Technische Universität Wien
Institut für Softwaretechnik und Interaktive Systeme

Emotionen in deutschen Texten: Ein quantitativer Ansatz mit GATE

Seminar (mit Bachelorarbeit)
von

Elisabeth Weigl

17. Oktober 2008

Betreuer: Ao.Univ.Prof. Dipl.-Ing. Dr.techn. A. Rauber

Eidesstattliche Erklärung

Ich erkläre eidesstattlich, dass ich die Arbeit selbständig angefertigt, keine anderen als die angegebenen Hilfsmittel benutzt und alle aus ungedruckten Quellen, gedruckter Literatur oder aus dem Internet im Wortlaut oder im wesentlichen Inhalt übernommenen Formulierungen und Konzepte gemäß den Richtlinien wissenschaftlicher Arbeiten zitiert, durch Fußnoten gekennzeichnet bzw. mit genauer Quellenangabe kenntlich gemacht habe.

Abstract

Diese Arbeit behandelt die Suche nach der Emotionalität eines Textes. Nicht jeder Text, dessen Absicht es sein sollte, so sachlich und subjektiv wie möglich zu sein, erfüllt auch diese Anforderungen. Eine einfache Messung, die Feedback darüber liefert, ob das Geschriebene der Intention des Schreibers entspricht, wäre wohl die einfachste Methode. Ab wann kann man also von einem zu gefülsbetonten und daher nicht mehr sachlichen Text sprechen? Wie kann man diese Eigenschaft messen, welche Kriterien müssen erfüllt sein? Mit Hilfe von GATE, einem Programm zur Verarbeitung natürlicher Sprache, werden ausgewählte Texte analysiert und so eine Skala geschaffen, durch die man ablesen kann, welche Tendenz ein Text aufweist.

Inhaltsverzeichnis

1 Einführung

1.1 Motivation

Kurz bevor ich mich zu dem Thema der Arbeit entschloss, habe ich in einer Internetbewertungsplattform meine Meinung über ein Produkt und das Geschäft, in dem ich dieses erworben hatte, kundgetan. Dazu habe ich meine Bewertungen mehrmals gelesen, um sie möglichst objektiv zu halten und nicht meine Emotionen über das eher schlecht gelaufene Geschäft einfließen zu lassen. Es stellte sich mir öfters die Frage, ob ich, trotz meiner Gefühlsausbrüche vor dem Bildschirm eine möglichst aussagekräftige Beurteilung geschrieben hatte. Schließlich soll nicht jemand, der diese Bewertung liest, denken, sie wäre im Affekt geschrieben. Es sollte klar sein, dass ich über meine Worte nachgedacht habe. Andererseits sollte meine Meinung dennoch beim Leser ankommen. Da kam mir ein Thema mit dem Inhalt, ob eine solche Bewertung zu emotional ist oder nicht, gerade recht.

Im Internet mag eine objektive Bewertung eine gute Sache sein. Im wissenschaftlichen Umfeld ist es hingegen zwingend erforderlich, Texte so sachlich wie möglich zu verfassen um den Leser nicht zu beeinflussen. Dazu könnte man versuchen, so wenige Eigenschaftswörter wie möglich zu verwenden, jedoch kann man sich nie ganz sicher sein, nicht doch die eigene Ansicht durch die Verwendung von genau dem einen oder anderen Wort einfließen zu lassen. Und gerade dann, wenn es darum geht, seine eigene Meinung zu vertreten, verfällt ein impulsiver Schreiber schnell in einen zu emotinalen Schreibstil, den er so nicht gewünscht hatte. Die Frage, ab wann es "zuviel" davon ist und wie man das objektiv messen kann, kommt auf.

1.2 Zielsetzung

Ziel der Arbeit ist es, herauszufinden, inwieweit diese Emotionalität in einem Text wirklich messbar ist und wie man sie einteilen kann. Die Frage wie soetwas objektiv messbar ist, lässt sich schnell klären: mit einem Programm, das Routinen befolgt und sich nicht vom täglichen Wohlbefinden beeinflussen lässt.

Die Herausforderung besteht zum einen darin, gefühlsbetonende Wörter ausfindig zu machen und zum anderen eine geeignete Bewertungsskala zu erstellen, die Aufschluss über die Emotionalität des Textes gibt.

1.3 Aufbau der Arbeit

Die vorliegende Arbeit beschreibt in Kapitel 2 (Grundlagen) verschiedene Seiten des Themas. Dazu werden zunächst die informationstechnischen Herangehensweisen an Textanalyse besprochen. Hier gehört Text Mining (2.2) dazu, dessen statistische und musterbasierte Methoden vorgestellt werden. Da dieses Kapitel viele grundlegende Themen enthält, basiert es auf dem Buch "Text Mining: Wissensrohstoff Text" von Gerhard Heyer [12]. Weiters wird versucht die Begriffe *Sentiment Analysis* und *Opinion Mining* in Kapitel 2.3 zu klären, sowie ein erster Zusammenhang zwischen diesen Ausdrücken und dem Thema dieser Arbeit zu ziehen. Im Kapitel 2.5 (Germanistik) wird auf Basis der Erkenntnisse von Silke Jahr [14] die germanistische Seite betrachtet. Ebenso wird erklärt, was das *Affektive Diktionär Ulm* und das *Dresdner Angstwörterbuch* sind. Das darauffolgende Kapitel 2.6 (Psychologie) beinhaltet ebenfalls Teile dieses Buches und soll ausschließlich psychologische Komponenten dieses Themas behandeln. Im Kapitel 3 (Umsetzung) wird zunächst auf die Theorie 3.1 der Emotionsanalyse eingegangen. Dazu werden die Ansätze von vorangegangenen, für die englische Sprache geschriebenen Publikationen erläutert um so auf den in dieser Arbeit verwendeten Ansatz zu gelangen. Danach wird die verwendete Architektur GATE in Kapitel 3.3 vorgestellt und im letzten Unterkapitel 3.4 (Praxis) die Anwendung des Ansatzes erklärt. Am Ende werden die Ergebnisse der Arbeit (4) erläutert, in dem Kapitel "Diskussion" (5) wird auf die Probleme der Textanalyse sowie die Möglichkeiten, die Verbesserungen einzelner Punkte bringen, eingegangen.

2 Grundlagen

2.1 Informatik

Der Begriff Informatik bezeichnet die "mathematische Lehre, die sich mit den Gesetzen bei der Übermittlung, Verarbeitung u. Wiedergewinnung von Informationen befasst" [20]. Im vielgenannten Informationszeitalter sind die Begriffe Information und Wissen zu alltäglichen Wörtern geworden. Information bezeichnet "Daten, die in einem Kontext interpretiert werden und somit eine Bedeutung für den Besitzer oder Empfänger haben", Wissen ist die "meist auf Erfahrung beruhende und objektiv nachprüfbare Kenntnis von Fakten und Zusammenhängen eines Weltausschnitts" [12]. Interessanter sind mittlerweile die Ausdrücke Wissensakquisition und Wissensmanagement. Der Umgang mit riesigen Datenmengen, die sich z.B. im Internet jeden Tag vervielfachen (Heyer [12] schreibt von ca. einer Million neuer [Text-] Dokumente pro Tag), ist entscheidend, um daraus Information zu gewinnen und in weiterer Folge Wissen zusammenzutragen.

2.2 Text Mining

Der Ausdruck *Text Mining* soll den Bezug zu *Data Mining* herstellen. Data Mining, zu Deutsch die "Datenschürfung" (von engl. mining - Bergbau), bezeichnet den "Entdeckungsprozess nützlicher Modelle oder nützlichen Wissens von Datenquellen" [15], die meistens Datenbanken mit verschiedensten Arten von Daten darstellen. Benötigt werden dazu Programme, die diese Aufgaben automatisch und schnell erledigen.

Im Text Mining geht es ausschließlich um Daten, die in einem Textformat vorliegen. Dieser "Wissensrohstoff Text", wie G. Heyer es immer wieder erwähnt, ist aber meistens unstrukturiert im Vergleich zu Daten aus einer Datenbank, die Relationen und damit Tabellen, Primarykeys, Foreignkeys, Tupeln etc. enthält. Ein menschlicher Leser bringt von selbst eine Struktur in den von ihm gelesenen Text um ihn zu verstehen und daraus Information zu erhalten, da er es so gelernt hat. Einem computerunterstütztem Verfahren jedoch muss das zuerst noch beigebracht werden. Charakteristische Grundkonzepte müssen "identifiziert, in einen systematischen Zusammenhang gebracht und instantiiert werden" [12]. Text Mining umfasst daher "Verfahren für die semantische Analyse von Texten, welche die automatische bzw. semi-automatische Strukturierung von Texten unterstützen" [12].

Diese Analysen werden mit Hilfe von statistischen oder musterbasierten Verfahren erstellt.

2.2.1 Statistische Verfahren

Hier werden sprachstatistische Gesetzmäßigkeiten auf einen zu analysierenden Text angewandt um ihm gewisse Merkmale zuschreiben zu können.

Zipfsche Gesetze

Die Zipfschen Gesetze setzen grundsätzlich den Rang eines Objekts mit dessen Wert in Verbindung. Im linguistischen Fall wird vom "Prinzip der geringsten Anstrengung" ausgegangen, nach dem der deutsche Wortschatz zumeist aus Funktionswörter wie *der*, *die* oder *und* (die drei häufigsten Wörter der deutschen Sprache) besteht. Der Rang r dieser Wörter multipliziert mit ihrer Häufigkeit n ist laut Heyer textabhängig konstant k.

$$r \cdot n \approx k \tag{2.1}$$

Die Häufigkeit selbst ist wiederum umgekehrt proportional zu ihrem Rang in der Häufigkeitsliste:

$$n \sim \frac{1}{r} \tag{2.2}$$

In der Praxis zeigt sich, dass diese Gesetze die Wirklichkeit nur annäherend repräsentieren. Eine Formel nach Mandelbrot liefert bessere Ergebnisse. In der Regel wird bei Analysen ohnehin von idealisierten Textstücken ausgegangen.

Anhand dieser Regeln lässt sich (bei gegebener Textlänge) die Anzahl der Wortformen, die darin n-Mal vorkommen bestimmen:

$$r_n = c \cdot \frac{N}{n} \tag{2.3}$$

r_n bezeichnet hier den größten Rang aller Wortformen, die genau n-Mal vorkommen, c ist eine Konstante, die von der Einzelsprache, aber nicht vom analysierten Text abhängt und $\frac{N}{n}$ ist die relative Häufigkeit der Wortform, wobei N die Anzahl der Wortformen eines Textes ist und n die Anzahl einer konkreten Wortform, die darin vorkommt.

Die Konstante c ist demnach allgemein und nach der Gleichung 2.1

$$c = \frac{k}{N} \tag{2.4}$$

Nach Heyer, dessen Zahlen aus dem "Projekt Deutscher Wortschatz"[1] der Universität Leipzip sind, beträgt $k = 18.000.000$ und $N = 222.000.000$ für die deutsche Sprache, wodurch sich eine Konstante $c \approx 0,08$ ergibt.

Interessant ist die in der Literatur berichtete Annahme, dass die Länge l eines Worts umgekehrt proportional zu ihrer Häufigkeit sein soll:

$$n \sim \frac{1}{l} \tag{2.5}$$

Heyer kann dies mit den Daten des deutschen Wortschatzs jedoch nicht bestätigen. [12]

[1] http://wortschatz.uni-leipzig.de/

Differenzanalyse

Die Differenzanalyse benötigt zwei grundlegende Objekte: einen Analysekorpus und einen Referenzkorpus. Ein Korpus besteht aus einer Menge von Texten, die im Fall des Analysekorpus die zu extrahierende Terme enthalten und im Fall des Referenzkorpus allgemeinsprachliche Dokumente sind. Die Verteilung von Wortformen im Text lässt sich in vier Klassen unterteilen:

- Klasse 1: Fachausdrücke, die im Referenzkorpus nicht vorkommen
- Klasse 2: Wortformen, die mit einer gewissen Wahrscheinlichkeit Fachterme sind; kommen im Analysekorpus relativ häufiger vor
- Klasse 3: Stoppwörter ((un)bestimmte Artikel, Präpositionen, Konjunktionen); treten in beiden Korpora mit ungefähr der gleichen relativen Häufigkeit auf
- Klasse 4: keine Fachterme; treten im Analysekorpus seltener als im Referenzkorpus auf

Für die Terminologieextraktion sind Klasse eins und zwei interessant. Das Projekt Deutscher Wortschatz hat eine Formel zur Berechnung der Häufigkeitsklasse HKL einer Wortform w mit $|w|$ als die Anzahl der Vorkommen von w entwickelt:

$$HKL(w) = ganzerAnteil\left(\log_2\left(\frac{|'\,der'\,|}{|\,w\,|}\right)\right) \tag{2.6}$$

Die Wortform *der* ist wie bereits erwähnt die häufigste Wortform in der deutschen Sprache. Das Ergebnis der Formel drückt aus, dass *der* ungefähr $2^{HKL(w)}$-Mal häufiger vorkommt, als w. Mit Hilfe dieser Häufigkeitsberechnung lassen sich die Häufigkeiten von Worten in einem zu analysierenden Fachtext mit denen aus einem allgemeinsprachlichen Text vergleichen. Die im Analysekorpus relativ häufiger vorkommenden Wortformen sind damit die gesuchten Fachausdrücke. Mit Filteroperationen (z.B. ausschließliche Selektion von Nomen) lassen sich die Ergebnisse noch weiter verbessern. [12]

Probabilistisches Sprachmodell

Das probabilistische Sprachmodell beschäftigt sich mit der Frage, ob ein Satz S ein korrekter Satz einer Sprache L ist. Dazu wird eine Grammatik benötigt, die den Satz auf syntaktische Korrektheit überprüft. Aber der Satz muss auch semantisch sinnvoll sein, was zur Zeit noch nicht ausreichend berechnet werden kann.

