Friedrich-Schiller-Universität Jena
Philosophische Fakultät
Institut für Germanistische Sprachwissenschaft

Automatische Analyse von Emotionen in Geschäfts- und Nachhaltigkeitsberichten

Arbeit zur Erlangung des akademischen Grades
Bachelor of Arts (B.A.)

vorgelegt von Sven Eric Büchel
Matrikelnummer: 130453
geboren am 11.09.1990 in Hamburg

Erstgutachter: Prof. Dr. Udo Hahn
Zweitgutachter: Prof. Dr. Peter Walgenbach

Jena, den 4. Februar 2016

Inhaltsverzeichnis

1 Einleitung

Die Computerlinguistik bietet Methoden, mit denen sich Text in Größenordnungen verarbeiten lässt, die durch manuelles Lesen niemals zu bewältigen wären. Durch Anwendungen dieser und vergleichbarer rechnerischer Methoden in anderen Wissenschaftsbereichen ergeben sich mitunter völlig neue Forschungsparadigmen, wie etwa die Digital Humanities in den Geisteswissenschaften oder jüngst die Computational Social Sciences in den Sozialwissenschaften (DiMaggio, 2015) zeigen. Obwohl Meinungen, Gefühle und Emotionen für viele Fragestellungen dort höchst relevant sind, bot die Computerlinguistik lange Zeit nur Verfahren zur Verarbeitung von faktischer Sprache an. Dies änderte sich Anfang der 2000er Jahre durch das Aufkommen der Sentiment Analysis (Liu, 2012, S. 5). Hiermit war es möglich, positive und negative Bewertungen in Texten zu erkennen. Erst seit relativ kurzer Zeit widmet sich die Computerlinguistik der Erkennung von Emotionen und anderer affektiver Phänomene, was ihren Messbereich erneut erweitert. Damit macht sie sich hoch interessant für einen neuen Zweig der Organisationsforschung, der sich mit der anthropomorphen Modellierung von Organisationen als *soziale Akteure* befasst. Die Kernfrage dieses Forschungszweigs lautet *Who are we as an Organization?* (Whetten, 2006). Nach dem Konzept der *Orginaziational Identity* (King, Felin & Whetten, 2010) tragen Organisationen dauerhafte Merkmale, die sie fundamental von anderen Organisationen unterscheiden. Offen ist dabei, welche bzw. wie viele Merkmale von dieser Anthropomophisierung abgedeckt sind. Eine empirische Untersuchung dazu, ob auch Emotionen zu diesen gehören, ist nicht bekannt. Die Frage ist insofern hochrelevant, als dass die sinnvolle Annahme weitreichender „Menschlichkeit“ auch Implikationen für die – z.B strafrechtliche – Sanktionierbarkeit von Organisationen hätte (Beyer et al., 2014; Ortman, 2010). Mit der Fähigkeit große Textmengen, schnell und intersubjektiv überprüfbar zu verarbeiten, eröffnet die Computerlinguistik einen Zugang zu Organisationen als Ganzes und macht diese Frage somit bearbeitbar.

Ziel der vorliegenden Arbeit ist daher eine Methode zu entwickeln und diese in einer Software-Anwendung zu implementieren, die in der Lage ist Emotionen in Texten zu messen. Diese soll auf ein Korpus aus englischsprachigen Geschäfts- und Nachhaltigkeitsberichten großer, börsennotierter Unternehmen angewandt werden. Die resultierenden Daten werden einer explorativen statistischen Analyse unterzogen. Entscheidend ist dabei, die Methode auf die für die Computerlinguistik neue Anwendungsdomäne anzupassen. Daher wird zunächst tiefergehend auf psychologische Modelle und formale Repräsentationsmöglichkeiten von Emotionen eingegangen und diese bezüglich ihrer Verwendbarkeit diskutiert.

In der Emotionserkennung, als junges, sich noch entwickelndes Forschungsfeld, wird ein Mangel an theoretischer Grundlagenarbeit von zahlreichen Autoren beklagt (Munezero, Montero, Sutinen & Pajunen, 2014; Calvo & Mac Kim, 2013). Aufgrund unterschiedlicher Emotionsmodelle und Repräsentationsformen sind emotionserkennende Systeme selten direkt in ihrer Performanz vergleichbar. Der Mangel an geteilten theoretischen Grundlagen und anderen Konventionen kann daher als ausgeprägtes Forschungshindernis betrachtet werden. Erschwerend hinzu kommt, dass neben der Computerlinguistik viele weitere akademische Disziplinen und Teildiziplinen für dieses Feld relevant sind, darunter Bildinformatik, Human-Computer Interaction, Cognitive Science und Psychologie. Dabei kann nur vermutet werden, dass weite Teile der für alle relevanten Veröffentlichungen – zum Beispiel zu Fragen der formalen Repräsentation – nur die jeweils eigene Community erreicht. Damit hängt zusammen, dass die Emotionserkennung natürlich nicht auf Text beschränkt bleibt, sondern es ebenfalls umfangreiche Forschung zu Emotionen in anderen „Kanälen“ wie etwa gesprochener Sprache, Mimik, Gestik und Körperhaltung sowie Biosignalen gibt (Calvo & D'Mello, 2010; Gunes & Schuller, 2013).

Die weitere Arbeit gliedert sich folgendermaßen: In Abschnitt 2 werden wesentliche Begriffe und Konzepte geklärt. Dabei wird insbesondere unter Rückgriff auf psychologische Kategorien die Emotionserkennung von der Sentiment Analysis abgegrenzt. Abschnitt 3 dient der Klärung des Forschungsstands und der Diskussion unterschiedlicher psychologischer und formaler Grundlagen für die entwickelte Methode. In Abschnitt 4 werden daraufhin die wesentlichen Eckpunkte für diese Methode mit Rücksicht auf die erwarteten Besonderheiten der Anwendungsdomäne festgelegt. Anschließend wird die entwickelte Formalisierung, die entworfene Architektur und ihre konkrete Implementierung vorgestellt. Abschnitt 5 gibt einen Überblick über die bestehenden Ressourcen, wovon eine zur Anwendung in der vorgestellten Architektur ausgewählt wird. Diese wird beschrieben und ihre weitere Verarbeitung erläutert. Im Abschnitt 6 wird anhand des bekannten RCV1-Korpus eine Plausibilitätsprüfung der durch das entwickelte Werkzeug erhobenen Daten durchgeführt. Eine quantitative Evaluation konnte in Ermangelung eines Testkorpus nicht durchgeführt werden. In Abschnitt 7 wird schließlich die explorative Datenanalyse des Unternehmenskorpus vorgestellt. Es folgen eine allgemeine Diskussion sowie das Fazit und ein Ausblick auf weiterführende Forschungsperspektiven in Abschnitt 8 bzw. 9.

2 Terminologische Bestimmungen

Für die Erkennung unterschiedlicher subjektiver menschlicher Zustände – wie Emotionen, Meinungen und Sentiments – in Text wird in dieser Arbeit zusammenfassend der Begriff „Subjectivity Analysis“ verwendet (Pang & Lee, 2008). Ein Problem der Subjectivity Analysis ist das Fehlen einer gemeinsamen Terminologie (Munezero et al., 2014). Da solche Zustände primär Gegenstand der Psychologie sind, sollten die verwendeten Begriffe einen eindeutigen Bezug auf psychologische Konzepte haben. Als erster Schritt dieser Arbeit werden daher wesentliche Terme aus diesem Bereich diskutiert und einer psychologischen Typologie affektiver Zustände eingeordnet.

Unter *emotion*[1] werden in der Literatur meist Phänomene wie Wut, Freude oder Trauer verstanden (Calvo & D'Mello, 2010; Strapparava & Mihalcea, 2008; Staiano & Guerini, 2014). Synonym dazu wird häufig der Begriff *affect* verwendet (Munezero et al., 2014). Seltener kommt auch *mood* zum Einsatz um die gleichen Phänomene zu bezeichnen (Bollen, Mao & Zeng, 2011; Acerbi, Lampos, Garnett & Bentley, 2013). Meistens werden diese Terme nicht genau oder nur durch Beispiele definiert.[2] Entsprechend divers sind die Bezeichnungen für das dazugehörige wissenschaftliche Feld. Am weitesten verbreitet scheint hierfür der Ausdruck Emotion Detection (ED) zu sein, aber ebenso werden eine Vielzahl anderer Terme benutzt.[3] Diese Arbeit verwendet daher den Ausdruck ED oder synonym dazu seine deutsche Übersetzung „Emotionserkennung“.

Auf der einen Seite stellen die Begriffe *emotion*, *affect* und *mood* also eine Gruppe mehr oder weniger synonym verwendeter Begriffe dar. Auf der anderen Seite stehen die Begriffe *sentiment* und *opinion*. Auch wenn diese beiden teilweise in einzelnen Formalisierungen voneinander abgegrenzt werden (Liu, 2012), ist in der Literatur doch anerkannt, dass die dazugehörigen Terme Sentiment Analysis (SA) bzw. Opinion Mining das gleiche Forschungsfeld beschreiben (Pang & Lee, 2008; Liu, 2012).[4]

Sentiments bzw. *opinions* können unterschiedlich komplex repräsentiert werden. Im einfachsten Fall ist Sentiment die semantische Orientierung einer Äußerung, also deren Eigenschaft, eine positive oder negative Wertung auszudrücken. Sentiments

[1] In diesem Abschnitt werden die englischen Begriffe aus der Literatur übernommen um durch zwangsläufig ungenaues Übersetzen nicht noch zusätzliche Verwirrung zu erzeugen.

[2] Liu (2012, S. 28) definiert Emotionen etwa als „our subjective feelings and thoughts.“

[3] Folgende Autoren verwenden den Ausdruck ED: Calvo und Mac Kim (2013); Gupta, Gilbert und Fabbrizio (2013); Canales und Martínez-Barco (2014); Agrawal und An (2012); Desmet und Hoste (2013); Lei, Rao, Li, Quan und Wenyin (2014). Andere mögliche Begriffe sind Affect Sensing, Affect Detection, Emotion Prediction (Munezero et al., 2014), Emotion Tagging (Das & Bandyopadhyay, 2009), Affect Recognition (Snow, O'Connor, Jurafsky & Ng, 2008) und Emotion Analysis (Staiano & Guerini, 2014).

[4] Für das Nebeneinander beider Begriffe werden historische Gründe genannt: Während „Sentiment Analysis“ eher in NLP-Kontexten verwendet wird, ist „Opinion Minig“ im Bereich des Information Retrieval (IR) stärker verbreitet (Pang & Lee, 2008, S. 6).

können auch als Wertungen auf einer mehrstufigen Skala – zum Beispiel von eins bis fünf – repräsentiert werden. Noch komplexer sind Repräsentationen die weitere Größen als die semantische Orientierung erfassen. Liu (2012) definiert *opinion* formal als ein 5-Tupel bestehend aus der bewerteten Entität, der bewerteten Eigenschaft dieser Entität, der Wertung (*sentiment*), dem Bewerter (*opinion holder*) und dem Zeitpunkt der Bewertung.

Zum Verhältnis von Sentiment Analysis und Emotion Detection gibt es unterschiedliche Standpunkte in der Literatur. Zahlreiche Autoren vertreten die Ansicht, dass Sentiment Analysis eine Teilaufgabe von Emotion Detection wäre, bzw. die semantische Orientierung eine vereinfachte Darstellung von Emotionen ist (Staiano & Guerini, 2014; Danisman & Alpkocak, 2008; Calvo & Mac Kim, 2013). Nach der gegenteiligen Meinung wird der Begriff Sentiment weiter gefasst, sodass dieser auch Emotionen miteinbezieht (Lei et al., 2014; Liu, 2012). Teilweise werden die Begriffe auch austauschbar benutzt (Rao, Li, Mao & Wenyin, 2014) oder nicht weiter differenziert (Palmer & Xue, 2010).

Die definitorische Trennung solcher Phänomene ist ebenso ein andauerndes Problem in der psychologischen Literatur. Scherer (2000, S. 140f.) stellt aus diesem Grund seiner Arbeit zu verschiedenen Emotionsmodellen eine Typologie bei, die Emotionen und vier weitere davon abzugrenzende Phänomene erfasst und unter dem Oberbegriff „affektive Zustände“(*affective states*) bündelt.

- Emotionen (*emotions*): „relatively brief episodes of synchronized responses by all or most organismic subsystems to the evaluation of an external or internal event as being of major significance (e.g, anger, sadness, joy, fear, shame, pride, elation, desperation).“

- Stimmungen (*moods*): „diffuse affect state, most pronounced as change in subjective feeling, of low intensity but relatively long duration, often without apparent cause (e.g., cheerful, gloomy, irritable, listless, depressed, bouyant).“

- Zwischenmenschliche Haltungen (*interpersonal stances*): „affective stance taken toward another person in a specific interaction, coloring the interpersonal exchange in that situation (e.g., distand, cold, warm, supportive, contemptous).“

- Einstellungen (*attitudes*): „relatively enduring, affectively colored beliefs, preferences, and predispositions toward objects or persons (e.g., liking, loving, hating, valuing, desiring).“

- Persönlichkeitsmerkmale (*personality traits*): „emotionally laden, stable personality dispositions and behavior tendencies, typical for a person (e.g., nervous, anxious, reckless, morose, hostile, envious, jealous).“

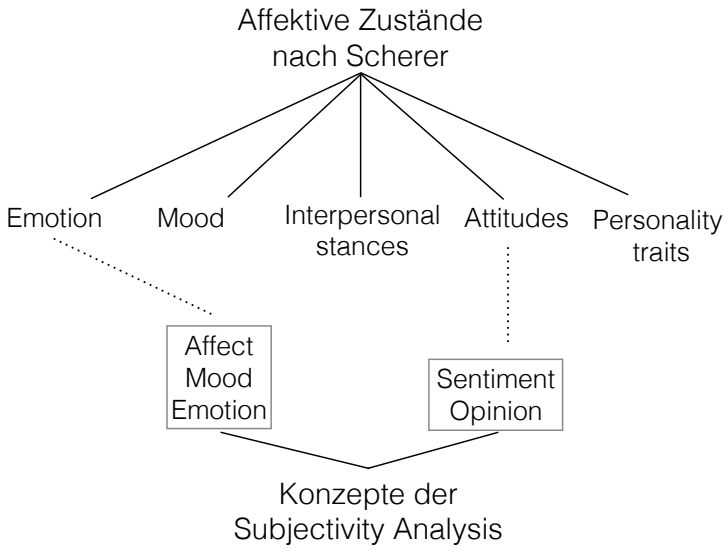

Abbildung 1: Verhältnis von Scherers Typologie affektiver Zustände zu wesentlichen Konzepten der Subjectivity Analysis.

Aus den gegebenen Beispielen geht deutlich hervor, dass nach Scherers Typologie affektiver Zustände Emotionen dem entsprechen, was in der Subjectivity Analysis meist als *emotion*, *affect* oder *mood* bezeichnet wird. *Sentiments* bzw. *opinions* dagegen drücken Wertungen aus und sind daher – zumindest implizit – auf ein bewertetes Objekt bezogen. Diese Objektbezogenheit findet sich auf Seiten von Scherers Typologie bei zwischenmenschlichen Haltungen und Einstellungen. Das Objekt einer zwischenmenschlichen Haltung kann nach Scherer jedoch nur eine Person sein. Darüber hinaus ist ihre Dauer auf eine konkrete Kommunikationssituation beschränkt. Beides passt nicht auf typische Anwendungsschwerpunkte der Sentiment Analysis zum Beispiel Produktbewertungen (Pang & Lee, 2008). Am besten entspricht *opinion* und *sentiment* daher der Einstellung (*attitude*) nach Scherers Typologie. Das Verhältnis der Konzepte der Subjectivity Analysis zu Scherers Typologie ist in Abbildung 1 dargestellt.

3 Theorie und Forschungsstand

Dieser Abschnitt gibt einen Überblick über die wesentlichen Möglichkeiten, wie Emotionserkennung konzeptioniert werden kann, indem er auf psychologische Emotionsmodelle und deren formale Repräsentation eingeht. Er gibt einen Überblick über bestehende methodische Ansätze und verweist auf einschlägige interdisziplinäre Anwendungen.

3.1 Emotionsmodelle

In der Psychologie gibt es keinen Konsens über eine gemeinsame Emotionstheorie. Stattdessen existiert eine Fülle unterschiedlicher Modelle und Ansätze dazu (Scherer, 2000). In der ED wurden vor allem *diskrete* und *dimensionale* Emotionsmodelle verwendet (Canales & Martínez-Barco, 2014), wobei dimensionale Modelle eher eine untergeordnete Rolle spielen (s. Tabellen 2 und 19). Daneben gibt es in der Psychologie noch zahlreiche weitere Emotionsmodelle und -Theorien, die jedoch größtenteils noch keinen Zugang in die ED gefunden haben (Calvo & Mac Kim, 2013).

3.1.1 Diskrete Emotionsmodelle

Diskrete Modelle (Scherer, 2000) gehen davon aus, dass sich die Vielfalt aller möglichen emotionalen Zustände auf eine geringe Anzahl eindeutig unterscheidbarer Emotionskategorien oder -klassen zurückführen lässt. Basisemotionen sind die in der Psychologie am weitesten verbreitete Auslegung solcher diskreten Emotionskategorien. Theoretiker der Basisemotionen gehen davon aus, dass sich diese im Laufe der Evolution als wichtige Überlebensstrategien herausgebildet haben, um sich an die Umwelt anzupassen und auf diese zu reagieren. Dementsprechend hat jede Basisemotion spezifische Auslösungsmechanismen und Reaktionsmuster. Zum Beispiel wird Furcht durch eine Bedrohungssituation ausgelöst und führt zu einem Fluchtverhalten. Ein wesentlicher Teil dieser Reaktionsmuster sind spezifische Gesichtsausdrücke, die als Beweis für das Konzept von Basisemotionen angeführt werden (Ekman, 1992). Ausgehend von der Tatsache, dass bei bestimmten Auslösern alle Menschen unabhängig von ihrem sprachlichen und kulturellen Hintergrund mit sehr ähnlichen Gesichtsausdrücken reagieren, sei es plausibel darauf zu schließen, dass sie über einen gemeinsamen Satz von Basisemotionen verfügen.

In der Emotionserkennung sind die sechs Basisemotionen nach Ekman (1992) mit Abstand am verbreitetsten (Calvo & Mac Kim, 2013). Er identifiziert nach der

oben skizzierten Methode Wut, Furcht, Trauer, Freude, Ekel und Überraschung.[5] Bei Anwendungen in der Emotionserkennung werden häufig verschiedene Abwandlungen dieser sechs Emotionsklassen verwendet. Calvo und Mac Kim (2013) verzichten zum Beispiel auf Überraschung und fassen Wut und Ekel zu einer Kategorie zusammen. Basisemotionen nach anderen Autoren werden dagegen nur sehr vereinzelt verwendet. Neviarouskaya, Prendinger und Ishizuka (2011) benutzen etwa die neun Basisemotionen von Carroll Izard.

3.1.2 Dimensionale Emotionsmodelle

Dimensionale Modelle (Scherer, 2000) gehen davon aus, dass sich emotionale Zustände durch ihre Position in einem mehrdimensionalen Raum eindeutig beschreiben lassen. Die Dimensionen sollten voneinander unabhängig sein, sodass sich ihre Werte frei miteinander kombinieren lassen. Diese Eigenschaft wird in der psychologischen Literatur als „Orthogonalität“ bezeichnet (Warriner, Kuperman & Brysbaert, 2013). Anspruch solcher Modelle ist es, eine möglichst kleine Anzahl von Dimensionen zu finden, die zusammen alle denkbaren emotionalen Zustände erfassen können (Russell & Mehrabian, 1977). Die Komponenten nehmen dabei numerische Werte[6] aus einem gegebenen Intervall an. Am weitesten in der ED verbreitet sind das dreidimensionale Modell von Russell und Mehrabian (1977) und das zweidimensionales Modell von Russell (1980).

Nach dem dreidimensionalen Modell setzen sich Emotionen aus den Komponenten Valenz, Erregung und Dominanz (*valence, arousal, dominace*) zusammen.[7] Daher wird es nachfolgend „VAD-Model“ genannt.[8] Der Valenzwert einer Emotion gibt dabei an, als wie angenehm sie empfunden wird. Er reicht von extrem negativer bis zu extrem positiver Bewertung. Diese Komponente entscheidet daher zwischen Emotionen wie Glück und Unglück oder Zufriedenheit und Unzufriedenheit. In der Literatur der ED herrscht, soweit dazu Stellung genommen wird, Einigkeit darüber, dass Valenz des VAD-Modells der semantischen Orientierung der SA entspricht (Calvo & Mac Kim, 2013). Die Erregungskomponente bezieht sich auf das Niveau der mentalen Aktivität. Sie unterscheidet daher zwischen Zuständen wie Schläfrigkeit einerseits und Ekstase oder Panik andererseits. (Bakker, van der Voordt, Vink &

[5] Darüber hinaus gibt es eine Reihe von Emotionen, die ebenfalls Basisemotionen sein könnten, darunter Scham, Ehrfurcht und Verachtung.

[6] Für Dezimalbrüche wird in dieser Arbeit durchgängig auf die angelsächsische Notation zurückgegriffen, sodass ein Punkt statt einem Komma als Dezimaltrennzeichen verwendet wird.

[7] Eine piktographische Interpretation dieser Komponenten stellt das Self-Assessment-Manikin (SAM) dar, das zum Messen von Emotionen verwendet wird (s. Abbildung 14 im Anhang).

[8] Die erste Komponente wird in der Literatur häufig mit *pleasure* bezeichnet. In diesen Fällen wird das Modell entsprechend *PAD-Model* genannt. Die vorliegende Arbeit hält sich mit der Verwendung des Begriffs „Valenz“ an den Gebrauch der später verwendeten Ressource (Warriner et al., 2013).

de Boon, 2014). Die Dominanzkomponente gibt an, wie sehr man das Gefühl der Kontrolle über die Situation hat bzw. wie stark man in seinen Handlungsspielräumen eingeengt oder sich von der Situation beeinflusst fühlt. Sie reicht vom Gefühl völliger Kontrolle bis zum Gefühl völligen Ausgeliefertseins (Bakker et al., 2014). In der Vergangenheit ist anhand zahlreicher Probandenstudien deutlich geworden, dass Dominanz weniger intuitiv verständlich ist, als Valenz und Erregung.[9]

Ein weiteres Problem der Dominanzkomponente ist, dass ihre Erklärungskraft in vielen Studien hinter der von Valenz und Erregung zurückbleibt. Damit hängt zusammen, dass sie häufig eine deutliche Korrelation mit Valenz aufweist (Russell & Mehrabian, 1977).[10] Dies erscheint plausibel, da sich Kontrolle mit Sicherheit assoziieren lässt und dadurch positiv bewertet wird. Daher hat Russell (1980) später sein 2D-Modell – auch „Circumplex-Modell“ genannt – entwickelt. Es unterscheidet sich von seinem und Mehrabians 3D-Modell vereinfacht ausgedrückt durch das Nichtvorhandensein der Dominanzkomponente.

In der neueren Forschung mehren sich jedoch die Hinweise, dass Dominanz genauso wichtig ist wie Valenz und Erregung (Bakker et al., 2014). In der Literatur zur Emotionserkennung wird darauf hingewiesen, dass diese Komponente – unabhängig von ihrer Güte als psychologisches Konstrukt – eine entscheidende Rolle in verschiedenen technischen Anwendungsszenarien spielt (Broekens, 2012). Auch wenn sie weniger gut fundiert ist als die anderen Komponenten, sei es ohne sie doch unmöglich, die ganze Bandbreite menschlicher Emotionen abzubilden (Broekens, 2012; Bakker et al., 2014).

3.1.3 Verhältnis von Basisemotionen und VAD-Modell

Ein wichtiger Unterschied zwischen Ekmans Basisemotionen und Russells und Mehrabians VAD-Modell ist, dass man sich nach dem VAD-Modell jederzeit in irgendeinem emotionalen Zustand befindet (Russell & Mehrabian, 1977). Bei Ekman (1992) sind Emotionen hingegen als kurze Episoden gedacht. Dies entspricht der Tatsache, dass die unterschiedlichen Theorien jeweils andere Aspekte von Emotionen besonders betonen (Scherer, 2000). Diskrete Modelle setzen ihren Fokus auf Reaktionsmuster, vor allem Gesichtsausdrücke. Dementsprechend finden aus dieser Perspektive Emotionen in erster Linie dann statt, wenn spezifische Reaktionen zu beobachten sind. Eine neutrale Emotion ist bei Ekman daher nicht vorgesehen.[11] Dimensionale

[9] Broekens (2012) setzt sich daher tiefergehend mit der Interpretation dieser Komponente und ihrer Bedeutung für die ED auseinander.

[10] Diese häufig beobachtete Korrelation macht die Annahme der Orthogonalität der Dimensionen des VAD-Raums zweifelhaft. Zwar gibt es auch Fälle, in denen Dominanz und Valenz entgegengesetzte Merkmalsausprägungen haben. Dies ist jedoch selten (Warriner et al., 2013).

[11] Dagegen können Anwendungen in der ED auch die Emotionsklasse *neutral* enthalten (Balahur, Hermida & Montoyo, 2012).

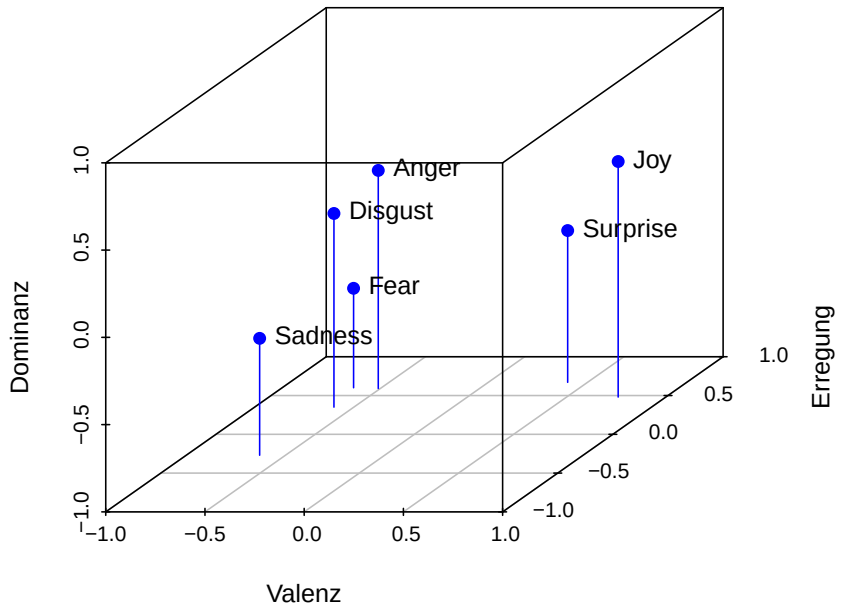

Abbildung 2: Einfache Abbildung von Ekmans Basisemotionen in den VAD-Raum nach Russell und Mehrabian (1977).

