarz“ in Klasse „rot“ (FP) oder in Klasse (TN) „schwarz“ gelangen kann.

Tabelle 2: Potentielle Ergebnisse einer binären Textklassifikation. [vgl. Schabel 2012: 22]

	Erlernte und angewandte Klassifizierung (Testmenge)	
Tatsächliche Klassifizierung (Trainingsmenge)	true positiv (TP)	false positiv (FP)
	false negative (FN)	true negative (TN)

Daraus lässt sich berechnen, wie treffsicher (engl. *accuracy*) ein Modell arbeitet, indem alle korrekten Klassifizierungen der Gesamtmenge an Klassifizierungen gegenübergestellt werden [Schabel 2012: 22].

$$Accuracy\ A = \frac{TP + TN}{TP + TN + FP + FN}$$

(4.4 - 1 [Schabel 2012: 23])

Soll nun betrachtet und beurteilt werden, wie präzise die Ergebnisse sind, kann die *Precision P* herangezogen werden. Sie gibt den Anteil aller richtigen Ergebnisse – je betrachteter Klasse – aus allen richtig klassifizierten Dokumenten wider, wohingegen der *Recall R* Aussagen zur Vollständigkeit der Ergebnisse zulässt (vgl. 4.4 - 2)[Ferber 2003: 52; Schabel 2012: 22f.].

$$Precision\ P = \frac{TP}{TP + FP} \quad Recall\ R = \frac{TP}{TP + FN}$$

(4.4 - 2 [Schabel 2012: 22f.])

Ein Modell mit (nahezu) 100 % Accuracy ist entweder sehr gut oder reproduziert lediglich die Ergebnisse der Trainingsmenge [vgl. Evert et al. 2010: 154; McGuigan 2013: 208]. Dieses Phänomen wird in der Literatur als Überanpassung (engl. *overfitting*) benannt [Felden 2006: 6]. Zum Lernen einer Klassifikation wird lediglich ein Teil der Dokumente zu Grunde gelegt, anhand derer Merkmale die Zuordnung weiterer Dokumente abgeleitet werden soll. Diese Teilmenge wird Trainingsmenge genannt, alle verbleibenden Dokumente werden als Testmenge bezeichnet. Mittels einer Kreuzvalidierung (engl. *cross validation*) werden alle Daten in *n* gleichgroße, disjunkte Teilmengen gegliedert, wobei *n-1* dieser Teilmengen als Lernmenge herangezogen werden [Evert 2010: 154f.]. Die verbleibende Teilmenge dient dann zum Testen des erlernten Modells. Dieser Vorgang wird daraufhin *n* mal wiederholt, sodass jeder Teil einmal als Testmenge fungiert und die erzielten Ergebnisse zur Berechnung der Evaluation dienen können [ebd.: 155].

Die Prozessgestaltung in RapidMiner beinhaltet, ähnlich den vorangegangen Prozessen, als ersten Schritt das Preprocessing der Dokumente. Anders als bisher wurde im Vorfeld der Verarbeitung eine Ordnerstruktur entsprechend der erfolgten Codierungen angelegt, sodass je

Code ein Ordner mit allen zugehörigen Dokumenten vorliegt[5]. So erhalten die Dokumente bei der Verarbeitung mit RapidMiner ein Label mit ihrer jeweiligen Codezuweisung. Das Label repräsentiert die Klasse, um dessen Zuweisung es im Folgenden gehen soll.

Darauf folgend wurden die Texte in RapidMiner auf Abschnitte mit *nachhaltig** tokenisiert und gefiltert. Die verbliebenen Sätze werden in ihre Wörter zerlegt und – analog zum Prozess zur Assoziationsanalyse – um Stoppwörter sowie Groß- und Kleinschreibung reduziert. Nachdem die Wörter auf ihren Wortstamm zurückgeführt wurden, werden alle Dokumente mit identischem Text herausgefiltert.

Um im weiteren Verlauf des Prozesses die Rechenzeit zu verringern, wurde die entstandene Dokumenten-Term-Matrix abgespeichert (vgl. Anhang A, 04_ExampleSet_ohneDublikate). Bevor der eigentliche Klassifizierungs-Algorithmus (hier Naive Bayes)[6] zur Anwendung kommt, ist es notwendig die einzelnen Klassen auszubalancieren, um ein *overfitting* zu vermeiden [vgl. McGuigan 2013: 206]. Die Dokumentenverteilung je Code ist in der Grundlagenstudie sehr unausgeglichen, sodass z.B. die alltagssprachliche Verwendung des Terms *nachhaltig** mit fast zwei Dritteln zu Buche schlägt [Fischer/Haucke 2015: 9]. Andere Codierungen, wie Such- oder Lernprozess, sind hingegen sehr unterrepräsentiert [ebd.][7].