Heyer beschreibt einen Ansatz, mit dem sich die Wahrscheinlichkeit für die Zugehörigkeit von S zu L berechnen lässt. Dazu existiert ein Trainingskorpus bestehend aus unterschiedlichen Wortformen. Die bedingte Wahrscheinlichkeit für das Auftreten des Worts w_j direkt vor dem Wort w_i ist

$$p(w_j \mid w_i) = \frac{p(w_i, w_j)}{p(w_i)} \tag{2.7}$$

Er erweitert diese Formel für n aufeinanderfolgende Worte, also einen Satz S der Länge n. Da dies ein zu komplexes und dabei starres Konstrukt wird, wird ein approximativer Weg eingeschlagen: Wortformen werden zu Phrasen gruppiert, es werden deshalb nur höchstens zwei Vorgänger berücksichtigt. Mit diesen Bi- bzw. Trigrammen wird die Komplexität reduziert, da nur mehr Dreierkombinationen untersucht werden müssen. Heyer führt weiters noch zwei Marker für den Satzanfang ein, um auch beim ersten Wort eines Satzes mit Trigrammen arbeiten zu können. Dem Problem von seltenen Wortformen entgeht er mit Smoothing, das noch Gewichtungsfaktoren hinzufügt. Letztendlich führen nur noch Wortformen, die nicht im Trainingskorpus vorkommen, zu einem unerwünschten Ergebnis.

Auch bei der Aufgabe, ob einer Folge S von Symbolen (einer Symbolmenge M_1) eine Folge T von Symbolen (der Symbolmenge M_2) zugeordnet werden kann, wird das probabilistische Sprachmodell verwendet. Dabei gibt es drei Problemstellungen:

- **maschinelles Übersetzen**: Ein Satz S der Quellsprache wird einem Satz T der Zielsprache zugeordnet
- **Part-of-Speech (POS) Tagging**: Wörter werden in Kategorien eingeteilt, T ist dann die zu bestimmende Folge von Kategorien
- **Spracherkennung**: S ist eine Folge von akustischen Signalen, denen eine Folge T von Wortformen zugeordnet wird

Für die Berechnung der Wahrscheinlichkeit, dass S aufgetreten und T zugeordnet wird, wird die Bayes'sche Formel für die bedingte Wahrscheinlichkeit herangezogen:

$$p(T \mid S) = \frac{p(T) \cdot p(S \mid T)}{p(S)} \tag{2.8}$$

Für die Wahrscheinlichkeit, dass die Aussprache einer Wortform w zu einer Lautfolge l (z.B. [fi:l] könnte "fiel", "viel" oder sogar "fehl" meinen) führt, nimmt Heyer folgende Formel:

$$w' = \arg\max(p(w) \cdot p(l \mid w)) \quad w \in W \tag{2.9}$$

Dem Problem beim POS-Tagging, dass Wortformen nicht immer in die richtige Kategorie eingeteilt werden (z.B. *das* kann sowohl Artikel als auch Relativpronomen sein), wird ebenfalls mit bedingter Wahrscheinlichkeit begegnet. Die wahrscheinlichste Kategorie k' aus der Menge der Kategorien K, der die Wortform w angehört ist $k' = \arg\max p(k \mid w), \quad k \in K$. Die Wahrscheinlichkeiten dafür, dass w zur Kategorie K gehört wird wiederum mit einem händisch getaggten Trainingskorpus berechnet. Dies kann jedoch, wie im Fall von "das", das im Beispiel von Heyer nur zu einer Wahrscheinlichkeit von $\frac{2}{3}$ ein Artikel ist, zu einem falschen Tagging führen.

Deshalb muss der Kontext, in dem die Wortform auftritt, betrachtet werden. Hierzu muss die Wahrscheinlichkeit berechnet werden, mit der im Trainingskorpus z.B. ein Relativpronomen bzw. ein Artikel auf einen Eigennamen folgt. Da im angegebenen Trainingskorpus kein Relativpronomen auf einen Eigennamen, sehr wohl aber ein Artikel auf einen Eigennamen folgt, führt die Anwendung der Bayes'schen Formel diesmal zur richtigen Kategorisierung. [12]

Signifikante Kookkurrenzen

Kookkurrenz ist das gemeinsame Auftreten zweier Wortformen in einem Satz. Ist dieses Vorkommen statistisch auffällig, wird von *signifikanten Kookkurrenzen* gesprochen. Dabei unterscheidet man

- **Nachbarschaftskookkurrenzen**, falls die beiden Wortformen direkt aufeinanderfolgen und
- **Satzkookkurrenzen**, wenn sie in einem Satz auftreten.

Nach Heyer kann auch von Textfenstern gesprochen werden, wenn z.B. fünf Worte vor oder nach einer gewissen Wortform nach einem Kookkurrenten gesucht werden.

Das Ergebnis der Signifikanz soll die menschliche Intuition in Bezug auf die Zusammengehörigkeit von Wortformen widerspiegeln. Dazu wird jedem Wortpaar ein Signifikanzwert zugeordnet. Liegt dieser Wert über einer festgelegten Schwelle, gilt das Paar als signifikant.

Zur Berechnung der Signifikanz benutzt Heyer die Poisson-Verteilung. Zunächst gibt es vier Größen:

- a bezeichnet die Anzahl der Sätze, in denen die Wortform A vorkommt,
- b die Anzahl, in denen die Wortform B vorkommt,
- k ist die Anzahl, in der sowohl A als auch B vorkommen und
- n ist die Gesamtanzahl aller Sätze.

Der Wert λ ist $\frac{ab}{n}$.

Heyer gibt an, dass typischerweise $\frac{(k+1)}{\lambda} > 2,5$ und $k > 10$ ist, wodurch sich die Signifikanz auf folgende Formel verkürzen lässt:

$$sig(A,B) \approx \frac{k \cdot (\log k - \log \lambda - 1)}{\log n} \tag{2.10}$$

Visualisieren lassen sich signifikante Kookkurrenzen durch Wortnetze. Diese enthalten einen zentral gelegenen Ausgangsbegriff, der mit anderen Wörtern durch eine Kante verbunden ist, sofern die Signifikanz der beiden Wörter einen festgelegten Schwellenwert überschreitet. Weiters wird die Linienstärke dicker, je größer der Signifikanzwert zweier Wörter ist. Mit dem Verfahren "simulated annealing" wird die Anordnung der Knoten bzw. Wörter optimiert, sodass inhaltlich zusammenhängende Wortformen möglichst nah beieinander stehen. Werden dann die entsprechenden Kanten für die Signifikanz eingefügt, entsteht ein ungerichteter Graph.

Interessante Anwendungsgebiete der signifikanten Kookkurrenzen sind z.B. die Suche nach Subsprachen in Texten. Ein einzelnes Wort z.B. in Mundart tritt häufig mit anderen Worten in Mundart auf. So lassen sich mit Hilfe der Kookkurrenzen ganze Wortfelder oder Textteile in einer anderen Sprache lokalisieren. Als zweites Anwendungsbeispiel nennt Heyer die Polysemie (Mehrdeutigkeit von Worten) wie z.B. beim Wort **Bank**, das für ein

Geldinstitut oder eine Sitzgelegenheit stehen kann. Hier werden die signifikanten Kookkurrenten zuerst vermischt aufgeführt. Wird aber für jede der Bedeutungen eine andere typische Wortform ausgewählt, erhält man aus der Durchschnittsmenge der Kookkurrenten dieser Wortform mit denen der gemeinsamen Bedeutungen das gewünschte Ergebnis. [12]

Anwendung auf andere Sprachen

Wenn ein Problem nicht typisch für eine spezielle Sprache ist, dann lässt es sich mittels einzelsprachenunabhängiger Methoden lösen. Beispiele dafür sind Rechtschreibkontrollen oder Spracherkennung. Die signifikanten Kookkurrenten sind ebenso in indoeuropäischen Sprachen oder im Koreanischen verwendbar. In den übrigen asiatischen Sprachen ist durch die Tatsache, dass keine Leerzeichen zwischen Wortformen benutzt werden, ein Vorverarbeitungsschritt notwendig, der die Worte für die weitere Bearbeitung trennt. [12]

Disambiguierung

Disambiguieren ist "ein sprachliches Zeichen einer Mehrdeutigkeit entheben, indem man es in bestimmte syntaktische u. semantische Kontexte einordnet" [20]. Es lässt sich in zwei Stufen einteilen:

1. ein Nachschlagewerk, das die verschiedenen Bedeutungen von Wortformen kennt
2. durch den Kontext in dem eine Wortform auftritt muss seine Bedeutung erkannt werden

Für die englische Sprache gibt es dazu das frei verfügbare semantische Lexikon "WordNet"[2] und für Deutsch das lizenzpflichtige "GermaNet"[3] (siehe auch Kapitel 3.2.1). Weiters gibt es noch das sprachenübergreifende "EuroWordNet", das laut eigenen Angaben eine mehrsprachige Datenbank für Wordnets aus verschiedenen europäischen Sprachen darstellt [5].

Ein Verfahren zum Erkennen einer Wortform ist z.B. in WordNet vorhandene semantische Hierarchien auszunutzen. Dabei wird die Bedeutung eines Worts auch mit Hilfe der Definition von über- oder untergeordneten Wortformen bestimmt. Eine weitere, sehr einfache Möglichkeit mit durchaus sehr guten Ergebnissen stellt auch das Baseline-Verfahren dar: Dabei wird lediglich die in WordNet verfügbare Häufigkeitsangabe einer Bedeutung herangezogen und immer die am häufigsten auftretende Relevanz ausgewählt.

Will man solche Nachschlagewerke erstellen, kann man diese manuell oder automatisch generieren. Bei der ersten Art stellt sich häufig die Frage, ob sich der große Aufwand lohnt, da nicht bei jedem Wort alle möglichen Bedeutungen relevant sein müssen. Heyer nennt dazu das (englische) Beispiel "space", das die (deutsche) Bedeutung "Weltraum", "Raum, der in einem Büro vermietet werden kann" und "Luftraum" haben kann. Diese

[2] http://wordnet.princeton.edu/
[3] http://www.sfs.uni-tuebingen.de/lsd/

Bedeutungen sind sehr ähnlich und eine Differenzierung muss vielleicht nicht getroffen werden. Die automatische Erstellung eines Wörterbuchs hingegen ist mit weniger Arbeit verbunden. Hier werden statistische Mittel gebraucht, wie z.B. die Analyse der signifikanten Satzkookkurrenten. Dabei wird zu jedem Wort eine Menge von Wortformen, mit denen das Wort auffällig häufig zusammen auftritt, berechnet. Aus dieser Menge wird dann ein Graph erstellt. Die Auswertung der Knotenmenge erfolgt dann mit einem Standard-Cluster-Verfahren, dessen Ergebnisse (Cluster) die Einträge im Nachschlagewerk ergeben. Diese Resultate haben jedoch den Nachteil, dass sie die Bedeutungsunterscheidungen nicht benennen können bzw. erkennen sie solche falsch oder auch gar nicht. [12]

Clustering

Das Ziel von Clustering ist, "Daten zu natürlichen Gruppen (Clusters)" [17] zusammenzufassen. Dazu ist eine möglichst große Homogenität innerhalb eines Clusters und Heterogenität zwischen den Clustern maßgebend. Cluster lassen sich in

- hart/disjunkt (jedes Element steht in genau einem Cluster) und
- soft/überlappend (jedes Element kann in mehreren Clustern stehen)

einteilen. Für Clusteringalgorithmen gibt es folgende Einteilung:

- nicht-hierarchisches (z.B. k-means) und
- hierarchisches Verfahren
  - bottom-up/agglomerativ: Ausgehend von vielen Clustern werden diese solange nach Ähnlichkeit zusammengefasst bis ein großer Cluster existiert.
  - top-down/divisiv: Ein großer Cluster wird so lange geteilt, bis viele kleine Cluster entstehen.

Bei den hierarchischen Verfahren ist die Wahl der Ähnlichkeitsfunktion ausschlaggebend. Diese sollte so gewählt werden, dass eine möglichst kleine Clusteranzahl bei möglichst großer Homogenität innerhalb der Cluster existiert. Die gebräuchlichsten Ähnlichkeitsmaße sind

- single linkage: die Ähnlichkeit der beiden ähnlichsten Punkte oder
- complete linkage: die Ähnlichkeit der beiden unähnlichsten Punkte.