Modelle hingegen konzentrieren sich vor allem auf die subjektive Wahrnehmung einer Emotion durch das Individuum bzw. dessen verbalen Ausdruck. Auf diese Weise lässt sich zu jedem Zeitpunkt der emotionale Zustand messen, der dann folglich auch *neutral* sein kann.

Trotz der oben skizzierten Unterschiede gibt es in der psychologischen Forschung Bemühungen beide Ansätze zusammenzuführen. Dadurch würden Studien untereinander vergleichbar und die Verwendbarkeit vorhandener Ressourcen erhöht. Stevenson, Mikels und James (2007) entwickelten zu diesem Zweck Regressionsmodelle um die beiden unterschiedlichen Repräsentionsmöglichkeiten von Emotionen ineinander umzurechnen. In der ED stellen Calvo und Mac Kim (2013) ein einfaches Verfahren vor um Emotionen im VAD-Raum auf Basisemotionen abzubilden (s. Abschnitt 3.4.4). Abbildung 2 zeigt die Position von Ekmans Basisemotionen im VAD-Raum. Die Werte dafür stammen von Russell und Mehrabian (1977), die Probanden aufgefordert haben, bestimmte Emotionen durch Valenz, Erregung und Dominanz zu bewerten. Aus den Antworten mehrerer Probanden wurde jeweils der Durchschnitt gebildet. Aus der Grafik ist ersichtlich, dass Ekmans Basisemotionen im Emotionsraum, den Russell und Mehrabian entwerfen, sehr ungleich verteilt sind. Nur eine der sechs Basisemotionen weist negative Erregung und nur zwei weisen positive Valenz auf. Im

Bereich niedriger Valenz und hoher Erregung sind dagegen die Hälfte der Basisemotionen gehäuft. Dies entspricht Ekmans Konzeption von Emotionen als durch die Evolution entstandene Reaktionsmuster auf wesentliche Umwelteinflüsse wie zum Beispiel Bedrohungen. Durch diesen Fokus erfassen sie jedoch nur einen geringen Teil der möglichen emotionalen Zustände des VAD-Modells.

Das bedeutet, dass die Verwendung von Basisemotionen in ED-Kontexten den Nachteil hat, dass die für eine Anwendungsdomäne besonders relevanten Emotionsausprägungen ggf. nicht erfasst werden können. Zum Beispiel ist die Emotion Ekel in Unternehmensberichten wahrscheinlich nicht vertreten. Von Ekman nicht abgedeckte Emotionen wie Zufriedenheit, Geborgenheit und Ruhe können dagegen für empirische Fragestellungen viel relevanter sein (Bollen et al., 2011). Das VAD-Modell hat somit den Vorteil, für alle Domänen gleichermaßen geeignet zu sein, da *alle* emotionalen Zustände erfasst werden. Dagegen muss die Auswahl der verwendeten Basisemotionen je nach theoretischem Hintergrund und Anwendungsfall variieren (s. Abschnitt 3.1.4).

3.1.4 Andere Modelle

Nicht alle Autoren verwenden Emotionsmodelle aus der Psychologie. Stattdessen orientieren sich manche an technischen Gegebenheiten oder den Anforderungen eines konkreten Anwendungsfall. Lei et al. (2014) nutzen die Funktion einer Nachrichten-Website, mit der Nutzer Nachrichten emotional bewerten können, als Ausgangspunkt für die Entwicklung eines umfangreichen Trainingskorpus für maschinelle Lernverfahren (s. Abschnitt 3.4). Daher übernehmen sie die Emotionskategorien dieser Seite. Desmet und Hoste (2013) entwickeln ein System zur Suizid-Prävention, das entsprechende Indikatoren in Online-Foren erkennen soll. Ihr System klassifiziert Nutzer-Beiträge nach 15 Emotionskategorien, die besonders geeignet sein sollen, beabsichtigte Selbstmorde vorherzusagen.

3.2 Formale Repräsentation

Die vorherzusagende Emotion einer linguistischen Einheit ist die Zielvariable der Emotionserkennung. Emotionen können jedoch unterschiedlich formal repräsentiert werden. Aus dem letzten Abschnitt geht hervor, dass Basisemotionen implizieren eine diskrete Zielvariable zu verwenden. Das 3D- und 2D-Modell legen dagegen nahe, dass die Zielvariable ein Vektor aus reellen, stetigen Komponenten ist. In der ED wird mit diesen Implikationen der psychologischen Theorie freier umgegangen, sodass teilweise auch Basisemotionen als Vektoren repräsentiert werden, wobei die Werte der Komponenten die Übereinstimmung mit der jeweiligen Emotionsklasse oder deren Wahrscheinlichkeit ausdrücken. Ebenso gibt es Arbeiten, die die theoretisch

	Zielvariable	
Emotionsmodell	diskret	stetig
diskret	[Label: JOY]	[JOY: 0.6 ANGER: 0.0 SADNESS: 0.0 DISGUST: 0.0 FEAR: 0.1 SURPRISE: 0.4]
dimensional	[VALENZ: + ERREGUNG: + DOMINANZ: −]	[VALENZ: 0.6 ERREGUNG: 0.5 DOMINANZ: -0.1]

Tabelle 1: Exemplarische Darstellung der möglichen Emotionsrepräsentationen nach Form der Zielvariable und dem Emotionsmodell, hier vereinheitlicht als Merkmalsstrukturen notiert.

unbegrenzte Anzahl emotionaler Zustände eines dimensionalen Modells auf wenige diskrete Klassen reduzieren.

Zur Veranschaulichung werden in Tabelle 1 alle vier daraus resultierenden Möglichkeiten exemplarisch dargestellt. Dabei wurden an dieser Stelle zur besseren Übersicht Merkmalsstrukturen verwendet, sodass alle Repräsentationen einheitlich notiert werden können. Beispiele für die jeweilige Verwendung folgen in Abschnitt 3.4. Den Repräsentationen der ersten Zeile liegen Ekmans Basisemotionen zugrunde. In der linken Zelle wird lediglich aus den sechs Emotionskategorien eine als Label zugewiesen, während in der rechten Zelle jeder Kategorie ein Zustimmungswert zwischen 0 und 1 zugeordnet wird. Die zweite Zeile verwendet das VAD-Modell. In der linken Zelle ist jede Komponente mit dem Wert „positiv“ oder „negativ“ besetzt. In der rechten Zelle werden jeweils Werte zwischen -1 und 1 zugewiesen. Jede der vier Merkmalsstrukturen soll dabei die entsprechenden Repräsentation des gleichen Satzes bzw. der gleichen Emotion darstellen. Die Kombination aus einer stetigen Zielvariable und Basisemotionen ist zwar in der Literatur mehrfach eingesetzt worden, beinhaltet jedoch möglicherweise einige noch nicht berücksichtigte Probleme bezüglich der empirischen Adäquatheit.[12] Die Zuordnung einer diskreten Zielvariable

[12]Das Grundproblem scheint zu sein, dass die Basisemotionen nicht in der gleichen Weise voneinander unabhängig sind, wie die Dimensionen des VAD-Modells. Dies wird deutlich, wenn man die Emotionen Trauer, Furcht und Freude betrachtet. Intuitiv liegen Trauer und Furcht offensichtlich näher zusammen als Trauer und Freude, was auch durch das VAD-Modell bestätigt wird (s. Abbildung 2). In einer Vektordarstellung der Basisemotionen würde dies jedoch nicht gelten. Insofern ist es höchst zweifelhaft, ob die Berechnung der üblichen Abstandsmaße dort sinnvoll ist. Darüber hinaus scheint es keine *eineindeutige* Beziehung zwischen den empirischen emotionalen

heißt *Klassifikation*. Bei einer stetigen Zielvariable heißt sie *Regression* (Bishop, 2007, S. 3).

3.3 Evaluation emotionserkennender Systeme

Die Entscheidung, Emotionserkennung als Klassifikation oder Regression zu betrachten hat auch Einfluss auf die Maße, die zur Beurteilung der Leistungsfähigkeit eines fertigen Systems herangezogen werden. Die Evaluation wird anhand eines *Testkorpus* durchgeführt. Im Fall der Emotionserkennung handelt es sich dabei um eine Kollektion linguistischer Einheiten, für die bereits Emotionsbewertungen vorliegen. Testkorpora sind i.d.R von menschlichen Annotatoren manuell beurteilt worden – in diesem Fall heißen sie „Goldstandard“ (Carstensen et al., 2010, S. 489f.). Sie müssen dies jedoch nicht zwingend sein.[13] Je besser ein bestehendes System die manuellen Bewertungen reproduzieren kann, desto höher ist seine Performanz. In der Emotionserkennung wird die Performanz bei Klassifikationsaufgaben durch die Maße Precision (P, Genauigkeit), Recall (R, Abdeckung) und F-Maß evaluiert (Strapparava & Mihalcea, 2008; Calvo & Mac Kim, 2013). Precision beruht auf dem Verhältnis der richtig klassifizierten Fälle zu allen Zuweisungen einer Klasse. Umgekehrt beruht Recall auf dem Verhältnis der richtig klassifizierten Fälle zu den Fällen die richtigerweise dieser Klasse zuzurechnen gewesen wären. Das F-Maß ist das gewichtete harmonische Mittel aus Precision und Recall.[14] Bei Regressionsaufgaben wird zur Evaluation stattdessen die Pearson-Korrelation r von den meisten Autoren benutzt (Lei et al., 2014; Strapparava & Mihalcea, 2008; Staiano & Guerini, 2014; Neviarouskaya et al., 2011; Katz, Singleton & Wicentowski, 2007).[15]

Im Falle manuell annotierter Korpora sind i.d.R. für jede Einheit mindestens zwei Wertungen von verschiedenen Annotatoren unabhängig voneinander abgegeben worden. Die Übereinstimmung der verschiedenen Wertungen wird als Inter-Annotator

Zuständen und ihrer formalen Repräsentation zu geben: ein Vektor, der gleiche Werte in der Freude- und Trauerkomponente hat, und einer, bei dem beide Komponenten null sind, repräsentieren wahrscheinlich den gleichen empirischen Zustand. Darüber hinaus scheint das Durchführen der üblichen Vektoroperationen nicht immer empirisch sinnvolle Ergebnisse zu liefern: Wut und Ekel sind laut dem VAD-Raum sehr ähnliche Emotionen. Insofern scheint es unplausibel, dass der Durchschnitt aus zwei Vektoren, die jeweils maximale Wut bzw. maximalen Ekel darstellen, ein Vektor mit mittlerer Wut und mittlerem Ekel ist. Diese Probleme können im Rahmen der Arbeit leider nur angedeutet werden. Offensichtlich verdienen sie jedoch eine tiefergehende Auseinandersetzung in der Literatur.

[13] Hasan, Rundensteiner und Agu (2014) nutzen etwa Hashtags, um automatisch Tweets zu labeln.

[14] Da Emotionen ein Merkmal mit mehr als zwei Ausprägungen sind, handelt es sich hierbei um ein Multiklassen-Problem. Daher ist die formale Definition der Maße P, R, und F komplexer als im einfachen Fall der binären Klassifikation. Eine Übersicht über die unterschiedlichen Varianten dieser Performanzmaße liefern Sokolova und Lapalme (2009).

[15] In einem Fall wird statt dem Korrelationskoeffizienten ein Rangkorrelationskoeffizient benutzt (Baccianella, Esuli & Sebastiani, 2010). Auf die Motivation, nur die Richtigkeit der Reihenfolge als Gütekriterium für die Regression zu verwenden, wurde jedoch nicht eingegangen.

Agreement (IAA) bezeichnet. Für Klassifikationsaufgaben wird die IAA häufig durch Cohens Kappa (κ) gemessen (Manning, Raghavan & Schütze, 2008, S. 165). Bei Regressionsaufgaben wird in der Emotionserkennung die Übereinstimmung als Durchschnitt der Pearson-Korrelation zwischen den Annotatoren angegeben (Strapparava & Mihalcea, 2007; Snow et al., 2008).[16] Die IAA dient häufig als Vergleichsmaßstab für die Leistung eines automatischen Annotators (*human ceiling*) (Jurafsky & Martin, 2008, S. 189).

Wie zahlreiche Autoren übereinstimmend berichten ist das Bewerten von Emotionen in Texten keine einfache Aufgabe für menschliche Annotatoren (Danisman & Alpkocak, 2008; Lei et al., 2014; Katz et al., 2007). Das Testkorpus des Emotionserkennungs-Wettbewerbs des Workshops SemEval-2007 (Strapparava & Mihalcea, 2007) bestand aus Nachrichtenüberschriften. Diese wurden mit Ekmans Basisemotionen annotiert, wobei jeweils eine Skala zwischen 1 und 100 verwendet worden ist. Die IAA betrug je nach Emotionsklasse zwischen $r = 0.36$ (Überaschung) und $r = 0.68$ (Trauer) und liegt damit deutlich unter der IAA für die semantische Orientierung ($r = 0.78$), die ebenfalls bewertet wurde. Snow et al. (2008) erreichen bei einer daran angelehnten Erhebung nur geringfügig bessere Übereinstimmungswerte. Einen Hinweis darauf, warum die IAA so niedrig ist, liefern Katz et al. (2007), die die Annotatoren nach der beendeten Aufgabe interviewt haben. Diese drückten Unsicherheit darüber aus, wessen Emotionen – die eigenen oder die des Subjekts eines Satzes – beurteilt werden sollte. Die Autoren illustrieren dies am Beispiel (1).

(1) Italy defeats France in World Cup Final.

Hier zeigt sich deutlich ein Problem der Emotionserkennung: Emotionen und ihre Auslöser sind subjektiv. Ob ein Ereignis für den Einzelnen Grund zur Freude oder zur Trauer ist, hängt vor der individuellen Bewertung dieses Ereignisses ab (Scherer, 2000; Balahur et al., 2012).[17] Ein weiteres Problem für die Evaluation ist, dass die ohnehin wenigen verfügbaren Goldstandards meist unterschiedliche Emotionsmodelle verwenden. Dadurch wird die Auswahl der verwendbaren Ressourcen weiter reduziert. Hinzu kommt, dass die Performanz von ED-Systemen sehr stark vom jeweiligen Testkorpus abhängt (Calvo & Mac Kim, 2013).

[16] Hierfür wird zunächst für jeden Annotator die Korrelation seiner Bewertungen mit dem Durchschnitt der Bewertungen der anderen Annotatoren berechnet, sodass ein Korrelationswert pro Annotator vorhanden ist. Aus diesen Korrelationswerten wird wieder der Durchschnitt gebildet.

[17] Eine methodische Implikation dessen ist, dass in der Emotionserkennung zwischen *Leser*- und *Schreiber*-Emotionen unterschieden werden können. Diese Überlegung wurde in einigen Arbeiten angeführt (Rao et al., 2014; Leveau, Jhean-Larose, Denhière & Nguyen, 2012; Calvo & Mac Kim, 2013). Allerdings hat eine weitergehende Auseinandersetzung mit diesem Aspekt soweit bekannt noch nicht stattgefunden.

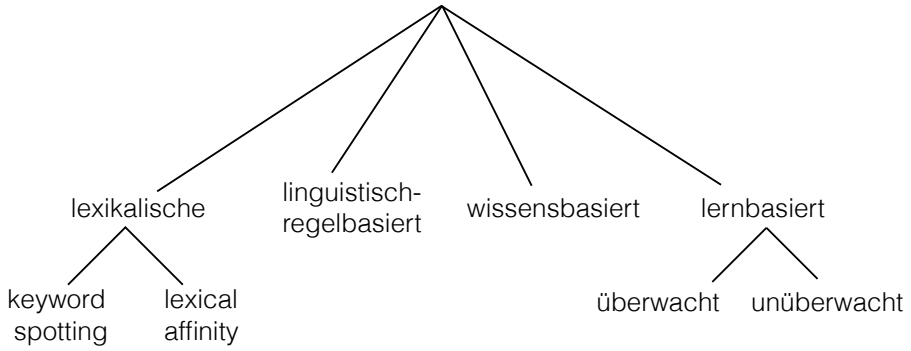

Abbildung 3: Vorgeschlagene Typologie methodischer Ansätze in der Emotion Detection.

	Diskrete Modelle		Dimensionale Modelle	
	Klassifikation	Regression	Klassifikation	Regression
Keyword-Spotting	✓	✓		
Lexical Affinity	✓	✓		
Linguistisch-Regelbasiert	✓	✓		
Wissensbasiert	✓			
überwachtes Lernen	✓	✓	✓	
unüberwachtes Lernen	✓	✓		

Tabelle 2: In der Literatur zur Emotionserkennung vorhandene Kombinationen aus Emotionsmodell, Form der Zielvariable und methodischem Ansatz; angefertigt nach Tabelle 19.

3.4 Methodische Ansätze

Nachfolgend werden die wichtigsten Ansätze der ED exemplarisch an ausgewählten Studien erläutert und diskutiert. Ein aktueller Survey findet sich bei Canales und Martínez-Barco (2014). Die hier vorgeschlagene Typologie orientiert sich lose an Canales und Martínez-Barco (2014), Agrawal und An (2012) sowie Cambria, Schuller, Xia und Havasi (2013). Sie wird in Abbildung 3 dargestellt. Die für diesen Abschnitt gesichtete Literatur ist in Tabelle 19 angegeben (s. Anhang).[18] Ihr Inhalt wird in Tabelle 2 zusammengefasst.

3.4.1 Lexikalischer Ansatz

Lexikalische Ansätze bilden die einfachste Form von Emotionserkennung. Sie bedienen sich als Indikatoren lediglich der in einer zu bewertenden linguistischen Einheit auftretenden, einzelnen Wörter. Dabei werden Informationen zum Kontext der Wörter und zu deren Abfolge nicht berücksichtigt. In der SA ist es üblich solche

[18] Daneben sind auch in der Psychologie Arbeiten veröffentlicht worden, in denen automatisch die Emotion von Texten gemessen wird (Leveau et al., 2012).

Verfahren nochmals in *Keyword-Spotting*- und *Lexical-Affinity*-Ansätze zu unterteilen (Cambria et al., 2013). Für die Emotionserkennung wird vorgeschlagen, die gleiche Differenzierung vorzunehmen.

Keyword Spotting verwendet zur Vorhersage einer Emotion eine Liste eindeutiger Hinweiswörter, die sich meist direkt auf die entsprechende Emotion beziehen, zum Beispiel *happy, sad* oder *afraid*. Beim Auftreten eines dieser Wörter wird die entsprechende Emotion zugewiesen. Dabei handelt es sich um den einfachsten Ansatz, der im Vergleich zu komplexeren Ansätzen nur geringe Performanz aufweist. Er wird jedoch aufgrund seiner leichten Umsetzbarkeit trotzdem vielfach in der SA eingesetzt (Cambria et al., 2013). Auch bei interdisziplinären Anwendungen von Emotionserkennung werden solche Verfahren häufig eingesetzt (Acerbi et al., 2013; Hajek, Olej & Myskova, 2014). Durch diese Eigenschaften eignet er sich zudem gut als Vergleichsmaßstab für die Beurteilung fortgeschrittener Verfahren (*baseline*). Strapparava und Mihalcea (2008) nutzen hierfür die Ressource WordNet Affect (Strapparava & Valitutti, 2004) um Hinweiswörter im Satz zu identifizieren.

Keyword Spotting hat durch seine Einfachheit zahlreiche Nachteile. Das Ignorieren von Kontext und Reihenfolge macht dieses Verfahren zum Beispiel unsensibel gegenüber Negation.

(2) Heute bin ich sehr glücklich.

(3) Heute bin ich überhaupt nicht glücklich.

(4) Gestern hatte ich einen Verkehrsunfall.

(2) und (3) würden durch das Auftreten des Wortes „glücklich“ mit dem Label *joy* versehen werden. Die Negation in (3) würde unerkannt bleiben. Darüber hinaus ist die Anwesenheit eindeutiger Hinweiswörter ein sehr oberflächliches Merkmal, da auch ohne solche Wörter Emotionen transportiert werden können, wie (4) zeigt.[19]

Der *Lexical-Affinity*-Ansatz bietet genau hier Vorteile. Anders als beim Keyword Spotting werden nicht nur Wörter, die sich direkt auf Emotionen beziehen, betrachtet, sondern eine i.d.R wesentlich größere Ressource (nachfolgend „Emotionslexikon“ oder kurz „Lexikon“ genannt) speichert den emotionalen Gehalt, den ein Wort unabhängig von seinem Kontext hat. So würde etwa in (4) das Wort „Verkehrsunfall“ erkannt werden und dem Satz entsprechend eine negative Emotion zugeordnet werden.

Die Performanz von Lexical-Affinity-Ansätzen übertrifft i.d.R. die von Keyword-Spotting (Cambria et al., 2013), ist jedoch stark von der Qualität des verwendeten Lexikons abhängig. Staiano und Guerini (2014) zeigen mit einem umfangreichen, durch Crowdsourcing erstellten Lexikon, dass der Lexical-Affinity-Ansatz durchaus

[19] In der Alltagssprache und sogar in besonders emotionsreicher poetischer Sprache beziehen sich weniger als 5% der Wörter direkt auf Emotionen, was die Leistungsfähigkeit dieses Ansatzes stark beschränkt (Pennebaker, Mehl & Niederhoffer, 2003).

performanter sein kann als Systeme, die fortgeschrittenere Methoden wie zum Beispiel maschinelles Lernen einsetzen. Sie erreichen mit ihrer Regression je nach Emotion zwischen $r = 0.54$ und $r = 0.21$, was die Performanz aller Systeme des SemEval-2007 (Strapparava & Mihalcea, 2007) und von Strapparava und Mihalcea (2008) übertrifft.

Trotz seiner Überlegenheit gegenüber Keyword-Spotting ist der Lexical-Affinity-Ansatz immer noch naiv, da er alleine auf der Wortebene arbeitet. Dadurch können verschiedene Wortbedeutungen nicht unterschieden werden und Negationen wie in (3) bleiben weiterhin unerkannt. Außerdem ist die Erstellung eines Lexikons mit viel Aufwand verbunden, wenn sie von menschlichen Ratern geleistet wird. Darüber hinaus führt dessen Übertragung in eine anderen Domäne meist zu deutlichen Performanzeinbußen (Cambria et al., 2013).

3.4.2 Linguistisch-regelbasierter Ansatz

Linguistisch-regelbasierte Ansätze erweitern lexikalische Ansätze dahingehend, dass die zugewiesene Emotion nicht mehr einzig und allein von den vorkommenden Wörtern abhängt. Stattdessen werden die Emotionen, die den Wörtern zugeordnet werden anhand linguistisch motivierter Regeln modifiziert. Der Vorteil dieses Ansatzes ist, dass die Performanz weiter verbessert werden kann, da ein gewisses Maß an Kontextinformation mit berücksichtigt wird. Allerdings ist das Aufstellen dieser Regeln sehr zeitaufwendig und bedarf darüber hinaus linguistischen Expertenwissens.

Agrawal und An (2012) verwenden etwa zur Modifikation der Wortemotionen Regeln, die auf dem Strukturbaum des Satzes beruhen und zum Beispiel bei Negationen und Adjektiv-Nomen-Verbindungen angewandt werden. Die resultierende Satzemotion wird dann als Mittelwert der modifizierten Wortemotionen berechnet.

3.4.3 Wissensbasierter Ansatz

Wissensbasierte Ansätze werden in der Emotionserkennung nur selten eingesetzt (s. Tabelle 19). Sie greifen auf außersprachliches Wissen zurück, wie es insbesondere in Ontologien repräsentiert sein kann. Der Aufbau solcher Wissensressourcen kann sich dabei sehr genau an Erkenntnissen aus der – im Fall der Emotionserkennung – Psychologie orientieren, sodass das Verfahren eine gute theoretische Fundierung ermöglicht. Ihr großes Hindernis ist jedoch der Arbeitsaufwand und das Expertenwissen, das zu ihrer Erstellung benötigt wird.

Balahur et al. (2012) extrahieren aus dem ISEAR-Korpus (International Survey of Emotional Antecedents and Reactions) Wissen darüber welche Ereignisse – abhängig von der semantischen Rolle der betroffenen Personen im Satz – welche Emotionen hervorrufen, und erstellen daraus eine Ontologie. Die Evaluation auf einem bis dahin

zurückgehaltenem Subkorpus zeigt vielversprechende Ergebnisse. Insbesondere ist die Übertragbarkeit auf andere Domänen größer als bei anderen Ansätzen. Allerdings ist die Abdeckung der Wissensressource noch stark eingeschränkt.

3.4.4 Lernbasierter Ansatz

Lernbasierte Ansätze beruhen auf der Anwendung von statistischen Lernalgorithmen. Sie lassen sich in überwachte und unüberwachte Verfahren einteilen (Bishop, 2007). *Überwachte Lernverfahren* sind Algorithmen, die über gelabelte Trainingsdaten, zum Beispiel Sätze, für die schon Bewertungen ihres Emotionsgehaltes vorliegen, generalisieren. Das heißt diese Algorithmen finden Funktionen, die von den Eingabewerten – den Sätzen – möglichst gut auf die Zielwerte – die Label – abbilden. Ist so eine Funktion gefunden, kann sie auf neue Sätze angewendet werden um die Bewertungen der menschlichen Rater möglichst gut zu imitieren (Bishop, 2007, S. 3). Solche Verfahren erreichen oft eine gute Performanz. Allerdings sind sie dafür von großen Mengen annotierter Trainingsdaten abhängig. Häufig eingesetzt werden Naive-Bayes- (NB) und k-Nearest-Neighbor-Klassifikatoren (kNN) sowie Support Vector Machines (SVM) (Roberts, Roach, Johnson, Guthrie & Harabagiu, 2012; Hasan et al., 2014; Strapparava & Mihalcea, 2008).

Hasan et al. (2014) umgehen die manuelle Annotation von Trainingsdaten, indem sie durch Hashtags Tweets automatisch einer von vier Emotionsklassen zuordnen. Sie verwenden eine Modifikation von Russells 2D-Modell, in der jeder Quadrant der Valenz-Erregungs-Ebene einer Emotionsklasse entspricht. Unter den betrachteten Lernalgorithmen weisen SVMs und kNN-Klassifikatoren die beste Performanz auf. Beide erreichen 90% F-Maß. Unklar bleibt, ob diese in Bezug auf frühere Arbeiten mit überwachten Lernverfahren sehr guten Ergebnisse (Strapparava & Mihalcea, 2008) darauf zurückzuführen sind, dass Tweets deutlich einfacher zu klassifizieren sind als andere Textgattungen, dass die Klassen des verwendeten Emotionsmodells einfacher zu erkennen sind oder dass durch das automatische Verfahren ein sehr großes Trainingskorpus erhoben werden konnte (134100 Tweets).

