Gottfried Wilhelm
Leibniz Universität Hannover
Fakultät für Elektrotechnik und Informatik
Institut für Praktische Informatik
Fachgebiet Software Engineering

Erstellung eines deutschen Datensatzes zur Stimmungsanalyse von Entwickleraussagen

Development of a German Dataset for Sentiment Analysis of Developer Statements

Bachelorarbeit

im Studiengang Informatik

von

Raymond Ochsner

Prüfer: Prof. Dr. rer. nat. Kurt Schneider
Zweitprüfer: Dr. rer. nat. Jil Ann-Christin Klünder
Betreuer: M.Sc. Martin Obaidi

Hannover, 25.08.2022

Erklärung der Selbstständigkeit

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit selbständig und ohne fremde Hilfe verfasst und keine anderen als die in der Arbeit angegebenen Quellen und Hilfsmittel verwendet habe. Die Arbeit hat in gleicher oder ähnlicher Form noch keinem anderen Prüfungsamt vorgelegen.

Hannover, den 25.08.2022

Raymond Ochsner

Zusammenfassung

Die Stimmungsanalyse ist ein wichtiges Verfahren, um die Stimmungen in Entwicklerteams zu untersuchen und darauffolgend die Produktivität und den Erfolg des Teams zu sichern. Bisher arbeiten Stimmungsanalysetools im Bereich des Software Engineerings nur auf Basis von englischen oder nichtdeutschen Goldstandard-Datensätzen. Um ein breiteres Spektrum an Daten für die Stimmungsanalysetools zu bieten, wurde in dieser Arbeit ein deutscher Datensatz mit 5.949 verschiedenen Entwickleraussagen, die aus dem deutschen Entwicklerforum Android-hilfe.de[1] extrahiert wurden, erstellt. Diese Aussagen wurden dann anhand des Emotionsmodells von Shaver et al. [55] von vier deutschsprachigen Informatikstudenten mit Kenntnissen der Softwareentwicklung in die sechs Basisemotionen gelabelt. Eine Evaluation des Labelprozesses wurde durchgeführt. Diese kam zu dem Ergebnis, dass der Datensatz hohe Übereinstimmungs- und Reliabilitätswerte aufweist. Basierend auf diesen Werten ist der erstellte Datensatz valide und aussagekräftig genug, um mit ihm im deutschsprachigen Raum zu arbeiten. Auswertungen in vorhandenen deutschen Stimmungsanalysetools zeigen, dass ein Tool, dass auf die Domäne des Software Engineering spezialisiert ist, fehlt. Auf Möglichkeiten, das Labeln des Datensatzes zu optimieren, wird eingegangen. Ebenso werden weitere Anwendungsfälle vorgestellt.

[1] https://www.android-hilfe.de/

Abstract

Development of a German Dataset for Sentiment Analysis of Developer Statements

Sentiment analysis is an important method for examining the moods in development teams and subsequently ensuring the productivity and success of the team. Until now, sentiment analysis tools in software engineering have only worked on the basis of English or non-German gold standard datasets. To provide a broader range of data for sentiment analysis tools, this work created a German dataset with 5.949 different developer statements extracted from the German developer forum Android-hilfe.de[2]. These statements were then labeled by four German-speaking computer science students with knowledge of software development into the six basic emotions using the emotion model of Shaver et al. [55]. An evaluation of the labeling process was conducted. This concluded that the data set had high agreement and reliability values. In conclusion, the data set created is valid and meaningful enough to work with in German-speaking domains. Evaluations in existing German sentiment analysis tools show that a tool specialized in the domain of software engineering is missing. Possibilities to optimize the labeling of the data set are discussed. Likewise, further use cases are presented.

[2]https://www.android-hilfe.de/

Inhaltsverzeichnis

Kapitel 1

Einleitung

1.1 Motivation

Stimmungen haben Einfluss auf unsere Denkweise und dementsprechend auch auf unser Verhalten [7] [11] [16]. Vor allem in Entwicklerteams im Bereich des Software Engineerings (SE), also dort, wo viel Kommunikation und soziale Interaktion stattfindet, kann eine ausgedrückte Stimmung einen gewissen Effekt bei den Mitmenschen auslösen [16]. So kam eine Studie, die von Graziotin et al. [16] durchgeführt wurde zu dem Entschluss, dass positive Stimmungen in Entwicklerteams die Produktivität der Entwickler steigern. Um diese Stimmungen genauer deuten oder beispielsweise negative Stimmungen frühzeitig erkennen zu können, werden Stimmungsanalysetools verwendet [31] [50] [57] [62].

Dabei wird zwischen solchen, die lexikonbasiert arbeiten und den Stimmungsanalysetools, die Prinzipien des Machine-Learnings nutzen, unterschieden [1]. Ein großes Problem dabei ist, dass diese Tools nicht auf die Domäne des SE angepasst sind und zu ungenauen Ergebnissen führen [5] [31] [41]. 2018 warnten Lin et al. [31] Forscher davor, sich auf die Ergebnisse der Anwendung von SE-spezifischen Aussagen in Stimmungsanalysetools zu verlassen, da die Tools noch nicht reif genug sind, um sie für weitere Tätigkeiten, wie beispielsweise das Vorschlagen von Softwarebibliotheken basierend auf den Meinungen von Entwicklern, zu nutzen. Zwar existieren für die deutsche Sprache mehrere Goldstandard-Datensätze für die allgemeine Stimmungsanalyse [8] [37], jedoch keines spezifisch für den SE-Bereich. Ebenfalls wurden neue Stimmungsanalysetools entwickelt, die an die Domäne des SE angepasst sind [2] [5] [27], jedoch existiert nach bestem Wissen noch keines, das mit einem deutschen Datensatz trainiert wurde. Aufgrund dessen kann ein deutscher Goldstandard-Datensatz bei der Entwicklung eines solchen spezifischen Tools genutzt werden. Eine Möglichkeit, so einen deutschen Datensatz zu erstellen wäre, die

vorhandenen englischen Datensätze mit Hilfe von Machine Translation zu übersetzen. Aufgrund der Komplexität von natürlichen Sprachen, falscher Grammatik in den Entwickleraussagen oder falscher Interpretation von bestimmten Redewendungen kann dies aber das Ergebnis verfälschen, weswegen die Verwendung von Entwickleraussagen in der Originalsprache zu bevorzugen ist [43] [49]. Aufgrund dieser Tatsachen wird in dieser Arbeit ein Goldstandard-Datensatz mit Hilfe von original deutschsprachigen Entwickleraussagen aus dem deutschen Android-App-Entwicklungs-Forum Android-Hilfe[1] erstellt.

1.2 Herangehensweise

Da es für die Stimmungsanalyse noch kein deutsches SE-spezifisches Tool gibt, benötigt es im Falle der Anwendung von Machine-Learning-Prinzipien einen Goldstandard-Datensatz, der im Rahmen dieser Bachelorarbeit erstellt werden soll. Um dies zu verwirklichen, wird in erster Instanz die Quelle für den Datensatz analysiert und ein Konzept für das Extrahieren der Entwickleraussagen erstellt. Sofern dies möglich ist, wird dieser Datensatz vorab mit einem deutschen Stimmungsanalysetools vorsortiert, um eine ähnliche Verteilung der Polaritäten *Positiv, Negativ* und *Neutral* auszuwählen und ihn dem Goldstandard gerecht werden zu lassen. Eine konzipierte Guideline, die sich auf ein Emotionsmodell stützt, dient dabei als Grundlage für das Labeln des Datensatzes in die zugehörigen Emotionen durch den Autor und vier weiteren Informatikstudenten. Abschließend sollen die Ergebnisse in vier deutschen Stimmungsanalysetools evaluiert und diskutiert werden, um sie ihrer Validität einordnen zu können.

1.3 Struktur der Arbeit

In Kapitel 2 wird auf die Grundlagen der Stimmungsanalyse und von Emotionen eingegangen. Dabei werden verschiedene Stimmungsanalysetools und Emotionsmodelle genannt. Außerdem werden Goldstandard-Datensätze genauer betrachtet und relevante Evaluationsmetriken vorgestellt.
Kapitel 3 enthält verwandte Arbeiten im Bereich der Erstellung von Datensätzen, sowohl innerhalb als auch außerhalb des SE-Bereichs und Arbeiten, die sich mit der Stimmungsanalyse befassen.
Daraufhin wird in Kapitel 4 das Konzept und die Erstellung des deutschen Datensatzes beschrieben.
In Kapitel 5 wird die Guideline und die Vorgehensweise beschrieben, wie die Teilnehmenden des Workshops anhand eines vorgestellten Emotionsmodells die einzelnen Entwickleraussagen labeln.

[1] `https://www.android-hilfe.de/`

Kapitel 6 befasst sich mit der Auswertung des Labelprozesses und des zuvor in Kapitel 5 erstellten Datensatzes in vorhandenen Stimmungsanalysetools. Anschließend werden in Kapitel 7 die Ergebnisse evaluiert und diskutiert. Dabei wird auch auf mögliche Validity Threats eingegangen.
Im letzten Kapitel 8 werden die Ergebnisse dieser Arbeit zusammengefasst. Ein Ausblick geht auf weitere Anwendungsmöglichkeiten ein.

Kapitel 2

Grundlagen

In diesem Kapitel werden die Grundlagen, die für die Arbeit relevant sind, erklärt. Zum einen wird die Stimmungsanalyse und explizit die im Bereich des SE konkretisiert. Dabei werden verschiedene Stimmungsanalysetools, sowohl aus dem allgemeinen als auch dem SE-Bereich genannt. Um ein tieferes Verständnis über die im späteren Verlauf festgesetzte Guideline für das Labeln des Datensatzes zu bieten, werden die Grundlagen von Emotionen und verschiedene Emotionsmodelle erwähnt. Anschließend wird eine Definition für den Goldstandard erörtert, um eine Orientierung bei der Erstellung des Datensatzes zu bieten. Im letzten Teil dieses Kapitels werden relevante Evaluationsmetriken vorgestellt.

2.1 Stimmungsanalyse

Die Stimmungsanalyse bezeichnet die Analyse von Meinungsäußerungen in Texten und ist in den letzten Jahren zu einem der begehrtesten Forschungsfelder der Informatik geworden [12] [56]. Stimmungen und die daraus entstehenden Meinungen sind oft hauptsächlich verantwortlich für das Verhalten von Menschen, weswegen die Forschung in diesem Bereich große Vorteile mit sich bringen kann [32]. So können beispielsweise Dienstleistungsunternehmen mit Hilfe der Stimmungsanalyse die Bedürfnisse ihrer Kunden frühzeitig erkennen und somit negative Trends ermitteln oder weitere Marketingschritte planen [56]. Umständliche Auswertungen von vielen Kundenmeinungen durch beispielweise Umfragen oder Kunden-E-Mails sind sehr zeitintensiv und kostspielig, wodurch eine automatische Analyse Abhilfe schaffen kann [56]. Daher wurden Stimmungsanalysetools entwickelt, die beispielsweise lexikonbasiert oder mit Machine-Learning-Ansätzen arbeiten, Texte analysieren und ihnen verschiedene Polaritäten wie *Positiv, Negativ* und *Neutral* oder verschiedene Emotionen wie *Freude, Trauer*, etc. zuordnen [35]. Ein Beispiel für eine Klassifizierung von Aussagen in Emotionen ist in Tabelle 2.1 zu sehen.

Aussage	**Label**
Excellent! This is exactly what I needed. Thanks!	Love
Horray for Android!	Joy
Why does the test return false?	Surprise
I've been tackling the same issue. It's a pain!	Anger
this link is dead :(	Sadness
@IgnacioOcampo, I gave up after a while I am afraid :(	Fear

Tabelle 2.1: Beispielaussagen aus einem Stack Overflow Datensatz[40]

Da die meisten Stimmungsanalysetools mit Polaritäten arbeiten, wäre eine Möglichkeit, einen solchen Datensatz, der nach einem Emotionsmodell in Emotionen gelabelt wurde, in Polaritäten zu übersetzen, die Emotionen *love* und *joy* als *Positiv* und *anger, sadness* und *fear* als *Negativ* zu werten [39]. In vielen Fällen wurden Aussagen, die als *Neutral* klassifiziert wurden, keiner Polarität oder Emotion zugeordnet werden [5] [39].

Lexikonbasierte Stimmungsanalysetools arbeiten dabei mit einem nach Polaritäten klassifiziertem Lexikon, das von den Entwicklern manuell erstellt wurde. Einzelne Wörter werden mit „+1" für die Polarität *Positiv* und „-1" für die Polarität *Negativ* versehen, wobei die Stärke der Zugehörigkeit zu dieser Polarität dem Intervall angepasst und im tieferen Nachkommastellenbereich liegen kann. Andernfalls ist es auch möglich, einen anderen Intervall wie „-5" bis „+5" zu nutzen, wie es beispielsweise bei SentiStrength angewendet wird [27]. Wird der Wert „0" einem Wort zugeschrieben, so wird dieser als *Neutral* gewertet. Bei der Analyse von Texten wird dann eine Gesamtstimmung auf Basis der Werte aller Wörter ermittelt.
Es ist ersichtlich, dass die Qualität dieses Lexikons ausschlaggebend für die Bewertung durch die Stimmungsanalysetools ist [56]. Ein bestehendes Problem ist auch, dass Sprachen doppelte Verneinungen nutzen, wodurch Stimmungsanalysetools die Stimmung daraufhin falsch erkennen können [29].