Damit alle Klassen gleichermaßen zum Lernen herangezogen werden können, ist eine geschätzte mittlere Menge der mit am wenigsten besetzten Kategorien gewählt worden. Mit sechs Klassen im Bereich von 300-500 Codierungen [vgl. ebd.], entsteht so eine Trainingsmenge von 300 Dokumenten je Klasse. Lediglich die beiden Klassen mit darunter liegender Anzahl an zugewiesenen Codes wurden auf die entsprechende Anzahl herabgesetzt, da das Modell zu ungenau würde, wenn tatsächlich der Umfang der kleinsten Klasse zum Ausbalancieren genommen worden wäre [vgl. McGuigan 2013: 206].

Testhalber ist der beschriebene Prozess ohne einen Klassenausgleich durchgeführt worden (vgl. Anhang A, Klassifikation02), wobei die Ergebnisse ein *overfitting* bestätigen: unterrepräsentierte Klassen sind nicht erkannt worden, während lediglich die stark ausgeprägte Klasse sehr gute Ergebnisse erzielen konnte. Insgesamt erreichte dieses Testmodell eine Genauigkeit von über 60 %, wobei die Aussagekraft wegen des genannten *overfittings* sehr gering ist.

Im eigentlichen Prozess ist für die Kreuzvalidierung die Anzahl gleich großer, disjunkter Teilmengen auf fünf und auf zehn gesetzt worden. Die Anwendung des Modells erfolgt,

---

[5] Zuvor waren die Dokumente auf Ordner je Jahr (2001, 2007, 2013) verteilt.

[6] Auf den Einsatz weiterer Algorithmen wurde verzichtet, da Naive Bayes einer der Standardalgorithmen für diesen Zweck ist und bereits eine Reihe von Arbeiten die Qualität der Verfahren miteinander vergleicht und abwägt.

[7] Die Reihenfolge der Codes in der Tabelle des Papers ist eine andere als in der vorliegenden Untersuchung. Auf Grund dessen sind die Label im Prozess, zusätzlich zur Nummerierung nach der Legende zu den Dokumenten, benannt worden.

entgegen der reduzierten Datenmenge zum Ausbalancieren der Klassen, auf die gesamte Datengrundlage, um eine vollständige Klassifizierung durchzuführen.

accuracy: 42.16% +/- 0.23% (mikro: 42.16%)

	true 0_Eigenr	true 1_Alltags	true 2_Oekok	true 3_Suchpı	true 4_Vernet	true 5_Verant	true 6_Sozio-	true 7_Oekon	true 8_Kritik	true 9_Unklar	true 10_Lernp	class precisio
pred. 0_Eiger	2823	14029	1753	9	1342	34	111	26	25	136	8	13.91%
pred. 1_Allta	17	31265	877	3	540	15	67	13	11	88	0	95.04%
pred. 2_Oekc	35	7788	6027	9	1253	21	52	23	31	114	12	39.23%
pred. 3_Such	0	153	15	414	48	1	4	0	0	10	0	64.19%
pred. 4_Vern	59	10835	2225	10	4562	44	66	49	39	109	10	25.33%
pred. 5_Vera	11	2047	361	3	250	1739	47	4	11	14	1	38.75%
pred. 6_Sozic	35	11203	945	7	833	25	2909	13	18	134	6	18.04%
pred. 7_Oekc	4	1857	232	1	148	0	5	1414	2	15	0	38.44%
pred. 8_Kritik	4	2218	340	0	184	10	14	2	1509	27	1	35.02%
pred. 9_Unkl	21	11364	1085	4	718	21	93	16	24	2792	1	17.30%
pred. 10_Ler	11	51	10	0	2	0	2	0	0	1	351	82.01%
class recall	93.48%	33.69%	43.45%	90.00%	46.17%	91.05%	86.32%	90.64%	90.36%	81.16%	90.00%	

Abbildung 7: Leistungsfähigkeit des Naive Bayes Klassifikationsmodells bei einer 10-fachen Kreuzvalidierung.