Interessante Anwendungen von Clusterverfahren bei Text Mining sind z.B. das Clustering von Wörtern mit semantischer Nähe oder von Texten mit Inhalt über ein gleiches Fachgebiet. [12, 17]

2.2.2 Musterbasierte Verfahren

Hier werden in Texten allgemeingültige und relevante Muster extrahiert um diese in anderen Texten identifizieren zu können.

Reguläre Ausdrücke

Reguläre Ausdrücke (Regular Expressions) bestehen aus Atomen die die Einheit dieser Sprache darstellen, und Operatoren, die eine Art Sprachgrammatik zulassen. Mit Hilfe der Operatoren können Atome verknüpft werden, um so nach Mustern in Texten zu suchen. So findet z.B. der Ausdruck *un[a-z]** alle Worte in einem Text, die mit der Silbe *un-* beginnen. Wildcards, wie . * ? helfen dabei, Gruppen von Zeichen zuzulassen oder die Anzahl von Zeichen anzugeben. Genaueres dazu findet sich in der Literatur [6].

Anwendungsgebiete im Text Mining finden sich z.B. bei der Suche nach Mustern in einem POS-getaggten Korpus um alle Nomen zu finden, bei der Suche nach Eigennamen in einem Text oder allgemein beim Information Retrieval.

Syntaktische Muster

Hier werden Muster gesucht, die sich an der syntaktischen Struktur eines Satzes orientieren. Besonders gut geht das bei POS-getaggten Texten. Hier kann z.B. nach Unterbegriffen eines Worts gesucht werden. Heyer nennt als Beispiel

> Studienrichtungen *wie* Medizin *und* Pharmazie

dessen Muster "[NN] wie [NN] und [NN]" ist (Anmerkung: [NN] steht für Normalnomen). Das erste Nomen steht dabei für den Ober-, die beiden anderen für Unterbegriffe. Ein weiteres interessantes Beispiel ist die Suche nach Definitionen:

> *Ein* Recherchebericht *ist ein* Bericht zur ...

Das Muster ist "[ART] [NN] ist [ART]" (Anmerkung: [ART] steht für Artikel).

2.3 Sentiment Analysis und Opinion Mining

Eine Suche nach *Sentiment Analysis*, dessen deutsche Übersetzung am besten mit *Gefühlsanalyse* zu wählen ist, hat oft ein Ergebnis, dessen Schlüsselwort *Opinion Mining*, zu deutsch die *Meinungsschürfung*, ist. Umgekehrt tritt dies ebenfalls häufig auf. Das deutsche Pendant zum englischen Wikipedia-Eintrag für *Sentiment Analysis* ist *Sentiment Detection*, in der Literatur ist ebenfalls oft von *Opinion Detection* die Sprache. Eine wirklich differenzierte Definition der beiden Ausdrücke zu finden gestaltet sich dadurch reichlich schwierig.

Geht man von den deutschen Übersetzungen aus, könnte Sentiment Analysis der Oberbegriff und Opinion Mining eine Kategorie davon sein. In diesem Kapitel wird mehr auf die angewandten Techniken der Meinungsschürfung eingegangen.

Opinion Mining ist in einer Zeit, in der Internetzugang und Onlinebestellungen alltäglich sind, sehr gefragt. Z.B. wollen Hersteller wissen, auf welcher der unzähligen Webseiten ihr Produkt eine gute Rezension erhalten hat, um dort ein Werbefenster für eben dieses einzublenden. Falls die Meinung über das Produkt jedoch nicht gut ausfällt, wäre das vielleicht keine so gute Idee. Ebenso möchten Kosumenten wissen, wo positive und negative Meinungen über ein für sie interessantes Produkt zu finden sind.

Eine grundlegende Technik des Opinion Mining ist die *Sentiment classification*. Dazu muss der vorliegende Text als positiv, negativ oder manchmal auch neutral klassifiziert werden. Für gewöhnlich passiert diese Einteilung auf Dokumentebene; das heißt, das gesamte Dokument wird einer Klasse zugewiesen. Ebenso kann die Einteilung aber auch auf Satzebene erfolgen. Die Methoden dazu sind vielfältig. Da die Forschung auf diesem Gebiet noch jung ist, kann man eher von Problembeschreibungen und aktuellen Untersuchungen reden, weniger von ausgereiften Techniken, die diese Probleme lösen. [15]

Es gibt drei Arten von Klassifizierung [15]:

1. Klassifizierung basierend auf **Sentiment Phrases**: Dieser Algorithmus benutzt das in Kapitel 2.2.1 vorgestellt POS-Tagging. Zuerst extrahiert er Adjektive oder Adverbien. Diese stellen gute Indikatoren für Subjektivität und Meinung dar. Da ein Adjektiv alleine aber sowohl positiv als auch negativ benutzt werden kann (z.B. "unvorhersehbare Probleme" oder "die unvorhersehbare Handlung eines Buchs"), muss dazu auch der Kontext, in dem das Wort auftritt, berücksichtigt werden. Dazu extrahiert der Algorithmus zwei aufeinanderfolgende Wörter, wenn deren POS-Tags einem bestimmten Muster entsprechen, wie z.B. Adjektiv-Nomen. Danach berechnet er die semantische Ausrichtung (semantic orientation - SO) der Phrasen. Zuletzt wird die durchschnittliche SO aller Phrasen errechnet.

2. Klassifizierung mit Hilfe von **Textklassifizierungs-Methoden**: Dieser einfache Ansatz behandelt das Problem als ein topic-based (themenbasiertes) Klassifizierungs-Problem. Dazu kann jeder Textklassifizierungs-Algorithmus benutzt werden.

3. Klassifizierung mit Hilfe einer **Auswertungsfunktion**: Dieser Algorithmus bewertet zunächst jeden Term des Trainingssets anhand einer Gleichung und vergibt Werte zwischen -1 und +1. Um ein ganzes Dokument auszuwerten, summiert er die Werte aller Terme und benutzt das Vorzeichen des Ergebnisses um die Klasse zu bestimmen.

2.4 Emotion Mining

Der Ansatz des Opinion Mining, einen Text nach meinungsäußernden Wörtern und Phrasen zu durchsuchen um ihn dann zu klassifizieren, stellt eine interessante Basis für die Suche nach Emotionalität in Texten dar. Dazu sollen zunächst Gefühlskategorien erstellt werden, auf die in Kapitel 2.6 noch genauer eingegangen wird. Anhand von Wortlisten, wie sie z.B. das *Affektive Diktionär Ulm* (2.5.2) beinhaltet, wird entweder ein Dokument oder werden Teile eines Dokuments nach ihrer Emotionalität, wie Ängstlichkeit, Liebe, Wut, etc. bewertet. So wird eine Kategorisierung des Textes erreicht. Dazu mehr in Kapitel 3.

2.5 Germanistik

In diesem Abschnitt wird der Germanistik die Aufgabe der manuellen Herangehensweise an die Emotionsanalyse von Texten zu Teil. Dazu sollen die quantitativen und qualitati-

ven Komponenten aus Silke Jahrs Arbeit [14] erläutert werden.

2.5.1 Emotionsbegriff und Sachtexte

> *Zur conditio humana gehören Emotionen. Sie spielen eine zentrale Rolle in der menschlichen Existenz, da alle Erfahrungen von Emotionen durchdrungen sind. Gefühlssysteme bestimmen als Grundschwingung das Wahrnehmen, das Denken, das Erleben und das Verhalten jedes einzelnen Menschen. Sie sind ein Modus, sich die Welt anzueignen und bilden eine wichtige Quelle für die Entscheidungen des Lebens, auch wenn sich Individuen dessen nicht bewußt sind.* [14]

Die Erscheinung "Emotion" wird seit einigen Jahren von unterschiedlichen wissenschaftlichen Disziplinen erforscht. Vor allem die Psychologie, die die Emotionsforschung als Kernbereich sieht, aber auch die Kommunikationsforschung, die Kognitionsforschung oder die Soziologie interessieren sich für die Gefühlsregungen von Menschen. Um das Gebiet der Emotionsanalyse so gut wie möglich zu begreifen, ist es notwendig die interdiszipliären Forschungen und Erkenntnisse zu betrachten.

Die emotionale Gesamtreaktion kann in verschiedene Teilreaktionen getrennt werden:

- physiologische Reaktionen (Herzschlag- und Atemfrequenz)
- tonische Haltungsreaktionen (An- und Entspannung)
- instrumentelle motorische Reaktionen
- expressive motorische Reaktionen (Gestik, Mimik)
- expressive sprachliche Reaktionen (syntaktische und lexikalische Selektion)
- subjektive Erfahrungskomponenten (Gefühle im eigentlichen Sinne)

Für die Arbeit von Silke Jahr ist zum einen der Begriff der Bewertung wichtig. Genauer gesagt sind "Gefühle Bewertungsreaktionen auf Ereignisse, auf das Tun und Lassen von Urhebern oder auf Personen und Objekte" [14]. Emotionales Erleben findet nicht ohne diese Bewertung und ohne reflexiven Selbstbezug statt. Zum anderen ist die hohe Ich-Beteiligung, die Selbstbetroffenheit, ein wichtiger Aspekt. Bei der Untersuchung auf Emotionalität in Texten sind für die Autorin als psychologische Dimension die subjektive Erfahrungskomponente sowie die expressive sprachliche Relation bedeutend. Ebenso wichtig ist die soziale Ebene, während die biologische irrelevant ist.

Die Begriffsdefinitionen und -differenzierungen der Ausdrücke werden in der Literatur oft gleich genutzt: So werden *Emotionen* und *Gefühle* oft synonym verwendet, ebenso wird häufig *Affekt* den beiden Worten gleich gesetzt. Im Unterschied dazu sind *Stimmungen* länger dauernd, diffuser und ihre Intensität ist geringer. Ein weiterer Begriff ist *Empfindung*, die eher das Körperliche zum Ausdruck bringen soll.

Jahr bezeichnet als Norm für einen Sachtext dessen Unemotionalität. Daher können manipulative Texte, die beim Leser gezielt Emotionen bewirken sollen, wie z.B. Texte

für die Wahlkampagne eines Politikers, nicht herangezogen werden. Sie behandelt ausschließlich Schriftwerke, die von sich aus sachlich gehalten werden sollten. Weiters sollte zwischen der Schreiber- und Leserperspektive unterschieden werden. Aus Untersuchungen ist hervorgegangen, dass die Emotion des Schreibers zwar oft beim Leser vorzufinden war, dass aber auch völlig unterschiedliche Gefühlsschichten aufgetreten sind. In Jahrs Arbeit wird diese Untersuchung nicht überprüft, da sie sich nur mit den Emotionen des Textverfassers beschäftigt. In Bezug auf die vorhin genannten manipulativen Texte wäre das allerdings ein interessantes Thema, wie z.B. Leser unterschiedlicher Meinung auf Texte von Politikern reagieren.

Aus dem Genannten kann man von drei Möglichkeiten der Emotionalität bei Texten sprechen:

- Der Sprecher/Schreiber drückt ein Gefühl aus
- Beim Hörer/Leser werden durch spachliche Mittel Gefühle geweckt
- Der Sprecher/Schreiber drückt Gefühle aus, die auch beim Hörer/Leser Gefühle ausdrücken

Weiters lässt sich festhalten, dass die wertende und emotionale Komponente von Sätzen schwer zu ermitteln ist. Das heißt nicht, dass ein Satz nicht emotional eingestuft werden kann, sondern vielmehr, dass anhand des grammatischen Aufbaus eines Satzes schwer auf seinen emotionalen Wert geschlossen werden kann. Das Problem ist, dass es keine klaren, formal feststellbaren Indikatoren dafür gibt. Andere Meinungen sagen, dass besonders syntaktische Muster zur Emotionsanalyse verwendet werden können. Als Beispiel wird unter anderem

> " Wie groß Peter geworden ist! "

genannt. Es fehlt jedoch eine Erläuterung, wie die Zuweisung dieses Satzes zu einer Menge emotionaler Sätze geschehen kann.

Im Gegenteil dazu sagt der Stil eines Satzes durch gewisse Ausdrucksverstärkung, Expressivität, sehr viel über dessen emotionalen Grad aus. [14]

2.5.2 Affektive Wörterbücher

In Jahrs Arbeit werden Wörterbücher, die emotionale Wörter beinhalten, angesprochen. Dazu zählt unter anderem eine Arbeit von Debus aus dem Jahr 1988 und von Hölzer das "Affektive Diktionär Ulm" aus 1991. Beide beziehen sich auf ein Werk von Dahl aus 1978, in dem er ein Kategorienschema mit acht Emotionsbereichen vorstellt.