In Experimenten konnte gezeigt werden, dass Stimmungsanalysetools implementiert mit Hilfe von Prinzipien des Machine Learnings denen, die lexikonbasiert arbeiten, performancetechnisch überlegen sind [2] [24] [63].
Die besten Ergebnisse für Machine Learning Tools erzielen solche mit linearen Support Vector Machines (SVM), da diese sehr präzise Texte analysieren und sie einer Polarität zuordnen können [5].

2.1.1 Stimmungsanalyse im Software Engineering

Die Stimmungsanalyse in der Domäne des SE wird seit der letzten Dekade genauer erforscht [5] [39] [44]. Da Beteiligte in der Softwareentwicklung oft

in Teams arbeiten oder beispielsweise Softwarenutzer im Internet textuelle Rezensionen zum Produkt abgeben, herrscht viel Kommunikation in diesem Bereich [18]. Dort, wo menschliche Kommunikation auftritt, sind auch Stimmungen und Emotionen identifizierbar [18]. Eine Studie von Graziotin et al. [16] hat ergeben, dass eine positive Stimmung in Entwicklerteams zu besseren kognitiven Leistungen sowie Kreativität und analytischem Denken führt. Um eine positive Stimmung zu bewahren, kann es also von Vorteil sein, Stimmungen zu analysieren und mögliche Trends ausfindig zu machen [16].
Will man eine Stimmungsanalyse durchführen, so kann man bestehende Stimmungsanalysetools verwenden [58]. Untersuchungen haben jedoch ergeben, dass bei Anwendung der Tools mit Datensätzen aus dem Bereich des SE schlechte Ergebnisse erzielt werden [31]. Ein Grund dafür war, dass das Lexikon nicht an das Vokabular der Softwareentwicklung angepasst sind und SE-spezifische Wörter einer falschen Polarität zugeordnet werden [31]. Aufgrund dieser Unterschiede wurden SE-spezifische Stimmungsanalysetools entwickelt [2] [5] [27].

2.1.2 Stimmungsanalysetools

Im Folgenden werden verschiedene Stimmungsanalysetools betrachtet, sowohl domänenunspezifische als auch solche, die an den SE-Bereich angepasst sind.

2.1.2.1 SentiStrength

SentiStrength ist ein von Thelwall et al. [58] entwickeltes Stimmungsanalysetool, das lexikonbasiert arbeitet und den einzelnen Wörtern einen Wert zwischen „-5" und „+5" entsprechend für die Polarität *Negativ* bzw. *Positiv* zuteilt, wobei „0" für *Neutral* steht. Aus dem Aufsummieren der einzelnen Werte leitet sich dann die Gesamtpolarität ab. Als Grundlage bei der Erstellung des Tools dienten hier 2.600 manuell gelabelte MySpace[1]-Texte.

2.1.2.2 SentiStrength-SE

Islam et al. [27] entwickelten SentiStrength-SE, ein an die Domäne des SE angepasstes Stimmungsanalysetool, das auf Grundlage von 5.600 manuell gelabelten issue Comments aus Jira[2] konzipiert wurde. Das Tool liefert bei Anwendung von SE-spezifischen Datensätzen bessere Ergebnisse, als das bereits vorhandene Tool SentiStrength [58], da Änderungen wie ein angepasstes Lexikon, Einbindung kontextbezogener Bedeutungen von spezifischen Wörtern oder Reinterpretationen von bestimmten Zeichen vorgenommen wurden

[1] `https://myspace.com/`
[2] `https://www.atlassian.com/de/software/jira`

[25]. Außerdem wurde darauf Rücksicht genommen Fehlermeldungen, die in die Aussagen eingefügt wurden, keinem Sentiment zuzuschreiben [25]. Das Tool erzielte bei Anwendung des Jira-Datensatzes von Ortu et al. [44] einen Gesamt-Durchschnitts-Accuracy-Wert von 77.48% [25].

2.1.2.3 SentiStrength_DE

SentiStrength_DE[3] ist ein von der Austrian Research Institute for Artificial Intelligence (OFAI)[4] entwickeltes deutsches Lexikon für SentiStrength[27]. Es arbeitet dabei mit modifizierten Lookup-Tables, um eine möglichst präzise Anwendung des Tools für die deutsche Sprache zu ermöglichen. So wird auf konkrete Negationen, Idiome oder Schlüsselwörter geachtet. Eine angepasste Liste an Boosterwörtern wie „extrem" oder als Beispiel für die negative Verstärkung von Wörtern „kaum" unterstützt dabei die Bewertung durch das Tool.

2.1.2.4 Senti4SD

Das Stimmungsanalysetool Senti4SD [5] arbeitet ebenfalls mit dem englischen, modifizierten Lexikon von SentiStrength [58], wurde aber mit einem Goldstandard-Datensatz von Stack Overflow trainiert. Es implementiert zusätzlich noch Keywordbasiertheit, die Texte in N-gramme aufteilt und separat bewertet, um die Keywords zu bestimmen und weitere semantische Merkmale, wie die Addition von unterschiedlichen Vektoren, die die verschiedenen Polaritätsklassen beinhalten [5]. Bei einem Vergleich zwischen SentiStrength, SentiStrength-SE und Senti4SD, bei dem ein Trainings-zu-Testdatensatz-Verhältnis von 70% zu 30% verwendet wurde, schneidet Letzteres am besten ab, wie in der Tabelle 2.2 zu sehen ist. Eine Erklärung

Tool	**Recall**	**Precision**	**F1-score**
SentiStrength	.82	.82	.82
SentiStrength-SE	.78	.78	.78
Senti4SD	.87	.87	.87

Tabelle 2.2: Vergleich zwischen den Gesamtperformanzen von Sentistrength/-SE und Senti4SD in Bezug auf Micro-avg [5]

dafür, wieso SentiStrength-SE bei diesem Goldstandard-Datensatz von Stack Overflow[5] schlechter als die domänenunspezifische Version SentiStrength abschneidet sei, dass das Tool ad-hoc-Heuristiken und Wortpolaritätsscores verwendet, die bei einem kleinen unausgewogenen Datensatz von 400 Entwickleraussagen in Jira beobachtet wurden [5].

[3] `https://github.com/OFAI/SentiStrength_DE`
[4] `https://www.ofai.at/research-areas`
[5] `https://stackoverflow.com/`

2.1.2.5 SentiCR

Ebenfalls als Vorlage diente SentiStrength bei der Erstellung von SentiCR [2]. Anstelle eines veränderten Lexikons, setzten Ahmed et al. hier auf Maschinelles Lernen, indem sie 2.000 code reviews zufällig auswerteten, manuell labelten und verschiedene Algorithmen anwendeten. Der „Gradient Boosting Tree"-Algorithmus schnitt dabei am besten ab, denn sie erzielten damit eine höhere Genauigkeit als SentiStrength oder andere Stimmungsanalysetools [2].

2.1.2.6 DEVA

Islam und Zibran [26] erstellten außerdem das SE-spezifische Stimmungsanalysetool DEVA, das in der Lage ist emotionale Zustände wie *Aufregung, Stress, Depressivität* und *Entspannung* zu berücksichtigen, wodurch es sich von anderen Tools unterscheidet. Für die Evaluation ihres Tools erstellten sie einen Datensatz aus 1.795 issue comments aus Jira [26].

2.1.2.7 GerVADER

GerVADER [59] ist ein deutsches Stimmungsanalysetool, das für die Domäne des Social Media als eine Art deutsches Gegenstück zu dem englischsprachigen Stimmungsanalysetool VADER [22] entwickelt wurde. Zum Testen haben Tymann et al. eine Teilmenge des SCARE-Korpus [54] verwendet und erzielten dabei eine hohe Genauigkeit. Ein großes Problem habe das Tool jedoch noch mit längeren negierten Sätzen, weil bei größeren Abständen zwischen der Negation und dem so genannten Boosterwort, das einen starken Einfluss auf die Einstufung der Polarität hat, keine Zuordnung stattfindet und die Polarität somit nicht negiert wird [59]. Vorteilhaft an GerVADER ist, dass es schnelle Ergebnisse produziert und kein Training benötigt, da kein Machine-Learning angewandt und somit die Konsistenz der Ergebnisse bewahrt wird [59].

2.1.2.8 BertDE

Guhr et al. [17] trainierten ein auf BERT [10] basierendes Stimmungsanalysetool für die Anwendung von deutschen Texten. Sie verwendeten bestehende deutsche Datensätze, die sie aber um einen Großteil erweitert haben, um nicht nur die Sparte des deutschen Social Media als Trainingssatz zu beinhalten. So bestehen ca. 66% der 5.355.043 Aussagen aus dem deutschen Reiseportal Holidaycheck.de [6].

[6] `https://www.holidaycheck.de/`

2.1.2.9 TextBlob-DE

TextBlob-DE[7] ist eine von Killer et al. erstellte Erweiterung des englischen Stimmungsanalysetools TextBlob[8]. Es arbeitet lexikonbasiert und wird als Pythonbibliothek importiert. Für die einzelnen Aussagen verwendet es Polaritäten im Intervall *-1* bis *1* entsprechend den Polaritäten *Negativ, Neutral* und *Positiv*.

2.2 Emotionen und Stimmungen

Psychologen teilen keine klare Definition von dem Begriff Emotion und sehen ihn als ein „hypothetisches Konstrukt" an [11]. Es sind jedoch Veränderungen von verschiedenen Komponenten erkennbar, so beispielsweise bei der physiologischen Komponente, die eine Zustandsänderung von „Herzrate, Hautleitfähigkeit oder Muskeltonus" beschreibt oder der kognitiven Komponente, die je nach Emotion mit verschiedenen Gedankenformen einhergeht [15]. Zudem bemerkbar sind die expressive Komponente, die den Mitmenschen die eigenen Emotionen erst bemerkbar macht und die motivationale Komponente, bei der adaptives Verhalten basierend auf Emotionen wahrnehmbar ist [15]. Eine weitere Komponente ist das affektive Erleben, das durch Gefühlsregungen, die sich durch eine kurze Dauer charakterisieren, bestimmt wird [15] [60]. Diese aktuellen Gefühlsregungen können mehrfach gleichzeitig auftreten, klingen aber nach kurzer Zeit wieder ab, anders als Stimmungen, bei denen man von „Dauertönungen des Erlebnisfeldes" spricht und diese sich eher schwach äußern [60].

Stimmungen können dabei als angenehm oder unangenehmes Erleben wahrgenommen werden und unterscheiden sich von Emotionen unter dem Aspekt, dass bei ihnen der Grund des affektiven Erlebens nicht unbedingt ersichtlich ist [11]. Russell [52] spricht von dem so genannten Kernaffekt, der den Begriffen Stimmung, Affekt, Aktivierung oder Gefühl nahe kommt.

So beschreibt Abbildung 2.1 ein dimensionales Emotionsmodell, das beim bewussten Erleben die Entstehung des Kernaffekts als einen Punkt, der durch die Mischung von den zwei Dimensionen, die hier als vertikale Linie für die Erregung und als horizontale Linie für den Intervall von unangenehm zu angenehm repräsentiert werden, darstellt. Zwischen den Achsen sind Gefühlsbeispiele abgebildet, die charakteristisch ihrer Region zugeordnet sind. Der Kernaffekt lässt sich durch verschiedene Ursachen beeinflussen, beispielsweise reizunabhängige Aspekte wie Hunger oder Durst, durch andere Erregungszustände, durch physische Faktoren wie „Wetter, Gerüche,

[7] `https://github.com/markuskiller/textblob-de`
[8] `https://github.com/sloria/textblob`

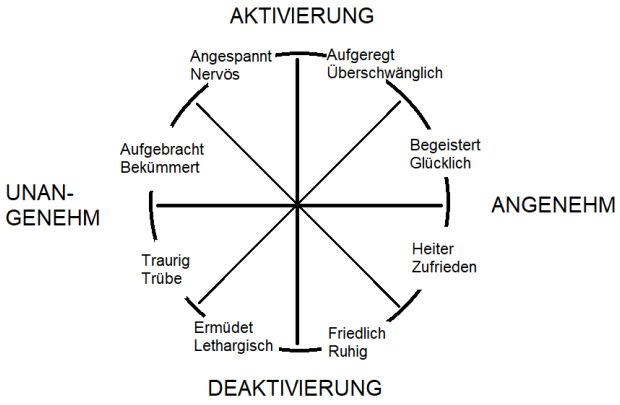

Abbildung 2.1: Deutsche Übersetzung des Kernaffektmodells nach Russell [52]

Lärm, ästhetische Qualität", durch unbewusste Ursachen oder auch durch andere Emotionen [52]. Aufgrund dieser Vielzahl von abhängigen Faktoren ist die Ursachenforschung von Stimmungen sehr komplex [52]. Bekannt ist jedoch, dass Stimmungen die Informationsverarbeitung beeinflussen, denn positive Stimmungen führen zu einer oberflächlicheren und negative Stimmungen zu einer analytischeren Form der Informationsverarbeitung [11].

2.2.1 Emotionsmodelle

Für die Entstehung von Emotionen wurden verschiedene Emotionsmodelle konzipiert, die sich zwar ähnlich sind, aber sich doch voneinander unterscheiden. Diese kann man in Basisemotions-, Klassifikations- und dimensionale Modelle unterscheiden [60].