Unüberwachte Lernverfahren greifen nicht auf Trainingsdaten zurück. Stattdessen sollen Strukturen in nicht gelabelten Eingabewerten ausfindig gemacht werden. Darunter fallen etwa Clustering- und Dimensionsreduktionsverfahren (Bishop, 2007, S. 3). Solche Verfahren werden relativ häufig in der Literatur vorgestellt (Strapparava & Mihalcea, 2007, 2008; Calvo & Mac Kim, 2013).[20] Um die Emotion eines Dokuments zu bestimmen, wird bei diesem Ansatz häufig auf die *Pseudo-Dokumenten-Methode*

[20] Dies geht vermutlich auf den Einfluss der Emotionserkennungsaufgabe des SemEval-2007 zurück, bei der durch das explizite Nichtbereitstellen von Trainingsdaten unüberwachte Verfahren bevorzugt wurden (Strapparava & Mihalcea, 2007).

zurückgegriffen (Strapparava & Mihalcea, 2008): Die Pseudo-Dokumente bestehen aus der Aneinanderreihung vieler Keywords, die jeweils direkten Bezug auf eine Emotionsklasse haben und so zusammen diese Klasse repräsentieren. Die Dokumente und Pseudo-Dokumente werden als Dokument-Term-Vektoren (s. Abschnitt 4.2) repräsentiert. Die Nähe eines zu bewertenden Dokuments zu den Pseudo-Dokumenten wird mit einem Abstands- bzw. Ähnlichkeitsmaß bestimmt. Häufig wird vorher ein geeignetes Dimensionsreduktionsverfahren eingesetzt. Bei Klassifikationsproblemen wird dem Dokument die Klasse des nächsten Pseudo-Dokuments zugewiesen. Bei Regressionsaufgaben wird die Nähe zu den Pseudo-Dokumenten in Übereinstimmungswerte umgerechnet. Der große Vorteil dieses Verfahrens ist die Unabhängigkeit von Trainingsdaten und Emotionslexika. Dadurch können bestehende Systeme schnell auf andere Domänen oder Emotionsmodelle angepasst werden. Hierbei sind lediglich die Pseudo-Dokumente zu bearbeiten. Der kritische Schritt ist dabei die Auswahl der Keywords aus denen diese erstellt werden (Strapparava & Mihalcea, 2008).

Calvo und Mac Kim (2013) vergleichen die Performanz verschiedener Dimensionsreduktionsverfahren auf vier verschiedenen Testkorpora miteinander und verwenden dabei die oben geschilderte Methode. Daneben berechnen sie mit einem Lexical-Affinity-Ansatz die Emotion der Dokumente im VAD-Raum. Anschließend verwenden sie auch hier die Pseudo-Dokumenten-Methode, um wieder auf eine Abwandlung von Ekmans Basisemotionen abzubilden. Diese Arbeit verwendet daher beide Emotionsmodelle. Die Zielvariable ist aber diskret und geht in ihren Ausprägungen auf Ekman zurück. Die beiden performantesten Verfahren sind in dieser Untersuchung die Dimensionsreduktion durch Nicht-negative Matrixfaktorisierung (NMF) und der Lexical-Affinity-Ansatz. Ein unmittelbarer Vergleich ist jedoch schwierig, da die Performanz sehr stark vom Testkorpus abhängt. Während NMF je nach Korpus zwischen 17% und 73% durchschnittliches F-Maß über alle Emotionen erreicht, schwankt die Performanz des Lexical-Affinity-Ansatzes weniger. Sie liegt zwischen 36% und 42%.

3.5 Wirtschafts- und sozialwissenschaftliche Anwendungen

In diesem Abschnitt werden drei Arbeiten vorgestellt, die Subjectivity Analysis auf Gegenstände der Wirtschafts- und Sozialwissenschaften anwenden. Sie sind für diese Arbeit relevant, weil sie entweder Geschäftsberichte zum Gegenstand haben oder Emotionserkennung interdisziplinär anwenden. Bestehende Arbeiten, die Emotionen, wie Freude, Trauer oder Wut in Geschäfts- oder Nachhaltigkeitsberichten untersuchen sind nicht bekannt. Auch in den Wirtschaftswissenschaften ist dieser Frage bislang nicht nachgegangen worden. Dort beschäftigt sich die Literatur zwar mit Emotionen *in*

Unternehmen, nicht aber mit Emotionen in deren externer Kommunikation (Elfenbein, 2007).

Hajek et al. (2014) wenden Sentiment Analysis auf Geschäftsberichte US-amerikanischer Unternehmen an. Ausgangspunkt für ihre Untersuchung ist die Feststellung, dass zur Vorhersage der weiteren finanziellen Entwicklung eines Unternehmens die quantitativen Inhalte eines Geschäftsberichts nicht ausreichen. Stattdessen wird ein Großteil der Information ausschließlich im Fließtext kodiert. Dieser soll deshalb automatisch ausgewertet werden. Sie betrachten elf sog. Sentiment-Kategorien, die neben *postiv* und *negativ* auch andere Label wie *activity* und *realism*, die sich schlecht theoretisch einordnen lassen, enthalten. Diese werden einem allgemeinen Wörterbuch und einem Finanz-Wörterbuch entnommen. Die Berichte wurden als Dokument-Term-Vektoren repräsentiert (s. Abschnitt 4.2), wobei zur Gewichtung der Relevanz der Terme für ein Dokument das *tf-idf*-Maß (s. Fußnote 24) verwendet wurde. Wörter die derselben Sentiment-Kategorie angehören wurden zusammengefasst und der Mittelwert ihrer *tf-idf*-Gewichte gebildet. Dieser Mittelwert dient als Maß der Übereinstimmung eines Berichts mit einer Kategorie. Die so resultierenden Daten werden benutzt, um Vorhersagemodelle für die finanzielle Entwicklung eines Unternehmens, die alleine auf den quantitativen Daten der Geschäftsberichte beruhen, zu verbessern.

Acerbi et al. (2013) untersuchen das Google Ngram-Korpus[21] nach Häufigkeiten bestimmter Emotionswörter im 20. Jahrhundert. Die Emotionswörter stammen aus WordNet Affect (Strapparava & Valitutti, 2004) und sind in sechs Kategorien, nämlich Ekmans Basisemotionen, gruppiert. Sie stellen in einer Zeitreihenanalyse fest, dass erstens die relative Entwicklung der Kategorien *happy* und *sad* sozio-politische Ereignisse und Zustände plausible abbildet. So werden während des Zweiten Weltkriegs besonders negative und in den 20er und 60er Jahren besonders positive Episoden gemessen. Zweitens zeigt ihre Analyse einen generellen Rückgang in der Verwendung von Emotionswörtern im Laufe des 20. Jahrhunderts. Drittens belegen sie die geläufige Annahme, dass amerikanisches Englisch emotionsreicher ist als britisches. Sie stellen fest, dass diese Entwicklung in den 60er Jahren begann und Teil einer allgemeineren Differenzierung der beiden Formen des Englischen ist.

Bollen et al. (2011) untersuchen, inwiefern subjektive Sprache auf Twitter geeignet ist, Entwicklungen auf dem Aktienmarkt vorherzusagen. Sie verwenden dafür zwei unterschiedliche Software-Werkzeuge. Das eine misst die semantische Orientierung auf einer zweistufigen Skala (*positiv* und *negativ*). Das andere misst die sechs Emotionen *calm*, *alert*, *sure*, *vital*, *kind* und *happy*. Den stärksten Zusammenhang mit dem Aktienkurs erreicht dabei nicht etwa die semantische Orientierung, sondern die

[21] `https://books.google.com/ngrams`

Emotionsklasse *calm*. Mit dieser lassen sich – einfache – Vorhersagemodelle der Kursentwicklung signifikant verbessern. Dieser Befund ist hochrelevant für die vorliegende Arbeit, da er zeigt, dass in interdiziplinären Anwendungen mit Emotionserkennung bessere Ergebnisse erzielt werden können als mit Sentiment Analysis.

4 Methoden

Nachfolgend wird das hier entwickelte Verfahren zur textuellen Emotionserkennung vorgestellt. Dafür werden zunächst unter Rückgriff auf die vorangegangen Abschnitte grundlegende Entscheidungen der Konzeption wie das verwendete Emotionsmodell und der verwendete methodische Ansatz diskutiert. Anschließend wird die ausgearbeitete Formalisierung, die daraus resultierende Architektur sowie deren Implementierung vorgestellt.

4.1 Grundlegende Konzeptionsentscheidungen

Aus dem letzten Abschnitt geht hervor, dass Arbeiten zur Emotionserkennung anhand von drei wesentlichen Eigenschaften charakterisiert werden können: dem zugrundeliegenden Emotionsmodell, der Form der Zielvariable und dem methodischen Ansatz.

In dieser Arbeit wird Emotionserkennung als ein Regressionproblem behandelt. Der Grund hierfür ist die erwartete Eigenschaft der Geschäfts- und Nachhaltigkeitsberichte, sich untereinander emotional sehr ähnlich zu sein,[22] sodass sie bei der Verwendung diskreter Emotionsklassen überwiegend derselben zugeordnet werden würden. Die angestrebte statistische Untersuchung verlangt dagegen nach einer möglichst genauen Messung. Eine Alternative wäre eine Klassifizierung auf Satzebene, über die dann der Durchschnitt berechnet wird. Aber auch hier ist davon auszugehen, dass durch eine grobe Klassifikation der – ohnehin als eher sachlich geltenden – Geschäftsberichte viele Sätze als neutral eingestuft werden würden und so Information verloren ginge.

Tabelle 2 zeigt, dass in der Emotionserkennung bisher in keiner Untersuchung ein dimensionales Emotionsmodell mit stetigen Zielvariablen kombiniert worden ist. Die vorliegende Arbeit versucht diese Forschungslücke zu füllen und verwendet daher Russells und Mehrabians VAD-Modell. Darüber hinaus liegt nach der psychologischen Theorie ein dimensionales Modell bei der Benutzung von stetigen Zielvariablen näher als diskrete Modelle (s. Abschnitt 3.2). Ferner ist vor Kurzem ein sehr umfangreiches Emotionslexikon für das VAD-Modell entwickelt worden (s. Abschnitt 5). Es ist deutlich umfangreicher als die bislang normalerweise hierfür verwendete Ressource und ist soweit bekannt in der Emotionserkennung noch nicht zum Einsatz gekommen. Auch dies macht die Verwendung dieses Modells attraktiv. Ein weiterer Grund hierfür ist, dass es in seiner Adäquatheit nicht von der Anwendungsdomäne abhängig ist, da jeder mögliche emotionale Zustand repräsentiert werden kann. Dies ist aufgrund des

[22] Der Vergleich der Streuung des Unternehmenskorpus mit dem RCV1 bestätigt diese Vermutung im Nachhinein (s. Abschnitte 6 und 7.2).

explorativen Charakters der späteren Korpusuntersuchung zu bevorzugen. Darüber hinaus geht aus Abschnitt 3.5 hervor, dass gerade der in den Basisemotionen nicht abgedeckte Bereich hoher Valenz und niedriger Erregung für Anwendungen mit wirtschaftswissenschaftlichem Gegenstand hoch relevant sein kann.

Die Entscheidung für den methodischen Ansatz in dieser Arbeit beruht in erster Linie auf einer Abwägung zwischen der erwarteten Performanz und der zeitlichen Realisierbarkeit. Da keine nach dem VAD-Modell annotierten Korpora verfügbar sind, scheiden überwachte Lernverfahren aus. Linguistisch-regelbasierte und wissensbasierte Methoden sind aus zeitlichen Gesichtspunkten ebenfalls nicht anwendbar. Daher wurde der Lexical-Affinity-Ansatz ausgewählt, da er eine gute Performanz bei entsprechend hochwertigen und umfangreichen Lexika aufweist, andererseits aber einfach implementiert werden kann. Die Entwicklung des neuen Emotionslexikons bietet hierfür besonders gute Voraussetzungen.

4.2 Formalisierung

Ein *Alphabet* Σ ist definiert als eine endliche, nicht-leere Menge von Symbolen. Die kleenesche Hülle Σ^* des Alphabets Σ ist geben durch

$$\Sigma^* = \bigcup_{i \in \mathbb{N}_0} \Sigma^i \,. \tag{1}$$

Σ^* heißt *Menge aller Wörter über dem Alphabet* Σ.
Das Vokabular V ist eine beliebige endliche, nicht-leere Teilmenge von Σ^*.

$$V := \{\omega_1, \omega_2, \omega_3, \ldots \omega_n\} \subset \Sigma^* \tag{2}$$

Dann heißt jeder Vektor D_j mit

$$D_j = (c_{\omega_1}, c_{\omega_2}, c_{\omega_3}, \ldots c_{\omega_n}) \tag{3}$$

Dokument bezüglich des Vokabulars V. Dokumente werden daher, entsprechend des in der Information Retrieval (IR) gebräuchlichen Vektorraum-Modells, als Dokument-Term-Vektoren repräsentiert (Manning et al., 2008). Der Wert der Komponente c_{ω_i} entspricht dabei der Häufigkeit des Wortes ω_i im unprozessierten Volltext-Dokument.[23] Beim Überführen eines Volltext-Dokuments in die formale Repräsentation als Dokument-Term-Vektor geht die Information über die Reihenfolge der Wörter verloren. Insofern spricht man in der Literatur von einem *Bag-of-Words*-Modell (BOW) (Manning et al., 2008, S. 117). Dies entspricht genau der in Abschnitt

[23] Auf der Implementierungsebene gilt dies natürlich erst nach dem Lemmatisieren des Volltexts.

3.4.1 skizzierten Charakteristik des lexikalischen Ansatzes. Am verbreitesten in der IR ist die Gewichtung der Wörter durch das *tf-idf*-Maß[24] anstatt durch die absolute Häufigkeit. In der vorliegenden Arbeit wird jedoch von dessen Verwendung Abstand genommen, da frühere Arbeiten keine Vorteile daraus Nachweisen konnten. Im Gegenteil zeigen Staiano und Guerini (2014), dass die Verwendung von absoluten oder relativen Häufigkeiten für fast alle der von ihnen erfassten fünf Emotionskategorien bessere Ergebnisse liefert. Die Verwendung von *tf-idf*-Gewichtung ist auch aus theoretischen Gründen fragwürdig. Anders als bei der Relevanz von Suchanfragen (*queries*), erscheint es unintuitiv anzunehmen, dass ein Wort geringeren Einfluss auf den Emotionswert eines Dokuments hat, wenn es auch in vielen anderen Dokumenten der Kollektion vorkommt. Zum Beispiel wird das Wort „furchtbar“ in einer Filmbewertung nicht weniger emotional relevant dadurch, dass auch viele andere Filmbewertungen dieses Wort enthalten.[25]

Eine Emotion e ist definiert als ein Vektor

$$e := (v, a, d) \in \mathbb{R}^3. \tag{4}$$

Seine drei Komponenten v, a, d heißen *Valenz*, *Erregung* und *Dominanz*. Obwohl das VAD-Modell vorsieht, dass die Werte der Emotionskomponenten aus einem bestimmten Intervall stammen, wurde darauf verzichtet, dies in die Definition einer Emotion zu integrieren. Diese Einschränkung ist hier nicht nötigt, da in einem konkreten Anwendungsfall das verwendete Emotionslexikon (s.u.) durch seine Einträge implizit Intervallgrenzen festlegt.

Ein *Emotionslexikon* L ordnet Wörtern aus Σ^* Emotionen zu. Es ist definiert als ein Tupel

$$L := (W, E) \quad \text{mit} \tag{5}$$

- $W = \{l_1, l_2, l_3, \ldots l_m\} \subset \Sigma^*$,
- $E = \{e_{l_1}, e_{l_2}, e_{l_3}, \ldots e_{l_m}\}$,

[24] Das *tf-idf*-Maß gewichtet die Relevanz von Termen für ein bestimmtes Dokument durch zwei Faktoren: zum einen durch die Vorkommenshäufigkeit des Wortes im Dokument tf (*term frequency*), zum anderen durch die inverse Dokumentenhäufigkeit *idf* (*inverse document freqency*), die höher ist, in je weniger Dokumenten der Kollektion das entsprechende Wort vorkommt. Somit ist die Relevanz eines Wortes für ein Dokument hoch, wenn es in diesem Dokument häufig vorkommt und in allen übrigen Dokumenten der Kollektion selten.

[25] Vor diesem Hintergrund ist es überraschend, dass das *tf-idf*-Maß trotzdem häufig in der Literatur eingesetzt wird (Strapparava & Mihalcea, 2008; Danisman & Alpkocak, 2008; Calvo & Mac Kim, 2013; Staiano & Guerini, 2014).

wobei e_{l_i} die Emotion ist, die dem Wort l_i zugeordnet ist.

Sei V ein Vokabular und L ein Emotionslexikon, dann ist die $(n, 3)$-Matrix A_{VL} die *Vokabular-Emotions-Matrix bezüglich V und L*. Sie enthält für jedes der n Wörter des Vokabulars die Emotion, die das Lexikon ihm zuordnet, oder die Emotion $(0, 0, 0)$, falls das Lexikon dem Wort keine Emotion zuordnet. Formal ist sie zeilenweise definiert als

$$\begin{pmatrix} a_{i1} & a_{i2} & a_{i3} \end{pmatrix} := \begin{cases} (e_{\omega_i})^T , & \text{wenn } \omega_i \in W \\ (0, 0, 0) , & \text{sonst.} \end{cases} \tag{6}$$

Diese Matrix erlaubt später die effiziente Berechnung von Dokumentenemotionen, da die Emotion der i-ten Komponente des Dokuments in derselben Zeile der Vokabular-Emotions-Matrix festgehalten ist. Dadurch muss jedes Wort des Vokabular nur einmal im Lexikon „nachgeschlagen“ werden (*dictionary look-up*) unabhängig davon, wie viele Dokumente in der Kollektion sind.

P_{VL} ist die *Lexikonprojektion bezüglich eines Vokabulars V und eines Lexikons L*. Sie setzt jede Komponente eines Dokuments D_j auf null, wenn dem Wort, dessen Häufigkeit durch sie repräsentiert wird, keine Emotion durch das Emotionslexikon zugeordnet wird. Alle anderen Komponenten werden unverändert gelassen. Formal ist sie komponentenweise definiert als

$$P_{VL}(D_j)_i = {c_{\omega_i}}' := \begin{cases} c_{\omega_i}, & \text{wenn } \omega_i \in W \\ 0, & \text{sonst.} \end{cases} \tag{7}$$

Damit entspricht die Summe über alle Komponenten von $P_{VL}(D_j)$ genau der Anzahl der Wörter im unprozessierten Volltext-Dokument, denen das Lexikon eine Emotion zuordnen kann. Diese wird nachfolgend *Anzahl der emotionsrelevanten Wörter* genannt. Sie wird im nächsten Schritt als Normalisierungsparameter verwendet.

Die *Dokumentenemotion e_{D_jVL} eines Dokuments D_j bezüglich eines Vokabulars V und eines Lexikons L* ist die Emotion, die dem Dokument D_j zugeordnet ist. Sie wird modelliert als die Summe aller Wortemotionen der emotionsrelevanten Wörter im Dokument geteilt durch einen Normalisierungsparameter. Der Normalisierungsparameter bewirkt zum einen, dass die Dokumentenemotion unabhängig von der Länge

des Dokuments ist. Zum anderen liegen die Werte von v, a, d so wieder in dem vom Lexikon vorgesehenen Bereich. Formal ist sie definiert als

$$e_{D_j VL} = \begin{cases} \frac{\sum_{i=1}^{n} {c_{\omega_i}}' \cdot (a_{i1}, a_{i2}, a_{i3})}{\sum_{i=1}^{n} {c_{\omega_i}}'}, & \text{wenn } \sum_{i=1}^{n} {c_{\omega_i}}' > 0 \\ (0,0,0), & \text{sonst} \end{cases} \quad \text{mit } {c_{\omega_i}}' = P_{VL}(D_j)_i \ . \tag{8}$$

Die Fallunterscheidung bewirkt lediglich, dass die Dokumentenemotion $e_{D_j VL}$ auch dann noch definiert ist, wenn im ersten Fall der Nenner null werden würde, also wenn das Dokument keine emotionsrelevanten Wörter enthält. Der erste Fall wird folgendermaßen erläutert: Tritt ein Wort ω_i im Dokument D_j nicht auf, ist c'_{ω_i} null. Ist das Wort ω_i nicht im Lexikon enthalten, ist sowohl c'_{ω_i} als auch a_{i1}, a_{i2} und a_{i3} null. Die Lexikonprojektion wird also im Zähler nicht zwingend benötigt – ebenso gut hätte man dort c_{ω_i}, also die Komponente des Urbilds, schreiben können. Im Nenner benötigt man die Lexikonprojektion jedoch um den Normalisierungsparameter zu bilden.

Die Konstruktion der Dokumentenemotion entspricht in der obigen Formalisierung der Summe aller Wortemotionen der emotionsrelevanten Wörter geteilt durch deren Häufigkeit. Dies entspricht genau dem komponentenweisen arithmetischen Mittel, wobei Wörter, die nicht im Lexikon verzeichnet sind, ignoriert werden. Diese Konzeption ist sowohl in der psychologischen Literatur als auch in der Literatur zur Emotionserkennung bekannt (Calvo & Mac Kim, 2013; Katz et al., 2007; Agrawal & An, 2012; Leveau et al., 2012). Daneben gibt es alternative Normalisierungsparameter. Naheliegend ist insbesondere die Anzahl aller Wörter im Dokument – unabhängig davon, ob diese im Lexikon verzeichnet sind oder nicht. Die Verwendung dieses Normalisierungsparameters würde bedeuten, dass Wörter, die nicht im Lexikon sind, implizit mit der neutralen Emotion $(0, 0, 0)$ bewertet werden. Dagegen spricht, dass empirisch betrachtet, die Emotion eines durchschnittlichen Wortes *nicht* der neutralen Emotion zu entsprechen scheint (s. Abschnitt 5). Daher würde dieses Vorgehen höchstwahrscheinlich zu verzerrten Ergebnissen führen.

4.3 Architektur

Dieser Abschnitt erläutert die Architektur die dem entwickelten Software-Werkzeug zugrunde liegt. Sie ist schematisch in Abbildung 4 dargestellt. Ausgangspunkt ist eine Kollektion von Eingabedokumenten. Diese werden zunächst einer Vorverarbeitung unterzogen. Die Vorverarbeitung besteht aus mehreren Teilschritten, wobei die wichtigsten eine lexikalische Normalisierung durch einen Stemmer oder Lemmatizer sowie das Entfernen von Stopwörter (*stopword removal*) sind. Grundsätzlich lassen

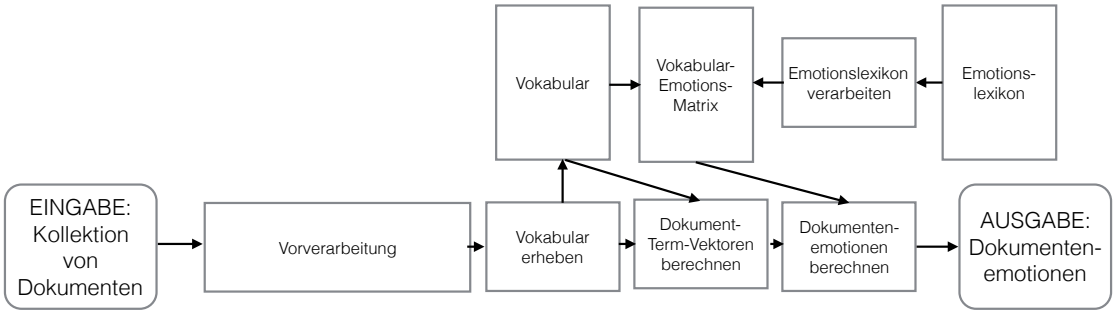

Abbildung 4: Architektur

sich sowohl Stemmer als auch Lemmatizer in dieser Architektur verwenden. Wichtig ist jedoch, dass dieselbe Verarbeitung auch auf das Emotionslexikon angewandt wird, da es sonst zu Fehlern während des *dictionary look-ups* kommen kann. Auf Grundlage eines Experiments fiel die Entscheidung auf einen Lemmatizer. Eine detaillierte Beschreibung des Problems, des Experiments sowie daran anknüpfender Analysen liefert Abschnitt 5.3. Da die Details der Vorverarbeitung vom ausgewählten Lemmatizer abhängen, werden diese im Abschnitt 4.4 weiter besprochen.

Das Ergebnis des ersten Verarbeitungsschritts ist eine Kollektion von Dokumenten, die nur noch lexikalisch normalisierte und im Sinne eines BOW-Modells relevante Token enthalten. Darauf aufbauend wird im nächsten Schritt das Vokabular der Kollektion erhoben, also die Menge jedes in der Kollektion mindestens einmal vorkommenden Wortes. Anschließend wird aus dem Vokabular und dem Emotionslexikon die im Abschnitt 4.2 definierte Vokabular-Emotions-Matrix berechnet, die jedem Wort des Vokabulars eine Emotion, also einen Vektor aus $\mathbb{R}^3$, zuweist. Das verwendete Emotionslexikon bestimmt das Intervall in dem die Komponenten der Emotionsvektoren liegen. Parallel dazu werden die normalisierten Dokumente der Kollektion in ihre formale Repräsentation als Dokument-Term-Vektoren überführt.[26] Im nun folgenden Schritt wird wie in Abschnitt 4.2 beschrieben, zunächst die Lexikon-Projektion auf die Dokument-Term-Vektoren angewendet. Anschließend wird aus dem Bild der Dokument-Term-Vektoren und der Vokabular-Emotions-Matrix die Dokumentenemotion nach der beschriebenen Formel berechnet. Diese wird im letzten Schritt – zusammen mit einigen weiteren während der Verarbeitung erhobenen Daten – ausgegeben.

[26] In der konkreten Implementierung wird dieser und alle folgenden Schritte zunächst für *ein* Dokument durchgeführt, bevor das nächste Dokument in einen Dokument-Term-Vektor überführt wird. Das liegt daran, dass bei einem großen Vokabular und und sehr vielen Dokumenten – wie zum Beispiel im RCV1 (s. Abschnitt 6) – verhältnismäßig viel Speicherplatz für die Dokument-Term-Vektoren benötigt wird.