Das affektive Diktionär Ulm - ADU

Entstanden ist dieses Diktionär[13] in den 90er Jahren des letzten Jahrhunderts. Es ist "ein inhaltliches Sprachanalyseverfahren, mit dem Affekte quantitativ bestimmt werden. Es wird dabei in Objektemotionen und Selbstemotionen unterschieden, die wiederum positiv oder negativ sein können" [7]. Das Diktionär umfasst acht bzw. zwölf Kategorien. Diese sind

1. Liebe
2. Begeisterung
3. Zufriedenheit und Erleichterung
4. Freude und Stolz
5. Zorn
6. Furcht
7. Depressivität und Schuld
8. Ängstlichkeit und Scham

Genaueres zur Einteilung findet sich in der Literatur wie [11]. Das Diktionär enthält ausschließlich Substantive und Adjektive aber keine Verben, da sich diese als zu stark kontextabhängig erwiesen [14]. Als Beispiel wird der Satz:

> Das Essen ist verbrannt.

genannt. Dieser kann sowohl als Feststellung und auch als emotionales Werteurteil aufgefasst werden. Adjektive können also genauso wie Verben zu einer falschen Interpretation führen [14].

Im Zuge dieser Arbeit wurde versucht, das ADU in irgendeiner Weise, sei es gedruckt oder als elektronische Wortliste, zu finden. Erst die Kontaktaufnahme per E-Mail mit einem der Co-Autoren des ADU, Horst Kächele, führte zum Erfolg. Dan Pokorny übermittelte daraufhin eine Liste mit 26.000 deutschen Vollformwörtern, die auf rund 2.000 Grundformen affektiver Wörter beruhen. An dieser Stelle ein großes Dankeschön dafür. Dan Pokorny schrieb, dass das ADU nur deshalb Substantive und Adjektive enthält, weil sie "das Problem der grammatisch/morphologisch widerspenstigen deutschen Verben [...] trotz etlicher Mühe mit diesem Ansatz nicht befriedigend lösen können".

Das Dresdner Angst Wörterbuch - DAW

Zwar wird das DAW in dieser Arbeit nicht verwendet, da es aber teilweise vergleichbar zum ADU ist, wird es kurz vorgestellt. Das DAW[4] ist eine deutschsprachige Computerversion der Gottschalk-Gleser-Angstskalen. Dieses Wörterbuch ist nach langer Internetrecherche das einzige, das online eine Wortliste zur Verfügung stellt. Das DAW benutzt CoAn[5], eine Software zur Text- und Inhaltsanalyse für Windows. Mit dem für die Installation notwendigen Zip-File kommt auch eine *daw2000.dic* Datei mit, in der rund 4000 angstbeschreibende Worte enthalten sind.

Louis A. Gottschalk und Goldine C. Gleser entwickelten seit den 60er Jahren des letzten Jahrhunderts in den USA ein Verfahren zur Sprachinhaltsanalyse. Das Gottschalk-Gleser-Verfahren ist das einzig weltweit verbreitete Verfahren dieser Art. Sie unterscheiden dabei

[4] http://rcswww.urz.tu-dresden.de/ berth/daw/daw.html
[5] http://www.coan.de/

sechs verschiedene Angstformen: Todesangst, Verletzungsangst, Trennungsangst, Schuldangst, Angst vor Scham/Schande, diffuse oder unspezifische Angst. Diese Kategorien werden jeweils noch unterteilt, ob die Angst erlebt wurde oder aufgetreten ist bei

- dem Sprechenden,
- einem anderen Lebewesen,
- einem unbelebten Objekt (nicht bei allen Kategorien möglich) oder
- ob es sich um Verneinung/Verleugnung handelt.

Das Gottschalk-Gleser Verfahren ist jedoch nicht fehlerfrei. Die Sprache wird nur auf ihre Symptomfunktion (Sprecher bezieht sich auf sich selbst) reduziert, die Symbol- (Sprecher bezieht sich auf Gegenstände und Sachverhalte) und Signalfunktion (Sprecher bezieht sich auf den Hörer) werden außer Acht gelassen.

Die Auswertung geschieht folgendermaßen: Jede der Angstformen wird einzeln ausgezählt, die Angst beim Sprecher z.B. wird mit 3 gewichtet, bei anderen Personen mit 2 und bei unbelebten Objekten mit 1. Für jede Angstkategorie wird mit folgender Formel ein Angstscore errechnet:

$$S = \sqrt{\frac{100}{WZ} \cdot (R + 0,5)} \tag{2.11}$$

S ist dabei der Angstscore, WZ die Anzahl der Worte und R der Rohwert, das ist die Summe der oben genannten Gewichtungen. Der Gesamtangstscore wird mit folgender Formel berechnet:

$$ZS = \sqrt{\frac{100}{WZ} \cdot (R_1 + R_2 + ... + R_n + 0,5)} \tag{2.12}$$

Eine rein manuelle Ausarbeitung pro Text dauert etwa 30-60 Minuten. Ein für den englischsprachigen Raum entwickelte Computerprogramm ist da natürlich um einiges schneller. Es geht dabei vom Verb eines Satzes bzw. dessen Grundform aus. Wird dieses Verb in einem Angst-Wörterbuch gefunden, wird versucht das zugehörige Subjekt und Objekt zu finden. Dann wird ermittelt, ob es sich wirklich um einen angstbeschreibenden Ausdruck handelt. Danach wird der Ausdruck gewichtet, Normwerten gegenübergestellt und das Ergebnis ausgegeben. Im deutschsprachigen Raum ist bislang keine überzeugendes Programm geschrieben worden. [23]

Wird das vorher erwähnte CoAn benutzt, ist es laut Readme-Datei notwendig, eine lizenzierte Version und SPSS, ein Statistikprogramm, zu haben. Hat man dieses nicht, erhält man eine nichtsaussagende Zeichenkette mit Zahlenwerten.

2.5.3 Analyse von Texten

Grundsätzlich sind kurze Texte nicht für die Analyse auf Emotionalität geeignet. Jahr meint dazu, dass der Text mindestens 70 Worte haben sollte.

Qualitative Analyse

Um die Emotion der Textautoren nachvollziehen zu können muss der Verfasser eines Schriftstücks Bewertungen vornehmen und diese müssen eine Ich-Beteiligung oder Selbstbetroffenheit beinhalten. Ich-Beteiligungen lassen sich mit Hilfe von I-Variablen (siehe Kapitel 2.6.2) ausfindig machen. Weiters wird

- die Präsenz von I-Variablen,
- die Intensität der I-Variablen,
- die Präsenz von Bewertungen und Bewertungskriterien sowie
- die Intensität der Bewertungen

betrachtet.

Quantitative Analyse

In einem quantitativen Zusammenhang betrachtet, kommt Jahr zu folgender Formel:

$$E_I = \frac{B(\sum SM + F_{Ex}) \sum Va}{W} \tag{2.13}$$

E_I ist die emotionale Intensität, B der Betroffenheitsfaktor, der die Präsenz der I-Variablen 1 beinhaltet (hat den Wert 1; liegt keine Ich-Beteiligung vor ist $B = 0$ und damit $E_I = 0$), $\sum SM$ ist die Summe der wertenden sprachlichen Mittel, F_{Ex} ist der Expressivitätsfaktor, W die Anzahl der Wörter des Textes und in $\sum Va$ wird ein Wert für die I-Variablen 2-8 eingesetzt. Genaueres dazu ist [14] zu entnehmen. Interessant bleibt der Expressivitätsfaktor F_{Ex}. Mit diesem wird die Ausdrucksstärke gemessen. So sind Modalworte wie *sozusagen* oder *eher* weniger expressiv wie *faszinierend* oder *ultimativ*. Mit diesem Faktor können Worte einen positiven oder negativen Wichtungswert erhalten. Die Ungenauigkeit hierbei ist natürlich die subjektive Meinung des Bewertenden. Jahr meint dazu, dass die "undifferenzierte Behandlung einzelner wertender sprachlicher Mittel dem Anliegen, die Intensitätsstärke von Emotionen zu bestimmen, weniger gerecht [wird] als mit der Einführung von Wichtungswerten" [14].

Diese Formel liefert bestimmt keine exakten Werte. Ob diese überhaupt reproduzierbar sind, da sie von der subjektiven Auffassung des Bewerters abhängen, ist ein eigenes Problem. Dennoch lässt sich die emotionale Intensität eines Textes abschätzen und mit der anderer Texte vergleichen. Weiters kann ein Grenzbereich bestimmen werden, über dem von einem emotionalen Text und unter dem von einem nicht-emotionalen Text gesprochen werden kann. Liegt ein berechneter Wert innerhalb dieses Bereichs, ist der Text nicht eindeutig zuordbar.

"Die Frage, ob einem Autor jedoch tatsächlich die rekonstruierte emotionale Betroffenheit beim Verfassen des Textes zuzuschreiben ist oder emotionale Betroffenheit über den emotionalen Stil nur vorgetäuscht wird, kann mit linguistischen Mitteln letztlich nicht entschieden werden." [14]

2.6 Psychologie

2.6.1 Struktur von Emotionen

Es existieren verschiedene Strukturierungen von Emotionen. Eine sehr allgemeine ist die Unterteilung von Gefühlen in *angenehm* und *unangenehm*. Eine weitere, spezifischere, teilt die Emotionen in zehn Basistypen ein: Freude, Zuneigung, Überraschung, Unruhe, Abneigung, Ärger, Traurigkeit, Verlegenheit, Schuld und Angst. Es gibt aber auch Kategorisierungen mit sieben bis 15 unterschiedlichen Grundemotionen. Grundsätzlich kann nur gesagt werden, dass es nicht *die* Klassifikation von Emotionen gibt, sondern immer eine nach den Zielen der jeweiligen Untersuchung sinnvolle Einteilung. [14]

2.6.2 Intensität von Emotionen

Hier benutzt Jahr die Intensitätsvariablen (I-Variablen) von Mees [16], die die Intensität von Emotionen unabhängig voneinander beeinflussen. Sie werden in globale (können alle Emotionsgruppen beeinflussen) und lokale (nur für bestimmte Emotionsgruppen relevant) Variablen unterteilt. Genaueres hierzu findet sich in [14] oder [16]. Die acht I-Variablen sind:

1. Psychologische Nähe mit der im Text behandelten Thematik
2. Wichtigkeit der thematisierten Sachverhalte für die Menschheit bzw. die Gesellschaft
3. Wichtigkeit für die eigene Person
4. Erwartung oder Nichterwartung
5. Tadelnswürdigkeit oder Verdiensthaftigkeit
6. Grad der sozialen Zustimmung
7. Grad der Überzeugtheit der eigenen Person
8. gesteigerte Betroffenheit

Um die die I-Variablen signalisierenden Elemente im Text zu finden, muss dieser unter drei Aspekten untersucht werden: der situative Rahmen (wer schreibt den Text, in welcher Situation befindet sich der Autor), die Inhaltsseite (Abschnittsgliederung) und die sprachliche Ausdrucksseite (inhaltliche Analyse der einzelnen Abschnitte). Weiters werden noch zwei Emotionsschemata, eines davon von Mees, vorgestellt. Für Details siehe [14].

3 Umsetzung

Für die Umsetzung des Themas, die Suche nach emotionalen Komponenten eines Textes, werden im Theoriekapitel beispielhaft vier Arbeiten beschrieben, die eine ähnliche Problemstellung gelöst haben um daraus den Ansatz dieser Arbeit abzuleiten. Allgemein finden sich dazu in der Literatur meistens Arbeiten, die die natürliche Sprache analysieren oder sich mit Mensch-Computer-Dialogen beschäftigen [1]. Zur Textanalyse existieren nicht ganz so viele Werke und diese beschreiben Vorgehensweisen für die englische Sprache. Da deutsche Literatur dazu nur spärlich wenn überhaupt vorhanden ist, wird dieser Teil vor allem die Herangehensweisen behandeln, die einzelsprachunabhängig sind.

3.1 Theorie

Grundsätzlich unterscheiden sich die Papers fast ausschließlich dadurch, welche Texte/Textkorpora sie analysiert haben und nach welchem Kriterium analysiert wurde. Diese Kriterien lassen sich in negative/positive Wertigkeit (vgl. Kapitel 2.3)[1] [4] oder in die Suche nach unterschiedlichen Emotionen [2] [19] einteilen.

Die erste Einteilung, ob ein Text positiv oder negativ gemeint ist, findet sich bei Opinion Mining Ansätzen. Hier reicht es oft zu wissen, wie die Bewertung eines Produkts ausfällt und weniger, welche Art von Emotionalität aufkommt (liebevoll, zornig, etc.). Oft kommt bei diesen Ansätzen noch eine neutrale Bewertung zum Einsatz, um eine bessere Differenzierung zu erhalten. Grundsätzlich ist diese Art der Einteilung sicher eine der einfacheren, da bei diesen Methoden lediglich die Anzahl der positiven und negativen Meinungswörter, die sich z.B. um ein Produktmerkmal finden lassen, gezählt werden müssen und dann der höhere Wert die Einstellung des Textverfassers zu dem Produktmerkmal bestimmt [4]. Doch muss zu dieser völlig lexikalischen Herangehensweise jedenfalls der Kontext betrachtet werden, um zu bestimmen ob etwas z.B. "nicht schlecht", "zu klein" oder "wunderbar klein" ist. Es kann auch vorkommen, dass Texte zwar mit acht Basisemotionen (Ärger, Ekel, Angst, Freude, Traurigkeit, positive/negative Überraschung, Neutral) annotiert werden, da aber die vorhandenen Trainingsdaten keine feinere Abstufung erlauben, wird schließlich nur die positive oder negative Wertigkeit einer Textpassage bestimmt [1].