Während das in Abbildung 2.1 vorgestellte Emotionsmodell von Russell [52] ein Zweidimensionales ist und sich die Emotionen anhand der Extrema von diesen zwei Dimensionen klassifiziert werden, existieren noch andere mehrdimensionale Modelle, so wie das dreidimensionale Emotionsmodell von Mehrabian und Russell [36], das die zwei Dimensionen zusätzlich durch den Aspekt der Dominanz ergänzt. Osgood et al. [45] verwendeten auch ein dreidimensionelles Modell, jedoch werden dabei die Dimensionen Bewertung, Aktivierung und Stärke benutzt.

Basisemotionsmodelle stützen sich auf eine bestimmte Anzahl an Basisemotionen, wobei durch Mischung dieser andere Sekundäremotionen enstehen, so wie Izard [28] 1992 sich auf die zehn Basisemotionen *happiness, suprise, sadness, fear, disgust, anger, interest-excitement, distress-anguish,*

shame und *guilt* festlegte [11]. Dabei sind diese Basisemotionen seit der Geburt als eine Art Grundstruktur vorhanden und kommen mit der Reife zur Entfaltung [60]. Plutchik verwendete bei seinem Basisemotionsmodell die acht Emotionen *aggressiveness, optimism, love, submission, awe, disappointment, remorse* und *contempt* [48].

Klassifikationsmodelle kategorisieren die Emotionen hierarchisch in Cluster, indem sie Kategorien samt Über- und Unterkategorien festlegen, wie beispielsweise nach dem Emotionsmodell von Shaver et al. [55], das in einer vereinfachten Form in Abbildung 2.2 zu sehen ist. Zu erkennen ist

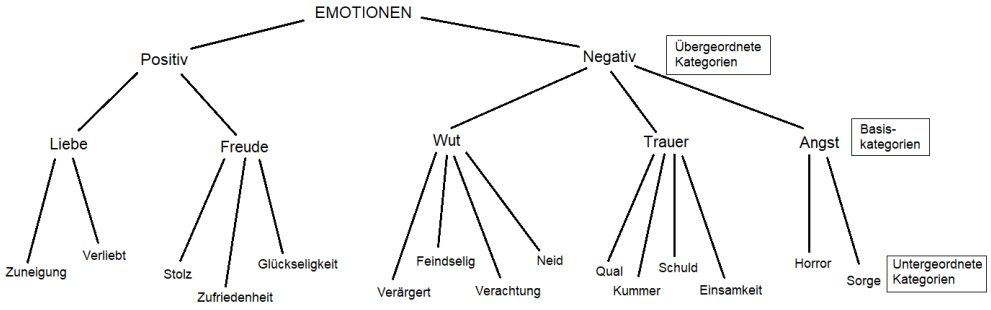

Abbildung 2.2: Deutsche Übersetzung der vereinfachten Version des Emotionsmodells nach Shaver et al. [13]

die grundlegende Aufteilung in *Positiv* und *Negativ*, die die übergeordneten Kategorien darstellen, gefolgt von den Basisemotionen *Liebe, Freude, Wut, Trauer* und *Angst*. Als Unterkategorien werden dort zu den Basisemotionen zugehörige Emotionen aufgelistet.

In der Standardversion dieses Emotionsmodells wurde die Emotionen *surprise* noch aufgenommen, da einige Emotiontheoretiker sie als Basisemotion gewertet haben, jedoch kamen Shaver et al. zu dem Entschluss, *surprise* nicht mit aufzunehmen, da dieser Cluster viel undifferenzierter als die anderen ist und in Studien diesem Cluster viel seltener eine Emotion zugeordnet wurde [55]. Ein Grund dafür könnte sein, dass „Überraschung" sowohl positiv als auch negativ gewertet werden kann und bei einer Zuordnung intuitiv andere Basisemotionen bevorzugt werden.

Parrott [46] unterteilte das Emotionsmodell von Shaver et al. [55] in eine weitere Unterkategorie für eine detailliertere Zuordnung. Dabei bezeichnet Parrott die sechs Basisemotionen als Primäremotionen, die von Shaver et al. als untergeordnete Emotionen als Sekundäremotionen und seine detailliertere Stufe als Tertiäremotionen. Die genaue Unterteilung ist in Abbildung 2.3 zu sehen. Dabei ist ersichtlich, dass den Sekundäremotion

Primäremotionen	Sekundäremotionen	Tertiäremotionen
Liebe	Zuneigung	Mitgefühl, Sympathie, Fürsorge, ...
	Begierde/sex. Lust	Verlangen, Leidenschaft, Verliebtheit
	Sehnsucht	
Freude	Heiterkeit	Unterhaltung, Genuss, Glück, ...
	Elan	Begeisterung, Eifer, Aufregung, Heiterkeit
	Zufriedenheit	Vergnügen
	Optimismus	Eifrigkeit, Hoffnung
	Stolz	Triumph
	Begeisterung	Begeisterung, Entzückung
Überraschung	Überraschung	Erstaunen, Verwunderung
Wut	Reizbarkeit	Ärger, Aufregung, Ärger, Mürrisch, ...
	Verzweiflung	Frustation
	Zorn	Empörung, Feindseligkeit, Bitterkeit, Hass,...
	Abscheu	Empörung, Verachtung
	Neid	Eifersucht
	Qual	Qual
Trauer	Leiden	Agonie, Qual, Schmerz
	Traurigkeit	Depression, Verzweiflung, Melancholie, ...
	Enttäuschung	Entsetzen, Unmut
	Schande	Schuld, Reue, Gewissensbisse
	Nachlässigkeit	Peinlichkeit, Demütigung, Unsicherheit, ...
	Mitleid	Mitleid
Angst	Entsetzen	Schock, Schreck,Panik, Hysterie, ...
	Ängstlichkeit	Spannung, Furcht, Sorge, Bedrängnis, ...

Abbildung 2.3: Deutsche Übersetzung des von Murgia et al. [38] erstellten Emotionsmodells nach Parrott [46]

teils mehrere Tertiäremotionen zugeordnet wurden, wie beispielsweise *Optimism* die Emotionen *Eagerness* und *Hope*.

2.3 Goldstandard-Datensätze in der Stimmungsanalyse

Verschiedene Forscher sprechen bei ihrer Erstellung von Datensätzen von einem Goldstandard, wobei sich die Verfahren dafür meist unterscheiden. Da es somit keine festgesetzten Regeln für die Erstellung von Goldstandard-Datensätzen gibt, wird sich im Folgenden an bestehenden Arbeiten orientiert, die von einem Goldstandard-Datensatz im Bereich der Stimmungsanalyse sprechen, um eine Definition zu erörtern.

Bei der Entwicklung von Senti4SD haben Calefato et al. [5] einen Goldstandard-Datensatz erstellt. Dabei setzt sich ihr Vorgehen aus vier Teilschritten zusammen. Im ersten Schritt haben sie sich mit Literatur über Emotionsmodelle befasst, auf die sie sich daraufhin stützen, um eine Guideline für das Labeln zu erstellen. Sie wählten dabei das Emotionsmodell von Shaver et al. [55], da es leicht verständlich und mit nur sechs verschiedenen Emotionen sehr kompakt sei. Der zweite Schritt befasst sich damit, die Entwickleraussagen zu extrahieren und unnötige Zeichenketten und Informationen zu entfernen, um den reinen relevanten Text zu besitzen. Daraufhin wendeten sie SentiStrength [27] an, um folgend einen balancierten Datensatz mit einem gleichen Anteil an positiven,

negativen und neutralen Aussagen zu haben. Damit der Datensatz manuell gelabelt wird und möglichst unvoreingenommene Ergebnisse erzielt werden, haben sie zwölf Entwickler aus dem Informatikstudium ausgewählt und ihnen in einer zweistündigen Sitzung die Guideline samt Beispielen erklärt. So wurden sie beispielsweise unterrichtet, dass wenn sie für eine Aussage zwei Emotionen unterschiedlicher Polarität festgestellt hätten, sie das Label „gemischt" vergeben sollen. Im folgenden Schritt, nach Berechnung des Cohens-Kappa-Wertes von 0.74, der die Einstimmigkeit zwischen zwei Ratern angibt, haben sie sich jedoch darauf geeinigt, die Aussagen mit den gemischten annotierten Emotionen nicht in den Goldstandard-Datensatz aufzunehmen, obwohl sie diese Aussagen durch ein Mehrheitsvotum zuvor einem einzelnen Label unterzogen. Der letzte Schritt befasste sich mit der Einbindung für die Erstellung des Stimmungsanalysetools Senti4SD [5].

Ortu et al. [44] sind bei der Erstellung ihres Goldstandard-Datensatzes aus Jira ähnlich vorgegangen, wobei sie sich auf das Emotionsmodell von Parrott [46] gestützt haben, das eine Erweiterung des Emotionsmodells von Shaver et al. [55] darstellt. Sie ließen den Datensatz von 16 Studenten und Wissenschaftlern labeln, während sie jedes mal in Kleingruppen das Label besprachen. Dabei wurde jedoch nur ein kleiner Teil des Datensatzes so gelabelt. Die zweite Gruppe bestand aus drei Ratern und annotierte 1.600 Aussagen mit drei der sechs Basisemotionen und die dritte Gruppe, die den größten Teil der Aussagen labelte, nutzte vier Basisemotionen, da diese einen deutlich höheren Übereinstimmungswert erzielten. Anzumerken ist hierbei auch, dass der Datensatz mit 67% neutralen und nur 19% positiven und 14% negativen Aussagen nicht ausgeglichen ist [39].

Saif et al. [53] erstellten 2013 für die Social-Media-Plattform Twitter[9] einen Goldstandard-Datensatz, indem sie zunächst acht bereits existierende Twitter-Datensätze evaluierten um Mängel festzuhalten und diese bei der Erstellung des eigenen Datensatzes zu vermeiden. Sie stellten fest, dass bei manchen Datensätzen keine strikte Annotationsmethodik existierte, wobei die Rater beispielsweise einen Teil der Aussagen mit den Polaritäten *Positiv* und *Negativ* bewerteten und den anderen Teil zusätzlich mit *Neutral, Irrelevant* oder *Sonstiges*. Oft wurde kein Interrater-Reliabilität bestimmt oder es fehlte schlicht die Angabe über die Anzahl der Rater. Ein anderes Problem sei, dass die anderen Datensätze die Aussagen nicht auf einem Entitäten-Level betrachteten, da manche Aussagen auch mehrere verschiedene Dinge mit unterschiedlichen Sentimenten beinhalten können und diese separat behandelt werden müssten. Dementsprechend einigten sie sich bei ihrer Erstellung des Goldstandard-Datensatzes auf die Polaritäten *Positiv, Negativ, Neutral, Gemischt* und *Sonstiges*, wobei *Gemischt* vergeben

[9] `https://twitter.com/`

wurde, wenn es verschiedene Stimmungen in der Aussage gab und *Sonstiges*, wenn es schwer war, sich auf eine Polarität festzulegen. Die Ratergruppe, die aus Studenten bestand, erhielten eine Broschüre die als Guideline für die Bewertung der 3.000 Aussagen galt. Anschließend bestimmten sie die Interrater-Reliabilität und nahmen nur die Aussagen mit einem hohen Wert und kamen so auf knapp 2.200 manuell gelabelte Aussagen. Diesen Datensatz evaluierten sie um ihn seiner Genauigkeit einzustufen [53].

2014 erstellten Malo et al. [33] einen Datensatz aus 10.000 zufällig ausgewählten Finanzartikeln. Diesen bereiteten sie vor, indem sie alle Sätze, die keine Wörter aus dem angepassten Lexikon beinhalteten, entfernten. Somit kamen sie auf 53.400 Sätze, die sie in ca. 5.000 Entitätssequenzsätze unterteilten. Um diesen Datensatz zu labeln, wurde den 16 Ratern gesagt, dass sie sich von der Perspektive eines Investors sehen sollen. 13 von den Ratern sind Masterstudenten gewesen, drei davon Forscher aus dem Finanzbereich. Dabei sollten sie die drei Polaritäten *Positiv, Neutral* und *Negativ* so vergeben, wie sie denken, welchen Einfluss dieser Satz auf den Börsenkurs haben würde. Durchschnittlich bewertete jeder Rater ca. 1.500 Sätze, womit jeder Satz von ca. fünf bis acht Ratern klassifiziert wurde. Es gab eine grobe Guideline, bei der festgesetzt wurde, dass es keine fixe Regeln gab, wie bestimmte Wörter annotiert werden sollen. Außerdem sollten sie bei dem Prozess des Labelns konsistent agieren und die Sätze voreingenommen ohne ihrer bisherigen Kenntnisse zu vorkommenden Themen bewerten. Nachdem das Labeln abgeschlossen war, haben sie die Interrater-Reliabilität berechnet und kamen auf gute Ergebnisse. Abschließend folgte eine Evaluation in fünf verschiedenen lexikonbasierten Stimmungsanalysetools des Finanzbereiches [33].

Es ist dementsprechend schwierig, einen Goldstandard zu definieren, da verschiedene Forscher dabei ihrer eigenen Auffassung nachgehen. Jedoch werden einige Gemeinsamkeiten deutlich.
Novielli et al. [39] stellten fest, dass ein subjektives Vergeben der Labels ohne Guideline zu schwammigen Goldstandard-Datensätzen führt. Somit ist es wichtig, eine feste Guideline, die sich auf ein theoretisches Emotionsmodell stützt festzulegen und die Rater hinreichend gut zu schulen [39] [53]. Außerdem soll ein Datensatz nicht nur von einer Person gelabelt werden, um die Subjektivität zu verhindern, wobei Ortu et al. [44] feststellten, dass bei mehr als zwei Ratern die Einstimmigkeit gleich blieb. Novielli et al. [39] merkten auch an, dass eine hohe Interrater-Reliabilität ein wichtiger Faktor für die Validität des Goldstandard-Datensatzes sei, weswegen die Berechnung dieser wichtig ist. Eine abschließende Evaluation in verschiedenen Tools kann mögliche Fehler und Genauigkeiten des Datensatzes hervorheben [33].