In Abbildung 7 sind die Ergebnisse der Kreuzvalidierung mit 10 Teilmengen dargestellt. Insgesamt konnte mit einer 10-fachen Klassifizierung eine Accuracy von 42,16 % erreicht werden. Mehrere Gründe sind für diese relativ schlechte Quote verantwortlich. Zum einen ist die Trainingsmenge zur Vermeidung eines *overfittings*, im Verhältnis zur gesamten Dokumenten- und Testmenge, relativ gering (ca. 1 : 4,6). Zum anderen ist die Anzahl der Attribute zur Klassifizierung sehr klein, da die Dokumente auf die Abschnitte mit *nachhaltig** gekürzt wurden. Es ist davon auszugehen, dass mit kleiner werdenden Teilmengen zum Trainieren (bzw. mit immer höherem *cross-fold*) die Accuracy steigt, wobei das Risiko eines *overfittings* proportional dazu steigt. Dies bestätigt die 5-fache Kreuzvalidierung, dessen Genauigkeit mit 40% unterhalb der der 10-fachen Kreuzvalidierung liegt. Aus diesem Grund wurde auf eine Darstellung verzichtet.
Ein weiterer Grund der geringen Treffsicherheit, kann auf die zum Teil mangelhafte OCR zurückgeführt werden. Wie bereits im Abschnitt zur Sentiment-Analyse erläutert (vgl. Abschnitt 4. 2), sind Texte mit Spaltenaufteilung zuweilen nicht korrekt eingelesen worden, sodass Worte und Zusammenhänge verloren gegangen sind. Werden Worte, wie in diesem Fall, als Attribute zur Klassifizierung genutzt, kann die Erkennung von Mustern nur so gut sein, wie die zu Grunde gelegten Dokumente verarbeitet werden können.
Des Weiteren kann das Modell mit einer Standardabweichung von +/- 0,23 % als stabil betrachtet werden, da für jeden Durchlauf der 10 Wiederholungen beinahe die gleiche Accuracy erzielt wurde. Bei der Betrachtung der einzelnen Klassen lässt sich erkennen, dass die alltagssprachliche Verwendung des Begriffs die genauste Klassifizierung erreicht (ca. 95 % Precision), im Gegensatz dazu allerdings sehr unvollständig ist (Recall ca. 34 %). Erklären lässt sich dies durch die Größe bzw. den Umfang der Klasse. Wie in der Grundlagenstudie zu erkennen, sind annähernd zwei Drittel aller codierten Abschnitte dieser Verwendung zugeschrieben [Fischer/Haucke 2015: 9] und somit hat diese Codierung eine hohe Treffer-Wahrscheinlichkeit. Die Unvollständigkeit hingegen ist möglicherweise auf die

Bandbreite der potentiellen Kontexte zurückzuführen. So kann beispielsweise selbst der Geschmack von Wein als „nachhaltig“ beschrieben werden [vgl. Stollenwerk 2015].
Ein genau gegenteiliges Ergebnis erzielt die Klasse der Eigennamen. Sie ist mit über 90 % Recall und lediglich knapp 14 % Precision, die vollständigste und gleichzeitig ungenauste Klasse. In diesem Zusammenspiel liegt die Vermutung nahe, dass während der Klassifizierung eine Zuweisung dieses Codes oft probiert wurde und dadurch zwar alle Dokumente erkannt, jedoch genauso viele falsch erkannt wurden.
Das beste Verhältnis aus Precision und Recall konnten tatsächlich die kleinsten Klassen erreichen (Code 3 und Code 10). Mit diesen Codes werden Textpassagen beschrieben, die betonen, dass das Konzept der Nachhaltigkeit stetig spezifiziert werden soll und keine Idée fixe ist. Somit scheinen die Kontexte dieser Klassen recht spezifisch zu sein.
Interessant für den Begriff selbst, ist die Verwendung im Sinne der einzelnen drei Dimensionen der Nachhaltigkeit (Ökologie, Ökonomie und Soziales) [vgl. Michelsen/ Adomßt 2014: 29], sowie im Sinne eines integrativen Verständnisses [vgl. ebd.: 28]. Während für die einzelnen Dimensionen die Vollständigkeit der Klassen *Soziales* und *Ökonomie* sehr hoch ist (ca. 86 % und ca. 90 %) ist die Genauigkeit lediglich für die Klassen *Ökologie* und *Ökonomie* im moderaten Bereich (jeweils ca. 40 %). Auf den Kontext bezogen gibt es folglich mehrere aussagekräftige Begriffe, die im Zusammenhang der ökologischen oder ökonomischen Verwendung herangezogen werden und für eine Klassifizierung ausschlaggebend sind. Die Vollständigkeit der sozialen Dimension ist leider bloß eingeschränkt tauglich, da die Precision weniger als 20 % beträgt. Von allen drei Dimensionen kann die der Ökonomie am besten abgebildet werden, was in Anbetracht der Dokumente logisch erscheint, da Ökonomisches in Zeitungsartikeln oftmals eine große Rolle spielt. Die integrative Klasse erreicht ziemlich genau 46 % Recall und 25 % Precision. Diese mittleren Werte bestätigen, dass der Nachhaltigkeisbegriff im Sinne der Brundtland Definition als komplex zu betrachten ist.
Auf Grund der vorhergehenden Ergebnisse wird auf die Durchführung und Erläuterung eines Clusterings verzichtet. Da weder ausreichende Assoziationsregeln, noch präzise Klassenzuweisungen erreicht werden konnten, ist anzunehmen, dass in einem unüberwachten Lernvorgang ebenso spärliche Ergebnisse zu Tage gefördert würden. Mutmaßlich könnte ein Cluster für eine alltagssprachliche Verwendung des Begriffes erzielt werden, während jegliche andere Kontexte unerkannt blieben.