4.4 Implementierungsdetails und Verfügbarkeit

Die vorangehend beschriebene Architektur wurde im Kommandozeilen-Werkzeug JenEmo (Jena Emotion Analyzer) in Java implementiert. Es ist *open-source* gestellt und steht unter einer MIT-Lizenz.[27] Es verfügt über eine einfache Benutzerschnittstelle, bei der es ausreicht, einen Verzeichnispfad anzugeben, um alle Textdateien in diesem Verzeichnis zu verarbeiten.[28]

Das Lemmatisieren ist der erste Schritt der Vorverarbeitung. Hierfür wird ein Lemmatizer, der sich aus verschiedenen Annotatoren des Stanford CoreNLP zusammensetzt, verwendet. Er wird im Folgenden vereinfacht „Stanford Lemmatizer" genannt. Das Stanford CoreNLP (Manning et al., 2014) ist ein Framework für Annotions-Pipelines von NLP-Anwendungen, das heißt ein Software-Gerüst, das es erlaubt, verschiedene Annotatoren wie Tokenizer, Wortarten-Tagger und Syntax-Parser modular miteinander zu verbinden. Es unterscheidet sich von vergleichbaren Frameworks, wie etwa UIMA, durch seine einfache Programmierschnittstelle. Es ist *open-source* gestellt[29] und in Java implementiert. Allerdings existieren auch Wrapper für viele weitere Programmiersprachen. Es verfügt über Annotatoren und Modelle für sechs natürliche Sprachen, wobei die volle Funktionalität nur für Englisch zur Verfügung steht. Der Workflow der ersten Annotatoren folgt dabei einer weitgehend festgelegten Reihenfolge: Tokenzier, Sentence-Splitter, Wortarten-Tagger und Lemmatizer. Danach stehen weitere zur Verfügung. Für die Verwendung im Werkzeug wurde eine Wrapper-Klasse geschrieben, die diese ersten vier Annotatoren zusammenfasst und so gemeinsam als Lemmatizer fungiert. Daraus resultiert auch die Reihenfolge der Vorverarbeitungsschritte in der Implementierung: Der Lemmatizer muss hier vor der Stoppwort-Entfernung eingegesetzt werden. Andernfalls ist damit zu rechnen, dass die Performanz des Wortarten-Taggers des Stanford CoreNLP darunter leiden würde, dass schon Wörter entfernt worden sind.

Nach dem Lemmatisieren liegt das Dokument als verkettete Liste von Wörtern vor. Aus dieser werden alle Zahlenausdrücke durch reguläre Ausdrücke entfernt. Als solche werden Kardinalzahlen, Dezimalbrüche, Datums-, Währungs- und Prozentangaben erkannt. Darauf folgt die Entfernung aller *Nicht-Wörter*. Darunter zählen hier Interpunktion und Anomalien wie Zahlen-Buchstaben-Kombinationen. Als *echtes Wort* gilt hingegen jeder Token, der mit einem Buchstaben anfängt. Anschließend

[27] `https://github.com/buechel/JenEmo` – Zusätzlich ist es als JAR-Datei auf der beigelegten DVD enthalten. Die dadurch erhobenen Daten des RCV1 und des Unternehmenskorpus sowie die dazugehörigen Metadaten sind ebenfalls beigefügt.

[28] Eine einfache grafische Oberfläche wäre jedoch eine wünschenswerte Erweiterung, um insbesondere für Anwender aus den Digital Humanities mit begrenzten technischen Kenntnissen attraktiver zu werden.

[29] `https://github.com/stanfordnlp/CoreNLP`

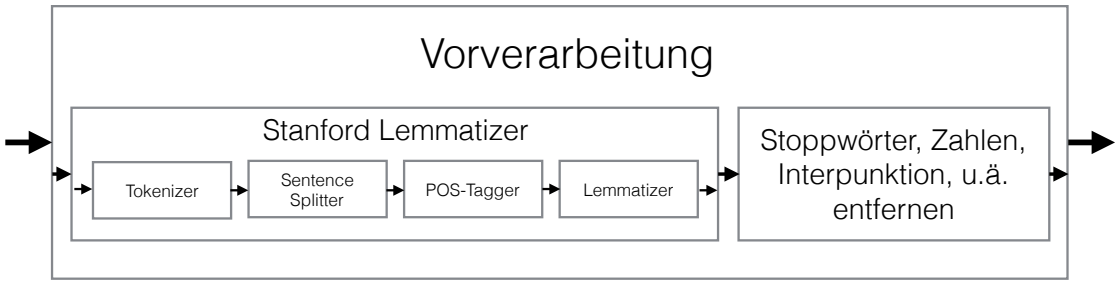

Abbildung 5: Komponenten und Schritte der Vorverarbeitung in der konkreten Implementierung.

werden die Stoppwörter entfernt. Die Entscheidung für eine bestimmte Stoppwortliste und deren Modifikation wird in Abschnitt 5.4 beschrieben. Die Komponenten und Schritte der Vorverarbeitung sind in Abbildung 5 schematisch dargestellt.

Die Ausgabe erfolgt im CSV-Format. Sie beinhaltet den Dateinamen als Identifikator des Dokuments und die Werte für Valenz, Erregung und Dominanz. Neben diesen obligatorischen Daten werden noch weitere Angaben gemacht, die einen zusätzlichen Einblick in die Beschaffenheit des Dokuments und die Verarbeitung erlauben. Dabei handelt es sich zum einen um die Standardabweichung von Valenz, Erregung und Dominanz innerhalb eines Dokuments. Dadurch kann beispielsweise unterschieden werden, ob ein Dokument einen neutralen Valenzwert hat, weil es durchgängig neutrale Wörter enthält, oder ob sehr negative und sehr positive Wörter zusammengenommen eine neutrale Emotion ergeben. Zum anderen werden Angaben darüber gemacht, wie viele Token jeweils nach den Schritten der Vorverarbeitung erhalten bleiben. Dabei handelt es sich erstens um die Anzahl aller lemmatisierter Token im Dokument, zweitens die der *echten Wörter*, drittens die der Nicht-Stoppwörter sowie viertens die der emotionsrelevanten Wörter. Darüber hinaus wird die Anzahl der Zahlenausdrücke mit angegeben, um zum Beispiel Fragestellungen nach dem Zusammenhang der Häufigkeit von Zahlen und der Emotion eines Dokuments behandeln zu können. Das Verhältnis der emotionsrelevanten Wörter zu den Nicht-Stoppwörtern eignet sich als Maß für die Erkennungsquote, da es sich bei Stoppwörtern i.d.R. um Funktionswörter handelt. Von diesen kann man annehmen, dass sie keinen oder einen vernachlässigbaren emotionalen Gehalt haben.[30] Das Verhältnis der emotionsrelevanten Wörter zu den Nicht-Stoppwörtern entspricht damit dem Anteil der Wörter, der das Werkzeug eine Emotion zuordnen konnte, an allen Wörtern, die einen Emotionsgehalt haben.

[30]Pennebaker et al. (2003) weisen jedoch darauf hin, dass unter Umständen die Verwendung von Pronomen Hinweise auf den emotionalen Zustand eines Sprechers bzw. Schreibers geben kann. Dies betrifft zum Beispiel das Verhältnis der Häufigkeiten von „ich“ und „wir“.

Neben dieser Ausgabe schreibt das Programm während der Laufzeit eine Reihe von Plain-Text-Dateien, auf die der Nutzer zugreifen kann. Dabei handelt es sich zum einen um eine Darstellung der normalisierten Dokumente, wobei jedes normalisierte Dokument einer Datei entspricht und in jeder Zeile der Dateien jeweils ein normalisiertes Wort steht. Die Reihenfolge der Wörter im Volltext-Dokument bleibt dabei erhalten. Zum anderen wird das Vokabular in eine Datei geschrieben, in der jede Zeile ein Wort des Vokabulars enthält. Ursprünglich sollten auch die Dokument-Term-Vektoren gespeichert werden. Da dies in einigen Fällen jedoch mit großem Speicherverbrauch verbunden sein kann (s. Fußnote 26), wurde diese Funktion bis auf Weiteres wieder entfernt. Bei zukünftigen Erweiterungen ist geplant, die Dokument-Term-Vektoren auf Wunsch dennoch speichern zu lassen. Darüber hinaus soll die Möglichkeit eingebaut werden statt Volltext-Dokumente direkt normalisierte Dokumente oder ein Vokabular und die dazugehörigen Dokumente-Term-Vektoren als Eingabe zu verwenden. Auf diese Weise kann bei wiederholten Durchläufen Rechenzeit gespart werden. Außerdem können durch die Dokument-Term-Vektoren mitunter auch solche Dokumente analysiert werden, für die – etwa aus rechtlichen Gründen – keine Volltexte zur Verfügung stehen.

Ressource	Einträge	Skala	Erhebung	Verfügbarkeit
M. Bradley und Lang (1999) (ANEW)	1034	9-Pkt	klassisch	beschränkt
Bestgen und Vincze (2012)	17350	9-Pkt	Bootstrapping	freier Download
Warriner et al. (2013)	13915	9-Pkt	Crowdsourcing	freier Download

Tabelle 3: Verfügbare Emotionslexika, die das VAD-Modell verwenden.

5 Ressourcen

Im Folgenden werden knapp die verfügbaren Ressourcen vorgestellt, die man als Emotionslexikon in der oben geschilderten Architektur verwenden kann. Allgemein ist die Anzahl der Ressourcen für die Emotionserkennung im Vergleich zu denen für Sentiment Analysis gering (Staiano & Guerini, 2014). Darüber hinaus verwenden sie unterschiedliche Emotionsmodelle, sodass deren Verwendbarkeit eingeschränkt ist. Für diskrete Modelle sind relativ viele geeignete Emotionslexika vorhanden. Nennenswert sind unter anderem WordNet-Affect (Strapparava & Valitutti, 2004) EmoLex (Mohammad & Turney, 2013) und DepecheMood (Staiano & Guerini, 2014).

Die Auswahl an Ressourcen für dimensionale Emotionsmodelle bzw. das VAD-Modells ist dagegen geringer. Soweit bekannt wurden in der Computerlinguistik und in verwandten Disziplinen bislang keine solchen Ressourcen angelegt. Allerdings gibt es drei Emotionslexika, die für die Psychologie entwickelt wurden. Sie werden dort in der Forschung zur Emotionsverarbeitung, zum Erinnerungsvermögen und insbesondere in der Psycholinguistik benötigt (Warriner et al., 2013). Tabelle 3 gibt einen Überblick über ihre wesentlichen Merkmale. Die Einträge von Emotionslexika werden in der psychologischen Literatur auch „(affektive) Normen“ genannt.

5.1 Überblick über mögliche Emotionslexika

Jahrelang wurde von fast allen Studien in diesem Gebiet ANEW (Affective Norms for English Words) (M. Bradley & Lang, 1999) oder eine übersetzte Versionen davon verwendet (Warriner et al., 2013). Die 1034 Einträge dieses Lexikons wurden in einer klassischen Probandenbefragung erhoben. Den Probanden wurden Wörter als Stimuli präsentiert. Daraufhin gaben sie die durch dieses Wort bei Ihnen ausgelöste Emotion in den Dimensionen Valenz, Erregung und Dominanz auf einer jeweils neun-stufigen Skala an. Hierfür wurde das Self-Assessment-Manikin (SAM) (M. M. Bradley & Lang, 1994) verwendet (s. Abbildung 14 im Anhang). Dabei handelt es sich um eine Gruppe von Piktogrammen, durch die emotionale Zustände ohne den Einsatz von Sprache angegeben werden können.

Das arithmetische Mittel aller Bewertungen zu einem Wort wurde als Emotionswert festgehalten. Die Studie wurde in den USA durchgeführt, sodass davon

auszugehen ist, dass in erster Linie Muttersprachler des amerikanischen Englischen teilgenommen haben. Die Weitergabe der Daten erfolgt kostenfrei, jedoch nur unter strengen Auflagen und ausschließlich an bestimmte akademische Einrichtungen.[31]

Bestgen und Vincze (2012) haben die ca. tausend Einträge von ANEW durch Bootstrapping[32] auf den siebzehnfachen Umfang erweitert. Hierzu wurde Latente Semantische Analyse (LSA)[33] eingesetzt. Einem Wort wurde ein Emotionswert auf der Basis seiner Nachbarn im semantischen Raum zugewiesen. Der zugewiesene Wert ist der Mittelwert der Nachbarn, deren Emotion aus ANEW bekannt ist. Die Emotionswerte von Wörtern, die schon in ANEW vorhanden sind, wurden erneut erhoben. Dadurch konnte die Korrelation zwischen Teilen des neuen Lexikons und der vorhandenen Ressource berechnet werden. Diese Korrelation dient als Gütemaß der Bootstrapping-Methode. Sie liegt je nach Emotionsdimension zwischen $r = 0.56$ für Erregung und $r = 0.71$ für Valenz. Das so entstandene Emotionslexikon ist frei verfügbar.[34]

Warriner et al. (2013) haben kürzlich eine weitere Ressource erstellt, die die Autoren als Erweiterung von ANEW verstehen. Die Emotionen der Einträge von ANEW wurden erneut erhoben. Zudem wurde ein Vielfaches an neuen Einträgen aufgenommen, sodass das Lexikon insgesamt 13915 umfasst. Möglich wurde diese sehr umfangreiche Erhebung durch moderne Crowdsourcing-Techniken über die Plattform Amazon Mechanical Turk (AMT)[35]. Es gibt zwei Wege diese Plattform zu nutzen: Als „Requester“ stellt man Aufgaben online, die derzeit nur durch menschliche Intelligenz zu lösen sind. Als „Worker“ wählt man unter den so gestellten Aufgaben aus, erledigt diese und erhält dafür jeweils einen geringen Geldbetrag. Snow et al. (2008) weisen nach, dass die Ergebnisse von Annotationsaufgaben, die über AMT erfüllt wurden, qualitativ gut sind. Dies gilt insbesondere für Emotionsannotationen. Bei diesen kann die Inter-Annotator Agreement der Worker sogar über der von Experten liegen. Aufgrund ihres Anspruchs replizieren Warriner et al. (2013) im Wesentlichen die Methodologie von M. Bradley und Lang (1999). Allerdings verwenden sie das SAM nicht, sondern die Ratings werden jeweils als Zahlenwerte zwischen eins und neun angegeben. Als Rater wurde nur zugelassen, wer einen Wohnort in den USA angab.

[31] `http://csea.phhp.ufl.edu/media/anewmessage.html`

[32] Als „Bootstrapping“ werden verschiedene Verfahren bezeichnet, bei denen eine Ressource automatische „aus sich selbst heraus“ erweitert wird. In der Computerlinguistik kommen solche Methoden etwa zum Einsatz, wenn nur ein Teil eines Korpus annotiert ist. Auf Basis dieser Annotationen können dann durch statistische Lernalgorithmen weitere Annotationen erzeugt werden (Jurafsky & Martin, 2008, S. 684).

[33] Bei LSA wird die Vektorraum-Repräsentation einer Kollektion durch Dimensionsreduktion in einen *semantischen Raum*, der üblicherweise nur noch wenige hundert Dimensionen hat, überführt, sodass semantisch ähnliche Dimensionen zusammengefasst werden (Manning et al., 2008, S. 412ff.). Wörter sind semantisch ähnlich, wenn sie in ähnlichen Wortkontexten verwendet werden.

[34] `http://www.psor.ucl.ac.be/personal/yb/Doc/BestgenVincze.zip`

[35] `https://www.mturk.com/mturk/welcome`

	M	SD_{Rater}	$SD_{Eintrag}$	Min	Max
Valenz	5.06	1.68	1.27	1.26	8.53
Erregung	4.21	2.30	0.90	1.60	7.79
Dominanz	5.18	2.16	0.94	1.68	7.90

Tabelle 4: Lage- und Streuungsmaße für Warriners Emotionslexikon.

Daher ist auch bei dieser Ressource davon auszugehen, dass die Bewertungen in erster Linie von Muttersprachlern des amerikanischen Englischen stammen. Als Normen wurden wiederum die Mittelwerte aller Bewertungen jeweils eines Wortes festgehalten. Die Daten wurden auf Übereinstimmung mit in anderen Studien erhobenen Normen untersucht. Es zeigt sich, dass die Werte relativ konsistent zwischen verschiedenen Studien und Sprachen sind. Dies gilt jedoch nicht gleichermaßen für alle Dimensionen. Der Vergleich mit ANEW zeigt eine Korrelation von $r = 0.95$ für Valenz, $r = 0.76$ für Erregung und $r = 0.80$ für Dominanz. Die Ressource lässt sich frei herunterladen und steht unter einer Creative-Commons-Lizens.[36]

Da die Qualität und die Abdeckung des Lexikons entscheidend für die Performanz von Lexical-Affinity-Ansätzen sind, kommen die Ressourcen von Bestgen und Vincze (2012) und Warriner et al. (2013) zur weiteren Verwendung in dieser Arbeit infrage. Die Entscheidung fällt zugunsten letzterer aus. Zwar hat sie etwa 20% weniger Einträge, jedoch ist ihre Übereinstimmung mit ANEW, das in der Psychologie jahrelang Standard war, größer. Außerdem geht sie auf menschliche Annotatoren zurück.

5.2 Warriners Emotionslexikon im Detail

Dieser Abschnitt geht auf einige statistische Eigenschaften von Warriners Emotionslexikon ein, da sie für die spätere Interpretation der Daten aus dem Unternehmenskorpus wichtig sind. Tabelle 4 gibt jeweils für die drei Dimensionen Valenz, Erregung und Dominanz folgende Maßzahlen an: das arithmetische Mittel, die durchschnittliche Standardabweichung der Beurteilungen eines Eintrags durch mehrere Rater (SD_{Rater}), die Standardabweichung der Einträge[37] ($SD_{Eintrag}$) sowie Minimum und Maximum der Einträge. Die Mittelwerte von Valenz und Dominanz liegen geringfügig über der neutralen Wertung fünf. Der Mittelwert von Erregung liegt hingegen annäherend einen ganzen Wertungspunkt niedriger. Dies ist insofern interessant, als dass $(5, 5, 5)$

[36] `http://crr.ugent.be/archives/1003`

[37] Warriner et al. (2013) nennen die Einträge „Lemmata“. Dies entspricht jedoch nicht der in der Computerlinguistik üblichen Verwendung dieses Begriffs (Jurafsky & Martin, 2008, S. 120), da zum Teil mehrere Flexionsformen eines Lexems enthalten sind (z.B. *peanut* und *peanuts*). Außerdem ist die Wortart nicht mit verzeichnet und verschiedene Wortbedeutungen werden nicht unterschieden.

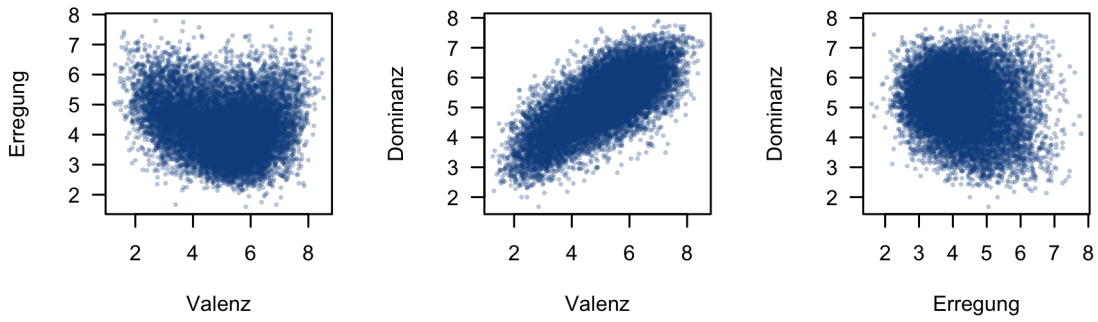

Abbildung 6: Streudiagramme der Emotionskomponenten in Warriners Emotionslexikon.

in diesem Fall zwar die neutrale Emotion ist, die Emotion eines durchschnittlichen Worts jedoch hiervon abweicht.

Valenz und Dominanz weisen eine linksschiefe Verteilung auf. Erregung ist hingegen rechtsschief. Das heißt, das die Mehrzahl der Wörter überdurchschnittliche Valenz- und Dominanz-, jedoch unterdurchschnittliche Erregungswerte aufweisen. Anders ausgedrückt sind viele Wörter angenehm und führen ein Gefühl von Kontrolle herbei, die meisten sind jedoch auch wenig erregend. Den Werten zur durchschnittlichen Standardabweichung zwischen den Ratern (SD_{Rater}) kann man entnehmen, dass die Streuung in den Valenzbewertungen wesentlich kleiner ist, als in den anderen Dimensionen. Interessanterweise deckt sich dies mit der Beobachtung, dass die Korrelation mit den Werten von ANEW in der Valenzdimension deutlich stärker ist als in den anderen Dimensionen. Ebenso verhält es mit den durch Bootstrapping erhobenen Normen von Bestgen und Vincze (2012).[38]

Darüber hinaus weisen die Autoren darauf hin, dass die Standardabweichung zwischen den Ratern geringer ist, je extremer die Werte des Wortes sind. Dies gilt für alle Dimensionen. Der Emotionswert neutraler Wörter scheint daher schwieriger bestimmbar zu sein, als der von stark emotionalen Wörtern. Die drei letzten Spalten zeigen, dass die Spannweite und die Streuung bei Valenz größer ist als in den anderen Dimensionen. Die Rater sind dort vermehrt und stärker von der neutralen Bewertung abgewichen bzw. haben die Skala im größeren Umfang ausgeschöpft.

Abbildung 6 zeigt alle Einträge des Lexikons in drei Streudiagrammen dargestellt, die jeweils zwei Dimensionen gegeneinander abtragen. Dies hat sich als eine praktische Form der Darstellung herausgestellt und wird in der vorliegenden Arbeit häufig benutzt. Tabelle 14 im Anhang zeigt die dazugehörige Korrelationsmatrix von Valenz,

[38]Dazu passt, dass laut Katz et al. (2007) die IAA bei der semantischen Orientierung höher ist als bei allen von Ekmans Basisemotionen.

Erregung und Dominanz. Beide Darstellungsformen zeigen eine starke Korrelation[39] zwischen Valenz und Dominanz. Dieses Phänomen ist schon in vielen vorherigen Studien beobachtet worden und ist ein wichtiges Argument gegen die Annahme der Orthogonalität der drei Komponenten (Warriner et al., 2013). Es alleine würde zudem gegen den Einbezug der Dominanzkomponente in der ED sprechen. Im weiteren Verlauf dieser Arbeit werden jedoch auch gegenteilige Befunde vorgestellt (s. Abschnitt 7). Das erste Diagramm in Abbildung 6 zeigt darüber hinaus einen deutlich ausgeprägten nicht-linearen Zusammenhang zwischen Valenz und Erregung. Dieser deutet darauf hin, dass Wörter dann besonders erregend sind, wenn sie stark positiv oder stark negativ sind. Wörter mittlerer Valenz sind dagegen eher „langweilig“.

5.3 Verarbeitung des Emotionslexikons

Zunächst werden die Werte von Valenz, Erregung und Dominanz im Lexikon dauerhaft transformiert, sodass alle im Intervall $[-4, 4]$ liegen anstatt im Intervall $[1, 9]$, wie es durch das Erhebungsverfahren zunächst der Fall war. Dadurch entspricht die neutrale Emotion, die neutrale Werte in allen Komponenten aufweist, gerade dem Nullvektor, also dem neutralen Element der Vektoraddition im Vektorraum der Emotionen. Gleichzeitig bleibt aber die Anzahl der Abstufungen erhalten, was die spätere Interpretation von Dokumentenemotionen erleichtert.[40]

In Abschnitt 4.3 wurde darauf hingewiesen, dass in der vorgestellten Architektur grundsätzlich sowohl die Verwendung eines Stemmers als auch eines Lemmatizers möglich ist. Da die Dokumente lemmatisiert bzw. gestemmt werden, muss die gleiche Verarbeitung auch auf das Lexikon angewandt werden, um sicherzustellen, dass der *dictionary look-up* wie gewünscht funktioniert. Bei der Verwendung eines Stemmers ist das offensichtlich.[41] Außerdem würden sonst Einträge im Lexikon stehen, die niemals einem der Wörter, die in den Dokument-Term-Vektoren repräsentiert sind, zugeordnet werden können. Zum Beispiel ist im Lexikon sowohl *relax* als auch *relaxed* eingetragen. In den normalisierten Dokumenten kann aber nur noch *relax* auftreten. Dies führt nicht nur zu einer unnötigen Verlängerung der Laufzeit des Programms. Darüber hinaus führt es auch dazu, dass jedem Auftreten von *relaxed* im Volltext, fälschlicherweise die Emotion von *relax* zugeordnet wird. Aus diesem Grund wird in solchen Fällen ein neuer, gemeinsamer Emotionswert durch Mittelwertbildung

[39] Zur Interpretation der Pearson-Korrelation richtet sich diese Arbeit nach folgender Konvention: schwache Korrelation ab $r = 0.1$, mittlere Korrelation ab $r = 0.3$, starke Korrelation ab $r = 0.5$ (Rasch, Friese, Hofmann & Naumann, 2010a, S. 90).

[40] Wenn sich zum Beispiel zwei Dokumente in der Valenzdimension um den Wert eins unterscheiden, lässt sich dies als identischer Abstand auf der Neun-Punkt-Skala, mit der das herangezogene Lexikon erhoben wurde, interpretieren.

[41] Ein Stemmer würde zum Beispiel ein im Volltext auftauchendes Wort *married* auf *marri* abbilden, wohingegen im Lexikon sonst nur ein Eintrag zu *marry* steht.

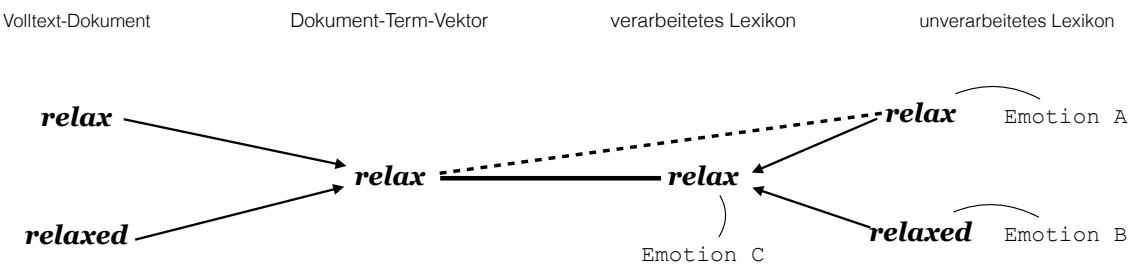

Abbildung 7: Beispielhafte Darstellung der Zusammenhänge der lexikalischen Normalisierung (Pfeile) eines Volltext-Dokuments und des Lexikons. Die Token *relax* und *relaxed* im Volltext-Dokument werden im Dokument-Term-Vektor beide als Auftreten von *relax* gezählt. Würde das Lexikon nicht ebenfalls normalisiert werden, würden fälschlicherweise auch dem Vorkommen von *relaxed* im Volltext die Emotion von *relax* zugeordnet werden (gestrichelte Linie). Aus diesem Grund wird in Fällen, in denen mehrere Einträge des Lexikons durch die Verarbeitung auf eine gemeinsame Form abgebildet werden, eine neue Emotion für den gemeinsamen Eintrag berechnet. Diese resultierende Emotion wird dann allen entsprechenden Einträgen im Dokument-Term-Vektor zugeordnet (breite durchgezogene Linie).

berechnet. Der Zusammenhang ist in Abbildung 7 schematisch dargestellt und gilt sowohl für Stemmer als auch für Lemmatizer. Der geschilderte Vorgang wird mit jedem Aufruf des Programms wiederholt, um die Möglichkeit zu erhalten unterschiedliche Stemmer und Lemmatizer mit dem gleichen Lexikon zu verwenden.