2.4 Relevante Evaluationsmetriken

Die Auswertungen des erstellten Datensatzes und der verschiedenen Stimmungsanalysetools erfordern bestimmte Metriken, die im Folgenden dargestellt werden. Notwendig dafür sind Precision, Recall, F1-Score, Accuracy, Macro-avg, Fleiss' Kappa und Cohens Kappa.

2.4.1 Performanz

Für die Evaluation der Stimmungsanalysetools sind Precision, Recall, F1-Score, Accuracy und Macro-avg wichtig. Dazu werden zunächst die Bestandteile erklärt.

True Positive (TP): Aussagen, die von dem Modell korrekt in die zugehörige Klasse eingestuft wurden.
False Positive (FP): Aussagen, die fälschlicherweise in die gewählte Klasse eingeordnet wurden.
True Negative (TN): Aussagen einer anderen Klasse, die korrekt in diese Klasse eingeordnet wurden.
False Negative (FN): Aussagen der zugehörigen Klasse, die fälschlicherweise in eine andere Klasse eingestuft wurden.

Damit können folgende Metriken definiert werden [34]:

Precision: Precision gibt die Genauigkeit wieder, indem es das Verhältnis von den korrekt klassifizierten Aussagen der zugehörigen Klasse zur Gesamtzahl der Aussagen dieser Klasse berechnet.

$$Precision = \frac{TP}{TP + FP}$$

Recall: Recall gibt das Verhältnis der korrekt klassifizierten Aussagen in Bezug auf alle dieser Klassen zugeordneten Aussagen wieder.

$$Recall = \frac{TP}{TP + FN}$$

F1-Score: Der F1-Score gibt Precision und Recall gewichtet zusammen wieder.

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

Accuracy: Accuracy gibt das Verhältnis zwischen den korrekt klassifizierten

Aussagen und aller Aussagen wieder.

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

Macro-avg: Macro-avg berechnet den Durchschnitt eines Messwertes über alle Klassen.

$$Macro - avg = \frac{\sum Messwert}{|Klassen|}$$

2.4.2 Interrater-Reliabilität

Zur Bestimmung von Übereinstimmungsraten können statistische Verfahren wie das Cohens Kappa [9] für zwei Rater oder Fleiss' Kappa [14] für zwei oder mehr Rater verwendet werden. Landis und Koch [30] erstellten eine Tabelle für die Kategorisierung der Kappa-Werte bezüglich des Maßes der Übereinstimmung. Diese ist übersetzt in Tabelle 2.3 einsehbar.

Kappa-Wert	**Maß der Übereinstimmung**
<0.00	Schwach
0.00-0.20	Leicht
0.21-0.40	Einigermaßen
0.41-0.60	Moderat
0.61-0.80	Erheblich
0.81-1.00	Fast perfekt

Tabelle 2.3: Übersetzte Kategorisierungstabelle nach Landis und Koch [30]

Dabei können die Kappa-Werte im bereich von -1 bis 1 liegen.

2.4.2.1 Cohens Kappa

Um Cohens Kappa [9], das für die Berechnung von Reliabilitäten zweier Rater angewendet wird, zu berechnen, wird folgende Formel verwendet:

$$k = \frac{p_0 - p_e}{1 - p_e}$$

Dabei steht p_0 für den Anteil der Fälle, in denen die Rater übereinstimmten und p_e für den erwarteten Anteil einer zufälligen Übereinstimmung [9].

2.4.2.2 Fleiss' Kappa

Fleiss' Kappa kann für zwei oder mehr Ratern verwendet werden und berechnet sich ähnlich wie Cohens Kappa. So kann die selbe Formel für die Bestimmung des Kappa-Wertes angewendet werden, jedoch ändern sich die Berechnungen für p_0 und p_e [14]:

$$p_0 = \frac{1}{Nn(n-1)} \left(\sum_{i=1}^{N} \sum_{j=1}^{k} n_{ij}^2 - Nn \right)$$

$$p_e = \sum_{j=1}^{k} p_j^2$$

Dabei stellt N die Gesamtzahl der Fälle und n die Anzahl der Rater dar.

Kapitel 3

Verwandte Arbeiten

In diesem Kapitel werden wissenschaftliche Arbeiten und Publikationen genannt, die sich entweder mit der Erstellung von Goldstandard-Datensätzen oder der Stimmungsanalyse befassen. Dabei wird zwischen Datensätzen inner- und außerhalb der Stimmungsanalyse und der Stimmungsanalyse aus dem SE-Bereich und aus anderen Domänen unterschieden.

3.1 Erstellung von Datensätzen

Mit der Erstellung von Datensätzen haben sich bereits viele Wissenschaftler befasst [39] [40] [44] [53]. Speziell im Bereich der Stimmungsanalyse haben Novielli et al. [40] einen englischen Goldstandard-Datensatz[1] aus 4.800 verschiedenen Fragen, Antworten und Kommentaren aus einer Q&A-Seite von Stack Overflow ertellt. Diese Aussagen wurden von je drei verschiedenen Doktoranden der Informatik in die sechs Basisemotionen *love, joy, anger, sadness, fear* und *surprise* nach dem Emotionsmodell von Shaver et al. [55] gelabelt. Dabei haben ca. 41% der Aussagen ein Emotionslabel bekommen, da die restlichen Aussagen neutral gewertet wurden [40].

2016 erstellten Ortu et al. [44] einen Datensatz[2] aus dem Projektmanagementtool Jira. Dabei wurden 392 issue comments mit den sechs Basisemotionen, 1600 issue comments mit *love,joy* und *sadness* und 4000 issue Sätze mit den Basisemotionen *love,joy,anger* und *sadness* gelabelt, wobei das Emotionsmodell von Parrott [46] als Vorlage galt. Nach dem ersten Durchgang des Labelns stellten sie fest, dass die Übereinstimmung bei der Vergabe eines Labels durch mehr als zwei Personen keine großen Änderungen mit sich brachten. Außerdem haben sie bemerkt, dass nur die Emotionen *love,joy* und *sadness* eine hohe

[1] `https://github.com/collab-uniba/EmotionDatasetMSR18/`

[2] `https://ansymore.uantwerpen.be/system/files/uploads/artefacts/alessandro/MSR16/archive3.zip`

Übereinstimmung hatten, weswegen sie bei der zweiten Gruppe nur diese Emotionen für die Labelvergabe nutzten. Bei der dritten Gruppe wurde noch zusätzlich die Basisemotion *anger* aufgenommen [44].

Novielli et al. [39] beschäftigten sich außerdem damit, eine Kombination aus drei plattformübergreifenden Datensätzen mit Hilfe vier verschiedener Stimmungsanalysetools zu untersuchen. Dabei haben sie einen Datensatz von Stack Overflow und den Jira-Datensatz von Ortu et al. [44] benutzt und zusätzlich einen Datensatz[3] mit 7000 pull-requests und commit-Kommentaren aus GitHub[4] erstellt und diese mit den Polaritäten *Positiv, Negativ* und *Neutral* versehen lassen [39]. Beim Jira-Datensatz wurden die sechs Basisemotionen in die drei Polaritäten übersetzt, um ihn für ihre Studie brauchbar zu machen, wobei *love* und *joy* als *Positiv* und *anger, sadness* und *fear* als *Negativ* gewertet wurden [39]. All diese SE-spezifischen Datensätze enthalten lediglich englische Aussagen.

Sänger et al. [54] haben einen deutschen Datenskorpus namens SCARE[5] mit Hilfe von App-Rezensionen aus dem Google Play Store[6] erstellt, wobei sie 1.760 verschiedene Aussagen sammelten, die sie wiederum in 3.953 subjektive Sätze unterteilten. Die Rezensionen stammten dabei von Apps aus elf verschiedenen Kategorien. Dabei wurde ein eher ungleichmäßig verteilter Datensatz erstellt, denn ca. 62% der Reviews wurden mit *Positiv*, ca. 36% mit *Negativ* und ca. 1,6% mit der Polarität *Neutral* versehen [54]. Dabei stützten sie sich nicht auf ein Emotionsmodell.

Ferner verwandt außerhalb des SE-Bereichs erstellten Boland et al. [4] einen deutschen Datensatz bestehend aus Rezensionen von Amazon[7], wobei über 63.000 einzelne Sätze von neun Personen mit den zugehörigen Polaritäten gelabelt wurden.

Saif et al. [53] analysierten 2013 acht verschiedene Datensätze von der Social-Media-Plattform Twitter, um einen eigenen Goldstandard-Datensatz zu erstellen, der im Gegensatz zu den anderen Datensätzen, die einzelnen Entitäten einer Aussage bewertet und somit Aussagen, die mehrere Entitäten mit verschiedenen Sentimenten beinhalten, einzeln evaluiert. Dabei stellten sie auch Mängel bei den anderen Datensätzen fest, um diese zu vermeiden.

2014 erstellten Malo et al. [33] einen Finanzdatensatz aus 10.000

[3] `https://doi.org/10.6084/m9.figshare.11604597.v1`
[4] `https://github.com/`
[5] `https://www.romanklinger.de/scare/`
[6] `https://play.google.com/store`
[7] `https://amazon.com/`

Finanznachrichten. Dabei sortierten sie die Nachrichten mit Hilfe eines Lexikons vor und ließen diesen Datensatz von 16 fachkundigen Personen mit einer Guideline labeln. Dabei wurden die drei Polaritäten *Positiv, Neutral* und *Negativ* vergeben [33].

Außerhalb der Stimmungsanalyse im Bereich der Radiologie haben Pawiro et al. [47] einen Goldstandard-Datensatz für die Validierung von 2D/3D-Bildregistrierungsalgorithmen erstellt. Dabei nutzten sie die modernsten bildgebenden Verfahren und verglichen ihren Datensatz mit bereits bestehenden. Ihr Datensatz zeigte bei der Evaluation Verbesserungen in der anatomische Detailgenauigkeit und der Qualität der Bilddaten auf [47].

Im Bereich der Wettervorhersage entwickelten Rasp et al. [51] einen Goldstandard-Datensatz[8], um eine einheitlich verwendbare Möglichkeit für den Vergleich verschiedener Studien zu haben. Der Datensatz soll als Benchmark für neue Machine-Learning-Ansätze im Gebiet der Wetterprognose dienen. Ein gängiger Vorgang in diesem Bereich ist es, einen Trainings-, einen Validierungs- und einen Testdatensatz zu erstellen, die zusammen den gesamten Datensatz repräsentieren [51].

3.2 Stimmungsanalyse

Die Stimmungsanalyse ist ein in den letzten Jahren viel an Popularität gewonnenes Forschungsfeld, wobei es im Bereich der Stimmungsanalyse innerhalb des SE ebenfalls einige Publikationen gab [2] [27] [31]. So erstellten Islam et al. [27] 2018 ein an die SE-Domäne angepasstes Stimmungsanalysetool namens SentiStrength-SE auf Grundlage von 5.600 manuell gelabelten issue comments aus Jira. Dabei soll SentiStrength-SE bei Anwendung von SE-spezifischen Datensätze bessere Ergebnisse liefern, als bereits vorhandene Tools wie SentiStrength [58], NLTK[9] und Stanford NLP[10].

Calefato et al. [5] entwickelten ein domänenspezifisches in Polaritäten klassifizierendes Stimmungsanalysetool mit dem Namen Senti4SD. Das Tool arbeitet englischsprachig und lexikonbasiert und wurde mit einem Goldstandard-Datensatz von Stack Overflow trainiert, wobei sie das Lexikon von SentiStrength [58] modifizierten [5].

Ebenfalls diente SentiStrength [58] bei der Erstellung von SentiCR von Ahmed et al. [2] als Vorlage. Anstelle eines veränderten Lexikons,

[8] `https://github.com/pangeo-data/WeatherBench`
[9] `https://www.nltk.org/`
[10] `https://github.com/stanfordnlp/CoreNLP`

setzten Ahmed et al. hier auf Maschinelles Lernen, indem sie 2.000 code reviews zufällig auswerteten, manuell labelten und verschiedene Algorithmen anwendeten. Damit erzielten sie eine höhere Genauigkeit als SentiStrength oder andere Stimmungsanalysetools [2].

Islam und Zibran [26] erstellten außerdem das SE-spezifische Stimmungsanalysetool DEVA, das in der Lage ist emotionale Zustände wie *Aufregung, Stress, Depressivität* und *Entspannung* zu berücksichtigen, wodurch es sich von anderen Tools unterscheidet.

Novielli et al. [39] führten eine Studie durch, bei der sie Datensätze aus drei verschiedenen Quellen zu einem kombiniert haben und diesen in vier verschiedenen Stimmungsanalysetools evaluierten, um das Verhalten der jeweiligen Tools beim Training mit Datensätzen aus anderen Quellen, als für die sie bereits trainiert worden sind, zu untersuchen. Die dabei verwendeten Tools SentiCR [2], Senti4SD [5], DEVA [26] und SentiStrength-SE [27] sind speziell auf die Domäne des SE angepasst und wurden bereits in Kapitel 2.1.1 genauer erklärt. Sie kamen zu dem Entschluss, dass beim Verwenden von Trainingsdatensätzen aus anderen Quellen die verschiedenen Tools eine schlechte Performanz aufweisen [39].