5. Bilanzierung der Methoden und Fazit

Zum Abschluss der durchgeführten inhaltsanalytischen Untersuchungen soll nun einerseits ein Resümee gezogen werden und andererseits die Vor- und Nachteile hinsichtlich des Zeitaufwandes der Methoden verglichen werden.

Um zu erwägen, ob eine automatische Verarbeitung von Texten zum Erzielen einer Inhaltsanalyse ergänzend oder gar ersetzend sein kann, sind der Aufwand und die Ergebnisse dieser Analyse mit denen der Grundlagenstudie zu vergleichen. Einerseits ist die computergestützte Bearbeitung einer Kontext bezogenen Aufgabenstellung in einem deutlich kürzeren Zeitrahmen durchführbar. Die vorliegenden Ergebnisse konnten im Rahmen einer Bachelorarbeit erzielt werden, bei der ein definiertes Zeitfenster von neun Wochen vorgegeben ist. Dem gegenüber bedurfte die Grundlagenstudie für die sowohl qualitative als auch quantitative Analyse ein halbes Jahr [Stollenwerk 2015]. Die beiden Zeiträume können jedoch nicht direkt miteinander verglichen werden, da in dieser Arbeit lediglich ein Teil der Grundlagenstudie nachempfunden wurde.

Weiterhin ist unklar, wie viel Zeit die ursprüngliche Datenbeschaffung und die Entwicklung des Codes in Anspruch genommen haben. Da beides für die automatische Analyse vorlag, bleibt nur der Vergleich der einzelnen Codierungen mit der Dauer der Prozesse in RapidMiner. Unabhängig von der tatsächlichen Dauer der Codierungen führt RapidMiner, da die Prozesse innerhalb von maximal vier Stunden Laufzeit beendet werden konnten.

Eine ergänzende oder ersetzende Funktion dieser Prozesse zum Zweck einer Inhaltsanalyse kann jedoch nur erfolgen, wenn die Ergebnisse qualitativ hochwertig und präzise sind. Durch die automatische Verarbeitung unterliegen sie den wissenschaftlichen Kriterien, da sie valide, reliabel, replizierbar und objektiv sind. Trotz allem muss jedoch vorausgesetzt werden, dass eine alleinige computerbasierte Auswertung nicht ausreichen kann, da es sich um statistische Methoden handelt. Die eigentlich qualitative Arbeit besteht somit in der Erstellung der Kategorien [Mayring 2008: 19].