Die Entscheidung für einen Lemmatizer ist auf der Basis folgenden Experiments gefallen. Das Lexikon wurde zum einen durch den Porter-Stemmer (Porter, 1980) zum anderen durch den Stanford Lemmatizer (s. Abschnitt 4.4) verarbeitet. Danach wird jeweils die Anzahl der resultierenden Einträge erhoben. Da die Leistungsfähigkeit des Lexical-Affinity-Ansatzes von der Abdeckung des Lexikons abhängt, mindert der Verlust von Termen die Performanz. Je mehr Einträge nach der Verarbeitung erhalten bleiben, desto eher ist also ein konkreter Stemmer oder Lemmatizer für den Einsatz mit dem hier entwickelten Tool geeignet.[42] Beim Porter-Stemmer handelt es sich um einen Algorithmus, der regelbasiert die Suffixe von Wortformen entfernt. Der Algorithmus arbeitet in sechs aufeinander folgenden Schritten, in denen jeweils eine Gruppe von Wortmanipulationen durchgeführt wird. Es ist einer der am weitesten verbreiteten Stemming-Algorithmen und wird durch seine Einfachheit und Effizienz vor allem bei IR-Anwendungen häufig verwendet (Jurafsky & Martin, 2008, S. 102).

Tabelle 5 zeigt die Ergebnisse des Experiments. Um auszuschließen, dass ein einziger Schritt des Porter-Stemmers für den Großteil der Verluste verantwortlich ist, wurde zusätzlich erhoben, wie viele Einträge verloren gehen, wenn jeweils fünf

[42] Angemessener aber nicht durchführbar wäre die Evaluation des Werkzeugs mit einem Testkorpus (s. Abschnitt 6).

Verarbeitung	n	%
ohne	13915	100
Porter	11891	85
Porter ohne Schritt 1	12520	90
Porter ohne Schritt 2	12078	87
Porter ohne Schritt 3	12205	88
Porter ohne Schritt 4	12073	87
Porter ohne Schritt 5	12832	92
Porter ohne Schritt 6	12223	89
Stanford Lemmatizer	13581	98

Tabelle 5: Absolute und relative Häufigkeit einzigartiger Terme in Warriners Emotionslexikon nach der Lemmatisierung bzw. dem Stemming.

$n_{Zuordnung}$	$n_{Häufigkeit}$	$\%_{\text{Häufigkeit}}$
1	13256	97.6
2	316	2.3
3	9	0.1

Tabelle 6: Häufigkeit der n-zu-1-Zuordnungen von Lexikoneinträgen auf Lemmata durch die Lemmatisierung.

statt sechs Schritte des Porter-Stemmers durchgeführt werden. Wäre bei einem dieser Versuche eine deutlich größere Zahl von Termen erhalten geblieben, wäre dies ein Anhaltspunkt gewesen, dass eine Modifikation des Porter-Stemmers sinnvoll sein kann. Die Ergebnisse zeigen, dass die Verwendung des Porter-Stemmers den Umfang des Emotionslexikons um 2000 Einträge reduziert. Der Stanford Lemmatizer reduziert dessen Größe hingegen um nicht einmal 350. Diese deutliche Diskrepanz verschwindet auch beim Weglassen einzelner Verarbeitungsschritte des Stemmers nicht. Am effektivsten ist das Entfernen des fünften Schrittes, was allerdings immer noch zu einem Verlust von über 1000 Einträgen führt. Die Entscheidung für den Stanford Lemmatizer fällt daher eindeutig aus. Der Grund für das vergleichsweise schlechte Abschneiden des regelbasierten Stemmers liegt vermutlich in dessen Konzeption. Durch das Abschneiden von Endungen werden auch Derivationssuffixe entfernt. Diese markieren jedoch unterschiedliche Wortarten, sodass zum Beispiel *computer* und *compute* beide zu *comput* umgewandelt werden. Der letzte Verarbeitungsschritt des Lexikons besteht darin, solchen Einträgen, die zu mehreren auf das gleiche Lemma abgebildet wurden, einen gemeinsamen Emotionswert zuzuordnen. Dies passiert hier durch komponentenweise Mittelwertbildung der Valenz-, Erregungs- und Dominanzwerte.

aid	aids	AIDS
batter	battered	battering
color	colored	coloring
confuse	confused	confusing
discourage	discouraged	discouraging
distinguish	distinguished	distinguishing
recycle	recycled	recycling
relax	relaxed	relaxing
vanquish	vanquished	vanquishing

Tabelle 7: Liste der Einträge von Warriners Emotionslexikon, die in einem 3-zu-1-Verhältnis lemmatisiert worden sind.

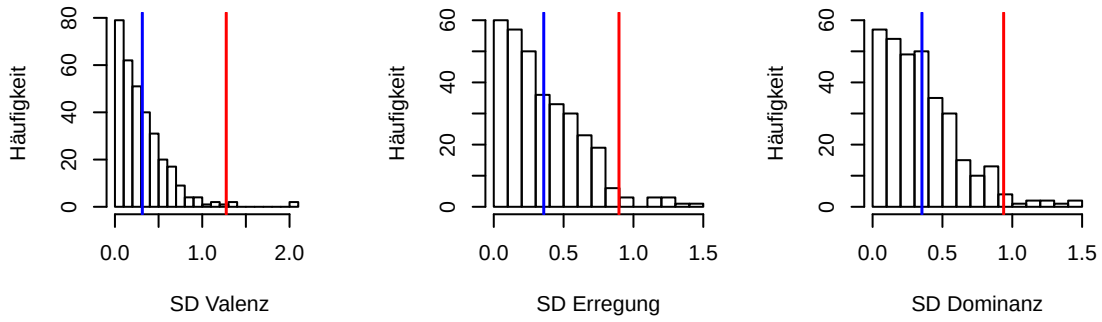

Abbildung 8: Histogramme der Standardabweichung von Valenz, Erregung und Dominanz von Lemmata des verarbeiteten Emotionslexikons, die aus mehreren Einträgen des unverarbeiteten Emotionslexikons gebildet wurden. Die blaue Linie gibt den Durchschnitt der Standardabweichungen aller dieser Lemmata an. Die rote Linie gibt zum Vergleich die Standardabweichung aller Einträge im unverarbeiteten Lexikon an.

In einem weiteren Experiment wurde die Verarbeitung von Lexikoneinträgen zu Lemmata genauer untersucht. Tabelle 6 gib an, wie häufig welche Anzahl von Einträgen auf ein Lemma abgebildet wurde. Sie zeigt, dass eine Eins-zu-Eins-Zuordnung die Regel ist, eine Drei-zu-Eins-Zuordnung sehr seltene Ausnahme und dass nicht mehr als drei Einträge zu einem Lemma umgewandelt wurden. Die wenigen Fälle einer Drei-zu-Eins-Zuordnung sind in Tabelle 7 dokumentiert. In acht der neun Fälle ist der erste Eintrag ein Nomen oder ein Verb und die anderen beiden Einträge sind die Partizipien des entsprechenden Verbs. Die erste Zeile zeigt eine Fehlleistung des Lemmatizers.[43]

[43]Das Krankheitsbild AIDS ist sowohl in Groß- als auch in Kleinbuchstaben im Lexikon enthalten. Dieses Phänomen lässt sich auch bei anderen Akronymen beobachten. Der Stanford Lemmatizer beachtet die Groß- und Kleinschreibung nicht, sondern interpretiert beides als Flexionsformen des Verbs *to aid*. Daher werden alle drei Einträge auf ein gemeinsames Lemma abgebildet.

Die Annahme, dass die Emotionen von Flexionsformen desselben Lemmas gleich bzw. sehr ähnlich sind, ist für die Verwendung von Emotionslexika zentral, da andernfalls nur ein Bruchteil der Wörter eines Dokuments als *emotionsrelevant* erkannt werden könnte. Diese Annahme lässt sich nach der Lemmatisierung von Warriners Emotionslexikon überprüfen: In 325 Fällen wurden mehrere Einträge jeweils einem Lemma zugeordnet. Die Einträge lassen sich daher als Flexionsformen desselben Lemmas betrachten. Zur Überprüfung der Annahme wurde jeweils die Standardabweichung ihrer Emotionskomponenten berechnet. In Abbildung 8 werden die Werte in drei Histogrammen dargestellt. Die Annahme der Generalisierbarkeit von Wortemotionen erhärtet sich daher, wenn die Streuung der Emotionswerte von Einträgen, die auf das gleiche Lemma abgebildet worden sind, deutlich geringer ist als die Streuung im gesamten Lexikon. In den Diagrammen müsste demnach jeweils der Großteil der Fälle deutlich links der roten Line liegen. Das ist hier der Fall. Die Streuung ist bei solchen Lemmata je nach Emotionskomponente durchschnittlich halb so groß wie die Streuung im gesamten Lexikon oder geringer (Lage der blauen zur roten Line).

Die wenigen Lemmata, deren interne Streuung größer als die Streuung im gesamten Lexikon ist (Fälle rechts der roten Linie), wurden manuell analysiert. Bei fast allen Fällen stimmen die Wortarten der Einträge nicht in allen syntaktischen Verwendungen überein (zum Beispiel *steer* und *steering*), sodass die Zuweisung desselben Lemmas nicht unbedingt richtig oder intuitiv ist. Die einzige Ausnahme bilden die Einträge *jaw* und *jaws*.[44] In den meisten Fällen liegt zudem eine offensichtliche Veränderung der Semantik vor, die über den Wortbildungsprozess durch Derivation hinausgeht, zum Beispiel *arm* und *armed* sowie *whale* und *whaling*. Diese Indizien sprechen dafür, dass eine überdurchschnittlich starke Streuung bei Einträgen desselben Lemmas in erster Linie durch Schwächen der maschinellen Lemmatisierung bzw. der fehlenden Wortarten- und Wortsinn-Markierung im Lexikon erklärt werden kann. Insgesamt deuten die Befunde also überzeugend darauf hin, dass der Emotionsgehalt von Flexionsformen des gleichen Lemmas stark unterdurchschnittlich streut.[45]

5.4 Stoppwortliste

Bei der Auswahl bzw. Erstellung der Stoppwortliste wurde Wert darauf gelegt, dass sie keine Einträge des Emotionslexikons enthält, da andernfalls durch das Entfernen

[44] In diesem Fall kann jedoch die Abweichung dadurch erklärt werden, dass *Jaws* der Originaltitel des Horrorfilms *Der weiße Hai* ist.

[45] Im Allgemeinen scheint dies jedoch nicht zu gelten. Danisman und Alpkocak (2008) weisen darauf hin, dass auch Tempus Einfluss auf die Wortemotion haben kann. Sie beobachten insbesondere, dass die Token *marry* und *love* meist in Sätzen der Kategorie *joy* auftreten, während die Token *married* und *loved* vermehrt in der Kategorie *sad* auftreten.

dieser Wörter Information für die Berechnung der Dokumentenemotion verloren gehen würde. Aus diesem Grund wurde zunächst eine eher kurze Stoppwortliste ausgewählt und diese dann weiter verarbeitet. Die Wahl fiel auf die Liste des Natural Language Toolkit (NLTK) (Bird, 2006). Diese umfasst 127 Einträge. Sie wurde lemmatisiert, da die Dokumente zum Zeitpunkt der Stoppwort-Entfernung nur noch aus Lemmata bestehen und andernfalls nach Wortformen gesucht würde, die in den Dokumenten gar nicht mehr vorkommen können. Dadurch verkürzte sich die Liste auf 106 Einträge, da – wie bei der Lemmatisierung des Lexikons – Wortformen zusammenfallen. Dies betraf bei der Stoppwortliste Flexionsformen von Pronomen, Hilfsverben und den Artikeln *a* bzw. *an*. Anschließend wurden alle Wörter entfernt, die auch im lemmatisierten Lexikon enthalten sind – *be*, *have*, *do*, *other*, *can* und *will* –, sodass die finale Stoppwortliste aus 100 Einträgen besteht.

	M	SD	Min	Max
Valenz	0.586	0.242	-1.691	2.86
Erregung	-1.009	0.187	-2.52	1.53
Dominanz	0.561	0.192	-1.262	2.62

Tabelle 8: Lage- und Streuungsmaße des RCV1.

6 Korpusvalidierung

Da das VAD-Modell in der textuellen Emotionserkennung noch nicht bei Regressionsaufgaben eingesetzt wurde, gibt es auch keine etablierte Testkorpora zur Evaluation des Werkzeugs. Allerdings gibt es in der Psychologie eine Ressource, die sich hierfür heranziehen ließe: ANET (M. M. Bradley & Lang, 2007). Dabei handelt es sich um eine Sammlung kurzer Sätze, die von Ratern auf ihren Emotionsgehalt hin bewertet worden sind. Es ist allerdings nur unter den selben strengen Auflagen wie ANEW erhältlich (s. Abschnitt 5).[46] Daher ist die Ressource für diese Arbeit leider nicht zugänglich gemacht worden.

Behelfsmäßig kann eine Korpusvalidierung durchgeführt werden. Dabei wird das Werkzeug auf ein in der Computerlinguistik schon bekanntes Korpus angewendet. Da die Eigenschaften des Korpus bekannt sind, können die Ergebnisse zumindest einer Plausiblitätsprüfung unterzogen werden. Außerdem dient dies als Vergleichsmaßstab um die Ergebnisse des Unternehmenskorpus interpretieren zu können. Darüber hinaus knüpft diese Arbeit durch das Einbeziehen vertrauter Korpora stärker an den Kernbereich der computerlinguistischen Forschung an.

Für das Vergleichskorpus fiel die Wahl auf das Reuters Corpus Volume 1 (RCV1), für das die bereinigte und mit weiteren Metadaten angereicherte Version von Lewis, Yang, Rose und Li (2004) maßgeblich ist. Es umfasst englischsprachige Nachrichten der Agentur Reuters aus den Jahren 1996 und 1997, insgesamt 804414 Dokumente. Diese sind manuell nach Themen (*Topics*) annotiert. Dabei wurde ein hierarchisches System aus 103 Topic-Kodes verwendet. Jedem Dokument sind so eine oder mehrere Kategorien zugeordnet. Daneben werden auch Kodes für den betroffenen Wirtschaftszweig und die Region verwendet. Dieses hochdifferenzierte System von Topic-Kodes erlaubt es, den Emotionsgehalt von Dokumenten unterschiedlicher Themen miteinander zu vergleichen und ihre Verteilung auf Plausibilität zu prüfen.

Für das ganze Korpus liegt die durchschnittliche Erkennungsrate durch das Werkzeug (Anzahl der emotionsrelevanten Wörter durch die Anzahl der Nicht-Stoppwörter) bei 64,8%. Tabelle 8 stellt wichtige Lage- und Streuungsmaße für den RCV1 dar. Wie schon bei der Untersuchung des Emotionslexikons (Tabelle 4) haben

[46] `http://csea.phhp.ufl.edu/media/anetmessage.html`

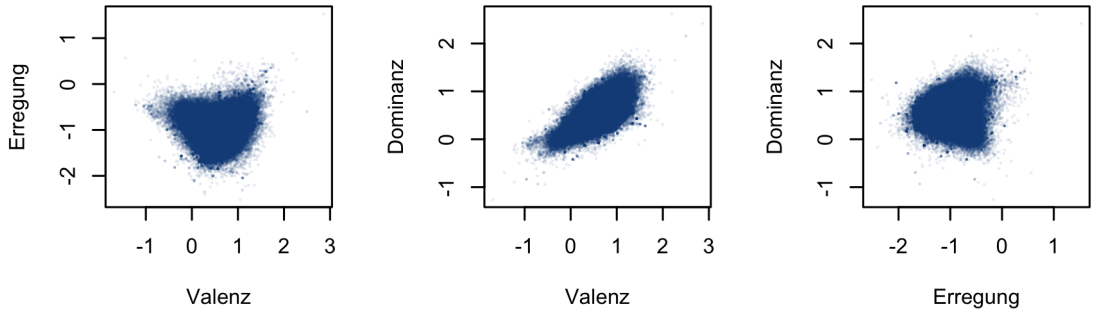

Abbildung 9: Streudiagramme der Emotionskomponenten im RCV1.

Valenz und Dominanz etwa gleiche Mittelwerte, während der von Erregung niedriger ist. Allerdings sind im RCV1 die Mittelwerte von Valenz und Dominanz deutlich höher als im Lexikon, während der Erregungsmittelwert fast mit dem des Lexikons identisch ist.

Tabelle 15 im Anhang dokumentiert die Korrelation von Valenz, Erregung und Dominanz. Die drei dazu gehörenden Streudiagramme finden sich in Abbildung 9. Das erste Diagramm zeigt einen deutlich ausgeprägten nicht-linearen Zusammenhang von Valenz und Erregung ($r = -0.04$). Dieser ist auch schon im Lexikon beobachtet worden.[47] Dies spricht dafür, dass für Nachrichten ebenso wie für Lexikoneinträge gilt, dass besonders positive und besonders negative als aufregend empfunden werden, durchschnittlich angenehme jedoch als eher langweilig bewertet werden. Das zweite Diagramm zeigt einen starken positiven linearen Zusammenhang von Valenz und Dominanz ($r = 0.68$) dieser ist ebenfalls schon im Lexikon beobachtet worden und ist konsistent mit dem Befund zahlreicher anderer Studien (Warriner et al., 2013). Im dritten Diagramm ist ein nicht-linearer Zusammenhang erkennbar, der in der Untersuchung des Lexikons nicht aufgefallen ist ($r = -0.05$). Erklärt werden kann diese nach rechts gekippte U-Form möglicherweise mit der starken Korrelation zwischen Valenz und Dominanz. Dies erklärt jedoch nicht, warum der Zusammenhang im Korpus stärker ausgeprägt ist als im Lexikon.

Von den 103 Topic-Kodes wurden 15 manuell ausgewählt, die eine möglichste große Vielfalt an unterschiedlichen Nachrichten repräsentieren sollen. Tabelle 9 listet die ausgewählten Kodes auf und stellt die Mittelwerte der damit markierten Dokumente dar. Da die Dokumente mit mehreren Topic-Kodes gelabelt sein können, werden manche davon in mehr als einer Zeile gezählt. Dies ist auch der Grund, warum an dieser Stelle kein Signifikanztest durchgeführt wird.

[47] Calvo und Mac Kim (2013) zeigen, dass dieser nicht-lineare Zusammenhang auch in anderen in der Emotionserkennung herangezogenen Korpora, zum Beispiel dem ISEAR-Korpus, zu beobachten ist.

Kode	Beschreibung	n	Valenz	Erregung	Dominanz
ECAT	Economy	119920	0.55	-0.99	0.54
CCAT	Corporate, Industrial	381327	0.63	-1.03	0.60
MCAT	Markets	204818	0.59	-1.06	0.55
GCRIM	Crime	32219	0.27	-0.86	0.36
GDEF	Defence	8842	0.48	-0.90	0.54
GDIP	International Relations	37739	0.51	-0.95	0.54
GDIS	Disasters, Accidents	8657	0.32	-0.88	0.29
GENV	Environment	6261	0.45	-0.95	0.44
GFAS	Fashion	313	0.71	-0.92	0.63
GHEA	Health	6030	0.38	-0.95	0.41
GPRO	Biographies	5498	0.54	-0.89	0.49
GREL	Religion	2849	0.46	-0.92	0.45
GSCI	Science, Technology	2410	0.63	-0.98	0.53
GSPO	Sports	35317	0.74	-0.93	0.65
GTOUR	Travel, Tourism	680	0.67	-0.97	0.60
GWEA	Weather	3878	0.56	-1.05	0.35

Tabelle 9: Beschreibung, absolute Häufigkeit und Mittelwerte von Valenz, Erregung und Dominanz für ausgewählte Nachrichtenkategorien des RCV1.

Die Tabelle verdeutlicht das große Gewicht, das Wirtschaftsnachrichten im RCV1 haben. Nicht nur, dass die Kategorien CCAT, ECAT und MCAT mit deutlichem Abstand am häufigsten vertreten sind, auch ihre Stellung im Kategoriensystem macht dies deutlich. Während es sich bei den drei Wirtschaftskategorien um Oberkategorien handelt, die sich in zahlreiche Unterkategorien aufgliedern, sind alle Nachrichten, die nicht Wirtschaft thematisieren, in der gemeinsamen Oberkategorie GCAT gruppiert.

Abbildung 15 (s. Anhang) zeigt drei Streudiagramme. Abgetragen werden darin die Mittelwerte der 15 ausgewählten Kategorien in jeweils zwei der drei Emotionsdimensionen. Die Kodes der Wirtschaftskategorie sind grün hervorgehoben. Die übrigen Kodes des RCV1 sind in Schwarz gehalten. Das blaue und rote Label wird in Abschnitt 7 eingeführt. Im Folgenden wird auf einige Details der Grafik eingegangen, um die Plausibilität der Ergebnisse zu prüfen. Laut der Daten sind die negativsten und erregendsten Nachrichten solche, die Kriminalität und Naturkatastrophen thematisieren (GCRIM und GDIS, siehe erstes Streudiagramm oben links). Jedoch weist GDIS einen deutlich niedrigeren Dominanzwert auf, was ein Hinweis auf die geringe Beherrschbarkeit von Unwettern und anderer Umweltereignisse sein kann. Im zweiten Diagramm liegen fast alle Kategorienmittelwerte auf einer Diagonalen, was durch die starke Korrelation von Valenz und Dominanz zu erklären ist. Nur der Wetterbericht (GWEA) schert durch deutlich verminderte Dominanz aus, was ebenfalls dafür spricht, dass sich die geringe Kontrollierbarkeit des Gegenstands

auf den Emotionsgehalt der dafür gebrauchten Sprache ausübt. Der Wetterbericht und Nachrichten über Marktlagen haben den geringsten Erregungswert. Dies scheint ebenfalls für die Mehrheit der Bevölkerung plausibel zu sein.[48] Die höchsten Valenzwerte weisen GSPO (Sport), GFAS (Mode) und GTOUR (Reisen) auf. Insbesondere Sport und Mode sind auf allen drei Diagrammen in geringem Abstand voneinander angeordnet. Dies entspricht der Alltagserfahrung, dass solche Berichte von der sonst üblichen, neutralen Nachrichtensprache abweichen. Die drei Wirtschaftskategorien fallen ebenfalls nahe zusammen, was weiter für die Plausibilität des Verfahrens spricht. Der Erregungswert von ECAT ist allerdings deutlich größer als der der anderen beiden Kategorien. Dies könnte ein Hinweis darauf sein, dass Konjunkturmeldungen unter den Wirtschaftsnachrichten diejenigen sind, die die größte Relevanz für den Bevölkerungsdurchschnitt haben, da gute bzw. schlechte Konjunkturmeldungen durchaus Einfluss auf den Alltag des Einzelnen haben können und so Erregung auslösen. Für die anderen Wirtschafts-Topics gilt dies wahrscheinlich nur bedingt.

Insgesamt sind die Ergebnisse plausibel und lassen sich gut in die Alltagserfahrung einordnen. Als Maß für die Güte des Werkzeugs ist diese Plausibilitätsprüfung jedoch nicht oder nur als Indiz geeignet. Der größere Nutzen der Untersuchung des RCV1 besteht darin, die nachfolgend präsentierten Ergebnisse aus der Analyse des Unternehmenskorpus vergleichend beurteilen zu können.

[48] Die Normen des Lexikons wurden aus der durchschnittlichen Bewertung vieler Rater gebildet. Daher sind auch die vom Werkzeug berechneten Werte als eine Durchschnittsbewertung zu betrachten.

7 Untersuchung des Unternehmenskorpus

Dieser Abschnitt widmet sich der explorativen statistischen Analyse des Korpus der Geschäfts- und Nachhaltigkeitsberichte. Insbesondere wird hier der Einfluss von Kontextmerkmalen wie dem Jahr und der Herkunft der Berichte auf deren Emotionsgehalt untersucht.

7.1 Korpusbeschreibung

Das hier untersuchte Korpus setzt sich aus jährlichen Geschäftsberichten und Nachhaltigkeitsberichten börsennotierter Unternehmen Deutschlands, der USA und Großbritanniens zusammen. Die genauen Auswahlkriterien für Berichte sind folgende: Zunächst wurden 90 Unternehmen ausgewählt. Dabei handelt es sich um alle Unternehmen, die 2014 im deutschen oder im US-amerikanischen Leitindex, dem DAX bzw. dem DJIA, notiert waren. Von den britischen Unternehmen wurden die ersten 30 im FTSE 100 Index ausgewählt. Als Ordnungskriterium diente hierbei die Marktkapitalisierung. Somit ist die gleiche Anzahl von Unternehmen aus allen Leitindizes im Korpus vertreten. Der FT 30 Index wurde nicht gewählt, da dort keine Unternehmen der Finanzbrache vertreten sind. Die Wahl fiel auf diese 90 Organisationen, da sie zusammengenommen einen großen Teil der Weltwirtschaft ausmachen (Goldenstein, Poschmann, Händschke & Walgenbach, 2015). Im Jahr 2014 entsprach die Summe ihrer Umsätze 7% des weltweiten BIP. Dieses lag in dem Jahr bei 77.3 Billionen USD. Darüber hinaus lag der Anteil der Volkswirtschaften Deutschlands, der USA und Großbritanniens hieran bei ca. 30%.

Von den so ausgewählten Unternehmen wurden sämtliche im Sommer 2015 online verfügbaren, auf Englisch abgefassten und voll digitalisierten Berichte unabhängig von deren Veröffentlichungszeitpunkt in das Korpus aufgenommen. Auf diese Weise wurden insgesamt 1676 ausgewählt, wovon 1087 auf Geschäfts- und 589 auf Nachhaltigkeitsberichte entfallen. Dazu kamen zwei Berichte in deren Dateien keine Textinformation hinterlegt war, sodass diese von der Analyse ausgeschlossen wurden. Zu jedem Dokument sind folgende Kontextmerkmale während der Erhebung festgehalten worden: Gattung (Geschäfts- oder Nachhaltigkeitsbericht), Referenzjahr, Unternehmen und Herkunft bzw. Aktienindex. Das Referenzjahr ist das Jahr auf das sich ein Bericht laut Titelseite bezieht. In manchen Fällen, insbesondere bei Nachhaltigkeitsberichten, werden dort zwei aufeinanderfolgende Jahren angegeben. In diesen Fällen gilt nur das erste der beiden als Referenzjahr. Es liegt für die Dokumente des Korpus zwischen 1992 (Henkel) und 2015 (BT Group und WalMart), wobei nur relativ wenige Berichte aus den 1990er Jahren stammen. Die Anzahl

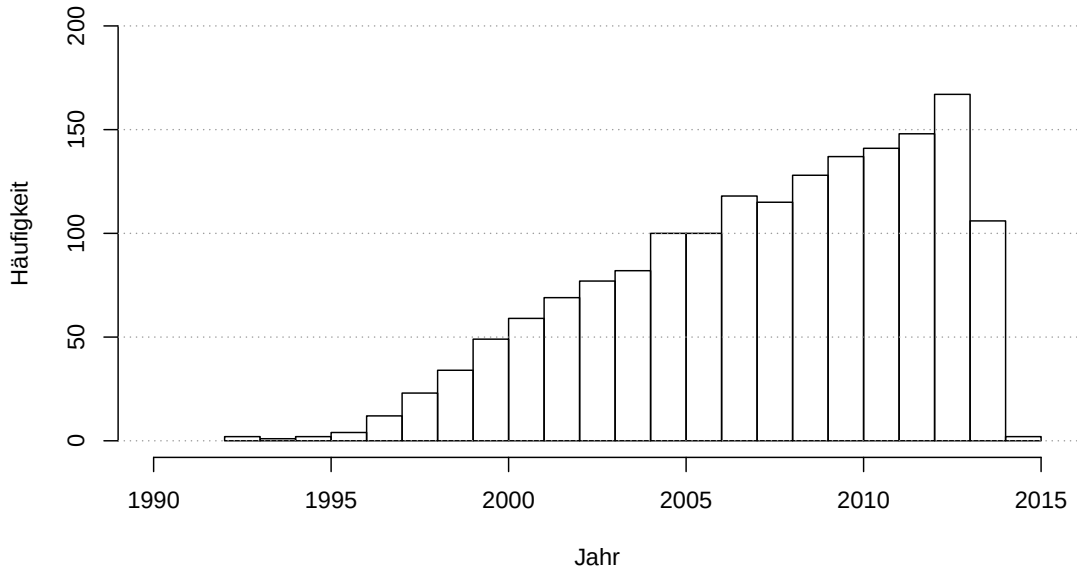

Abbildung 10: Histogramm der Anzahl der Dokumente pro Jahr im Unternehmenskorpus.

der Dokumente pro Jahr ist in Abbildung 10 dargestellt. Von den 180 theoretisch möglichen, relevanten Berichten pro Jahr[49] sind somit ab Mitte der 2000er Jahre über die Hälfte im Korpus vertreten mit Ausnahme des Jahres 2015.