Lin et al. [31] entwickelten 2018 ein eigenes Deep Learning Stimmungsanalysetool auf Basis von 40.000 Fragen und Antworten von Stack Overflow. Ihr Tool zeigte schlechte Ergebnisse auf, was eine mögliche Folge von der Auswahl des Trainingsdatensatzes sein könnte. Im Allgemeinen warnen sie andere Forscher im Feld der Stimmungsanalyse für den SE-Bereich davor, sich auf die Ergebnisse der Tools zu verlassen, da dieses Forschungsfeld noch zu jung und zu ungenau sei [31].

Im Jahr 2017 entwickelten Calefato et al. [6] das Toolkit EmoTxt, das die Basisemotionen nach dem Emotionsmodell von Shaver et al. [55] aus Texten erkennt. EmoTxt ist jedoch domänenunspezifisch [6].

GerVADER [59] ist ein deutsches Stimmungsanalysetool, das von Tymann et al. für die Social Media Domäne als eine Art deutsches Gegenstück zu dem englischsprachigen Stimmungsanalysetool VADER [22] entwickelt wurde. Zum Testen haben Tymann et al. eine Teilmenge des SCARE-Korpus [54] verwendet und erzielten dabei eine hohe Genauigkeit. Ein großes Problem habe das Tool jedoch noch mit längeren negierten Sätzen, weil bei größeren Abständen zwischen der Negation und dem so genannten Boosterwort, das einen starken Einfluss auf die Einstufung der Polarität hat, keine Zuordnung stattfindet und die Polarität somit nicht negiert wird [59].

Guhr et al. [17] entwickelten 2020 das auf BERT [10] basierende deutsche

Stimmungsanalysetool BertDE[11]. Der Trainingsdatensatz besteht unter anderem aus dem SCARE-Korpus [54] aber auch aus anderen Quellen wie dem deutschen Reiseportal Holidaycheck.de[12].

Herrmann [19] entwickelte im Rahmen seiner Bachelorarbeit ein Stimmungsanalysetool für Live-Meetings von Softwareprojekten. Zunächst war die Idee des Tools, Audiodateien zu verarbeiten und zu evaluieren. Später erweiterten Hermann et al. [21] das Tool, so dass es in der Lage war, auch textbasierte Aussagen zu betrachten. Dabei liest es für die deutsche Sprache den Datensatz ein und wertet es in den vier vorhandenen Stimmungsanalysetools GerVADER [59], BertDE [17], TextBlob-DE[13] und SentiStrength_DE[14] aus. Per Mehrheitsentscheid wird dann eine Gesamtpolarität berechnet. Für die englische Sprache werden die fünf Tools Senti4SD [5], SentiCR [2], TextBlob[15], SentiStrength [58] und SentiStrength-SE [27] angewandt [21].

[11] https://github.com/oliverguhr/german-sentiment
[12] https://www.holidaycheck.de
[13] https://github.com/markuskiller/textblob-de
[14] https://github.com/OFAI/SentiStrength_DE
[15] https://github.com/sloria/textblob

Kapitel 4

Erstellung des deutschen Datensatzes

In diesem Kapitel wird das Verfahren zur Erstellung des deutschen Goldstandard-Datensatzes erklärt. Dabei wird zuerst auf die Quelle des Datensatzes eingegangen und folgend auf die Funktionsweise des Crawlers, der verwendet wurde, um die Daten aus dem Entwicklerforum zu extrahieren. Außerdem wird auf die Zusammensetzung des ungelabelten Datensatzes eingangen.

4.1 Auswahl der Quelle

Da Entwicklerforen wie Stack Overflow oder GitHub fast nur englische Aussagen beinhalten und diese sich somit nicht für diese Arbeit geeignet hätten, wurde ein deutsches Entwicklerforum gesucht. Nach einer Recherche unter Berücksichtigung der Sprache, der SE-Domäne, der Extrahierungsmöglichkeit und der Menge der Daten kam der Bereich der Android-App-Entwicklung des deutschen Forums Android-Hilfe[1] in Frage. Mit aktuell 14.088 Themen und 74.946 Beiträgen (Stand 15.06.2022) bietet es eine Menge an deutschsprachigen Inhalt. Außerdem wird in den Forenregeln[2] vorausgesetzt, dass Beiträge ausschließlich in deutscher Sprache zu verfassen sind und sich die Nutzer des Forums bemühen sollen, auf eine korrekte Rechtschreibung zu achten. Da im Durchschnitt also ca. 3,4 Beiträge pro Thema geschrieben werden, herrscht in dem Forum eine hohe Kommunikation zwischen Entwicklern. Dementsprechend wurde es als geeignet eingestuft, um einen Datensatz zu erstellen.

[1] `https://www.android-hilfe.de/forum/android-app-entwicklung.9/`
[2] `https://www.android-hilfe.de/help/regeln/`

4.2 Funktionsweise des Crawlers

Der Crawler für das Entwicklerforum Android-Hilfe.de wurde mit Hilfe des Python-Frameworks Scrapy[3] erstellt. Scrapy ist ein Open-Source-Modul und für das Extrahieren von Daten aus dem Web zuständig. Dabei wird die erste Seite des Forums als Start gewählt. Daraufhin werden die einzelnen Threads der Seite durchgegangen, wobei jeder Thread in die Funktion „crawl_posts" geschickt wird. Dabei werden alle Beiträge der Seite durchgegangen, wobei aufgrund des Seitenaufbaus des Forums die einzelnen Zeilen der Beiträge zusammengesetzt werden müssen. Bilder, die für diese Arbeit irrelevant sind, haben den Text „Zum Vergrößern anklicken...", der für den Datensatz gefiltert wird. Damit die Aussage in den Datensatz aufgenommen wird, darf sie nicht zu lang sein. Von daher wurde eine Grenze von 200 Zeichen festgelegt. Außerdem wurden bestimmte Aussagen, die Zitate enthalten oder automatische Zusammenfügungen sind, rausgefiltert. Sobald alle Beiträge jedes Threads der Seite abgearbeitet sind, geht der Crawler mit Hilfe des „Nächste Seite"-Buttons weiter zur folgenden Seite. Der gesamte Prozess geschieht rekursiv, bis die letzte Seite durchgegangen wurde.

4.3 Zusammensetzung des Datensatzes

Der Datensatz, der mit Hilfe des in Kapitel 4.2 beschriebenen Android-Hilfe-Crawlers erstellt wurde, beinhaltet 20.380 verschiedene deutsche Entwickleraussagen. Um einen möglichst balancierten Datensatz zu haben, wurde eine Vorsortierung mit dem in Kapitel 2.1.2.7 beschriebenen Stimmungsanalysetools GerVADER [59] vorgenommen. Dieser kam zu dem Ergebnis, dass der Datensatz aus 10.560 positiven, 7.046 neutralen und 2.774 negativen Aussagen besteht. Da GerVADER Tendenzen zu einer Zugehörigkeit, die in einem Intervall von -1 für negativ und +1 für positiv liegen angibt, wurden für diesen Datensatz die Aussagen dementsprechend ab- oder aufsteigend sortiert. Dies ist wichtig, um möglichst wenige widersprüchliche Aussagen zu haben, die man später verschiedenen Emotionen bzw. Polaritäten zuordnen könnte. So wurden je 2.000 Aussagen mit den höchsten Werten aus den drei Polaritäten zu einem Datensatz addiert. Diese 6.000 Aussagen beinhalteten teilweise irrelevante Informationen wie „gesendet von meinem iPhone XR", wenn die User des Forums dies eingestellt hatten. Diese Signaturen wurden dann nachträglich manuell entfernt.

[3] `https://scrapy.org/`

Kapitel 5

Labeln des Datensatzes

Im Folgenden wird der gesamte Prozess des Labelns samt der Teilnehmenden des Workshops vorgestellt. Auch auf die konzipierte Guideline für das Labeln wird eingegangen.

5.1 Die Guideline für den Labelprozess

Als Guideline für das Labeln wurde eine kleine Legende angefertigt, die den Ratern neben der Tabelle mit dem Datensatz angezeigt wurde. Zur leichteren Einprägung der korrespondierenden Zahlen, die für die verschiedenen Basisemotionen stehen, wurden diese in verschiedenen Farben dargestellt, wie in Abbildung 5.1 zu sehen.

Abbildung 5.1: Grafische Guideline für die Labelvergabe

Die Basisemotionen leiten sich von dem vereinfachten Emotionsmodell 2.2 nach Shaver et al. [55] ab. Eben diese Grafik wurde neben der Legende angezeigt. Sie dient den Ratern als Grundlage für ihre Entscheidung beim Labelvorgang. So können sie beim Vorfinden einer Emotion der unteren Kategorie auf die darüberliegende Basisemotion schließen und dementsprechend das Label vergeben. Zusätzlich zu dem vereinfachten Emotionsmodell nach Shaver et al. 2.2 wurde die Basisemotion *Surprise* aufgenommen, da diese in anderen Goldstandard-Datensätzen auch beachtet

wurde [40] und Shaver et al. außerhalb des Diagramms *Surprise* weiterhin als eine Basisemotion betrachteten [55]. Ähnlich wie in der Guideline von Novielli et al. [40] sollten die Teilnehmer beim Auffinden mehrerer Emotionen in einer Aussage alle notieren, sich jedoch möglichst auf eine einigen. Sollten sie keine Emotion auffinden können, sollten sie die Aussage mit dem Label *Neutral* versehen.

5.2 Teilnehmende des Workshops

Neben dem Autor dieser Bachelorarbeit nahmen vier weitere Informatikstudenten an dem Workshop für das Labeln des Datensatzes teil. Alle Personen sind männlich und im Alter zwischen 20 und 25 Jahren. Während einer der Teilnehmenden die Bachelorarbeit bereits abgeschlossen hat, schrieben drei andere, inklusive des Autors, zum Zeitpunkt des Workshops ihre Bachelorarbeit. Alle fünf Teilnehmer hatten bereits Erfahrungen in der Softwareentwicklung und mit Entwicklerteams mit regelmäßiger Kommunikation gesammelt.

5.3 Durchführung des Workshops

Zur Durchführung des Workshops trafen sich die fünf Teilnehmer und gingen gemeinsam die in Kapitel 5.1 vorgestellte Guideline durch. Um mögliche Unklarheiten, die im späteren Verlauf hätten auftreten können zu verhindern, wurde ein Beispieldatensatz mit 20 Aussagen, die nicht im originalen Datensatz vorhanden sind, vorgestellt. Nachdem alle Unklarheiten beseitigt wurden, erhielten die verschiedenen Teilnehmer ihren Teildatensatz mit 3.000 Aussagen, wobei der Autor dieser Arbeit den kompletten Datensatz mit 6.000 Aussagen bekam. So entstanden zwei Gruppen. Die Teilnehmer haben als ersten Teilschritt 100 Aussagen gelabelt. Daraufhin haben sich die Rater der jeweiligen Gruppen getroffen und die Label verglichen. Dies sollte dazu dienen, Unstimmigkeiten vor dem Labeln des großen, restlichen Teils aufzudecken und zudem zu schauen, in welchen Bereichen man sich uneinig ist und aus welchen Gründen. Zusätzlich hatten die Teilnehmer die Möglichkeit, Fragen zu klären. Erste Unstimmigkeiten traten dabei auf, die anschließend besprochen wurden und im folgenden Kapitel 6.1.1 erläutert werden. Nach dem Ende des Workshops bekam jeder Teilnehmer vier Wochen Zeit, den gesamten Datensatz zu labeln, wobei jeder auch die ersten 100 vergebenen Label hinsichtlich der in Kapitel 6.1.1 beschriebenen Unstimmigkeiten überarbeitet hat. Ein Teilnehmer aus der zweiten Gruppe hatte aus Zeitgründen mit dem Labeln aufgehört, weswegen ein Teilnehmer aus der ersten Gruppe seinen vollen Teildatensatz übernahm. Nachdem alle fertig waren, trafen sich die einzelnen Gruppen, um Unstimmigkeiten und deren Lösungen zu besprechen. Diese werden in Kapitel 6.1.2 beschrieben.

Kapitel 6

Ergebnisse

Im Folgenden werden die Ergebnisse, die mit der Auswertung des Labelprozesses und der Auswertung in Stimmungsanalysetools einhergehen, vorgestellt. Dabei wird auch auf die Unstimmigkeiten während des Labelns eingegangen.

6.1 Auswertung des Labelprozesses

Um die Auswertung des Labelprozesses zu betrachten, wird zunächst auf die Unstimmigkeiten zwischen den Ratern bei der Labelvergabe eingegangen. Anschließend werden die Ergebnisse dieses Prozesses präsentiert.

6.1.1 Unstimmigkeiten nach dem ersten Durchgang

Nach dem ersten Durchgang des Labelns, bei dem die ersten 100 Aussagen bewertet wurden, konnte festgestellt werden, dass es Schwierigkeiten gab, die Emotion *Liebe* richtig einzuordnen. Dieses Problem konnte gelöst werden, indem sich an dem englischen Goldstandard-Datensatz von Novielli et al. [40] orientiert wurde. Dort wurde *Liebe* vergeben, wenn sich eine Person positiv an eine andere Person widmet, indem sie bspw. ihr Lob ausspricht oder sich bei ihr bedankt. Im letzteren Fall herrschte auch Unklarheit, da ein ausgesprochener Dank subjektiv von einigen Teilnehmern als *Freude* aufgenommen wurde. Diese Aussagen konnten als Paradebeispiel dafür dienen, dass man die Aussagen möglichst objektiv beurteilen soll, man also nur dann ein Label vergibt, wenn die aussagende Person dies explizit mit seiner Wortwahl äußert. Mit diesem Punkt konnten auch Unstimmigkeiten bei der Vergabe der anderen Emotionen gelöst werden, da einige Teilnehmer zu viel Subjektivität in die Bewertung einer Aussage einfließen lassen haben. Diese Auffälligkeit ist beispielsweise hier bemerkbar:

> „So klappt es auch nicht. Es kommt immer eine Fehlermeldung:
> wie soll ich es denn jetzt machen?"