Für die vorliegende Untersuchung wird deutlich, dass der alleinige Einsatz von Text Mining Verfahren noch nicht in der Lage ist, ebenso präzise Ergebnisse zu liefern, wie die der Referenzstudie. Demnach reicht es nicht aus, einen bestimmten Prozentsatz an Dokumenten händisch zu klassifizieren, um eine solche Menge in einem nächsten Schritt automatisch zu quantifizieren. Ebensowenig kann beispielsweise auf ein Clustering als Alternative zur qualitativen Inhaltsanalyse zurückgegriffen werden, da ein Erkennen feiner Unterschiede in Kontexten unwahrscheinlich ist. Der Begriff *nachhaltig** bzw. die Bedeutung von Nachhaltigkeit ist offenbar zu vielfältig und komplex, als dass Muster dazu erkannt werden können. In Abhängigkeit von dem vorliegenden Datenformat (.pdf), ist die Analyse nicht ausreichend, um beständige Methoden der sozialwissenschaftlichen Forschung zu ergänzen oder gar zu ersetzten. Weiterhin bleibt jedoch anzunehmen, dass mit fortschreitender

Verbesserung der vorbereitenden Maßnahmen zur Untersuchung unstrukturierter Daten (wie Texte), die Qualität der Ergebnisse in gleichem Maße zunehmen wird. Dies betrifft im vorliegenden Fall die optische Zeichenerkennung, die deutlich dazu beigetragen hat, die Ergebnisqualität zu beeinträchtigen.

Text Mining Methoden und Softwares wie RapidMiner liefern erhebliche Vorteile, da der Einsatz dieser Techniken dem Zeitaufwand zur Analyse stetig steigender Mengen an Dokumenten begegnet. Der Forschungsbedarf, um einen interdisziplinären Einsatz der vorgestellten Methoden stetig und zuverlässig nutzen zu können, bleibt jedoch nach wie vor bestehen, obwohl mit der vorliegenden Arbeit ein weiterer Schritt dahingehend unternommen wurde. Weiterhin ist der fachübergreifende Einsatz der Techniken nicht pauschal auszuschlagen, da beispielsweise die Extraktion von Schlagwörtern von wissenschaftlichen Texten bereits erfolgreich durchgeführt wurde [vgl. Lübbing/Osada 2014: 29].

Zusammenfassend bleibt zu sagen, dass eine qualitative Inhaltsanalyse mit RapidMiner trotz allem mit Verfahren wie denen einer Assoziationsanalyse, einer Klassifizierung oder eines Clusterings realisierbar ist.

6. Literatur

BÜTEFISCH, Siegfried/Petermann, Jörg (2014): Der rote Fisch 5 – Impulse für werbewirksame Gestaltung und Kommunikation - Leitfaden 5. Erfolg im Internet und in digitalen Medien. 2. Aufl., Norderstedt, Books on Demand.

EVERT, Stefan/Frötschl, Bernhard/Lindstrot, Wolf (2010): Statistische Grundlagen In: Carstensen, Kai-Uwe et al. (Hrsg.): Computerlinguistik und Sprachtechnologie. 3., überarb. und erw. Aufl., Heidelberg, Spektrum Akademischer Verlag.

FAYYADD, Usama/Piatetsky-Shapiro, Gregory/Smyth, Padhraic (1996): From data mining to knowledge discovery in databases. IN: American Association for Artificial Intelligence, 17/3, S. 37 - 54.

FELDEN, Carsten et. al (2006): Evaluation von Algorithmen zur Textklassifikation. Technische Universität Bergakademie Freiberg: Arbeitspapiere.

FERBER, Reginald (2003): Information Retrieval. Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web. Heidelberg.

FISCHER, Daniel/Haucke, Franziska (2015, Juni 11-14): "To sustain, or not to sustain…?". An empirical analysis of the usage of "sustainability" in German newspapers. Presented at Bridging Divides: Spaces of Scholarship and Practice in Environmental Communication The Conference on Communication and Environment, Boulder, Colorado.

HILDEBRANDT, Lutz / Boztuğ, Yasemin (2007): Ansätze zur Warenkorbanalyse im Handel. In: Schuckel, Marcus / Toporowski, Waldemar (Hrsg.): Theoretische Fundierung und praktische Relevanz der Handelsforschung. Wiesbaden, S. 217 - 234.

HIPP, Jochen (2003): Wissensentdeckung in Datenbanken mit Assoziationsregeln. Verfahren zur effizienten Regelgenerierung und deren Integration in den Wissensentdeckungsprozess. Eberhard-Karls-Universität Tübingen: Dissertation.