Die zweite Art der Einteilung, nach unterschiedlichen Emotionen ist da schon wesentlich differenzierter. Hier werden mitunter [19] nicht nur Korpora untersucht, ob darin Emotionen auftreten, sondern wie stark diese darin vorkommen. Dabei gibt es Ansätze mit nur drei Intensitätskategorien [2] und solche, die diese Skala in 100 Segmente teilen[19]. Es wird von sechs Basisemotionen ausgegangen (Ärger, Ekel, Angst, Freude, Traurigkeit und Überraschung) aber auch von "gemischten Emotionen" und "keinen Emotionen" gesprochen [2].

Die angesprochenen unterschiedlichen Textkorpora bestehen je nach Arbeit aus

- Nachrichten-Schlagzeilen [19],
- Blogeinträgen [2],
- Kundenrezensionen [4] oder
- Kindergeschichten [1].

Weitere Unterscheidungen bestehen in der zu analysierenden Ebene. Hierbei wird, wie in Kapitel 2.3 schon angesprochen, meistens zwischen Analyse auf Dokument- und Satzebene unterschieden. In zwei der erwähnten Arbeiten wird auf Satzebene eine Emotionskategorie [1] [2] vergeben. Nur beim Opinion Mining wird auf die Merkmalsebene eingegangen[4].

Interessant ist noch die unterschiedliche Herangehensweise an die manuelle Annotation von Trainingsdaten. Hier werden zwischen Arbeitsteams [1] und sechs unabhängig arbeitenden Personen [19], die Texte annotieren, verschiedene Anzahlen an "Annotatoren" verwendet. Ebenso wird mit dem Problem der Übereinstimmung bei den Annotationen umgegangen: Es wird ein gemeinsamer Konsens gesucht [4], es entscheidet ein Autor [1], es wird die Korrelation nach Pearson verwendet um zu Vereinheitlichen[19] oder es wird Cohens Kappa Statistik[1] benutzt um den Konsens zwischen den Annotatoren zu finden.

3.2 Andere Ansätze

3.2.1 WordNet Affect und GermaNet

WordNet und GermaNet wurden schon in Kapitel 2.2.1 vorgestellt. WordNet ist eine große lexikalische Datenbank für die englische Sprache. Nomen, Verben, Adjektive und Adverben werden in Mengen synonymer Wörter, sogenannten Synsets, gruppiert. Diese Synsets sind wiederum untereinander verlinkt, wenn sie semantische oder lexikalische Verbindungen aufweisen. Mit WordNet-Affect, das in [2] Erwähnung findet, werden Teilen dieser Synsets affektive Bedeutungen zugeschrieben. Diese werden in sechs Listen, entsprechend sechs Basisemotionen, eingeteilt.

Das kostenpflichtige GermaNet ist das deutsche Pendant zu WordNet. Es kann als online-Thesaurus gesehen werden[10]. Im Unterschied zu WordNet aber scheint es keine affektive Einteilung oder affektive Synsets zu geben.

Als freie Alternative zu GermaNet sei noch das deutsche Projekt "OpenThesaurus"[2] sowie das Projekt Deutscher Wortschatz (siehe auch Kapitel 2.2), das neben umfangreichen Angaben zu einem Wort auch dessen Synonyme aufzählt, erwähnt.

[1] Maß für die Verbesserung gegenüber Zufallsklassifikator[18]
[2] http://www.openthesaurus.de/

3.2.2 Statistical/Semantical Emotional Text Analyzer

Der *Statistical-* und *Semantical Emotional Text Analyzer* sind zwei interessante Textanalysetools[3] von Alexander Osherenko[4] von der Universität Augsburg. Ersterer stellt einen sprach- und domänenunabhängigen Ansatz für Opinion Mining dar, der *Semantic Text Analyzer* zeigt eine semantische Herangehensweise an die Textanalyse. Letzterer bebildert seine Auswertung sogar mit Fotos von Mimiken.

3.3 Architektur GATE

Das Akronym GATE steht für "General Architecture for Text Engineering"[3] und ist ein führendes Toolkit für Text Mining, geschrieben in Java. Es beinhaltet eine Architektur, ein opensource-Framework und eine graphische Entwicklungsumgebung. GATE kann für alle Arten von Sprachverarbeitung, wie Information-Extraction, verwendet werden. [9] Häufig wird das Programm zum Annotieren von Texten verwendet, wie es auch für diese Arbeit der Fall sein wird. Einen Screenshot von GATE zeigt Abbildung 3.1.

Es gibt zwei Hauptquellen, mit denen GATE umgeht:

- Language Resource (LR - Sprachquelle): Das sind Datenquellen wie Lexika, Korpora, Thesauri oder Ontologien.
- Processing Resource (PR - Verfahrensquelle): Hierzu zählen programmatische oder algorithmische Quellen, wie Lemmatiser, Übersetzer, Parser oder Spracherkennung. Typischerweise enthalten Processing-Resources häufig Language-Resources, wie z.B. ein Tagger (vgl. 2.2.1 und POS-Tagging) oft ein Lexikon hat.

Diese Quellen können in GATE einfach geladen werden, die Processing-Resources können auch in einer Pipeline geladen werden, um hintereinander ausgeführt zu werden. Dieser ganze Komplex bildet dann die Applikation.

Die für diese Arbeit relevanten GATE-Tools sind:

- der OntoGazetteer, eine PR, in der Instanzlisten von Ontologiekonzepten[5] geladen werden können. Dazu werden eine oder mehrere ".lst" Dateien, sowie eine "mappings.def" und eine "lists.def" Datei benötigt. Die .lst-Dateien enthalten in jeder Zeile eine Instanz eines Objekts, z.B. enthält die Datei *Liebe.lst* in jeder Zeile ein Wort aus der Gefühlskategorie Liebe des ADUs2.5.2. Die *mappings.def*-Datei beschreibt die Relationen zwischen den .lst-Dateien und den Ontologiekonzepten. Die *lists.def*-Datei beinhaltet die Relationen zwischen den verschiedenen .lst-Dateien und dem Annotationstool, das der Gazetteer erzeugen soll.
- der Jape-Transducer, ebenfalls eine PR, die zum Manipulieren von Annotationen dient. Mit Hilfe der Jape-Grammatik lässt sich beschreiben, wie die Annotationen

[3] http://emotion.informatik.uni-augsburg.de:8080/WebInterface/
[4] http://www.informatik.uni-augsburg.de/ osherenk/
[5] Ontologien dienen in verschiedenen Bereichen als Mittel der Strukturierung und zum Datenaustausch, um bereits bestehende Wissensbestände zusammenzufügen[21]

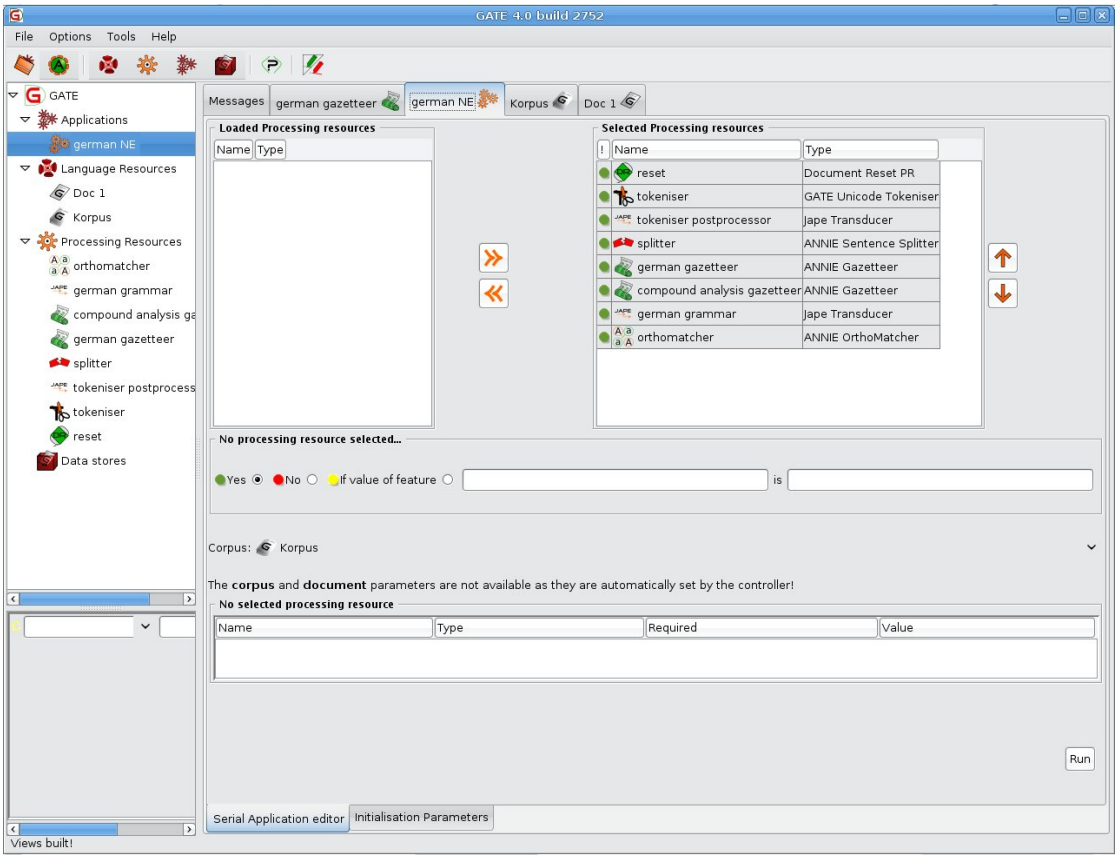

Abbildung 3.1: Screenshot von GATE

verändert werden sollen. Konkrete Beispiele dazu finden sich in 3.4 und natürlich im GATE-Manual[6].

3.4 Praxis

3.4.1 Grundlegendes

Im Gegensatz zu den genannten Vorgehensweisen in Kapitel 3.1 sind die Mittel für diese Arbeit nicht ganz so umfangreich. Dafür kann mit einer langjährig ausgearbeiteten Wortliste, dem ADU, eine sehr feine Granulierung bei der Zuordnung der Emotionswörter eines Textes vorgenommen werden. So soll versucht werden, einen Text nicht nur als emotional zu erkennen sondern auch zu ergründen, welche Art von Gefühlen vorrangig vorkommen. Hierbei wird, wie schon in Kapitel 2.5.2 beschrieben, ausgehend vom ADU eine Einteilung in zwölf unterschiedliche Gefühlskategorien vorgenommen. Da es den Rahmen dieser Arbeit sprengen würde, wie bei den genannten Papers Texte händisch zu annotieren um

[6] http://gate.ac.uk/sale/tao/index.html

Trainingsdaten zu erstellen, wurden elf emotionale und elf nicht-emotionale Texte ausgesucht, die persönlich und subjektiv so eingestuft wurden. Diese sollen helfen, eine Skala zu finden, mit der man bestimmen kann, ob ein Text emotional ist oder nicht.

Die ausgewählten emotionalen Texte sind unterschiedlichster Art, aber mit mindestens 70 Wörtern (vgl. Kapitel 2.5.3):

- Vier-Pfoten 1: Text über Hundemafia[7]
- Vier-Pfoten 2: Text über Kaninchenhaltung in Käfigen[8]
- Jugend-FPÖ: Standpunkte der Jugend-FPÖ Burgenland[9]
- Grüne-Jugend: Artikel über das Gedenkjahr 2005[10]
- GNE: Text über die Probleme einer GNE (Gaumennahterweiterung)[11]
- Yahoo Answer: Meinung eines Abtreibungsgegners[12] (hier wurden Umlaute wie "ue" zu "ü" geändert, um mit den Wortlisten arbeiten zu können)
- Widerstand.info: Artikel einer Organisation von Abtreibungsgegnern[13]
- Liebe: Liebestext[14]
- Das Kind spricht nicht: ein Vater regt sich über Gespräche auf[15]
- Microsoft: ein Brief an die Microsoft-GUI-Entwickler[16]
- Welt.de: über das Glücksgefühl eines Motorradfahrers in der Kurve[17]

Die elf ausgewählten nicht-emotionalen Texte sind vorrangig aus Onlinezeitungen bzw. sind sie häufig technischer Natur. Der Grund dafür ist, emotionalen Wörtern möglichst auszuweichen. Die sachlichen Texte sollen zeigen, wie häufig emotionale Wörter unbewusst oder auch unbeabsichtigt in den Sprachgebrauch einfließen. Die Themen selbst sind aber wieder so unterschiedlich wie bei den emotionalen Texten:

- Blog-Bedienungsanleitung: wie man einen Blog benutzt[18]
- Linux: der Wikipediaartikel zum Thema Linux[19]