So haben zwei Rater die Emotion *Trauer* detektiert, der dritte *Angst*. Als Gründe gaben die Rater an, dass sie in so einer Situation verzweifelt oder verängstigt wären. Die Person hinter dieser Aussage äußert jedoch keine Emotion, weswegen die Aussage als neutral gewertet werden muss.

Im Allgemeinen konnten sich die Teilnehmer darauf einigen, dass Beleidigungen der Emotion *Wut* zugeordnet werden. Außerdem sind Fragen nicht unbedingt der Emotion *Überraschung* zuzuordnen. Viel eher kommt es in diesen Fällen auf Signalwörter wie „merkwürdig" oder „komisch" an, die bei der Feststellung einer Emotion unterstützen können.

Auffällig war, dass teilweise vergessen wurde, die Emotion *Überraschung* zusätzlich mit *Positiv* oder *Negativ* zu versehen. Gründe dafür waren, dass aufgrund der Seltenheit des Vorkommens der Emotion *Überraschung*, dies schlicht vergessen wurde. Deswegen wurden alle Teilnehmer darauf hingewiesen, dies in ihren Datensätzen zu korrigieren und in Zukunft zu berücksichtigen.

6.1.2 Unstimmigkeiten nach dem letzten Durchgang

Nachdem alle Teilnehmer ihre Datensätze fertiggelabelt hatten, trafen sich die zwei Gruppen separat um die Unstimmigkeiten zu lösen und um am Ende einen kompletten Datensatz mit eindeutig vergebenen Emotionslabel zu erstellen. Eine erste Auswertung ergab, dass die Teilnehmer in 1.205 von 6.000 Fällen nicht einer Meinung waren, also mindestens ein Rater ein anderes Label vergab als die anderen zwei. Jedoch gab es von diesen 1.205 Fällen insgesamt nur 91, in denen alle drei Rater unterschiedliche Label vergaben. Diese Unstimmigkeiten wurden in den Gruppensitzungen besprochen, um ein eindeutiges Label zu bestimmen. Ein Beispiel für so eine Unstimmigkeit ist bei folgender Entwickleraussage beobachtbar:

> „Schade, aber ich hatte sowas schon befürchtet. Naja, dann muß ich mich halt doch mit endlos langen dateinamen rumplagen"

So hat Rater 1 die Emotion *Angst* vergeben, da aus seiner Sicht der Aussagende diese Emotion mit dem Schlüsselwort „befürchtet" signalisiert hat. Rater 2 sah in dieser Aussage die Emotion *Trauer* aufgrund des Wortes „Schade", da dies eine klare Gefühlsäußerung darstellt. Rater 3 hingegen erkannte *Wut*, da die Wahl des Wortes „rumplagen" Genervtheit und somit einen aggressiven Ton ausstrahlt. Nach einer fünfminütigen Besprechung konnten sich die Rater darauf einigen, diese Aussage mit dem Label *Trauer* zu versehen, da das Wort „befürchtet" mit der Vergangenheit in Verbindung gebracht wurde und nicht mehr der aktuellen Gefühlsempfindung dieser Aussage angehört. Zudem muss das Wort „rumplagen" nicht unbedingt der Emotion *Wut* zugeordnet werden. Somit überwog am Ende das Schlüsselwort

„Schade" und die Emotion Trauer wurde entsprechend zugeordnet.
Bei einigen Aussagen haben sich die Rater über ihr vergebenes Label gewundert und zügig einem anderen Rater zugestimmt, was sie meist selber damit begründet haben, dass sie aus Versehen die falsche Taste gedrückt und dies wohl übersehen haben müssen. Außerdem wurde sich darauf geeinigt Aussagen, die auf Englisch waren oder nur aus Code bestanden, aus dem Datensatz zu entfernen.

Schlussendlich entstand somit ein Datensatz mit 5.949 manuell gelabelten Entwickleraussagen. Genauere Ergebnisse werden im folgenden Unterkapitel 6.2 beschrieben.

6.2 Ergebnisse des finalen Datensatzes

Folgend wird die erste Auswertung des finalen Datensatzes vorgestellt. Abbildung 6.1 zeigt dabei die Zusammensetzungen der Polaritäten des Datensatzes zu den drei verschiedenen Phasen.

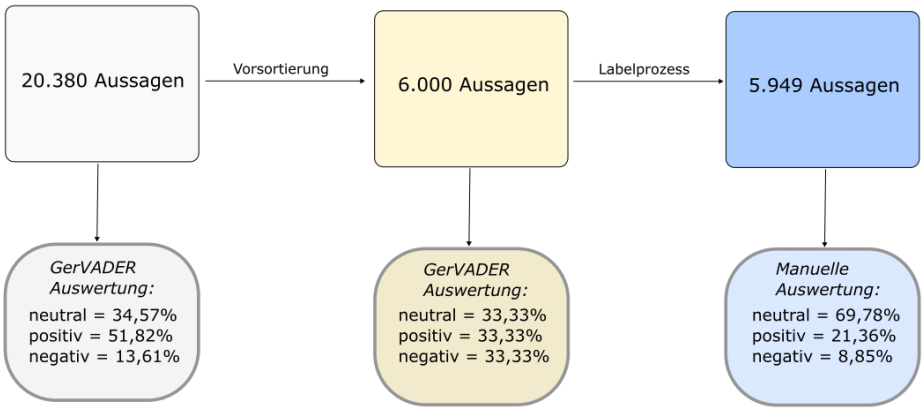

Abbildung 6.1: Zusammensetzung des Datensatzes zu verschiedenen Phasen

Die erste Auswertung des Labelprozesses hat ergeben, dass sich der Datensatz mit den 5.949 Aussagen aus 69.78% neutralen, 21.36% positiven und 8,85% negativen Entwickleraussagen zusammensetzt, was sich aus den einzelnen Angaben aus Tabelle 6.1 errechnen lässt, da für die Übersetzung der Emotionen in Polaritäten ebenso wie in der Arbeit von Novielli et al. [39] vorgegangen wurde. Dafür wurden *Liebe, Freude* und *positive Überraschung* in die Polarität *Positiv* und *negative Überraschung, Wut, Trauer* und *Angst* in die Polarität *Negativ* übersetzt.

Dabei ist ersichtlich, dass am häufigsten die Emotionen *Liebe* und *Trauer* mit einem Anteil von 19.06% und 6.45% und am seltensten *positive*

Aussagen	**Aussagen mit diesem Label**								**N**
	Neutral	Liebe	Freude	Pos. Über.	Neg. Über.	Wut	Trauer	Angst	
#	4.151	1.134	133	4	46	89	384	8	5.949
%	69.78%	19.06%	2.24%	0.07%	0.77%	1.5%	6.45%	0.13%	

Tabelle 6.1: Vorkommen der Emotionslabel im Datensatz

Überraschung und *Angst* mit 0.07% und 0.13% detektiert wurden.

Um die Validität des Labelprozesses zu prüfen und die Auswirkungen einer Zwischenbesprechung auf die Übereinstimmung und die Interrater-Reliabilität zu bestimmen, wurden in Tabelle 6.2 eben diese Werte berechnet.

Gang		**Neutral**	**Liebe**	**Freude**	**Pos. Üb.**	**Neg. Üb.**	**Wut**	**Trauer**	**Angst**	**Ges.**
1.	Überein.	0.63	0.86	0.71	0.88		0.84	0.9	0.92	0.43
	Fleiss' K	0.5	0.04	0.32	0.23		0.33	0.39	0.03	0.36
2.	Überein.	0.82	0.93	0.96	0.98	0.99	0.96	0.94	0.995	0.8
	Fleiss' K	0.71	0.85	0.47	0.28	0.1	0.37	0.67	0.34	0.71
Diff.	Überein.	+.19	+.07	+.25	-		+.12	+.04	+.08	+.37
	Fleiss' K	+.21	+.81	+.15			+.04	+.28	+.31	+.35

Tabelle 6.2: Reliabilitäten zwischen den Ratern für die Emotionen

Im ersten Durchgang ist erkennbar, dass aufgrund dessen, dass manche Rater vergessen haben, bei der Vergabe der Emotion *Überraschung* die Polarität mitanzugeben, nur ein gemeinsamer Übereinstimmungs- und Fleiss' Kappa Wert steht, weswegen auch eine Differenzrechnung zwischen den beiden Durchgängen nicht möglich war.
Außerdem ist ersichtlich, dass über die Emotionen hinweg recht hohe Übereinstimmung herrscht, die Fleiss' Kappa-Werte jedoch gering ausfallen. Sowohl die Übereinstimmungs- als auch Fleiss' Kappa-Werte aller Label waren nach dem zweiten Labeldurchgang höher. Sehr hohe Übereinstimmungswerte erzielten dabei die Emotionen *Angst, neg. Überraschung* und *pos. Überraschung*, wobei sie zugleich die niedrigsten Fleiss' Kappa-Werte besitzen.

Einen sehr großen Zuwachs verzeichnete der Fleiss Kappa-Wert von der Emotion *Liebe* mit einer Differenz von +0.81. Auch die Übereinstimmungs- und Fleiss' Kappa-Werte über die gesamten Emotionen konnten einen großen Zuwachs mit +0.37 bzw. +0.35 erzielen.

Novielli et al. [40] erzielten sehr ähnliche Übereinstimmungs- und Fleiss' Kappa-Werte bei den einzelnen Emotionen. Letztlich fallen die Kappa-Werte für *Angst* und *Liebe* dort mit 0.45 und 0.66 im Gegensatz zu 0.67 und 0.85

deutlich geringer aus, während *Wut* mit 0.62 im Gegensatz zu 0.37 hier deutlich geringer ausfällt. Die anderen Emotionen haben nur sehr leichte Abweichungen [40].

Tabelle 6.3 zeigt die Reliabilitäten für die Polaritäten, die aus den Emotionen übersetzt wurden. Auch dort verzeichnen alle Werte des zweiten Durchgangs einen Anstieg, auch wenn dieser eher moderat ausfällt, da die Werte im ersten Durchgang bereits hoch waren.

Gang		**Neutral**	**Positiv**	**Negativ**	**Gesamt**
1.	Überein.	0.67	0.83	0.78	0.66
	Fleiss' K	0.55	0.69	0.51	0.58
2.	Überein.	0.82	0.91	0.89	0.81
	Fleiss' K	0.71	0.81	0.59	0.73
Diff.	Überein.	+.15	+.08	+.11	+.15
	Fleiss' K	+.16	+.12	+.08	+.15

Tabelle 6.3: Reliabilitäten zwischen den Ratern für die Polaritäten

6.3 Auswertung in Stimmungsanalysetools

Für die Auswertung in Stimmungsanalysetools wurden vier deutsche Tools gewählt und nachfolgend deren Ergebnisse evaluiert.

6.3.1 Auswahl der Tools

Für die Evaluation des Datensatzes in deutschen Stimmungsanalysetools wurden die lexikonbasierten Tools GerVADER [59], SentiStrength_DE[1] und TextBlobDE[2] und das Machine-Learning-Tool BertDE [17] gewählt.

6.3.2 Ergebnisse der Tools

Um den erstellten Datensatz in den Stimmungsanalysetools zu evaluieren, wurden „classification reports" angefertigt, die die Precision, Recall und F1-Scores von den drei Polaritätsklassen und von Macro-avg und einen Accuracy-Wert, die allesamt im Grundlagenkapitel 2.4.1 erklärt wurden, berechnen.

Die Auswertung in Tabelle 6.4 zeigt deutlich, dass das Stimmungsanalysetool SentiStrength_DE die höchsten Macro-avg-F1-Scores und Accuracy-Werte aufweist. Man kann somit sagen, dass SentiStrength_DE mit einem

[1] `https://github.com/OFAI/SentiStrength_DE`
[2] `https://github.com/markuskiller/textblob-de`

Tool	Neutral			Positiv			Negativ			Macro-avg			Accuracy
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	
GerVader	.89	.57	.70	.53	.84	.65	.24	.58	.34	.55	.67	.56	.63
BertDE	.88	.22	.35	.60	.61	.61	.13	.90	.23	.54	.58	.39	.36
Senti-Strength_DE	.78	.85	.82	.67	.39	.49	.38	.47	.42	.61	.57	**.58**	**.72**
TextBlobDE	.72	.70	.71	.35	.37	.36	.15	.17	.16	.41	.41	.41	.58

Tabelle 6.4: Performanz des Datensatzes in Bezug auf die Tools

Accuracy-Wert von 0.72 die höchste Genauigkeit auf den erstellten Datensatz besitzt, dies aber immer noch keine gute Performanz darstellt. Dabei erzielt BertDE mit 0.36 den niedrigsten Accuracy-Wert.