LÜBBING, David/Osada, Sebastian (2014): Automatische Generierung und Verifizierung von Keywords für wissenschaftliche Publikationen. In: Rieger, Bodo/Benjamins, Axel (Hrsg.): Text Mining in wissenschaftlichen Publikationen. Osnabrück, S. 20-30.

MAYRING, Philipp (2008): Qualitative Inhaltsanalyse. Grundlagen und Techniken. 10., neu ausgestattete Aufl., Weinheim, Beltz.

MCGUIGAN, Neil (2013): Detection Text Message Spam. In: Hofmann, Markus/Klinkenberg, Ralf (Hrsg.): RapidMiner. Data Mining Use Cases and Business Analytics Applications. Boca Raton, S. 199-211.

MICHELSEN, Gerd/Adomßent, Maik (2014): Nachhaltige Entwicklung: Hintergründe und Zusammenhänge. In: Heinrichs, Harald/Michelsen, Gerd (Hrsg.): Nachhaltigkeitswissenschaften. Berlin, Springer Spektrum, S. 3 - 60.

PETERSOHN, Helge (2005): Data Mining. Verfahren, Prozesse, Anwendungsarchitektur. München, Oldenbourg.

RAPID-I-Wiki (2010): Regular expressions. Unter: https://rapid-i.com/wiki/index.php?title=Regular_expressions (Stand: 13.11.2015).

REMUS, Robert/Quasthoff, Uwe/Heyer, Gerhard (2010): SentiWS - a Publicly Available German-language Resource for Sentiment Analysis. In: Proceedings of the 7th International Language Ressources and Evaluation (LREC'10), S. 1168 - 1171.

RUNKLER, Thomas A. (2010): Data Mining. Methoden und Algorithmen intelligenter Datenanalyse. Wiesbaden, Vieweg + Teubner.

SCHABEL, Simon (2012): Aggregation von Produktbewertungen aus dem World Wide Web. Technische Universität Dresden: Diplomarbeit.

SCHARKOW, Michael (2011): Automatische Inhaltsanalyse und maschinelles Lernen. Universität der Künste Berlin: Dissertation.

SCHICKER, Edwin (2014): Datenbanken und SQL. Eine praxisorientierte Einführung mit Anwendungen in Oracle, SQL Server und MySQL. 4., überarb. Aufl., Wiesbaden, Springer.

SEIDEL, Ludwig Michael (2013): Text Mining als Methode zur Wissensexploration: Konzepte, Vorgehensmodelle, Anwendungsmöglichkeiten. Hochschule Wismar: Master-Thesis.

SHARAFI, Armin (2013): Knowledge Discovery in Databases. Eine Analyse des Änderungsmanagements in der Produktentwicklung. Wiesbaden, Springer.

SIEGMUND, Carsten (2006): Einführung in Text Mining. In: Witte, René/Mülle, Jutta (Hrsg.): Text Mining: Wissensgewinnung aus natürlichsprachigen Dokumenten. Karlsruhe, S. 41-58.

STARK, Susanne (1992): Stilwandel von Zeitschriften und Zeitschriftenwerbung. Analyse zur Anpassung des Medienstils an geänderte Kommunikationsbedingungen. Heidelberg, Physica-Verlag.

STOLLENWERK, Thomas (2015, 12. Mai): Studie zum Begriff Nachhaltigkeit | BIORAMA. Unter: http://www.biorama.eu/fischer-interview/ (Stand: 04.12.2015).

WORTMANN, Felix (2013): Rapid Miner - 9. Sentiment Analysis. Unter: https://vimeo.com/62985704 (Stand: 12.11.2015).

WITTE, René/Mülle, Jutta (2006): Text Mining: Wissensgewinnung aus natürlichsprachigen Dokumenten. Universität Karlsruhe: Interner Bericht.

Erklärung zur eigenständigen Arbeit

Hamburg, den 26.01.2016

„Ich erkläre hiermit an Eides statt, dass

• die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt wurden,

• alle Stellen der Arbeit, die wörtlich oder sinngemäß aus anderen Quellen übernommen wurden, als solche kenntlich gemacht wurden,

• und die Arbeit in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegt wurde.“

(Name, Matrikelnummer, Unterschrift)

Verzeichnis des Anhangs

A. Speichermedium mit allen in RapidMiner erstellten Prozessen und Ergebnissen sowie den zur Analyse herangezogenen Daten in komprimierter Form.

B. CD mit einer Kopie der vorliegenden Arbeit als .pdf.