7.2 Uni- und bivariate Analyse der Emotionsverteilung

Die Berichte wurden durch das vorangehend beschriebenen Werkzeug verarbeitet und so ihre Emotionswerte berechnet. Die durchschnittliche Erkennungsrate (siehe Abschnitt 4.4) beträgt 75%, was für die gute Abdeckung des gewählten Lexikons spricht spricht. Tabelle 10 gibt univariate Maßzahlen für Valenz, Erregung und Dominanz im gesamten Korpus an. Im Vergleich zum Reuters-Korpus (siehe Tabelle 8) zeigt sich ein leicht erhöhter Valenz- und Dominanzwert. Das Phänomen, dass Valenz und Dominanz ungefähr beide bei 0.6 liegen, während der Erregungswert etwa bei -1 liegt, lässt sich auch hier beobachten. Im Vergleich zum RCV1 ist die Standardabweichung im Unternehmenskorpus weniger als halb so groß, was darauf hinweist, dass die des Unternehmenskorpus sich emotional ähnlicher sind, als die Dokumente des Vergleichskorpus. Darüber hinaus ist auffällig, dass die Daten –

[49]Dies gilt unter der stark vereinfachenden Annahme, dass im jeweiligen Jahr dieselben Unternehmen in den drei Leitindizes vertreten waren wie 2015 und dass von jedem Unternehmen ein Geschäftsbericht und ein Nachhaltigkeitsbericht angefertigt worden ist.

	M	SD	Min	Max
Valenz	0.65	0.09	-0.40	1.18
Erregung	-1.01	0.05	-1.57	-0.19
Dominanz	0.62	0.07	-0.12	0.93

Tabelle 10: Lage- und Streuungsmaße der Emotionskomponenten für das Unternehmenskorpus.

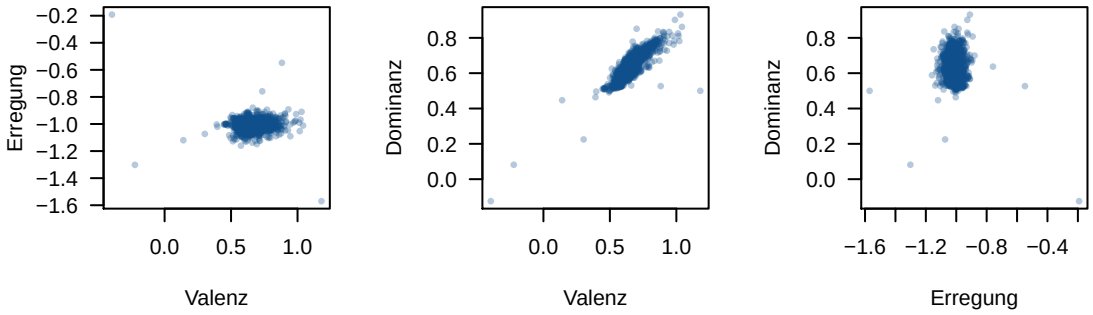

Abbildung 11: Streudiagramme der Emotionskomponenten im Unternehmenskorpus.

gemessen an der Standardabweichung – eine große Spannweite aufweisen. Dies deutet auf starke Ausreißer hin.

Als nächstes wird der Zusammenhang zwischen den Emotionskomponenten betrachtet. Abbildung 11 zeigt die Berichte des Unternehmenskorpus in drei Streudiagrammen dargestellten. Auffällig ist, dass anders als beim RCV1 das Valenz-Erregungs-Diagramm keine U-Form mehr aufweist. Das heißt, dass der dort bestehende nicht-lineare Zusammenhang hier nicht ausgeprägt ist. Dies könnte mit der geringeren Streuung im Unternehmenskorpus zusammenhängen: Da dessen Dokumente auf der Valenz-Erregungs-Ebene deutlich enger zusammenliegen als die Dokumente des RCV1, bilden sie möglicherweise nur einen Ausschnitt der U-Form, der jedoch nicht als solcher erkennbar ist. Dies könnte mit der impliziten Beschränkung des Korpus auf sehr erfolgreiche Unternehmen zusammenhängen.[50]

Der Zusammenhang von Valenz und Dominanz ($r = 0.87$) ist noch stärker ausgeprägt als im RCV1 ($r = 0.68$) oder im Emotionslexikon ($r = 0.72$). Das bedeutet, dass mehr Wörter verwendet werden, die einen hohen Valenz- *und* Dominanzwert haben, oder dass positive Wörter häufiger zusammen mit Kontrolle ausdrückenden Wörtern benutzt werden. Dies könnte ein Hinweis darauf sein, dass in Geschäfts-

[50]Eine mögliche Erklärung ist, dass Wörter generell dann erregender sind, wenn sie besonders positiv oder negativ sind. Die erfolgreichen Unternehmen des Korpus würden also positiv und daher auch mit hoher Erregung sprechen. Demnach würde das Streudiagramm einen Teil der U-Form, der rechts oberhalb der Mitte liegt, zeigen.

und Nachhaltigkeitsberichten Kontrolle implizit positiver bewertet wird als im Vergleichskorpus. Darüber hinaus veranschaulichen die Streudiagramme sehr gut, dass es einige wenige Extremwerte gibt, die sehr stark vom Rest der Datenpunkte abweichen. Zwischen Erregung und Dominanz sowie Erregung und Valenz besteht kein linearer Zusammenhang ($|r| < 0.1$ in beiden Fällen). Die vollständige Korrelationsmatrix ist in Tabelle 16 im Anhang dargestellt.

7.3 Untersuchung der Extremfälle

Nachfolgend sollen die stärksten Ausreißer ermittelt werden. Dazu wurden die Valenz-, Erregungs- und Dominanzwerte durch die z-Transformation standardisiert, sodass jeweils ihr Mittelwert null und ihre Standardabweichung eins ist. Anschließend wurde für jedes Dokument die euklidische Norm des Emotionsvektors berechnet. Im Dreidimensionalen handelt es sich dabei anschaulich um die Länge des Vektors, also den standardisierten Abstand eines Datenpunkts vom Zentroiden. Die Berichte wurden nach der berechneten euklidischen Norm absteigend sortiert und die ersten zehn Fälle – also die am stärksten abweichenden Dokumente – in Tabelle 18 im Anhang dokumentiert. Diese enthält neben den Metadaten der Dokumente die Emotionswerte (nicht-standardisiert), die euklidische Norm (standardisiert) sowie die Erkennungsrate. Die Emotionswerte wurden in nicht-standardisierter Form angeben, um einen Abgleich mit den Extremalwerten aus Tabelle 10 zu erlauben. Auffällig ist, dass in sieben der zehn Fälle die Erkennungsrate extrem niedrig ist (unter 15%), was höchstwahrscheinlich die Ursache für das starke Abweichen dieser Dokumente ist. Eine manuelle Überprüfung ergab, dass es sich bei allen sieben Fällen um Anomalien handelt. Der erste Fall, der Nachhhaltigkeitsbericht der Deutschen Bank 2002, ist in deutscher Sprache verfasst,[51] bei den anderen sechs Fällen treten Kodierungsprobleme auf, weswegen die Dokumente aus dem ursprünglich PDF-Format nicht korrekt in Plain-Text-Dateien umgewandelt wurden. Die verbliebenen drei Fälle weisen keine Anzeichen inkorrekter Verarbeitung auf. Nach der manuellen Überprüfung scheint es plausibel, dass diese extremen Emotionswerte tatsächlich dem Sprachgebrauch der Unternehmen entsprechen. Zu beachten ist, dass die euklidische Norm des ersten Falls, der eine unauffällige Erkennungsrate aufweist, nur ca. ein Viertel so groß ist wie die des stärksten Ausreißers. Die extremsten sechs Ausreißer sind also alle auf Anomalien im Korpus zurückzuführen.

[51] Wie dieses Dokument in den Korpus gelangt ist, muss hier leider offen bleiben, da dessen Zusammenstellung nicht Teil dieser Arbeit ist.

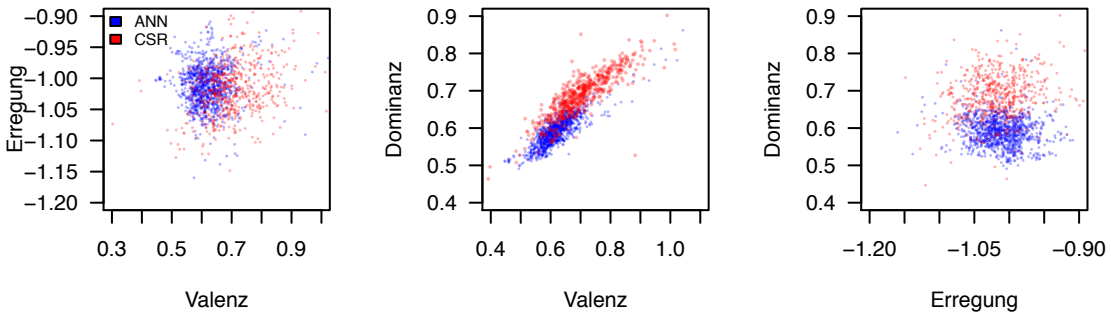

Abbildung 12: Streudiagramme der Emotionskomponenten im Unternehmenskorpus mit farblicher Hervorhebung der Gattung (Ausschnitt).

	M_{ANN}	M_{CSR}	SD_{ANN}	SD_{CSR}	p	d
Valenz	0.62	0.70	0.06	0.11	0.00	-0.98
Erregung	-1.01	-1.02	0.03	0.06	0.19	–
Dominanz	0.59	0.68	0.04	0.07	0.00	-1.70

Tabelle 11: Einfluss der Gattung auf den Emotionsgehalt der Dokumente im Unternehmenskorpus.

7.4 Einfluss der Gattung

Der Einfluss der Gattung wird in Abbildung 12 veranschaulicht. Die drei Streudiagramme stellen die einzelnen Berichte als Datenpunkte dar, wobei Geschäftsberichte blau und Nachhaltigkeitsberichte rot markiert sind. Um die Struktur besser sichtbar zu machen, wird ein Ausschnitt der Verteilung ohne Ausreißer gezeigt. Tabelle 11 gibt jeweils Mittelwert und Standardabweichung für Geschäftsberichte (ANN) und Nachhaltigkeitsberichte (CSR) an. Um den Einfluss der Gattung auch mit formalen Mitteln beurteilen zu können, wurde für jede Emotionskomponente ein zweiseitiger t-Test durchgeführt.[52] Darüber hinaus wurde Cohens d als Maß für die Effektstärke von signifikanten Zusammenhängen berechnet.[53]

Wie vor allem auf dem zweiten und dritten Diagramm gut zu erkennen ist, sind die beiden Gattungen klar auseinanderzuhalten. Zwischen den beiden Punktwolken gibt es nur wenig Überlappung. Die Erwartungswerte von Geschäfts- und Nachhaltigkeitsberichten unterscheiden sich höchst signifikant in der Valenz- und Dominanzkomponente. Nachhaltigkeitsberichte sind positiver und dominanter als

[52] Die unterschiedlichen Signifikanzniveaus werden in dieser Arbeit folgendermaßen bezeichnet: signifikant bei $0.05 \geq p > 0.01$, hoch signifikant bei $0.01 \geq p > 0.001$, höchst signifikant bei $0.001 \geq p$.

[53] Bei der Interpretation des Effektstärkemaßes Cohens d wird sich an folgende Konvention gehalten: kleiner Effekt ab $|d| = 0.2$, mittlerer Effekt ab $|d| = 0.5$, starker Effekt ab $|d| = 0.8$ (Rasch et al., 2010a, S. 68).

Geschäftsberichte. Insbesondere für Dominanz überschreitet d den doppelten Schwellenwert eines starken Effekts. Die beiden Gattungen unterscheiden sich in dieser Komponente deutlich stärker als in der Valenz, was im Gegensatz zu dem im Lexikon festgestellten starken Zusammenhang zwischen den beiden Komponenten steht. Darüber hinaus widerspricht der Befund den Zweifeln, die in zahlreichen Studien an der Wichtigkeit von Dominanz angemeldet worden sind (Warriner et al., 2013). Auf die Erregungskomponente hat die Gattung hingegen keinen signifikanten Einfluss. Die Standardabweichung von Nachhaltigkeitsberichten ist in allen Emotionskomponenten etwa doppelt so hoch wie die von Geschäftsberichten. Deren Emotionsgehalt streut daher sehr viel stärker.

7.5 Vergleich mit dem RCV1

Nach der Differenzierung von Geschäfts- und Nachhaltigkeitsberichten, werden die beiden Gattungen nun separat mit den Nachrichtenkategorien des Reuters-Korpus verglichen. Abbildung 15 (s. Anhang) zeigt drei Streugramme in denen die Mittelwerte der Nachrichtenkategorien des RCV1 abgetragen sind. Kategorien von Wirtschaftsnachrichten sind grün markiert, Geschäftsberichte blau und Nachhaltigkeitsberichte rot. Aus den Streudiagrammen geht hervor, dass Geschäftsberichte gut mit den Wirtschaftskategorien des RCV1 zusammenfallen. Dies gilt insbesondere für die Kategorie GCAT, die Nachrichten über Unternehmen und die Industrie zusammenfasst. Der Emotionsgehalt von Geschäftsberichten entspricht damit dem Emotionsgehalt von Nachrichtentexten über Unternehmen. Dies bedeutet, dass Unternehmen in Geschäftsberichten emotional den gleichen Sprachgebrauch haben, wie die Nachrichten, in denen über sie berichtet wird. Nachhaltigkeitsberichte weisen zwar in der Erregungsdimension nur einen kleinen Abstand zu den Wirtschaftskategorien und den Geschäftsberichten auf, in der Valenz- und der Dominanzdimension fallen sie jedoch mit der Gruppe der Mode- (GFAS) und Sportnachrichten (GSPO) zusammen. Der Dominanzwert der Nachhaltigkeitsberichte ist durchschnittlich größer als der aller Nachrichtenkategorien.

7.6 Einfluss der Herkunft

Da die Variable Herkunft drei Ausprägungen hat wurde zur Untersuchung ihres Einflusses eine Varianzanalyse statt eines t-Tests durchgeführt. Tabelle 17 (s. Anhang) gibt für jede Emotionskomponente und jeden Aktienindex den betreffenden Mittelwert und die Standardabweichung an. Der Varianzanalyse liegt der F-Test zugrunde. Dessen p-Wert wird ebenfalls angegeben. Als Effektstärkemaß wird Eta-Quadrat (η^2) verwendet. Es lässt sich als Anteil der aufgeklärten Variation an der Gesamtvariation,

bzw. als Fehlerreduktionsmaß (PRE-Maß) interpretieren. Die Tabelle ist vertikal in drei Abschnitte gegliedert. Der erste Abschnitt erfasst *alle* Berichte, der zweite und dritte Abschnitt jeweils nur Geschäfts- bzw. Nachhaltigkeitsberichte. Nachfolgend wird zunächst der erste Abschnitt besprochen.

Der Einfluss der Herkunft ist zwar bei allen Emotionskomponenten höchst signifikant, jedoch sind die Effekte klein.[54] Dies ist möglicherweise ebenfalls auf die Zusammenstellung des Korpus zurückzuführen. Es ist denkbar, dass sich die besonders erfolgreichen, weltweit agierenden Unternehmen verschiedener Volkswirtschaften stärker ähneln, als der Durchschnitt der Unternehmen dies tun würde. Im US-amerikanischen DJIA sind sowohl Erregungs- als auch Valenzwerte durchschnittlich höher als in den anderen Aktienindizes. Darüber hinaus ist die Streuung von Valenz und Dominanz im DJIA deutlich um ca. 70% bzw. 50% erhöht. Dies könnte darauf hindeuten, dass US-amerikanische Firmen, was den Emotionsgehalt der Berichte betrifft, individueller in der Gestaltung vorgehen und dort Gattungskonventionen weniger stark ausgeprägt sind als in Großbritannien und Deutschland. Dieser Befund deckt sich mit bisher unveröffentlichten Ergebnissen zur Semantik des Verantwortungsbegriff (Goldenstein et al., 2015). Auch in dieser Arbeit wurde zwischen den Dokumenten des DJIA eine relative große Streuung im Vergleich zu den anderen beiden Aktienindizes festgestellt.[55]

Bei der weiteren Analyse des Einflusses der Herkunft ist bei der Durchsicht zahlreicher Streudiagramme aufgefallen, dass die Verteilung der Berichte des DJIA eine Struktur aufweist, die deutlich zwei separate Punktwolken zeigt. Es hat sich herausgestellt, dass es sich dabei um eine Gruppierung in Geschäft- und Nachhaltigkeitsberichte handelt. Abbildung 13 zeigt drei Streudiagramme, wobei ein Datenpunkt einem Bericht des DJIA entspricht. Geschäftberichte sind blau markiert, Nachhaltigkeitsberichte rot. Insbesondere in dem zweiten, aber auch in dem dritten Diagramm sind die getrennten Punktwolken deutlich erkennbar. Auch im ersten Diagramm zeigt sich, dass sich beide Gruppen nur wenig überlappen, allerdings wird dies nur durch die farbliche Markierung und nicht durch die räumliche Struktur ersichtlich.

Diese Beobachtungen legen nahe, dass der Effekt, den die Herkunft auf den Emotionsgehalt der Berichte hat, durch den stärkeren Einfluss der Gattung überlagert werden könnte. Aus diesem Grund wurde der Herkunftseinfluss nochmals unter Kontrolle der Gattung analysiert. Der zweite und dritte Abschnitt von Tabelle 17

[54] Zur Interpretation von η^2 wird sich an folgende Konvention gehalten: kleiner Effekt bei $\eta^2 < 0.06$, mittlerer Effekt ab $\eta^2 = 0.06$, starker Effekt ab $\eta^2 = 0.14$ (Rasch, Friese, Hofmann & Naumann, 2010b, S. 38).

[55] Interessanterweise war dort der Einfluss der Herkunft deutlich stärker ausgeprägt. Dies könnte möglicherweise darauf hindeuten, dass Kultur und nationale Gemeinsamkeiten auf begrifflich-kognitiver Ebene einen stärkeren Einfluss haben, während auf emotionaler Ebene individualistische Einflüsse dominieren.

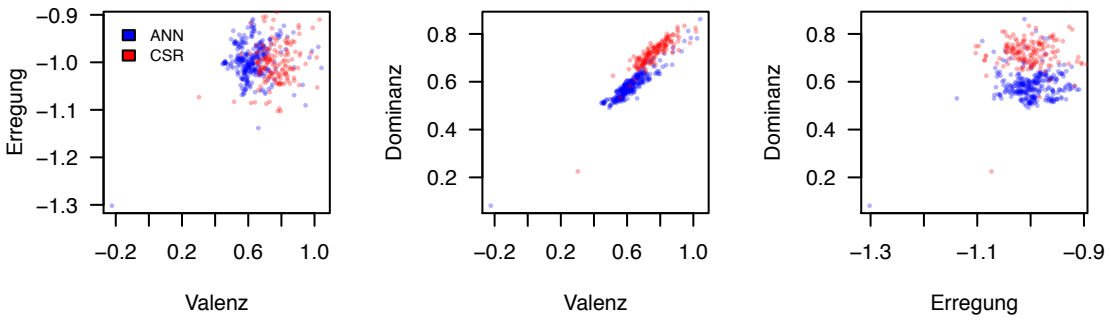

Abbildung 13: Streudiagramme der Berichte des DJIA mit farblich hervorgehobener Gattung.

zeigt den Einfluss der Herkunft unter Konstanthaltung der Gattung. Hierbei können deutlich stärkere Effekte als ohne Drittvariablenkontrolle festgestellt werden. Bei den Geschäftsberichten ist der Einfluss der Herkunft auf Erregung und Dominanz jeweils höchst signifikant. In der Valenzdimension wurde kein signifikanter Effekt festgestellt. Allerdings streuen die Geschäftberichte des DJIA mehr als doppelt so stark wie die der anderen Indizes. Der Einfluss auf Erregung ist von mittlerer Stärke. Geschäftsberichte des DJIA weisen eine höhere Erregung auf als Geschäftsberichte der anderen Indizes. Der Effekt auf den Dominanzwert ist nur schwach ausgeprägt. Auch hier ist die Streuung im DJIA doppelt so hoch, wie in den anderen Indizes.

Die Nachhaltigkeitsberichte betreffend lassen sich in der Valenz- und Dominanzkomponente höchst signifikante Erwartungswertunterschiede feststellen. Die Effektstärke ist jeweils mittelgroß, in der Valenzkomponente ist sie fast stark. Sowohl Dominanz als auch Valenz sind im DJIA deutlich höher als in den anderen Indizes. In der Erregungskomponente ist der Erwartungswertunterschied hoch signifikant, die Effektstärke jedoch sehr klein.

Während sich Geschäftsberichte zwischen den Indizes also vor allem in ihrer Erregungskomponente unterscheiden, sind es bei Nachhaltigkeitsberichten die anderen Komponenten, Valenz und Dominanz, die den stärksten Unterschied zwischen den Berichten des DAX, des DJIA und des FTSE machen. Wird die Gattung nicht konstant gehalten, wird der Einfluss der Herkunft vom stärkeren Gattungssignal überlagert und weist deutlich kleinere Effektstärken auf. Als Konsequenz aus dieser Feststellung wird für die nachfolgenden Analysen jeweils nach der Gattung der Berichte kontrolliert.[56]

[56]Die bisherigen uni- und bivariaten Analysen haben den Einfluss der Gattung bereits mit berücksichtigt. Dies gilt nur nicht für die Korrelation zwischen den Emotionskomponenten Valenz, Erregung und Dominanz. Eine nachträgliche Überprüfung hat jedoch keine Auffälligkeiten zwischen den Gruppen festgestellt.

7.7 Einfluss des Referenzjahres

Zur Untersuchung des Einflusses des Referenzjahres wurden ebenfalls Varianzanalysen durchgeführt. Diese wurden gegenüber der Pearson-Korrelation bevorzugt, da sie auch nicht-lineare Zusammenhänge erfassen. Dies ist angemessen, da mit Blick auf den langen Untersuchungszeitraum die Annahme linearer Zusammenhänge unplausibel wäre. Die Ergebnisse sind in Tabelle 12 dargestellt. Aufgrund der großen Anzahl der Merkmalsausprägungen dieser Variable wurde auf die Angabe der Mittelwerte und Standardabweichungen für jede Gruppe verzichtet. Insgesamt ist der Einfluss des

		p	η^2
alle	Valenz	0.4070	–
	Erregung	< 0.001	0.010
	Dominanz	0.4168	–
ANN	Valenz	< 0.001	0.062
	Erregung	< 0.001	0.020
	Dominanz	< 0.001	0.036
CSR	Valenz	< 0.001	0.025
	Erregung	0.0572	–
	Dominanz	0.0106	0.011

Tabelle 12: Einfluss des Referenzjahres auf den Emotionsgehalt der Dokumente im Unternehmenskorpus: ohne Drittvariablenkontrolle (alle), nur Geschäftsberichte (ANN), nur Nachhaltigkeitsberichte (CSR).

Referenzjahres auf den Emotionsgehalt der Dokumente im Vergleich zum Einfluss von Gattung und Herkunft eher klein. Ohne Drittvariablenkontrolle besteht nur in der Erregungskomponente ein (höchst) signifikanter Zusammenhang. Die Effektstärke ist jedoch sehr klein. Bei den Geschäftsberichten sind alle Zusammenhänge höchst signifikant, die Effektstärken sind ebenfalls klein. Die Effektstärke der Valenz ist allerdings fast mittelgroß und deutlich stärker ausgeprägt als alle anderen berechneten Effektstärken in dieser Tabelle. Dies könnte ein Hinweis auf eine Beeinflussung durch die konjunkturelle Entwicklung sein. Bei Nachhaltigkeitsberichten ist der Zusammenhang in der Valenzkomponente höchst signifikant. In der Dominanzkomponente ist er hoch signifikant. Die Effektstärken bleiben jedoch auch hier vernachlässigbar gering. Auf Nachhaltigkeitsberichte scheint das Referenzjahr also einen viel kleineren Einfluss zu haben als auf Geschäftsberichte. Die Zeitreihe der Mittelwerte der Emotionskomponenten ist in Abbildung 17 im Anhang dargestellt. Dabei werden Geschäfts- und Nachhaltigkeitsberichte getrennt betrachtet. Zur Interpetation der Grafik sei angemerkt, dass die Zahl der Berichte aus den 1990er Jahren vergleichsweise gering ist und eine erhöhte Streuung der Mittelwerte in dieser Zeit darauf zurückzuführen sein kann.

7.8 Einfluss des Unternehmens

Als letztes wird der Einfluss des Unternehmens untersucht. Auch hierfür wurden mehrere Varianzanalysen durchgeführt. Die Ergebnisse sind in Tabelle 13 dargestellt. Der Einfluss des Unternehmens auf den Emotionsgehalt der Berichte muss als sehr

		p	η^2
alle	Valenz	< 0.001	0.306
	Erregung	< 0.001	0.335
	Dominanz	< 0.001	0.264
ANN	Valenz	< 0.001	0.490
	Erregung	< 0.001	0.702
	Dominanz	< 0.001	0.556
CSR	Valenz	< 0.001	0.525
	Erregung	< 0.001	0.314
	Dominanz	< 0.001	0.477

Tabelle 13: Einfluss des Unternehmens auf den Emotionsgehalt der Dokumente im Unternehmenskorpus: ohne Drittvariablenkontrolle (alle), nur Geschäftsberichte (ANN), nur Nachhaltigkeitsberichte (CSR).

groß eingestuft werden. Auch wenn ein direkter Vergleich aufgrund unterschiedlicher Maßzahlen nicht möglich ist, ist er doch ähnlich beherrschend wie der Einfluss durch die Gattung (s. Abschnitt 7.4). Alle neun Varianzanalysen zeigen höchst signifikante Erwartungswertunterschiede für die jeweilige Emotionskomponente. Die Effektstärken sind durchgängig sehr groß – bis auf eine Ausnahme sind sie mehr als doppelt so groß wie der Schwellenwert zum starken Effekt.