Obaidi et al. [42] erstellten eine systematic mapping study (SMS) für Stimmungsanalysetools mit Anwendung im SE-Bereich, bei der sie die Macro-avg-F1-Scores und Accuracy-Werte verschiedener Quellen zusammenfassten und für diese einzelnen englischen Stimmungsanalysetools Durchschnittswerte bildeten. Zum Vergleich ist dabei ersichtlich, dass die Performanz englischer lexikonbasierter Tools wie SentiStrength [58] oder DEVA [26] mit einem durchschnittlichen Accuracy-Wert von 0.71 bzw 0.77 und einem durchschnittlichen Macro-avg-F1-Score von 0.61 bzw. 0.75 deutlich besser ist, als die der deutschen lexikonbasierten in Bezug auf den in dieser Arbeit erstellten Datensatz. Nur SentiStrength_DE hat mit einer Accuracy von 0.72 und einem Macro-avg-F1-Score von 0.58 ähnlich hohe Ergebnisse.

Um das Machine-Learning-Tool BertDE zu vergleichen, kann man die Performanzwerte englischer Machine-Learning-Tools wie Senti4SD [5] oder SentiCR [2] aus der SMS von Obaidi et al. [42] betrachten. Sie wurden ebenfalls mit Daten anderer Domänen trainiert. So erzielten sie 0.74 bzw. 0.77 Accuracy und 0.66 bzw. 0.69 Macro-avg-F1-Score im Durchschnitt. BertDE hat auf den hier erstellen deutschen Goldstandard-Datensatz im Vergleich sehr niedrige Werte mit einer Accuracy von 0.36 und einem Macro-avg-F1-Score von 0.39.

Bei dem Stimmungsanalysetool GerVADER ist erkennbar, dass der Precision-Wert für *Negativ* mit 0.24 auf einem tiefen Niveau liegt. Bei BertDE ist dieser Wert noch geringer, wobei der F1-Score für die neutrale Klasse ähnlich niedrig ausfällt. Bemerkenswert ist, dass die lexikonbasierten Tools für die neutrale Klasse mit F1-Scores zwischen 0.70 und 0.82 am besten performen.
Neben den bereits erwähnten Höchstwerten erzielte SentiStrength_DE die höchsten F1-Scores für die Klasse *Neutral* mit einem Wert von 0.82.

Zu beachten ist, dass die anderen lexikonbasierten Tools ebenfalls erhöhte Werte für diese Klasse aufweisen.

Mit einem F1-Score für *Negativ* von 0.16 hat TextBlobDE den niedrigsten Wert. Allgemein fallen die F1-Scores für *Negativ* bei allen Tools sehr gering aus. In gewissen Maße spiegeln sich diese Ergebnisse in den Reliabilitäten zwischen dem Datensatz und den einzelnen Stimmungsanalysetools, die in Tabelle 6.5 ablesbar sind, wieder.

Tool		**Neutral**	**Positiv**	**Negativ**	**Gesamt**
GerVADER	Cohens K	**0.37**	**0.52**	0.25	**0.38**
BertDE	Cohens K	0.10	0.50	0.09	0.18
Senti-Strength_DE	Cohens K	0.32	0.40	**0.36**	0.35
TextBlobDE	Cohens K	0.08	0.18	0.07	0.12

Tabelle 6.5: Reliabilitäten zwischen Mensch und Stimmungsanalysetool

Erkennbar ist, dass GerVADER und SentiStrength_DE mit Cohens Kappa-Werten von 0.38 und 0.35 die höchsten Übereinstimmungen erzielten, aber dennoch als minimal gelten.

Kapitel 7

Diskussion

Im Rahmen dieser Bachelorarbeit wurde ein deutscher Datensatz für die Stimmungsanalyse von Entwickleraussagen erstellt. Um die Ergebnisse aus Kapitel 6 korrekt einzuordnen, folgt eine Diskussion mittels Interpretation dieser. Auf mögliche Validity Threats wird eingegangen.

7.1 Interpretation der Ergebnisse

Aus den Ergebnissen aus Kapitel 6 lässt sich schließen, dass trotz einiger Unstimmigkeiten zwischen den Ratern nach den beiden Labeldurchgängen hohe Übereinstimmungs- und Reliabilitätswerte erzielt werden konnten, die anderen Goldstandard-Datensätzen aus dem SE-Bereich wie beispielsweise denen von Novielli et al. [39] [40] stark ähneln.

Die Differenzen dieser Werte zwischen den beiden Durchgängen machen deutlich, dass eine Zwischenbesprechung, um Unstimmigkeiten zu lösen und einen passablen Goldstandard-Datensatz zu erstellen, unverzichtbar ist. Vor allem der starke Sprung bei dem Kappa-Wert von der Emotion *Liebe* in Tabelle 6.2 zeigt, dass trotz einer gelehrten Guideline jeder Mensch einen gewissen Anteil an Subjektivität in die Vergabe von Emotionslabel einfließen lässt und eine Zwischenbesprechung solche Unterschiede minimieren kann. Betrachtet man die Polaritäten in Tabelle 6.3, so sind die Verbesserung durch eine solche Zwischenbesprechung nicht sehr hoch, da die Unterscheidung zwischen *Positiv* und *Negativ* intuitiver ist, als die Festlegung auf eine genaue Basisemotion. Jedoch ist sie auch dort sinnvoll, um Aussagen in die Klasse *Neutral* einzuordnen, damit mögliche Subjektivität des Raters von tatsächlicher existierender Stimmung innerhalb einer Aussage abgegrenzt wird.

Trotz Vorsortierung des Datensatzes mit GerVADER [59] ist mit einem Anteil von 69.78% (siehe Tabelle 6.1) ein unbalancierter Datensatz

entstanden. Die niedrigen F1-Scores für die Polaritäten in Bezug auf den von Menschen vergebenen Label in Tabelle 6.4 zeigen somit, dass GerVADER dazu neigt, in neutralen Entwickleraussagen eine Polarität zu entdecken. Vor allem ist dies bei der Klasse *Negativ* der Fall. Da SentiStrength_DE die höchste Genauigkeit mit einem Accuracy-Wert von 0.72 erzielt hat, ist dieses Tool für die Vorsortierung deutscher Sätze aus dem SE-Bereich eventuell besser geeignet als GerVADER. Dies kann jedoch je nach Auswahl der Datensatzquelle variieren, da die Entwickleraussagen subjektive Äußerungen von unterschiedlichen Individuen darstellen.

Zwar erzielt SentiStrength_DE ähnlich hohe Macro-avg-F1-Scores und Accuracy-Werte wie vergleichbare englische Tools wie SentiStrength bei Betrachtung des Durchschnitts mehrerer Datensätze [42] und würde sich somit am besten für die in dieser Arbeit benutzten Domäne eignen. Jedoch fallen diese und die Cohens Kappa-Werte schlecht aus (siehe Tabelle 6.5). Anzumerken hierbei ist, dass Imtiaz et al. [23] feststellten, dass kein Tool für die SE-Domäne zuverlässig ist und sich Mensch und Tool selten sehr einig sind. Deswegen sollte man sich nicht gänzlich auf dieses Tool verlassen, es könnte aber als Anfangspunkt neuer Evaluationen in diesem Bereich dienen.

Bei Betrachtung der Accuracy und F1-Scores des Machine-Learning-Tools BertDE fällt auf, dass dieses am schlechtesten performt. Anzumerken hierbei ist, dass BertDE nicht auf die Domäne des SE trainiert wurde. Das englische Tool Senti4SD [5] schneidet beispielsweise bei der Evaluation domänenunbekannter Datensätze ebenfalls schlecht ab [63], wohingegen es bei der Anwendung von Datensätzen der selben Domäne mit einem Accuracy-Wert von 0.89 ein deutlich besseres Ergebnis erzielt. Ben et al. [3] stellten fest, dass Domänenspezifität wichtig für Machine-Learning ist, um genaue Ergebnisse zu liefern. Dementsprechend wäre es möglich, dass BertDE nach einem Training mit dem in dieser Arbeit erstellten Datensatz ebenfalls gut performen würde.

Intention dieser Arbeit war es, einen deutschen Goldstandard-Datensatz für die Stimmungsanalyse von Entwickleraussagen zu erstellen, da nach bestem Wissen bisher kein deutscher Datensatz für den SE-Bereich existiert. Novielli et al. [39] merkten an, dass eine hohe Interrater-Reliabilität ein wichtiger Faktor für die Validität des Goldstandard-Datensatzes sei. Die Auswertungen zeigen, dass dies mit dieser Arbeit, wenn man die Ergebnisse mit vergleichbaren Erstellungen von Goldstandard-Datensätzen vergleicht, erfolgreich gelungen ist. Um die Validität des Datensatzes im Bereich des SE weiter zu bestätigen, hätte es sein können, dass die bereits vorhandenen deutschen Stimmungsanalysetools diese Domäne gut bewertet hätten. Da dies jedoch nicht der Fall ist, ist es wichtig, etablierte deutsche Stimmungsanalysetools für den SE-Bereich zu entwickeln, da solche im

englischen Bereich bessere Performanzen ausweisen [42]. Aufgrund dessen, das so ein deutsches Tool noch nicht existiert, könnte der erstellte Datensatz zur Erstellung dessen verwendet werden.

7.2 Validity Threats

Folgend werden mögliche Validity Threats benannt, wobei die Kategorisierung mittels der Threads nach Wohlin [61] erfolgt. Unterschieden wird dabei zwischen *Threats to Conclusion, Internal, External* und *Threats to Construct Validity.*

Bei der Labelvergabe wurde jede Aussage von nur drei Personen gerated. Dies könnte man als Kritikpunkt ansehen, da für valide Ergebnisse noch mehr Rater vorhanden sein sollen (Threat to Conclusion Validity). Ortu et al. [44] stellten jedoch fest, dass sich die Vergabe des Labels bei mehr als zwei Ratern nicht signifikant ändert und drei Personen somit ausreichen sollten.
Trotz Beachtung dieser Feststellung ist die Klassifizierung von Menschen nicht perfekt. So können, wie in Kapitel 5 festgestellt, menschliche Fehler passieren, indem man sich beispielsweise vertippt oder gedanklich abschweift, wenn man eine lange Zeit ohne Pause mit dem Labeln beschäftigt ist. Außerdem hat jeder Mensch ein individuelles Empfinden, weswegen trotz fester Guideline eine Subjektivität bei der Labelvergabe einfließt [20]. Da bei dem finalen Ergebnis eine Mehrheitsauswahl stattfand, könnte es also passieren, dass zu einer Aussage mindestens zwei Rater eine Fehlentscheidung getroffen hätten und dies am Ende mit in den fertigen Datensatz aufgenommen wurde (Threat to Internal Validity).

Die hohen Schwankungen der Übereinstimmungs- und Fleiss' Kappa-Werten zeigen, dass eine gewisse Beständigkeit zwischen den Ratern nicht gegeben war und weitere Zwischenbesprechungen noch weitere Veränderungen mit sich bringen könnten (Threat to Conclusion Validity). Andernfalls ist aber erkennbar, dass die Schwankungen der Werte sich größtenteils innerhalb der Polaritäten selbst bewegten.

Alle Rater waren männlich und im Alter von 20 bis 25 Jahren. Es müsste analysiert werden, welche Einflüsse das Geschlecht und das Alter auf den Labelprozess haben und ob eine größere Aufteilung der Teilnehmer sinnvoller wäre (Threat to External Validity).

Anders als ursprünglich geplant, hat ein Rater aus der ersten Gruppe zusätzlich 3.000 Aussagen der zweiten Gruppe gelabelt, da die Person aus der zweiten Gruppe seine Teilnahme abgelegt hat. Für die Auswertungen

des ersten Durchgangs wurden noch die Label des Ausgetreteten gewertet, was zu einer verzerrten Vergleichbarkeit der unterschiedlichen Durchgänge führt (Threat to Internal Validity). Trotzdem wurde insgesamt jede Aussage von drei unterschiedlichen Ratern bewertet und die positiven Auswirkungen einer Zwischenbesprechung behalten ihre Validität, da selbst wenn der Ausgetretene keine Verbesserungen in seiner Labelvergabe verzeichnet hätte, seine Reliabilitäten das finale Ergebnis nur marginal beeinflusst hätten.

Lin et al. [31] stellten fest, dass Machine-Learning eine Art „blackbox" sei und dass das Training vom englischen Entwicklerforum Stack Overflow negative Resultate mit sich brachte, da die Quelle eventuell nicht gut genug sei. Dieses Problem könnte trotz dessen, dass Android-hilfe.de als akzeptable Quelle scheint, hier ebenfalls bestehen (Threat to Construct Validity).

Bei dem Vergleich der Ergebnisse wurden englische Arbeiten herangezogen, was nur eine bedingte Validität, aufgrund diverser sprachlicher Unterschiede zwischen der deutschen und englischen Sprache, mit sich bringt. Andernfalls ist ein Vergleich mit deutschen Arbeiten nicht möglich, da bisher noch kein Datensatz mit deutschen Entwickleraussagen aus dem SE-Bereich erstellt wurde (Threat to Construct Validity).

Zusätzlich kann bemängelt werden, dass in dieser Arbeit zu wenige Stimmungsanalysetools zur Evaluation verwendet wurden, um die Auswertungen als repräsentativ anzusehen (Threat to Construct Validity). Jedoch sind dies nach bestem Wissen die gängigsten und etabliertesten frei verfügbaren Tools für die deutsche Stimmungsanalyse.

Kapitel 8

Zusammenfassung und Ausblick

Dieses Kapitel gibt einen abschließenden Überblick über die Ergebnisse dieser Arbeit. Anschließend folgt ein Ausblick vorgestellt, der auf weitere Anwendungsmöglichkeiten und offene Probleme eingeht.