[7] http://vierpfoten.at/website/output.php?id=1233&language=1, 16.9.2008
[8] http://vierpfoten.at/website/output.php?id=1226&language=1, 16.9.2008
[9] http://www.freiheitliche-jugend.at/standpunkte.php, 16.9.2008
[10] http://www.gajwien.at/suspect/suspect11/burnaustria.htm, 16.9.2008
[11] http://www.sanfte-zahnklammern.de/spangen/g_spreng/g_spreng.html, 16.9.2008
[12] http://de.answers.yahoo.com/question/index?qid=20070916093502AADByIa, 16.9.2008
[13] http://www.widerstand.info/meldungen/2885.html, 16.9.2008
[14] http://de.blog.360.yahoo.com/blog-bGoMabkpYrYnKfCWcodVQL4YHIE-?cq=1&p=2701, 16.9.2008
[15] http://blogs.taz.de/reptilienfonds/2008/09/16/das_kind_spricht_nicht/, 16.9.2008
[16] http://powerbook.blogger.de/2008/09/16/444050/liebe-microsoft-gui-verantwortliche/, 16.9.2008
[17] http://www.welt.de/print-welt/article210836/Kniet_nieder.html, 16.9.2008
[18] http://www.fockner.net/index.php?/pages/Blog-Bedienungsanleitung.html, 17.9.2008
[19] http://de.wikipedia.org/wiki/Linux, 17.9.2008

- Heise-Artikel: ein Artikel über den Eingriff in die E-Mail-Verschlüsselung[20]
- Zeit-Artikel 1: Artikel aus "Die Zeit" über den LHC[21]
- Donauau: Welche Pflanzen im September blühen[22]
- Die-Presse-Artikel: Google baut ein Satellitennetz[23]
- Zeit-Artikel 2: Artikel über aktuelle Forschung auf dem Gebiet der Krebsforschung[24]
- Schnapsherstellung: wie Schnaps erzeugt wird[25]
- Zeit-Artikel 3: Artikel über Neubesetzungen in Oxford[26]
- ORF: Artikel über dunkle Materie[27]
- MSN-Datenschutzbedingungen[28]

Die angegebenen Links zeigen die Quellen der Texte an. Dabei wird manchmal nur ein Teil des auf der Seite befindlichen Materials verwendet. Nicht zum eigentlichen Text zählende Teile (z.B. Publikationsort oder Autorenname) wurden herausgenommen um nicht zum Text nicht zugehörige Wörter ebenfalls in die Analyse miteinzubeziehen.

3.4.2 Konfiguration von GATE

Die vorliegende Wortliste des ADU ist in einer ".dex" Datei gespeichert, die einzelnen Wörter sind alphabetisch geordnet, wobei zuerst die Nomen und danach die Adjektive stehen. Vor jedem Wort steht die Emotionskategorie, wobei folgende Zuordnung getroffen wird:

- 1 Liebe
- 2 Begeisterung
- 3a Zufriedenheit
- 3b Erleichterung
- 4a Freude
- 4b Stolz

[20] http://www.heise.de/newsticker/Eingriff-in-E-Mail-Verschluesselung-durch-Mobilfunknetz-von-O2-/meldung/116073, 17.9.2008
[21] http://www.zeit.de/dpa/2008/9/15/iptc-bdt-20080910-480-dpa_18902676.xml?page=all, 17.9.2008
[22] http://www.donauauen.at/?area=news&story_id=804, 17.9.2008
[23] http://diepresse.com/home/techscience/internet/412957/index.do, 17.9.2008
[24] http://www.zeit.de/online/2008/36/direkte-reprogrammierung-von-zellen?page=all, 17.9.2008
[25] http://www.schnaps-vom-kaiserstuhl.de/schnaps-herstellung.htm, 17.9.2008
[26] http://www.zeit.de/2008/36/C-Oxford, 17.9.2008
[27] http://science.orf.at/science/news/145476, 17.9.2008
[28] http://privacy2.msn.com/de-at/fullnotice.aspx, 17.9.2008

- 5 Zorn
- 6 Furcht
- 7a Depressivität
- 7b Schuld
- 8a Ängstlichkeit
- 8b Scham

Für die Aufteilung in einzelne .lst Dateien, die vom Gazetteer benötigt werden, wird ein Bash-Script verwendet, das die Wörter nach ihrer Kategorie in einzelne Dateien aufteilt und die (alpha-)numerischen Zeichen löscht. Dieses Script für z.B. "Scham" sieht folgendermaßen aus:

```
#!/bin/bash
while read line
do
        if [ ${line:0:2} = "8b" ]; then
                echo ${line:3} >> Scham.lst
        fi
done
```

Nach dem Erstellen der einzelnen lst-Dateien sollte überprüft werden, ob deren Kodierung mit der in GATE eingestellten Kodierung übereinstimmt, um keine Probleme bei Wörtern mit Umlauten zu erhalten. Die Dateien werden im gleichen Ordner gespeichert, in dem auch die *lists.def* Datei liegt, die mit dem Gazetteer verwendet werden soll.

GATE wird mit der mitgelieferten Datei *german.gapp* initialisiert, um ein Grundgerüst für die Arbeit mit deutschen Texten zu erhalten. Der *Compound Analysis Gazetteer* wird jedoch aus der Liste der Processing-Resources herausgenommen, da dessen Ergebnisse nicht relevant für die Arbeit sind sondern nur Probleme bei der Negationssuche (siehe 3.4.3) bringen können. Weiters kann noch der *Orthomatcher* entfernt werden. Die übrigen PRs sind:

- reset - um das Dokument jedesmal vor dem Annotieren wieder rückzusetzen
- tokeniser - unterteilt den Text in einzelne Tokens
- tokeniser preprocessor - hilft z.B. dabei, korrekt numerische Aufzählungen zu erkennen (um sie nicht mit Satzenden zu verwechseln)
- splitter - unterteilt den Text in Sätze
- german gazetteer - greift auf die list.def Datei zu
- german grammar - für die Einteilung in beliebige Kategorien der gefunden, annotierten Wörter

Die (Initialisierungs-) Parameter werden einfach übernommen, lediglich beim German-Gazetteer wird darauf geachtet, dass auch wirklich nur ganze Wörter gefunden werden (wholeWordsOnly == true).

Um die Annotierung auf Texte anwenden zu können müssen diese in den "Language Resources" geladen und einem Korpus zugeordnet werden. Auf diesen Korpus können die PRs dann angewendet werden. Abbildung 3.2 zeigt die verwendeten PRs und die geladenen Texte.

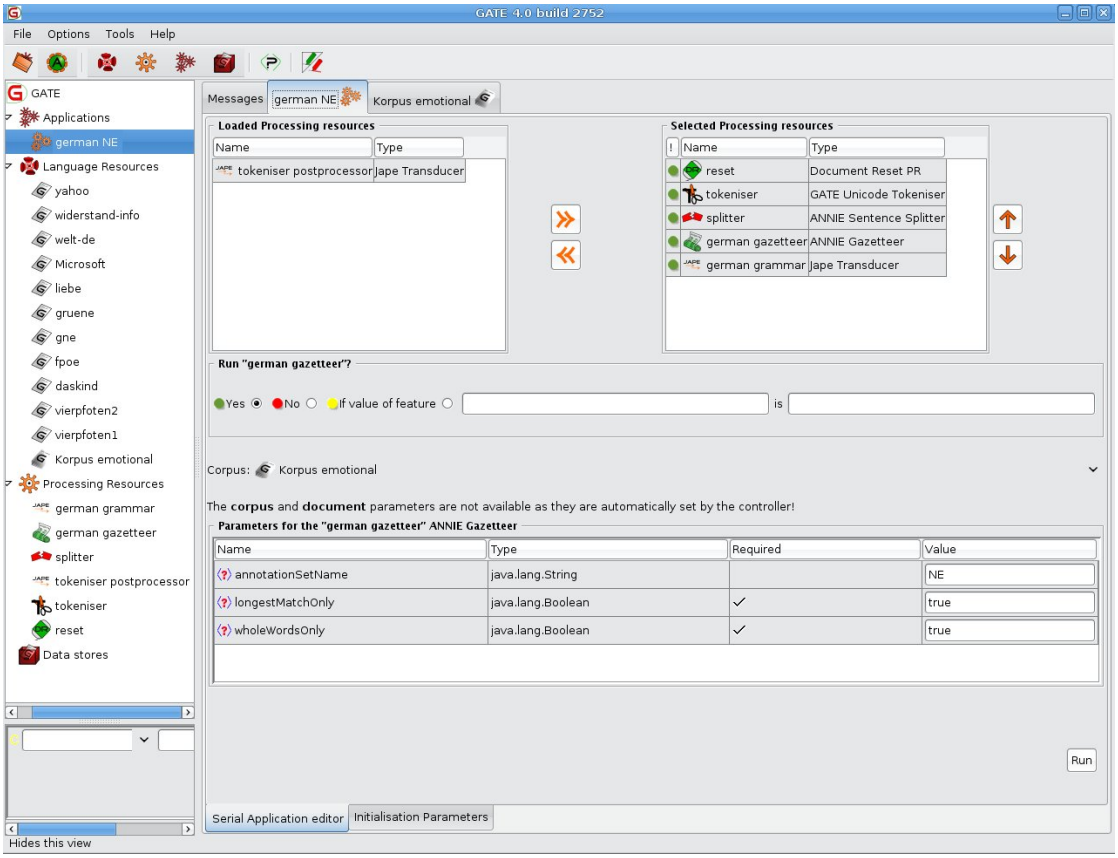

Abbildung 3.2: Konfiguration von GATE

3.4.3 Erforderliche Dateien

Die schon mitgelieferte *lists.def* Datei für den deutschen Sprachgebrauch, die im German-Gazetteer angegeben ist, wird durch eine *lists.def* Datei mit Einträgen für jede emotionale Kategorie ersetzt. Ein Beispieleintrag für die Kategorie *Erleichterung* sieht folgendermaßen aus:

```
Erleichterung.lst:emotion:erleichterung:german
```

Erleichterung.lst ist dabei die Datei, in der alle Wörter des Bereichs "Erleichterung" vorkommen. Die Doppelpunkte dienen als Trennungszeichen von *emotion*, dem "majorType" (Überbegriff) und *erleichterung*, dem "minorType" (Unterbegriff). Diese Zuordnung wird benötigt um einzelne Annotationen zu erhalten. *german* steht für die Sprache und dient nur der Vollständigkeit. Obligatorisch sind nur der Listenname und der majorType.

Weiters werden noch vier JAPE-Dateien, die üblicherweise in einem Verzeichnis namens *grammar* gespeichert werden, verwendet: *emotion.jape* und *allemotions.jape*, sowie *counter.jape* und *main.jape*. Die letzte Datei dient dazu, eine Referenz auf die drei anderen Dateien darzustellen, da man lediglich eine Datei im Gazetteer laden kann.

main.jape

```
MultiPhase:     TestTheGrammars
Phases:
emotion
counter
allemotions
```

counter.jape zählt alle Wörter eines Textes. Dabei werden keine Nummern oder Datums-Token mitgezählt, da nur Wort-Token emotional sein können und von der Summe an emotionalen Wörtern auf das Dokument geschlossen werden soll. Alle Tokens als Gesamtwortzahl zu nehmen wäre unzulässig, da diese auch Satzzeichen enthalten. Die Regel "countRule" sucht dabei Tokens, dessen Art ("kind") es ist, ein Wort zu sein. Die so gefundenen Tokens von der LHS (Left-Hand-Side) der Regel werden nach dem "–>" auf der RHS (Right-Hand-Side) der Gruppe "Wortanzahl" zugeordnet und damit annotiert. "rule = "countRule" ist eine Art der annotierten Wörter, dient hier aber lediglich zum Debuggen, um zu sehen, welche Regel beim Annotieren zum Einsatz kam.

counter.jape

```
Phase: Counter
Input: Lookup Token
Options: control = appelt

Rule: countRule
(
        ({Token.kind == word})

):counterlabel
-->
:counterlabel.Wortanzahl = {rule = "countRule"}
```

Ähnlich wie *counter.jape* sehen auch die Regeleinträge von *emotion.jape* aus. Als Beispiel wird die "AngstRule" angeführt. Hier wird der majorType und minorType jedes Worts überprüft und bei einem Treffer wird diesem Wort die Art "aengstlich" zugeordnet. Die übrigen elf Emotions-Regeln sind gleich aufgebaut.