Allerdings muss auch beachtet werden, dass durch die Einschränkung auf eine Gattung und ein Unternehmen die Größe der verglichenen Gruppen im Verhältnis zur Gesamtzahl der Dokumente im Korpus sehr klein ist. Bei Geschäftsberichten ist die Effektstärke bei Erregung besonders hoch. Der Anteil der durch das Unternehmen erklärten Variation an der Gesamtvariation liegt hier bei über 70%. Bei Nachhaltigkeitsberichten sind die Effektstärken für Valenz und Dominanz besonders stark ausgeprägt. Diese betragen jeweils ca. $\eta^2 = 0.5$. Im Vergleich dazu liegt die Effektstärke bei Erregung nur ca. $\eta^2 = 0.3$. Dies deckt sich mit dem Befund aus der Untersuchung des Einflusses der Herkunft: Beide Analysen zeigen, dass unter Geschäftsberichten die Erregungskomponente besonders gut durch die unabhängigen Variablen erklärt werden kann, während es bei Nachhaltigkeitsberichten die Valenz- und die Dominanzkomponenten sind. Diese Übereinstimmung ist allerdings zumindest teilweise darauf zurückzuführen, dass die Merkmale Herkunft und Unternehmen eng miteinander zusammenhängen, da jedes Unternehmen eindeutig einem Aktienindex zugeordnet ist.

Die Varianzanalysen haben gezeigt, dass das Unternehmen einen großen Einfluss auf den Emotionswert der Berichte hat. Um darüber hinaus Aussagen über die Art seines Einflusses treffen zu können, wurden mehrere hierarchische Clusteranalysen durchgeführt, um die Ähnlichkeiten zwischen den Unternehmen besser erfassen zu können. Abbildung 16 im Anhang zeigt das Dendrogramm einer Clusteranalyse aller Unternehmen, wobei für jedes Unternehmen der Mittelwert aus allen dessen Berichten gebildet wurde. Als Abstandsmaß wurde der euklidische Abstand verwendet. Die Cluster wurden mit Average-Group-Linkage verbunden. Leider ist im Dendrogramm kein deutliches Muster, wie etwa eine Gruppierungen nach Branchen, zu erkennen. Auch weitere Clusteranalysen, die sich auf einen Index oder eine Gattung beschränkt haben, blieben im Ergebnis unspezifisch. Damit kann zum jetzigen Zeitpunkt zwar gesagt werden, dass sich Unternehmen stark im Emotionsgehalt ihrer Berichte unterscheiden, allerdings ist noch kein Muster gefunden worden, dem diese Unterschiede zu folgen scheinen.

8 Diskussion

Dieser Abschnitt schätzt zunächst die Leistung des entwickelten Werkzeugs im Vergleich zu bestehenden Arbeiten ab, geht anschließend auf die Bedeutung der erzeugten Daten für die Organisationsforschung ein und erörtert vor diesem Hintergrund, Vor- und Nachteile der Verwendung des VAD-Modells.

8.1 Abschätzung der Performanz des verwendeten Tools

Sicherheit über die Messgüte der erhobenen Daten kann nur durch eine quantitative Evaluation erlangt werden. Da kein nach dem VAD-Modell annotierter Testkorpus vorliegt, konnte eine solche Evaluation nicht durchgeführt werden. Daher sind Zweifel an der Verlässlichkeit der erhobenen Daten zunächst angemessen. Die vergleichsmäßige Untersuchung des RCV1, zeigt jedoch Ergebnisse, die der Alltagswahrnehmung von verschiedenen Nachrichtenkategorien gut entsprechen. Die vom Werkzeug produzierten Daten sind in diesem Fall also plausibel, was ein wichtiger Hinweis auf dessen Verwendbarkeit in der weiteren Untersuchung ist. Darüber hinaus kann die Performanz des Werkzeugs im Vergleich mit schon bestehende Arbeiten grob abgeschätzt werden: Der Ansatz von Calvo und Mac Kim (2013) ähnelt der hier vorgestellten Methode sehr stark. Jedoch bilden sie im letzten Verarbeitungsschritt die im VAD-Raum gemessenen Emotionen mit der Pseudo-Dokumenten-Methode wieder auf Basisemotionen ab (s. Abschnitt 3.4.4). Am geläufigen Testkorpus des SemEval-2007 erreichen sie im Durchschnitt über alle Emotionsklassen 38,6% F-Maß. Es ist sehr wahrscheinlich, dass das in dieser Arbeit vorgestellte Tool beim gleichen Versuchsaufbau einen höheren Performanzwert erreichen würde, da das verwendete Lexikon ca. zehn mal so umfangreich ist. Könnten die im VAD-Raum berechneten Emotionsbewertungen direkt evaluiert werden, ist anzunehmen, dass der Wert der Pearson-Korrelation noch höher liegen würde, da die Abbildung mit der Pseudo-Dokumenten-Methode eine zusätzliche Fehlerquelle ist. Darüber hinaus zeigen die drei bekannten Arbeiten, bei denen mit dem lexikalischen Ansatz sowohl eine Klassifikation als auch eine Regression durchgeführt worden ist, dass der Performanzwert der Regression (gemessen durch die Pearson-Korrelation) deutlich über dem Performanzwert der Klassifikation (gemessen durch das F-Maß) liegt (Katz et al., 2007; Strapparava & Mihalcea, 2008; Staiano & Guerini, 2014). In Anbetracht der Tatsache, das die IAA bei Emotionsannotationen gering ist – im Durchschnitt der Emotionsklassen $r = 0.53$ bei der Regressionsaufgabe des SemEval-2007 – , deuten diese Überlegungen auf ein sehr respektables Ergebnis hin. Dies gilt insbesondere im Hinblick auf die Einfachheit des verwendeten Ansatzes.

Darüber hinaus sind die Dokumente des Unternehmenskorpus im Vergleich zu den Dokumenten der häufig verwendeten Testkorpora sehr lang. Zwar wäre eine auf Satzebene berechnete Emotion häufig ungenau, es ist jedoch davon auszugehen, dass sich die entstehenden Fehler auf Dokumentenebene zumindest in Teilen gegenseitig ausgleichen. Insgesamt kann daher festgehalten werden, dass die erhobenen Daten, insbesondere vor dem Hintergrund der geringen Übereinstimmungswerte menschlicher Rater, als valide betrachten werden können.

8.2 Ergebnisse für die Organisationsforschung

Mit der Entscheidung nur Unternehmen aus deutschen, US-amerikanischen und britischen Leitindizes aufzunehmen, beinhaltet das Unternehmenskorpus bewußt nur relativ lange bestehende und sehr erfolgreiche Unternehmen. Weniger erfolgreiche oder erst kürzlich entstandene Unternehmen sind dagegen nicht darin vertreten. Daher kann es überhaupt nur für einen geringen Teil aller Unternehmen repräsentativ sein. Inwiefern sich die Ergebnisse der Datenanalyse auch auf andere Unternehmen oder Formen von Organisationen übertragen lassen, ist ungeklärt. Hinzu kommt, dass das Korpus durch die Erhebungsmethode bzw. durch die eingeschränkte Onlineverfügbarkeit von Berichten weder eine Vollerhebung noch eine Zufallsstichprobe darstellt. Insofern sind die Ergebnisse der durchgeführten Signifikanztests nicht uneingeschränkt aussagekräftig. Andererseits ist nicht davon auszugehen, dass ein Zusammenhang zwischen dem Emotionsgehalt von Berichten und ihrer Onlineverfügbarkeit besteht. Daher ist es plausibel anzunehmen, dass das Korpus durch das Fehlen dieser Berichte *nicht* verzerrt wird. Darüber hinaus sind für den Zeitraum ab Mitte der 2000er Jahre bis 2014 über die Hälfte der erwarteten Berichte auch im Korpus vorhanden. Also kann zumindest für diese Jahre davon ausgegangen werden, dass das Korpus die Gesamtheit der Berichte der betrachteten Unternehmen angemessen repräsentiert.

Ein auffälliger Befund ist der hohe Dominanzwert der Nachhaltigkeitsberichte innerhalb des Unternehmenskorpus. Dies gilt insbesondere im Vergleich zum RCV1. Dass Nachhaltigkeitsberichte dadurch – anders als Geschäftsberichte – emotional eine größere Nähe zu Sport- und Modenachrichten statt zu Wirtschaftsnachrichten haben, ist überraschend, besonders vor dem Hintergrund, dass die Sprache des Managements intuitiv ohnehin schon als dominant eingeschätzt werden dürfte. Auf der anderer Seite ist die Tatsache, dass Nachhaltigkeitsberichte eine höhere Dominanz aufweisen als Geschäftsberichte hochplausibel, da sich CSR-Berichte genuin auf Aktivitäten des Unternehmens beziehen, die dieses freiwillig – also selbstbestimmt – durchführt. Dies geht aus der Definition des Managementkonzepts CSR hervor (Matten & Moon, 2008). Dagegen sind die Inhalte von Geschäftsberichten und insbesondere die Performance

eines Unternehmens von vielen externen und nicht beeinflussbaren Faktoren abhängig. Insofern ist eine niedrigere Dominanz hier plausibel. Ein weiterer Punkt ist, dass zwischen der Performance und der Verantwortungsübernahme eines Unternehmens nur ein sehr lockerer Zusammenhang besteht (Orlitzky, Schmidt & Rynes, 2003), was die deutliche Divergenz der beiden Gattungen in der Dominanzkomponente noch weiter nachvollziehbar macht. Auffällig ist auch, dass die Korrelation von Valenz und Dominanz erheblich stärker ausgeprägt ist als im Vergleichskorpus sowie im Lexikon. Hier könnte vermutet werden, dass Unternehmen Kontrolle noch positiver bewerten, als es das durchschnittliche Individuum tun würde. Unklar bleibt, warum zwischen Valenz und Erregung keine U-förmige Verteilung erkennbar ist. Auf Möglichkeit, dies durch die Zusammenstellung des Korpus zu erklären, ist bereits hingewiesen worden. Eine alternative Erklärung wäre, dass einzelne Teilkorpora, die sich zum Beispiel durch Herkunft und Gattung unterscheiden, jeweils diese U-Form aufweisen, sich diese aber so überlagern, dass sie im vollständigen Korpus nicht mehr erkennbar sind. In diesem Fall würde eine tiefergehende Datenanalyse das erwartete Ergebnis bringen.[57]

Der Einfluss des einzelnen Unternehmens auf den Emotionsgehalt seiner Berichte ist sehr stark ausgeprägt. Dass bis jetzt kein Muster gefunden wurde, dass die Art dieses Einflusses beschreibt, ist ein starkes Indiz dafür, dass Geschäfts- und Nachhaltigkeitsberichte eine für dieses Unternehmen spezifische Emotionalität aufweisen. Diese These wird dadurch weiter unterstützt, dass sich unabhängig von der Gattung sehr große Effektstärken nachweisen lassen. Demnach wäre es tatsächlich angemessen im Sinne einer *Organizational Identity* anzunehmen, dass Unternehmen als soziale Akteure auch über distinktive emotionale Merkmale verfügen. Diese erstaunliche Schlussfolgerung würde weiter untermauert werden, wenn sich auch nach einer umfassenden Drittvariablenkontrolle immer noch ein starker Einfluss des Unternehmens messen lässt. Diese Drittvariablenkontrolle sollte dann auch weitere als die hier erfassten Kontextmerkmale – wie etwa die Branche, die mittelfristige finanzielle Entwicklung oder das Gründungsjahr – miteinbeziehen.

Gleichzeitig liefert die Varianzanalyse bezüglich des Einflusses des Unternehmens ein deutliches Indiz dafür, dass Emotionen nicht nur ein distinktives, sondern auch ein dauerhaftes Merkmal im Sinne der *Organizational Identity* sind: Unter Kontrolle der Berichtart bestehen die Gruppen, deren Streuung jeweils mit der gesamten Streuung im Datensatz verglichen wird, aus höchstens einem Bericht pro Jahr. Dass der Einfluss der Firma auf den Emotionswert der Dokument so groß, bedeutet also zusätzlich zu dem vorherigen Absatz, dass die Berichte eines Unternehmens hinsichtlich ihrer

[57] Allerdings deutet die einzige dazu präsentierte Grafik (s. Abbildung 13) wenn überhaupt auf eine U-Form im Erregungs-Dominanz-Diagramm hin, wenn nur jeweils eine Gattung betrachtet wird.

Emotionen sehr viel weniger streuen als die Gesamtheit der Berichte des jeweiligen Teilkorpus. Sie sind also zeitlich relativ stabil.

Die Tatsache, dass die Streuung der Emotionen im DJIA stärker ist als in den anderen Aktienindizes, lässt vermuten, dass dies ebenso für die durchschnittliche Emotion eines dortigen Unternehmens gilt. Dies ist erst noch empirisch zu prüfen, würde jedoch bedeuten, dass sich diese Unternehmen stärker in ihren Emotionen unterscheiden als Unternehmen anderer Herkünfte. Dies spräche dafür, dass die Anthropomorphisierung US-amerikanischer Unternehmen im Bezug auf Emotionen besser empirisch gerechtfertigt ist, als bei anderen Unternehmen.

8.3 Eignung des VAD-Modells

Die Auswahl des VAD-Modells hat zahlreiche Vorteile während der Analyse des Unternehmenskorpus gezeigt. Besonders auffällig ist hier die einfache graphische Darstellbarkeit der Ergebnisse in Streudiagrammen. Durch die Repräsentation von Emotionen in einem dreidimensionalen Raum, ist ihre Verteilung zudem sehr gut begreifbar. Bei höherdimensionalen Daten, wie sie etwa bei der Verwendung von Ekmans Basisemotionen entstanden wären, wäre dies nicht unmittelbar gegeben.

In der Untersuchung des Unternehmenskorpus hat sich die Dominanzdimension als sehr wichtig herausgestellt, da sich in dieser Dimension Geschäftsberichte sehr gut von Nachhaltigkeitsberichten trennen lassen. Insofern war es angemessen, auf das 3D- und nicht auf das 2D-Modell zurückzugreifen. Die Frage nach der Wichtigkeit von Dominanz wird aktuell in psychologischen Literatur diskutiert (Bakker et al., 2014). Die vorliegende Arbeit liefert somit ein weiteres Argument zu deren Gunsten.

Die Orthogonalität von Valenz, Erregung und Dominanz erlaubt es zudem, übliche Vektoroperationen durchzuführen und Abstandmaße zu berechnen, was in anderen Repräsentationsformen fragwürdig ist (s. Abschnitt 3.2). Hierdurch lässt sich eine formale Repräsentation einer Emotion im VAD-Raum eindeutig einem empirischen Zustand zuordnen und umgekehrt. Diese Arbeit ist soweit bekannt die erste, die eine Dokumentenemotion als einen Vektor im VAD-Raum misst. Sie schließt damit eine Forschungslücke und und bietet Anknüpfungspunkte für weitere Untersuchungen. Dafür musste der Nachteil in Kauf genommen werden, keine quantitative Evaluation durchführen zu können.

9 Fazit und Ausblick

Die vorliegende Arbeit hat zunächst eine Präzisierung des Gegenstandsbereichs der Emotion Detection unter Rückgriff auf psychologische Konzepte vorgeschlagen und sie so von der Sentiment Analysis abgegrenzt. In einem zweiten Schritt wurden wesentliche Emotionsmodelle und -repräsentationen vorgestellt und kritisch diskutiert. Dies führte zur Entwicklung einer eigenen Formalisierung des Problems, die anders als bisherige Arbeiten vorsieht, Emotionen als Vektoren im VAD-Raum zu repräsentieren. Darauf aufbauend wurde eine quelloffene Software-Anwendung entwickelt. Diese verwendet ein aus der Psychologie stammendes Emotionslexikon. Anders als dessen Vorgänger wurde es noch nicht in der Computerlinguistik eingesetzt und verfügt über das Zehnfache an Einträgen. Da in Ermangelung eines Testkorpus keine quantitative Evaluation vorgenommen werden konnte, wurde das Werkzeug auf das bekannten RCV1-Korpus angewandt und die Ergebnisse positiv auf Plausibilität geprüft. In einem letzten Schritt wurde ein Korpus aus Unternehmens und Nachhaltigkeitsberichten großer, börsennotierten Unternehmen durch die Anwendung analysiert. Die Untersuchungsergebnisse sind ein starkes Argument dafür, dass Emotionen tatsächlich im Sinne der *Organizational Identity* ein dauerhaftes und distinktives Merkmal von Unternehmen als soziale Akteure sind.

Aufgrund ihres vielfältigen Inhalts eröffnet diese Arbeit zahlreiche Möglichkeiten für zukünftige Forschung. Für die weitere Entwicklung und Konsolidierung der Emotion Detection, wäre es wünschenswert, dass sich eine geringe Anzahl von Emotionsmodellen fest im Feld etabliert. Hierdurch würde Arbeiten in diesem Feld in einem größeren Maße miteinander vergleichbar werden. Aufgrund der dargelegten Vorteile des VAD-Modells wäre es ein Gewinn für die ED, wenn ein nach diesem Modell annotiertes Korpus zur freien Verfügung stehen würde. Hierfür bietet sich zum Beispiel an, das Testkorpus des SemEval-2007 zu annotieren. Da dieses schon entsprechend der Basisemotionen bewertet worden ist, ließe sich durch dessen Annotation nach dem VAD-Modell der Zusammenhang dieser beiden Emotionsmodelle weiter erforschen. Insbesondere könnten die Algorithmen, die diese beiden Repräsentationsformen ineinander überführen, deutlich verbessert werden. Da Emotionsannotation durch Crowdscourcing gute Ergebnisse liefert, wäre dieses Vorhaben auch mit vergleichsweise wenig Arbeit und finanziellen Mitteln umsetzbar. Ein weiteres Feld für anknüpfende Arbeit ist die Verbesserung des Werkzeugs durch Einbeziehung fortgeschrittener NLP-Methoden. Aufgrund der häufig guten Performanz, wäre hier vor allem die Verwendung überwachter Lernverfahren interessant. Dies würde allerdings ausreichend gelabelte Daten voraussetzen. Darüber hinaus bietet die Verwendung des VAD-Modells einen interessanten Vorteil zur weiteren Verbesserung des Tools: durch

die – zumindest unterstellte – Orthogonalität der Dimensionen lässt sich die Emotionserkennung in drei Teilprobleme aufteilen, nämlich das richtigen Bewerten von Valenz, Erregung und Dominanz. Dadurch, dass Valenz der semantischen Orientierung der Sentiment Analysis entspricht, lässt sich für die Lösung des ersten Teilproblems auf bestehende Methoden aus diesem Feld zurückgreifen. Durch die starke Korrelation von Valenz und Dominanz ist außerdem davon auszugehen, dass sich mit geringem Aufwand leistungsstarke Vorhersagemodelle für die Dominanz hierauf aufbauen lassen – auch wenn dieses Vorgehen in einem gewissen Spannungsverhältnis zur Orthogonalitätsannahme stehen würde. Im Ergebnis heißt das, dass sich wahrscheinlich für zwei der drei Teilprobleme mit relativ einfachen Mitteln performante Lösung finden lassen.

Für die Organisationsforschung wäre eine Ausweitung des Untersuchungsgegenstand auf finanziell erfolglose oder weniger lange bestehende Unternehmen interessant. Mögliche Fragestellungen hierbei wären, ob das Valenz-Erregungs-Diagramm die erwartete U-Form ausbildet bzw. an welcher Stelle die bisher untersuchten Unternehmen zu finden sein werden. Darüber hinaus muss geklärt werden, ob auch kleinere und weniger erfolgreiche Unternehmen distinktive und dauerhafte Emotionen aufweisen. Eine weitere interessante Fragestellung zielt darauf ab, die Wörter zu identifizieren, die den hohen Dominanzwert der CSR-Berichte ausmachen. Hierzu kann zum Beispiel die Schnittmenge der Wörter betrachtet werden, die einerseits in CSR-Berichten sehr viel häufiger vorkommen als etwa in Jahresberichten und andererseits einen hohen Dominanzwert aufweisen. Darüber hinaus sind große Teile der erhobenen Daten bisher kaum beachtet worden. Hier könnte etwa untersucht werden, wie sich die Streuung der Emotionswerte zwischen den Unternehmen unterscheidet und wie sie sich über die Zeit verändert. Ein denkbarer Anknüpfungspunkt wäre zum Beispiel die Beobachtung, dass die Dokumente des DJIA stärker streuen als die der anderen Aktienindizes. Offen ist bisher geblieben, ob dies für den gesamten betrachteten Zeitabschnitt gilt und ob im DJIA nur die Streuung der einzelnen Dokumente so hoch ist oder ob dies auch für die Mittelwerte der Dokumente eines Unternehmens gilt. Auch die erhobenen Daten zur internen Standardabweichung, mit der sich nachweisen lässt, wie unterschiedlich die Emotionen innerhalb eines einzelnen Dokuments sind, sind noch nicht beachtet worden.

Abschließend bleibt festzuhalten, dass diese Arbeit durch den Einsatz computerlinguistischer Verfahren einen wichtigen Beitrag zur Forschung zum Konzept *Organizational Identity* liefert. Gerade weil die hier eingesetzten Methoden verhältnismäßig einfach sind, verdeutlicht sie das enorme Potenzial solcher interdisziplinären Ansätze. Sie berechtigt daher zur Hoffnung, dass in der Zukunft zahlreiche fruchtbare Verbindungen von Computerlinguistik und Computational Social Science folgen.

Literaturverzeichnis

Acerbi, A., Lampos, V., Garnett, P. & Bentley, R. A. (2013). The expression of emotions in 20th century books. *PloS one*, *8* (3), e59030. doi: 10.1371/journal.pone.0059030

Agrawal, A. & An, A. (2012). Unsupervised Emotion Detection from text using semantic and syntactic relations. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT)* (Bd. 1, S. 346–353).

Baccianella, S., Esuli, A. & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)* (S. 2200–2204).

Bakker, I., van der Voordt, T., Vink, P. & de Boon, J. (2014). Pleasure, arousal, dominance: Mehrabian and Russell revisited. *Current Psychology*, *33* (3), 405–421.

Balahur, A., Hermida, J. M. & Montoyo, A. (2012). Building and exploiting EmotiNet, a knowledge base for emotion detection based on the appraisal theory model. *IEEE Transactions on Affective Computing*, *3* (1), 88–101.

Bestgen, Y. & Vincze, N. (2012). Checking and bootstrapping lexical norms by means of word similarity indexes. *Behavior Research Methods*, *44* (4), 998–1006.

Beyer, S., Bohn, S., Grünheid, T., Händschke, S., Kerekes, R., Müller, J. & Walgenbach, P. (2014). Wofür übernehmen Unternehmungen Verantwortung? Und wie kommunizieren sie ihre Verantwortungsübernahme? *Zeitschrift für Wirtschafts- und Unternehmensethik*, *15* (1), 57–80.

Bird, S. (2006). NLTK: The natural language toolkit. In *Proceedings of the COLING/ACL on Interactive Presentation Sessions* (S. 69–72).

Bishop, C. M. (2007). *Pattern recognition and machine learning*. New York: Springer.

Bollen, J., Mao, H. & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2* (1), 1–8.

Bradley, M. & Lang, P. (1999). Affective Norms for English Words (ANEW): Stimuli, instruction manual and affective ratings. *Technical report C-1, Gainesville, FL. The Center for Research in Psychophysiology, University of Florida.*.

Bradley, M. M. & Lang, P. J. (1994). Measuring emotion: the Self-Assessment Manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25* (1), 49–59.

Bradley, M. M. & Lang, P. J. (2007). Affective Norms for English Text (ANET): Affective ratings of text and instruction manual. *Techical Report. D-1, University*

of Florida, Gainesville, FL.

Broekens, J. (2012). In defense of dominance: PAD usage in computational representations of affect. *International Journal of Synthetic Emotions*, *3* (1), 33–42.

Calvo, R. A. & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, *1* (1), 18–37.

Calvo, R. A. & Mac Kim, S. (2013). Emotions in text: dimensional and categorical models. *Computational Intelligence*, *29* (3), 527–543.

Cambria, E., Schuller, B., Xia, Y. & Havasi, C. (2013). New avenues in Opinion Mining and Sentiment Analysis. *IEEE Intelligent Systems*, *28* (2), 15–21.

Canales, L. & Martínez-Barco, P. (2014). Emotion Detection from text: A survey. In *Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC).* Zugriff auf `https://www.aclweb.org/anthology/W/W14/W14-6905.pdf` (o.S.)

Carstensen, K.-U., Ebert, C., Ebert, C., Jekat, S., Klabunde, R. & Langer, H. (Hrsg.). (2010). *Computerlinguistik und Sprachtechnologie. Eine Einführung* (3. Aufl.). Heidelberg: Spektrum.

Danisman, T. & Alpkocak, A. (2008). Feeler: Emotion classification of text using vector space model. In *Proceedings of the AISB 2008 Symposium on Affective Language in Human and Machine* (S. 53–59).

Das, D. & Bandyopadhyay, S. (2009). Sentence level emotion tagging. In *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009* (S. 1–6).

Desmet, B. & Hoste, V. (2013). Emotion detection in suicide notes. *Expert Systems with Applications*, *40* (16), 6351–6358.

DiMaggio, P. (2015). Adapting computational text analysis to social science (and vice versa). *Big Data & Society*, *2* (2), 1–5.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6* (3-4), 169–200.

Elfenbein, H. A. (2007). Emotion in organizations. *Academy of Management Annals*, *1* (1), 315–386.

Goldenstein, J., Poschmann, P., Händschke, S. G. M. & Walgenbach, P. (2015). *The globally and locally embedded meaning of corporate responsibility.* (Unveröffentlichtes Arbeitspapier des Lehrstuhls für ABWL / Organisation, Führung und Human Resource Management der Universität Jena)

Gunes, H. & Schuller, B. (2013). Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision*

Computing, *31* (2), 120–136.

Gupta, N., Gilbert, M. & Fabbrizio, G. D. (2013). Emotion detection in email customer care. *Computational Intelligence*, *29* (3), 489–505.

Hajek, P., Olej, V. & Myskova, R. (2014). Forecasting corporate financial performance using sentiment in annual reports for stakeholders' decision-making. *Technological and Economic Development of Economy*, *20* (4), 721–738.

Hasan, M., Rundensteiner, E. & Agu, E. (2014). EMOTEX: Detecting emotions in Twitter messages. In *2014 ASE BIGDATA/SOCIALCOM/CYBERSECURITY Conference* (S. 27—31).