8.1 Zusammenfassung

Stimmungsanalyse im SE-Bereich wird unter anderem dafür verwendet, positive Stimmungen zu manifestieren [16] oder negative Stimmungen frühzeitig zu identifizieren [31] [50] [62]. Eine von Graziotin et al. [16] durchgeführte Studie kam zu dem Entschluss, dass positive Stimmungen in Entwicklerteams die Produktivität der Entwickler steigern. Um diese Stimmungen zu erkennen, wurden Stimmungsanalysetools entwickelt, die an die Domäne des Software Engineering angepasst sind [2] [5] [27]. Diese sind jedoch nur für den englischen Bereich anwendbar. Aufgrund dessen sollte im Zuge dieser Arbeit ein deutscher Goldstandard-Datensatz für die Stimmungsanalyse von Entwickleraussagen erstellt werden, da er dafür dienen könnte, ein deutsches Machine-Learning-Stimmungsanalysetool für die SE-Domäne zu trainieren.
Um dies zu verwirklichen, war es notwendig, sich auf eine qualitativ hochwertige Datenquelle festzulegen und anschließend eine feste Guideline für den Labelprozess zu erstellen, die sich an bestehenden Arbeiten orientiert. Auswertungen zeigten, dass neben dieser Guideline eine Zwischenbesprechung der Rater elementar ist, um gute Übereinstimmungs- und Reliabilitätswerte zwischen den Ratern zu erzielen.
Nach dem Labelprozess entstand ein Datensatz mit 5.949 deutschen Entwickleraussagen, bei dem jede Aussage durch drei verschiedene Rater in die sechs Basisemotionen nach Shaver et al. [55] klassifiziert wurde. Basierend auf den erzielten hohen Übereinstimmungs- und Reliabilitätswerten lässt sich sagen, dass am Ende ein repräsentabler Goldstandard-Datensatz entstanden

ist.
Evaluationen in deutschen Stimmungsanalysetools zeigen, dass diese für die Verwendung in der Domäne des Software Engineering nicht verlässlich sind, da sie schlechte Performanzwerte erzielten. Umso wichtiger ist es, ein deutsches Machine-Learning-Stimmungsanalysetool für die SE-Domäne zu entwickeln. Dafür könnte der in dieser Arbeit erstellte Goldstandard-Datensatz dienen.

8.2 Ausblick

Das Konzept der Erstellung des Datensatzes mitsamt seiner Ergebnisse kann für die Erstellung weiterer Datensätze behilflich sein.
Eine Möglichkeit, einen gelabelten Datensatz bei der Erstellung zu optimieren, wäre es zu testen, ob weitere Zwischenbesprechungen zu noch besserer Einigkeit bei der Labelvergabe führt.
Um das Risiko zu minimieren, dass sich Rater bei ihrer Labelvergabe vertippen, könnte ein neues Konzept oder eine spezielle Software dienen.
Bei der Auswertung in Stimmungsanalysetools müsste darauf geachtet werden, dass die Tools an die jeweilige Domäne angepasst sind. Außerdem ist eine höhere Anzahl an Stimmungsanalysetools bei der Evaluation von Vorteil, um repräsentable Ergebnisse zu liefern.
Um ein deutsches Machine-Learning-Stimmungsanalysetool zu konstruieren, könnte dieser Datensatz für das Training dienen. Dennoch erzielen diese eine höhere Genauigkeit, je größer die Goldstandard-Datensätze sind. Dementsprechend wäre es notwendig, weitere Datensätze aus anderen deutschsprachigen Quellen unter Beachtung der hier vorgeschlagenen Optimierungen zu erstellen.

Literaturverzeichnis

[1] M. Ahmad, S. Aftab, S. S. Muhammad, and S. Ahmad. Machine learning techniques for sentiment analysis: A review. *Int. J. Multidiscip. Sci. Eng*, 8(3):27, 2017.

[2] T. Ahmed, A. Bosu, A. Iqbal, and S. Rahimi. Senticr: a customized sentiment analysis tool for code review interactions. In *2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 106–111. IEEE, 2017.

[3] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

[4] K. Boland, A. Wira-Alam, and R. Messerschmidt. *Creating an Annotated Corpus for Sentiment Analysis of German Product Reviews*, volume 2013/05 of *GESIS-Technical Reports*. GESIS - Leibniz-Institut für Sozialwissenschaften, Mannheim, 2013.

[5] F. Calefato, F. Lanubile, F. Maiorano, and N. Novielli. Sentiment polarity detection for software development. *Empirical Software Engineering*, 23(3):1352–1382, 2018.

[6] F. Calefato, F. Lanubile, and N. Novielli. Emotxt: A toolkit for emotion recognition from text. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 79–80, 2017.

[7] L. G. Chepenik, L. A. Cornew, and M. J. Farah. The influence of sad mood on cognition. *Emotion*, 7(4):802, 2007.

[8] M. Cieliebak, J. M. Deriu, D. Egger, and F. Uzdilli. A Twitter corpus and benchmark resources for German sentiment analysis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 45–51, Valencia, Spain, Apr. 2017. Association for Computational Linguistics.

[9] J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[11] F. Dorsch and M. A. Wirtz. *Dorsch - Lexikon der Psychologie, Lexikon der Psychologie*. Hogrefe;, Bern, 2020. Lexikon der Psychologie.

[12] R. Feldman. Techniques and applications for sentiment analysis. *Commun. ACM*, 56(4):82–89, apr 2013.

[13] K. W. Fischer, P. R. Shaver, and P. Carnochan. How emotions develop and how they organise development. *Cognition and emotion*, 4(2):81–127, 1990.

[14] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[15] A. C. Frenzel, T. Götz, and R. Pekrun. Emotionen. In *Pädagogische Psychologie*, pages 205–231. Springer, 2009.

[16] D. Graziotin, X. Wang, and P. Abrahamsson. Happy software developers solve problems better: psychological measurements in empirical software engineering. *PeerJ*, 2:e289, Mar. 2014.

[17] O. Guhr, A.-K. Schumann, F. Bahrmann, and H. J. Böhme. Training a broad-coverage german sentiment classification model for dialog systems. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1627–1632, 2020.

[18] E. Guzman and W. Maalej. How do users like this feature? a fine grained sentiment analysis of app reviews. In *2014 IEEE 22nd international requirements engineering conference (RE)*, pages 153–162. Ieee, 2014.

[19] M. Herrmann. Automatische klassifikation von aussagen in meetings von entwicklungsteams. *Bachelorthesis, Leibniz Universität Hannover*, 2021.

[20] M. Herrmann, M. Obaidi, L. Chazette, and J. Klünder. On the subjectivity of emotions in software projects: How reliable are pre-labeled data sets for sentiment analysis? *Journal of Systems and Software*, 193:111448, 2022.

[21] M. Herrmann, M. Obaidi, and J. Klünder. Senti-analyzer: Joint sentiment analysis for text-based and verbal communication in software projects, 2022.

[22] C. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225, 2014.

[23] N. Imtiaz, J. Middleton, P. Girouard, and E. Murphy-Hill. Sentiment and politeness analysis tools on developer discussions are unreliable, but so are people. In *2018 IEEE/ACM 3rd International Workshop on Emotion Awareness in Software Engineering (SEmotion)*, pages 55–61, 2018.

[24] M. R. Islam, M. K. Ahmmed, and M. F. Zibran. Marvalous: Machine learning based detection of emotions in the valence-arousal space in software engineering text. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 1786–1793, 2019.

[25] M. R. Islam and M. F. Zibran. Leveraging automated sentiment analysis in software engineering. In *2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR)*, pages 203–214. IEEE, 2017.

[26] M. R. Islam and M. F. Zibran. Deva: sensing emotions in the valence arousal space in software engineering text. In *Proceedings of the 33rd annual ACM symposium on applied computing*, pages 1536–1543, 2018.

[27] M. R. Islam and M. F. Zibran. Sentistrength-se: Exploiting domain specificity for improved sentiment analysis in software engineering text. *Journal of Systems and Software*, 145:125–146, 2018.

[28] C. E. Izard. Basic emotions, relations among emotions, and emotion-cognition relations. 1992.

[29] O. Kolchyna, T. T. Souza, P. Treleaven, and T. Aste. Twitter sentiment analysis: Lexicon method, machine learning method and their combination. *arXiv preprint arXiv:1507.00955*, 2015.

[30] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[31] B. Lin, F. Zampetti, G. Bavota, M. Di Penta, M. Lanza, and R. Oliveto. Sentiment analysis for software engineering: How far can we go? In *Proceedings of the 40th international conference on software engineering*, pages 94–104, 2018.

[32] B. Liu. *Sentiment analysis: Mining opinions, sentiments, and emotions.* Cambridge university press, 2020.

[33] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796, 2014.

[34] C. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*. Cambridge University Press Location Cambridge, MA, 2008.

[35] W. Medhat, A. Hassan, and H. Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113, 2014.

[36] A. Mehrabian and J. A. Russell. *An approach to environmental psychology*. the MIT Press, 1974.

[37] S. Momtazi. Fine-grained german sentiment analysis on social media. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1215–1220, 2012.

[38] A. Murgia, P. Tourani, B. Adams, and M. Ortu. Do developers feel emotions? an exploratory analysis of emotions in software artifacts. *11th Working Conference on Mining Software Repositories, MSR 2014 - Proceedings*, 05 2014.

[39] N. Novielli, F. Calefato, D. Dongiovanni, D. Girardi, and F. Lanubile. Can we use se-specific sentiment analysis tools in a cross-platform setting? In *Proceedings of the 17th International Conference on Mining Software Repositories*, pages 158–168, 2020.

[40] N. Novielli, F. Calefato, and F. Lanubile. A gold standard for emotion annotation in stack overflow. In *2018 IEEE/ACM 15th International Conference on Mining Software Repositories (MSR)*, pages 14–17, 2018.

[41] M. Obaidi and J. Klünder. Development and application of sentiment analysis tools in software engineering: A systematic literature review. *Evaluation and Assessment in Software Engineering*, pages 80–89, 2021.

[42] M. Obaidi, L. Nagel, A. Specht, and J. Klünder. Sentiment analysis tools in software engineering: A systematic mapping study. *Information and Software Technology*, 151:107018, 2022.

[43] M. Okpor. Machine translation approaches: issues and challenges. *International Journal of Computer Science Issues (IJCSI)*, 11(5):159, 2014.

[44] M. Ortu, A. Murgia, G. Destefanis, P. Tourani, R. Tonelli, M. Marchesi, and B. Adams. The emotional side of software developers in jira. In *2016 IEEE/ACM 13th Working Conference on Mining Software Repositories (MSR)*, pages 480–483. IEEE, 2016.

[45] C. E. Osgood, G. J. Suci, and P. H. Tannenbaum. *The measurement of meaning*. Number 47. University of Illinois press, 1957.

[46] W. G. Parrott. *Emotions in social psychology: Essential readings*. psychology press, 2001.

[47] S. Pawiro, P. Markelj, F. Pernuš, C. Gendrin, M. Figl, C. Weber, F. Kainberger, I. Nöbauer-Huhmann, H. Bergmeister, M. Stock, et al. Validation for 2d/3d registration i: A new gold standard data set. *Medical physics*, 38(3):1481–1490, 2011.

[48] R. Plutchik. A psychoevolutionary theory of emotions. *Social sciences information*, 21(4-5):529–553, 1982.

[49] A. Poncelas, P. Lohar, A. Way, and J. Hadley. The impact of indirect machine translation on sentiment classification. *arXiv preprint arXiv:2008.11257*, 2020.

[50] R. Prabowo and M. Thelwall. Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(2):143–157, 2009.

[51] S. Rasp, P. D. Dueben, S. Scher, J. A. Weyn, S. Mouatadid, and N. Thuerey. Weatherbench: a benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203, 2020.

[52] J. A. Russell. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145, 2003.

[53] H. Saif, M. Fernandez, Y. He, and H. Alani. Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the sts-gold. 2013.

[54] M. Sänger, U. Leser, S. Kemmerer, P. Adolphs, and R. Klinger. SCARE — the sentiment corpus of app reviews with fine-grained annotations in German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1114–1121, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[55] P. Shaver, J. Schwartz, D. Kirson, and C. O'connor. Emotion knowledge: further exploration of a prototype approach. *Journal of personality and social psychology*, 52(6):1061, 1987.

[56] M. Siegel and M. Alexa. *Sentiment-Analyse deutschsprachiger Meinungsäußerungen: Grundlagen, Methoden und praktische Umsetzung*. Springer, 2020.

[57] M. Thelwall. The heart and soul of the web? sentiment strength detection in the social web with sentistrength. In *Cyberemotions*, pages 119–134. Springer, 2017.

[58] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment strength detection in short informal text. *Journal of the American society for information science and technology*, 61(12):2544–2558, 2010.

[59] K. Tymann, M. Lutz, P. Palsbröker, and C. Gips. Gervader-a german adaptation of the vader sentiment analysis tool for social media texts. In *LWDA*, pages 178–189, 2019.

[60] D. Ulich and P. Mayring. *Psychologie der Emotionen*. 2003.

[61] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén. *Experimentation in software engineering*. Springer Science & Business Media, 2012.

[62] L. Zhang, S. Wang, and B. Liu. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253, 2018.

[63] T. Zhang, B. Xu, F. Thung, S. A. Haryono, D. Lo, and L. Jiang. Sentiment analysis for software engineering: How far can pre-trained transformer models go? In *2020 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 70–80. IEEE, 2020.