Die “negationRule“ hat eine höhere Priorität als die übrigen Regeln, die eine Standardpriorität von -1 haben. Sie wird daher vor den anderen ausgeführt, um Negationsannotationen schon vorher zu ”belegen“, damit andere Regeln dort nicht mehr zutreffen können. Dabei wird nach einem Verneinungswort, das in einer *Negation.lst* Datei vorkommt gesucht, das einem Emotionswort innerhalb eines Satzes folgt (oder umgekehrt). Die Beschränkung “innerhalb eines Satzes” wird mit dem Input-Parameter “Split” sichergestellt. Da aber im Falle einer Negation nichts annotiert werden soll und die JAPE-Grammatik keinen Nicht-Operator wie ! kennt, wird dort einfach “Nichts”, was für JAPE dem Ausdruck {} entspricht, angegeben.

emotion.jape

```
Phase: Emotion
Input: Split Lookup
Options: control = appelt

Rule: negationRule
Priority: 30
( {Lookup.majorType == verneinung} {Lookup.majorType == emotion} )
|
( {Lookup.majorType == emotion} {Lookup.majorType == verneinung} )
:negation -->
        {}

//
// ANGST
//
Rule: angstRule
(
        ({Lookup.majorType == emotion, Lookup.minorType == aengstlich})

):emotionlabel
-->
:emotionlabel.Aengstlichkeit = {kind="aengstlich", rule = "angstRule"}
```

Die Datei *allemotions.jape* enthält, genauso wie *emotion.jape*, die negationRule und zusätzlich eine Regel um alle Emotionswörter zu annotieren.

allemotions.jape

```
 Rule: allEmotionRule
(
        {Lookup.majorType == emotion}
):allemotionlabel
-->
:allemotionlabel.AlleEmotionen = {kind="emotion", rule = "allEmotionRule"}
```

3.4.4 Annotieren der Texte

Das Annotieren der Texte geschieht unspektakulär. In der “Application german NE” wird nach der Konfiguration von GATE auf den Button “Run” geklickt und auf das Ergebnis

gewartet. Nach dem Öffnen eines zu annotierenden Textdokuments sind die gefundenen Kategorien im “Annotation Sets”-Panel rechts zu sehen. Klickt man die gewünschte Annotierung an, werden alle gefundenen Wörter in einer bestimmten Farbe unterlegt und im (durch Klick auf “Annotations” geöffneten) Annotations-Panel werden die Typen, Merkmale (“Features”) sowie der Start- und Endpunkt angegeben. Darunter wird noch die Gesamtanzahl der Annotationen dieser Kategorie angezeigt. Ein Beispiel dazu zeigt Abbildung 3.3.

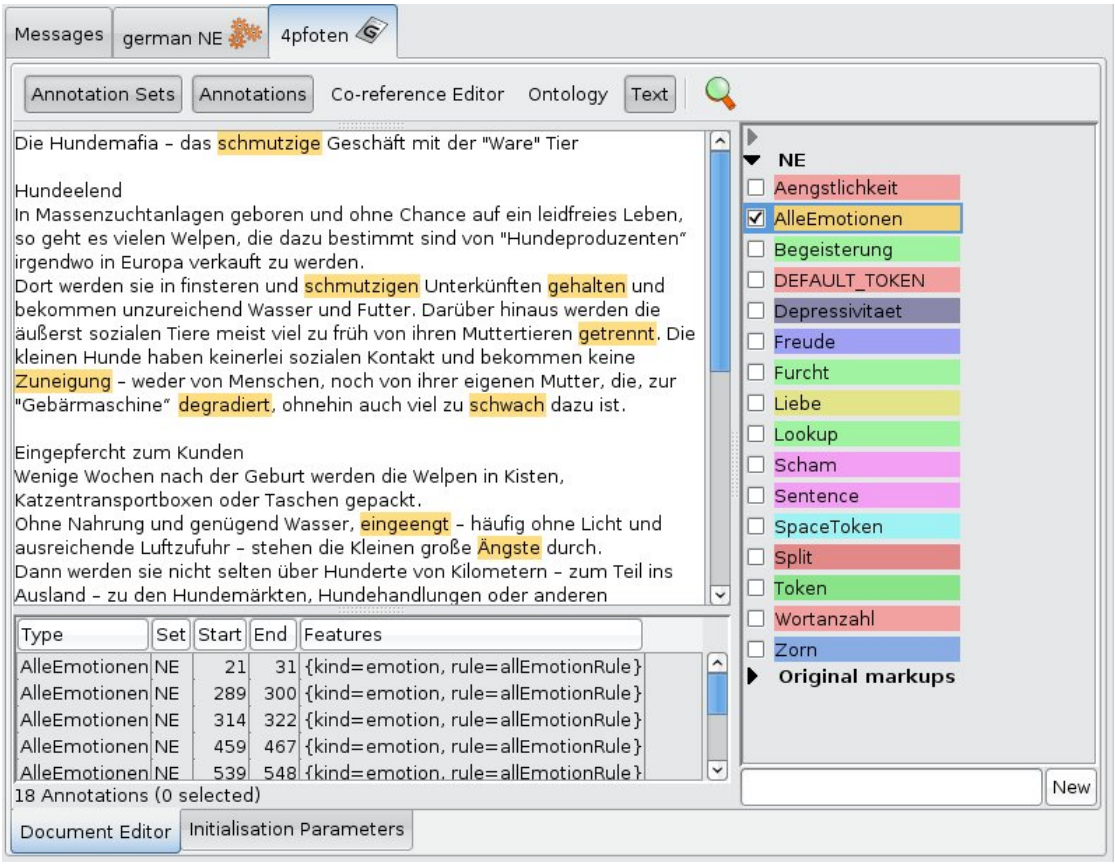

Abbildung 3.3: Annotation Sets für Vier-Pfoten 1 Dokument

4 Ergebnisse

Das angestrebte Ziel ist es zum einen mit der Gesamtanzahl der gefundenen emotionalen Wörter eine Skala zu erstellen, nach der ein Text als emotional oder nicht emotional unterschieden werden kann. Zum anderen soll aus der Summe der gefundenen Annotationen darauf geschlossen werden können, ob ein Text freundlich, aggressiv, liebevoll, etc. ist.

Die Tabelle enthält die Rohdaten für die Berechnungen, die für dieses Ziel notwendig sind. *Text* entspricht dabei den in 3.4.1 angeführten Kurznamen für die Texte, *Gesamt* ist die Gesamtwortzahl, das sind alle echten Worte, die im Text vorkommen (ein Datumswort kann nie emotional sein kann und würde somit nur die Statistik verfälschen). *alle E* ist die prozentuale Anzahl der emotionalen Wörter und die Spalten *meiste E 12* und *meiste E 8* sind die Emotionskategorien mit den meisten Annotierungen im Text nach der Einteilung in zwölf bzw. acht Kategorien. Die Einteilung in nur acht Bereiche erfolgt um eventuell eindeutigere Werte zu erhalten.

Text	Gesamt	alle E in %	meiste E 12	meiste E 8
Vier-Pfoten 1	287	6,272	Depressivität(7a)	Ängstlichkeit(8)
Vier-Pfoten 2	427	3,044	Begeisterung(2)	Begeisterung(2)
Jugend-FPÖ	915	2,404	Furcht(6)	Freude(4)
Grüne-Jugend	599	2,17	Depr.(7a),Freude(4a)	Depr.(7),Freude(4)
GNE	966	2,277	Ängstlichkeit(8a)	Ängstlichkeit(8)
Yahoo Answer	132	5,30	Depr.(7a),Freude(4a)	Depr.(7),Freude(4)
Widerstand.info	368	2,446	Furcht(6)	Furcht(6),Freude(4)
Liebe	355	1,972	Liebe(1)	Liebe(1)
Das Kind	933	1,179	Depressivität(7a)	Depressivität(7)
Microsoft	250	0,8	Liebe(1),Scham(8b)	Liebe(1),Ängstl.(8)
Welt.de	787	2,541	Freude(4a)	Freude(4)
Blog-Bedienung	825	1,09	Begeisterung(2)	Begeisterung(2)
Linux	4682	1,153	Freude(4a)	Freude(4)
Heise-Artikel	332	0,602	Liebe(1),Scham(8b)	Liebe(1),Ängstl.(8)
Zeit-Artikel 1	612	1,633	4a,7a,8a	Freude(4)
Donauau	457	1,97	Freude(4a)	Freude(4)
Presse-Artikel	483	1,035	4a,4b,7a,8a	Freude(4)
Zeit-Artikel 2	616	0,649	2,4a,6,7a	2,4,6,7
Schnapsherstellung	357	0,28	Freude(4a)	Freude(4)
Zeit-Artikel 3	412	0,243	Ängstlichkeit(8a)	Ängstlichkeit(8)
ORF	602	0,831	Freude(4a),Furcht(6)	Freude(4),Furcht(6)
MSN	2788	0,681	Begeisterung(2)	Begeisterung(2)

5 Schlussbetrachtungen

5.1 Diskussion und Interpretation der Ergebnisse

5.1.1 Einteilung in Gefühlskategorien

Ein Ziel dieser Arbeit soll die Einteilung der Texte in bestimmte Gefühlskategorien sein. Dazu wurde der, zugegeben sehr naive, Ansatz des Opinion Mining gewählt: Sind im Text mehr positive als negative Meinungswörter, ist die Meinung des Textes positiv, sonst negativ[4]. Bei der Betrachtung der Tabelleneinträge "meiste E 12" bzw. "meiste E8" auf Seite 29 verglichen mit lediglich den Textbeschreibungen, die in Kapitel 3.4.1 gegeben sind, fallen Widersprüchlichkeiten auf.

So behandelt der *Vier-Pfoten 2*-Text die Haltung von Kaninchen in Mastbetrieben, hat aber am meisten Begeisterungsworte enthalten. Dies könnte durch den positiv wirkenden Teil des Textes kommen, in dem von den Erfolgen des Projekts berichtet wird.

Ungewöhnlich sticht auch der *Grüne-Jugend*-Text ins Auge. Die Freude im Text ist in Textauszügen wie folgendem schwer zu finden:

> [...]Die bösen Alliierten! Haben "unser" Land geteilt! Genau das könnten die meisten Menschen nach dieser Aktion verstehen. Denn auch hier wird wieder verabsäumt zu sagen, warum das Ganze denn so war. Die Bundesregierung will damit zeigen, dass "unser" Land nicht immer frei war. Wieso zeigen sie dann nicht, wie eingeschränkt alles zur Zeit des Naziregimes war? Das war nun wirklich viel schlimmer. [...]

Hier wurde unter anderem dreimal das Wort "Befreiuung" annotiert, sowie einmal "Freiheit". Dass diese Wörter dabei aber nicht die persönliche Freiheit des Autors bezeichnen, konnte nicht berücksichtigt werden und führt so zu dieser merkwürdigen Zuweisung. Die Wörter "Freiheit" bzw. "freie" (Wahl) weisen auch dem Text *Yahoo Answer* die Kategorie "Freude" zu, obwohl sein Autor von "Mord auf brutalste Weise" in Bezug auf Abtreibung schreibt.

Der Text *Das Kind spricht nicht* ist laut Annotation depressiv, da der Autor fragt, ob "[...] es ein Verlust [ist], einmal weniger zu hören, dass das mit den Juden ja schon eine schlimme Sache war damals [...]" und die Worte "Verlust" und "schlimm" vorkommen.

Dass der Brief an die Microsoft-Entwickler wiederum "Liebe" und "Scham" ausdrückt, kommt daher, dass lediglich zwei Wörter annotiert wurden: das Wort "Liebe" in der Anrede und im Satz "Aber könnt ihr mir mal verraten, welcher Teufel Euch geritten hat, [...]" das Wort "verraten".

Die einzigen Zuordnungen, die als menschlicher Annotator vielleicht auch so getroffen werden würden, sind die der Texte *Vier-Pfoten 1* (Depressivität) und *Welt.de* (Freude).

Zusammenfassend lässt sich sagen, dass das Ziel dieser Arbeit, mit GATE Texten eine Gefühlskategorie zuzuweisen, mit dem hier gewählten Ansatz nicht erreicht wurde. Die gefundenen Annotationen waren aufgrund ihrer Anzahl oft nicht aussagekräftig genug. Eine Erklärung dafür ist, dass vor allem Texte ausgesucht wurden, in denen emotional über etwas geschrieben wurde, der Autor aber nicht immer über sich selbst reflektiert hat. Das *Affektive Diktionär Ulm* wurde allerdings auch nicht dazu erstellt, um diese Art der Emotionalität festzustellen, sondern um bei der Psychoanalyse dem Psychoanalytiker helfen zu können, den Inhalt von Gesprächen mit Patienten eine Gefühlskategorie zuweisen zu können. Solche Transkriptionen hätten vielleicht zu einem anderen Ergebnis geführt. Allerdings schreibt auch Silke Jahr über das ADU, dass "Wörter, die isoliert als Gefühlswörter bewertet wurden, im Kontext teilweise einer anderen Emotionskategorie zugeordnet oder als nicht-emotional eingestuft werden mußten" [14].

5.1.2 Einteilung in emotional/nicht-emotional

Im Gegensatz zur Einteilung nach Kategorien kann hier durch Betrachten der Anzahl an emotionalen Wörtern in den Texten eine Tendenz erkannt werden. Da das Ziel aber eine Skala sein soll, die wie in 2.5.3 beschrieben einen Wertebereich aufweist, wird versucht mit Hilfe einer statistischen Verteilung eine Lösung zu finden. Der Wertebereich soll zwischen den Werten für eindeutig emotionale und eindeutig nicht-emotionale Texte eine Grenze darstellen.