Jurafsky, D. & Martin, J. H. (2008). *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (2. Aufl.). Upper Saddle River: Prentice Hall.

Katz, P., Singleton, M. & Wicentowski, R. (2007). SWAT-MP: The SemEval-2007 systems for task 5 and task 14. In *Proceedings of the 4th International Workshop on Semantic Evaluations* (S. 308–313). Zugriff auf `http://dl.acm.org/citation.cfm?id=1621474.1621541`

King, B. G., Felin, T. & Whetten, D. A. (2010). Finding the organization in organizational theory. A meta-theory of the organization as a social actor. *Organization Science*, *21* (1), 290–305.

Lei, J., Rao, Y., Li, Q., Quan, X. & Wenyin, L. (2014). Towards building a social emotion detection system for online news. *Future Generation Computer Systems*, *37*, 438–448.

Leveau, N., Jhean-Larose, S., Denhière, G. & Nguyen, B.-L. (2012). Validating an interlingual metanorm for emotional analysis of texts. *Behavior Research Methods*, *44* (4), 1007–1014.

Lewis, D. D., Yang, Y., Rose, T. G. & Li, F. (2004). RCV1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, *5*, 361–397.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, *5* (1), 1–167. Zugriff auf `https://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf` (Online-Edition)

Manning, C. D., Raghavan, P. & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press. Zugriff auf `http://nlp.stanford.edu/IR-book/pdf/irbookprint.pdf` (Online-Edition)

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J. & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (S. 55–60). Zugriff auf `http://www.aclweb`

.org/anthology/P/P14/P14-5010

Matten, D. & Moon, J. (2008). ‚Implicit' and ‚explicit' CSR: A conceptual framework for a comparative understanding of Corporate Social Responsibility. *Academy of Management Review*, *33*, 404–424.

Mohammad, S. M. & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, *29* (3), 436–465.

Munezero, M., Montero, C. S., Sutinen, E. & Pajunen, J. (2014). Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Transactions on Affective Computing*, *5* (2), 101–111.

Neviarouskaya, A., Prendinger, H. & Ishizuka, M. (2011). Affect Analysis Model: novel rule-based approach to affect sensing from text. *Natural Language Engineering*, *17* (01), 95–135.

Orlitzky, M., Schmidt, F. L. & Rynes, S. L. (2003). Corporate social and financial performance: A meta-analysis. *Organization Studies*, *24*, 403–441.

Ortman, G. (2010). *Organisation und Moral. Die dunkle Seite*. Weilerswist: Velbrück.

Palmer, M. & Xue, N. (2010). Linguistic annotation. In A. Clark, C. Fox & S. Lappin (Hrsg.), *Handbook of Computational Linguistics and Natural Language Processing* (S. 238–270). New York: John Wiley & Sons.

Pang, B. & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and trends in Information Retrieval*, *2* (1-2), 1–135.

Pennebaker, J. W., Mehl, M. R. & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, *54* (1), 547–577.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14* (3), 130–137.

Rao, Y., Li, Q., Mao, X. & Wenyin, L. (2014). Sentiment topic models for social emotion mining. *Information Sciences*, *266*, 90–100.

Rasch, B., Friese, M., Hofmann, W. J. & Naumann, E. (2010a). *Quantitative Methoden. Einführung in die Statistik für Psychologen und Sozialwissenschaftler* (3. Aufl., Bd. 1). Berlin: Springer.

Rasch, B., Friese, M., Hofmann, W. J. & Naumann, E. (2010b). *Quantitative Methoden. Einführung in die Statistik für Psychologen und Sozialwissenschaftler* (3. Aufl., Bd. 2). Berlin: Springer.

Roberts, K., Roach, M. A., Johnson, J., Guthrie, J. & Harabagiu, S. M. (2012). EmpaTweet: Annotating and detecting emotions on Twitter. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12).*

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*, 1161–1179.

Russell, J. A. & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, *11* (3), 273–294.

Scherer, K. R. (2000). Psychological models of emotion. In J. Borod (Hrsg.), *The neuropsychology of emotion* (S. 137–162). New York: Oxford University Press.

Snow, R., O'Connor, B., Jurafsky, D. & Ng, A. Y. (2008). Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing* (S. 254–263).

Sokolova, M. & Lapalme, G. (2009, Juli). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *45* (4), 427–437. Zugriff auf `http://dx.doi.org/10.1016/j.ipm.2009.03.002` doi: 10.1016/j.ipm.2009.03.002

Staiano, J. & Guerini, M. (2014). Depeche Mood: A lexicon for emotion analysis from crowd annotated news. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Bd. 2, S. 427–433). Zugriff auf `http://www.aclweb.org/anthology/P/P14/P14-2070`

Stevenson, R. A., Mikels, J. A. & James, T. W. (2007). Characterization of the Affective Norms for English Words by discrete emotional categories. *Behavior Research Methods*, *39* (4), 1020–1024.

Strapparava, C. & Mihalcea, R. (2007). SemEval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations* (S. 70–74). Zugriff auf `http://dl.acm.org/citation.cfm?id=1621474.1621487`

Strapparava, C. & Mihalcea, R. (2008). Learning to identify emotions in text. In *Proceedings of the 2008 ACM Symposium on Applied Computing (SAC)* (S. 1556–1560).

Strapparava, C. & Valitutti, A. (2004). WordNet-Affect: an affective extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)* (S. 1083–1086).

Warriner, A. B., Kuperman, V. & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45* (4), 1191–1207.

Whetten, D. A. (2006). Albert and Whetten revisited: Strengthening the concept of Organizational Identity. *Journal of Management Inquiry*, *15*, 219-234.

Abkürzungsverzeichnis

AMT	Amazon Mechanical Turk
ANEW	Affective Norms for English Words (Emotionslexikon)
ANET	Affective Norms for English Text
ANN	jährliche Geschäftsberichte (Annual Reports)
BOW	Bag-of-Words(-Modell)
CSR	Nachhaltigkeitsberichte (Corporate Social Responsibility Reports)
DAX	Deutscher Aktienindex
DJIA	Dow Jones Industrial Average (US-amerikanischer Leitindex)
ED	Emotion Detection
FTSE	Financial Times Stock Exchange 100 Index (britischer Leitindex)
IAA	Inter-Annotator Agreement
IR	Information Retrieval
kNN	k-Nächste Nachbarn (k-Nearest Neighbors) (Klassifikator)
LSA	Latente Semantische Analyse
NB	Naive Bayes (Klassifikator)
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NMF	Nicht-negative Matrixfaktorisierung
RCV1	Reuters Corpus Volume 1
SA	Sentiment Analysis
SAM	Self-Assessment-Manikin
SVM	Support Vector Machine (Klassifikator)

n	absolute Häufigkeit (in Tabellen)
%	prozentuale relative Häufigkeit (in Tabellen)
Min	Minimum
Max	Maximum
M	arithmetisches Mittel
SD	Standardabweichung
r	Pearson-Korrelation (Zusammenhangsmaß)
p	p-Wert (Signifikanzwert)
d	Cohens d (Effektstärkemaß)
η^2	Eta-Quadrat (Effektstärkemaß)
tf-idf	Termhäufigkeit – inverse Dokumentenhäufigkeit (term frequency – inverse document frequency) (Gewichtungsfunktion in der IR)
P	Precision (Genauigkeit)
R	Recall (Abdeckung)
F	F-Maß (gewichtetes harmonisches Mittel aus P und R)
κ	Cohens Kappa (Maß für die IAA)

Abbildungsverzeichnis

Tabellenverzeichnis

Anhang: Zusätzliche Tabellen und Abbildungen

	Valenz	Erregung	Dominanz
Valenz	1	-0.19	0.72
Erregung		1	- 0.18
Dominanz			1

Tabelle 14: Korrelationsmatrix der Emotionskomponenten für Warriners Emotionslexikon (Pearson-Korrelation).

	Valenz	Erregung	Dominanz
Valenz	1	-0.04	0.68
Erregung		1	-0.05
Dominanz			1

Tabelle 15: Korrelationsmatrix der Emotionskomponenten für das RCV1 (Pearson-Korrelation).

	Valenz	Erregung	Dominanz
Valenz	1	-0.03	0.87
Erregung		1	-0.09
Dominanz			1

Tabelle 16: Korrelationsmatrix der Emotionskomponenten für das Unternehmenskorpus (Pearson-Korrelation).

		M_{Dax}	M_{DJIA}	M_{FTSE}	SD_{DAX}	SD_{DJIA}	SD_{FTSE}	p	η^2
alle	Valenz	0.64	0.67	0.64	0.07	0.12	0.07	< 0.001	0.027
	Erregung	-1.02	-1.00	-1.02	0.05	0.04	0.04	< 0.001	0.036
	Dominanz	0.61	0.63	0.63	0.06	0.09	0.06	< 0.001	0.020
ANN	Valenz	0.62	0.62	0.62	0.04	0.10	0.04	0.5426	–
	Erregung	-1.02	-1.00	-1.02	0.03	0.04	0.03	< 0.001	0.073
	Dominanz	0.59	0.58	0.60	0.03	0.06	0.03	< 0.001	0.049
CSR	Valenz	0.67	0.76	0.68	0.11	0.10	0.09	< 0.001	0.138
	Erregung	-1.03	-1.00	-1.02	0.07	0.04	0.06	0.0020	0.021
	Dominanz	0.66	0.72	0.68	0.08	0.07	0.06	< 0.001	0.110

Tabelle 17: Einfluss der Herkunft auf den Emotionsgehalt der Dokumente im Unternehmenskorpus: ohne Drittvariablenkontrolle (alle), nur Geschäftsberichte (ANN), nur Nachhaltigkeitsberichte (CSR).

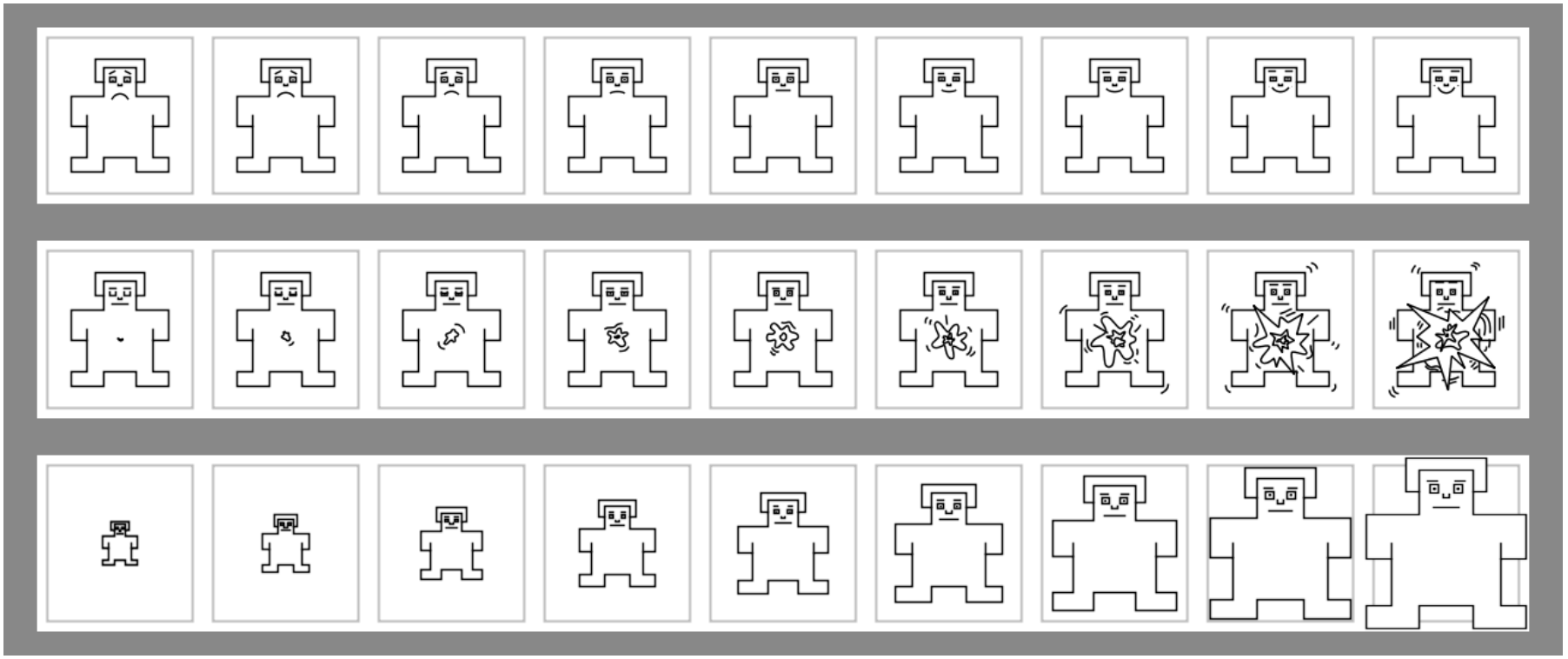

Abbildung 14: Self-Assessment-Manikin (SAM) (M. M. Bradley & Lang, 1994). Piktogramme zur sprachfreien Angabe von Emotionen im VAD-Raum. Die drei Reihen stellen jeweils eine neunstufige Skala für (von oben nach unten) Valenz, Erregung und Dominanz dar. Abgerufen am 31.01.2016 unter `http://irtel.uni-mannheim.de/pxlab/demos/index_SAM.html`.

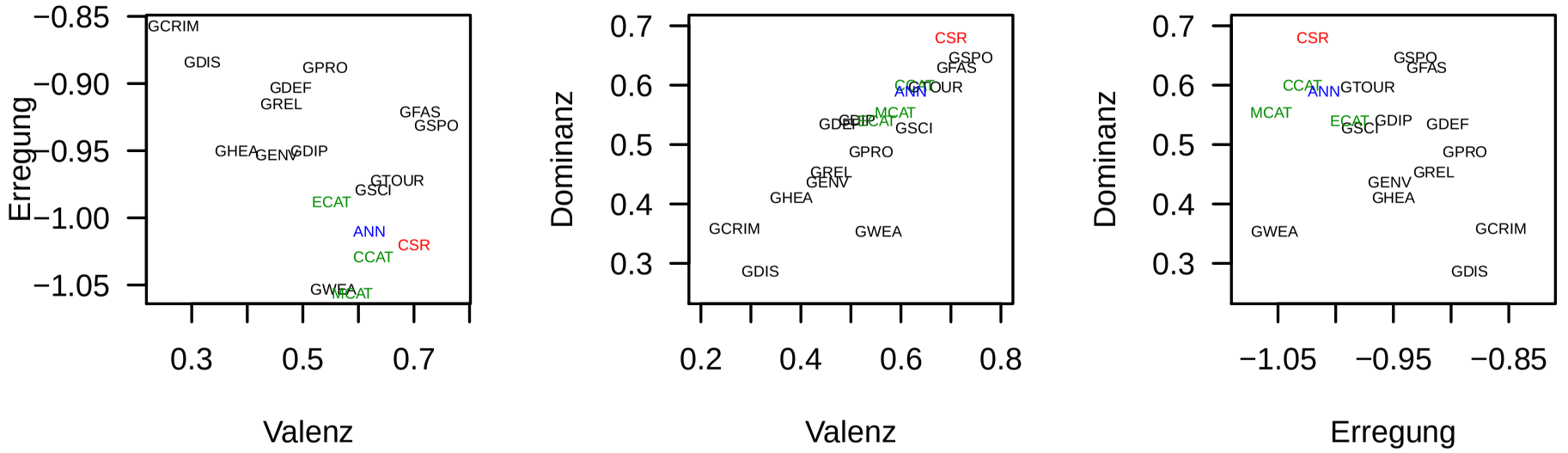

Abbildung 15: Streudiagramme der Kategorienmittelwerte des RCV1 zusammen mit dem Mittelwert der Geschäftsberichte (ANN) und der Nachhaltigkeitsberichte (CSR) des Unternehmenskorpus.

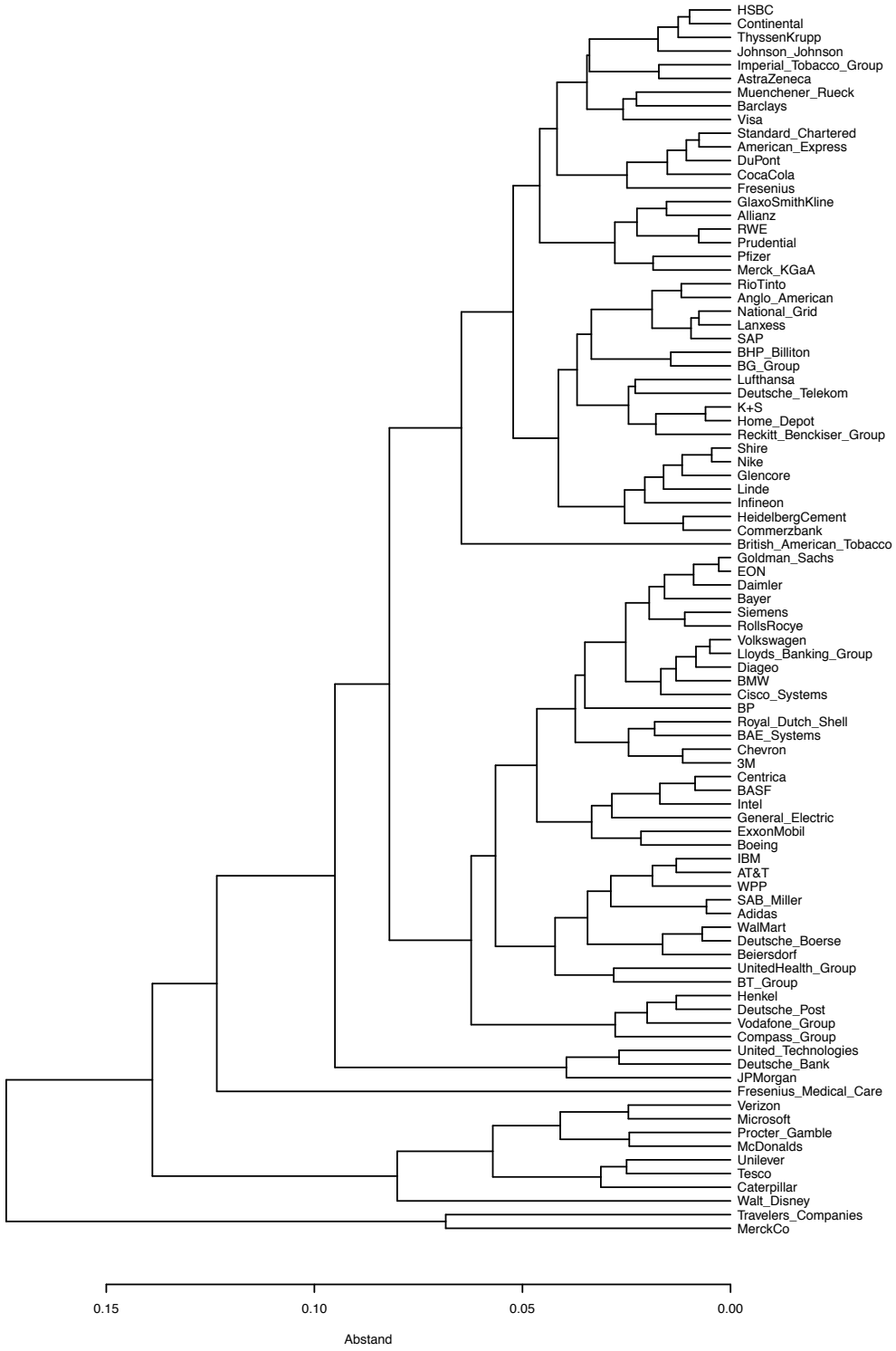

Abbildung 16: Dendrogramm der hierarischen Clusteranalyse der Unternehmensmittelwerte basierend auf euklidischem Abstand und Average-Group-Linkage.

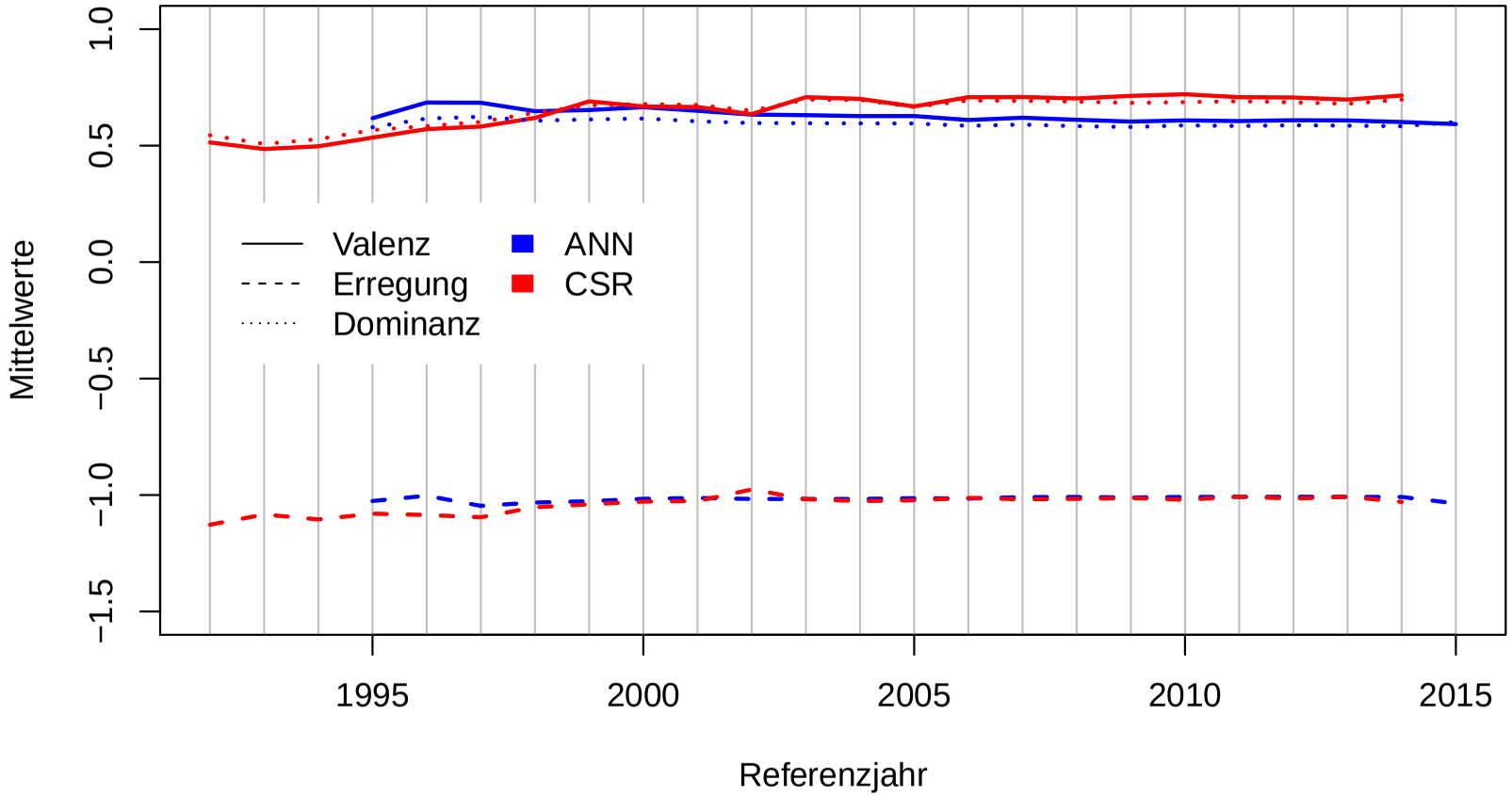

Abbildung 17: Zeitreihe der Jahresmittelwerte der Emotionskomponenten im Unternehmenskorpus aufgegliedert nach Gattung.

Gattung	Herkunft	Unternehmen	Jahr	Valenz	Erregung	Dominanz	Norm	Erkennung
CSR	DAX	Deutsche Bank	2002	-0.40	-0.19	-0.12	23.98	0.13
ANN	DJIA	McDonald's	2006	-0.22	-1.30	0.08	13.92	0.01
CSR	FTSE	BT Group	2010	1.18	-1.57	0.50	13.71	0.00
CSR	FTSE	BT Group	2009	0.88	-0.55	0.53	10.65	0.04
CSR	DJIA	Walmart	2005	0.30	-1.07	0.22	6.99	0.00
CSR	DAX	Linde	2005	0.14	-1.12	0.45	6.60	0.01
CSR	DJIA	Boeing	2013	1.03	-0.91	0.93	6.54	0.74
CSR	DJIA	Microsoft	2007	0.99	-0.93	0.90	5.82	0.80
CSR	DAX	Infineon	2011	0.73	-0.76	0.64	5.69	0.01
ANN	DJIA	McDonald's	2005	1.04	-1.01	0.86	5.56	0.74

Tabelle 18: Darstellung der zehn stärksten Ausreißer des Unternehmenskorpus gemessen an der euklidischen Norm der z-transfomierten Emotionswerte (Norm). Zusätzlich wird die Erkennungsrate (Erkennung) mit angegeben.

Arbeit	Zielvariable	Emotionsmodell	Ansatz
Roberts et al. (2012)	diskret	diskret	überwachtes Lernen
Agrawal und An (2012)	diskret	diskret	linguistisch-regelbasiert
Calvo und Mac Kim (2013)	diskret	diskret*	unüberwachtes Lernen
Danisman und Alpkocak (2008)	diskret	diskret	unüberwachtes Lernen
Desmet und Hoste (2013)	diskret	diskret	überwachtes Lernen
Balahur et al. (2012)	diskret	diskret	wissensbasiert
Hasan et al. (2014)	diskret	dimensional	überwachtes Lernen
Neviarouskaya et al. (2011)	beides	diskret	linguistisch-regelbasiert
Strapparava und Mihalcea (2008)	beides	diskret	Keyword Spotting
Strapparava und Mihalcea (2008)	beides	diskret	überwachtes Lernen
Strapparava und Mihalcea (2008)	beides	diskret	unüberwachtes Lernen
Staiano und Guerini (2014)	beides	diskret	lexical affinity
Katz et al. (2007)	beides	diskret	lexical affinity

* Calvo und Mac Kim (2013) verwenden in einem Zwischenschritt das VAD-Modell, aber bilden dies abschließend wieder auf ein Teilmenge von Ekmans Basisemotionen ab.

Tabelle 19: Literatur zur textuellen Emotion Detection nach Form der Zielvariable (diskret oder stetig), Emotionsmodell (diskret oder dimensional) und methodischem Ansatz.

Eigenständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegeben Hilfsmittel und Quellen angefertigt habe. Die eingereichte Arbeit ist nicht anderweitig als Prüfungsleistung verwendet worden oder in deutscher oder einer anderen Sprache als Veröffentlichung erschienen. Seitens des Verfassers bestehen keine Einwände, diese Arbeit für die öffentliche Benutzung zur Verfügung zu stellen.

Jena,

Sven Eric Büchel