Für die Berechnungen werden zunächst der Mittelwert[1] und die Standardabweichung[2] von beiden Textgruppen (emotional/nicht-emotional) und von beiden gemeinsam errechnet. Für die emotionalen Texte wird errechnet:

$$\overline{x_e} = 2,764 \qquad s_e = 1,853 \tag{5.1}$$

und für die nicht-emotionalen Texte:

$$\overline{x_{\neg e}} = 0,924 \qquad s_{\neg e} = 0,606 \tag{5.2}$$

Der Mittelwert und die Standardabweichung aller ausgewerteten Texte ist:

$$\overline{x_{e\neg e}} = 1,844 \qquad s_{e\neg e} = 2,453 \tag{5.3}$$

Aufgrund der von mir gewählten Vorauswahl (stark emotional/kaum emotional) sind Texte mit "normal viel" Emotionalität, die den Großteil einer Menge von zufällig gewählten Texten ausmachen würden, in den Datensätzen nicht vorhanden. Weiters ist die Stichprobengröße von elf sehr gering gewählt, somit kann der Wert für die Standardabweichung bereits durch einen einzigen Ausreißer (fast 10% der Stichprobe!) stark verzerrt werden. Darum kann keine statistisch einwandfreie Aussage über die Verteilung von Emotionalität in Texten und somit über die Schwellenwerte (Anfang und Ende des Wertebereichs von "normal emotional") getroffen werden.

[1] $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

[2] $s = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2\right) - \overline{x}^2}$

Dennoch zeigen die Werte eine recht eindeutige Tendenz: den höchsten Wert von Nicht-Emotionalität (1.97) und den niedrigsten Wert von Emotionalität (0, 8). Der letzte Wert ist der des *Microsoft*-Textes, in dem nur zwei Wörter annotiert wurden, und diese falsch (vgl. 5.1.1). Betrachtet man den Text daher als Ausreißer, so wäre der *Das Kind spricht nicht*-Text der mit dem niedrigsten Wert (1, 179). Ich würde daher einen Wertebereich aufstellen, der von $1,179 - 1.97$ geht, also ungefähr zwischen 1% und 2% liegt. Dies wäre der Bereich, in dem man keine eindeutige Zuordnung zu emotional oder nicht-emotional treffen kann. Liegt der errechnete Wert aller Emotionswörter in Bezug auf den Text unter 1, spreche ich von nicht-emotionalen Texten, liegt er über 2 ist der Text emotional.

Die Zuordnung in die Kategorien emotional/nicht-emotional kann daher als fast erreichtes Ziel gesehen werden. Zwar wurden auch hier viele Worte falsch annotiert oder gar emotionalere Texte als nicht sehr emotional eingestuft. Eine ungefähre Tendenz kann aber dennoch herausgelesen werden. Da im Zuge dieser Arbeit keine größere Stichprobengröße genommen werden konnte, kann keine statistische Aussage über die Werte getroffen werden. Es wurde aber ein Wertebereich vorgeschlagen, der grob auf die Verteilung der Werte eingeht.

5.2 Probleme der Textanalyse

Im Verlauf der Erstellung einer solchen Arbeit werden verschiedenste Probleme, die bei der Textanalyse auftreten können, sichtbar. So z.B. muss unterschieden werden zwischen gesprochener und geschriebener Sprache. Gesprochene Sprache hat Eigenheiten, auf die speziell eingegangen werden muss. So werden häufig Worte zusammengefügt: "Hast du's schon gekriegt?". Würde dieser Satz hochdeutsch in einem Text verfasst sein, würde wohl eher "Hast du es schon erhalten?" dort stehen.

Ein weiteres Problem, das während der Suche nach Texten aufgefallen ist, sind Tippfehler. Gerade Blog-Einträge und ähnliches, womöglich auch noch in einer sehr emotionalen Phase schnell verfasst, werden kaum noch einmal gelesen bevor sie online gestellt werden. Ebenso stellt die Verwendung von "ue, oe oder ae" für deutsche Umlaute eine Herausforderung dar (vlg. Text *Yahoo-Answer*). Da das ADU keine "Meinten Sie: ..."-Vorschlagsfunktion wie Google beinhaltet, können solche Fehler oder Eigenheiten nicht automatisch erkannt und berücksichtigt werden.

Dass die Bedeutung eines Wortes von dessen Kontext abhängt, ist eine Tatsache, die unabdinglich in einer Textanalyse behandelt werden muss. Problematisch wird aber auch diese Bewertung, wenn z.B. sarkastische oder zynistische Aussagen im Text vorkommen. Nicht jeder gibt diesen Äußerungen die gleiche Bedeutung. Probleme kann hier aber auch ein Text mit grammatikalischen Fehler verursachen. Worauf sich ein emotionales Wort bezieht, ist dann vielleicht nicht mehr eindeutig nachvollziehbar.

Unterschiedliche Textarten müssen unterschiedlich interpretiert werden. Für die Textanalyse ist wichtig zu wissen, ob es sich um einen politischen Text, eine Erzählung oder einen Fachtext handelt. Die Ergebnisse der Arbeit wären vielleicht eindeutiger gewesen, wären nur Texte einer bestimmten Art genommen worden. Diese sind einander ähnlicher und zeigen so vielleicht vergleichbare Muster, die eine besser zu analysierende statisti-

sche Verteilung ergeben hätten. Problematisch bleiben dabei aber immer noch Texte mit dichterischer Freiheit. Will jemand z.B. ein Gedicht über Liebe schreiben, kann ein zu "Liebe" synonymes Wort in fast jedem Satz vorhanden sein. Diese Eigenheiten, ebenso wie stilistische Ausprägungen eines Textes können die Ergebnisse stark verfälschen.

Weiters sind Worte wie "Freiheit" problematisch, da sie, wie man in der Analyse in 5.1.1 gesehen hat, in verschiedensten Arten von Texten erscheinen. Das Wort "frei" und Abwandlungen davon wurden z.B. im *Linux*-Text elf mal annotiert, und tragen daher zu der großen Anzahl an "Freude"-Wörtern bei. Das Problem ist, dass sich die Autoren hier nicht für ein synonymes Wort entscheiden konnten, da auf den Ausdruck "freie Software" hingewiesen werden wollte. Da solche Worte von der emotionalen Einstellung des Verfassers unabhängig sind, sollten sie als nicht-wertend interpretiert werden.[14]

Ebenfalls ein Problem bei der Textanalyse ist die Tatsache, dass das Erleben und Benennen von Gefühlen sehr unterschiedlich ist. Ein Autor muss nicht in jedem Text seine Gefühle gleich erleben und zu Papier bringen. Und nicht jeder Mensch hat die gleiche Art und Weise über seine Gefühle zu schreiben, manche sind verhaltener, manche offener.

Weiters lässt sich nicht klar zwischen Emotionalität und Nicht-Emotionalität unterscheiden. Das ADU kann dabei helfen, wie auch eine Skala, die nicht nur einen Grenzwert hat, über dem von Emotionalität gesprochen werden kann, sondern einen Wertebereich, in dem nicht klar ist, wie zu entscheiden ist. Eine eindeutige Lösung lässt sich nicht finden. Passend dazu kommt die Frage auf, ab wann von Zorn oder Aggresivität oder ab wann von Furcht oder Ängstlichkeit gesprochen werden kann. Nicht immer kann ein klarer Unterschied zwischen solchen Kategorien gezogen werden.

Und letztlich ist die Intention eines Autors ausschlaggebend. Ob ein Autor absichtlich Emotionalität miteinfließen ließ, um seine Leser zu beeinflussen, oder ob er sie völlig unbewusst benutzt hat, lässt sich schließlich nicht feststellen.[14]

5.3 Ausblick

Wie sich gezeigt hat, ist die Aufgabe der Analyse von Texten in Bezug auf ihren emotionalen Wert nicht gar so einfach wie anfangs gedacht. Der wichtigste und zugleich am schwierigsten zu realisierende Aspekt ist es, zu erkennen, in welchem Zusammenhang ein Wort gebraucht wird. Eine solch qualitative Lösung ist unumgänglich. Ebenso kann auf die Punktation eingegangen werden, da Rufzeichen ebenfalls dem vorangegangenen Satz einen emotionalen Charakter zuweisen können. Die besten Ergebnisse in der Literatur haben Machine-Learning-Verfahren erreicht[2, 19]. Hier könnte GATE mit geeigneten Wortlisten beim Annotieren helfen um emotionale Wörter schneller zu finden. Ebenso kann GATE zum maschinellen Lernen selbst eingesetzt werden. Genaueres dazu findet sich im User-Guide des Text-Engineering Programms.

Weiters wäre es wichtig, die von Silke Jahr angesprochene Ich-Bezogenheit emotionaler Konstrukte zu erkennen. Nur dann kann auf die Emotinalität des Autors wirklich eingegangen werden (vgl. 2.5.3). Ebenso könnte die Stärke eines Emotionswortes gemessen werden. Dazu kann der in Kapitel 2.5.3 angesprochene Expressivitätsfaktor zum Einsatz

kommen.

Wie schon im vorhergehenden Kapitel "Probleme der Textanalyse" (5.2) angesprochen, sollte beim Auswählen der Texte für die Trainingsset-Korpora darauf geachtet werden, Texte ähnlicher Art zu verwenden. Dadurch können vielleicht beim Testset, das wiederum Texte ähnlich zum Trainingsset beinhaltet, bessere Ergebnisse erzielt werden.

Abschließend kann gesagt werden, dass jede Verbesserung des hier gewählten Ansatzes vielleicht zu exakteren Resultaten führt, dass aber jede Methode dem Verfasser eines Textes die Eigenschaft der Emotionalität lediglich unterstellen und niemals beweisen kann.

Literaturverzeichnis

[1] Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. Emotions from text: machine learning for text-based emotion prediction. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 579–586, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

[2] Saima Aman and Stan Szpakowicz. Identifying expressions of emotion in text. In Václav Matousek and Pavel Mautner, editors, *TSD*, volume 4629 of *Lecture Notes in Computer Science*, pages 196–205. Springer, 2007.

[3] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, 2002.

[4] Xiaowen Ding, Bing Liu, and Philip S. Yu. A holistic lexicon-based approach to opinion mining. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 231–240, New York, NY, USA, 2008. ACM.

[5] EuroWordNet. Eurowordnet:building a multilingual database with wordnets for several european languages. `http://www.illc.uva.nl/EuroWordNet/`. Online-Quelle; zugegriffen am 6.9.2008.

[6] Jeffrey E. F. Friedl. *Reguläre Ausdrücke*. O'Reilley, 2006.

[7] Klinik für psychosomatische Medizin und Psychotherapie. `http://www.klinik.uni-frankfurt.de/zpsy/psychosomatik/pages/Forschung/3_1.htm`. Online-Quelle; zugegriffen am 21.9.2008.

[8] GATE. Annotationmanual. `http://gate.ac.uk/sale/am/annotationmanual.pdf`. Online-Quelle; zugegriffen am 14.9.2008.

[9] GATE. Gate information extraction. `http://gate.ac.uk/ie/index.html`. Online-Quelle; zugegriffen am 14.9.2008.

[10] GermaNet. Germanet. `http://www.sfs.uni-tuebingen.de/lsd/`. Online-Quelle; zugegriffen am 12.9.2008.

[11] J. W. Berry H. Dahl, M. Hölzer. *How to classify emotions for psychotherapy research.* Ulm: Ulmer Textbank., 1992.

[12] G. Heyer, U. Quasthoff, and T. Wittig. *Wissensrohstoff Text; Text Mining: Konzepte, Algorithmen, Ergebnisse*. w3l-Verlag, Bochum, 2005.

[13] M. Hölzer, N. Scheytt, and H. Kächele. Das 'Affektive Diktionär Ulm' als eine Methode der quantitativen Vokabularbestimmung. In *Textanalyse. Anwendungen der computerunterstützten Inhaltsanalyse. Beiträge zur 1. Textpack-Anwenderkonferenz*, pages 185–212, Opladen, Germany, 1992. Westdeutscher Verlag.

[14] Silke Jahr. *Emotionen und Emotionsstrukturen in Sachtexten. Ein interdisziplinaerer Ansatz zur qualitativen und quantitativen Beschreibung der Emotionalitaet von Texten*. Walter de Gruyter, 2000.

[15] Bing Liu. *Web Data Mining*. Springer, 2007.

[16] Ulrich Mees. *Die Struktur der Emotionen*. Hogrefe-Verlag für Psychologie, 1991.

[17] Dr Dieter Merkl. Clustering. `http://www.ec.tuwien.ac.at/~dieter/teaching/dm08-clustering.pdf`. Online-Quelle; zugegriffen am 6.9.2008.

[18] Dr Dieter Merkl. Evaluation. `http://www.ec.tuwien.ac.at/~dieter/teaching/dm08-evaluation.pdf`. Online-Quelle; zugegriffen am 12.9.2008.

[19] Carlo Strapparava and Rada Mihalcea. Learning to identify emotions in text. In *SAC '08: Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560, New York, NY, USA, 2008. ACM.

[20] Dr. Renate Wahrig-Burfeind. Fremdwörterlexikon, 2003.

[21] Wikipedia. Ontologie (informatik) — wikipedia, die freie enzyklopädie. `http://de.wikipedia.org/w/index.php?title=Ontologie_(Informatik)&oldid=50919175`, 2008. Online-Quelle; Stand 21.9.2008.

[22] WordNet. Wordnet. `http://wordnet.princeton.edu/`. Online-Quelle; zugegriffen am 12.9.2008.

[23] Dresdner Angst Wörterbuch. Daw - verfahren. `http://rcswww.urz.tu-dresden.de/~berth/daw/dawverfahren1.html`. Online-Quelle; zugegriffen am 10.9.2008.