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

Analyse von Einflüssen in der Sentimenterkennung von Entwicklern

Bachelorarbeit

im Studiengang Informatik

von

Jendrik Martensen

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

Hannover, 14. Februar 2022

Erklärung der Selbstständigkeit

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit selbstständig 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 14.02.2022

_____

Jendrik Martensen

Kurzfassung

Ein wichtiger Faktor für die erfolgreiche Durchführung eines Software Projekts ist die Stimmung im Entwicklerteam. Positive Stimmung sorgt für produktivere Arbeit und erhöht die Chancen auf ein erfolgreiches Softwareprojekt. Um die Stimmung aus der textbasierten Kommunikation abzuleiten, können Sentimentanalyse-Tools verwendet werden.

Bei der Verwendung dieser Tools kann es jedoch zu Fehlern und Uneinigkeiten kommen. In dieser Arbeit werden durch eine Umfrage unter Teilnehmenden des Softwareprojektes des Fachgebiets Software Engineering der Leibniz Universität Hannover mögliche Einflussfaktoren auf die Sentimentvergabe betrachtet und Zusammenhänge zwischen den Faktoren und der Sentimentvergabe untersucht. Dabei sind die untersuchten Faktoren die Stimmung der Entwickler, die mehrfache Umfrageteilnahme, und damit die mehrmalige Sentimentvergabe für die gleichen Aussagen, die Projektphase und die Gruppendynamik Entwicklerteam.

Die Ergebnisse lassen vermuten, dass die Stimmung keinen Einfluss hat, eine mehrfache Teilnahme zu einem veränderten Fokus und mehr Abweichungen bei der Sentimentvergabe führen und Konflikte innerhalb der Projektgruppe für eine negativere Wahrnehmung und Sentimentvergabe sorgen. Auch deuten die Ergebnisse auf einen Zusammenhang zwischen Projektart und Sentimentvergabe hin.

Abstract

An important aspect for a successful software project is the mood in the development team. A positive mood leads to greater productivity and increases the probability of a successful software project. Sentiment analysis tools can derive the mood from the text-based communication.

Using these tools can cause errors or disagreements between different tools. This work will analyze possible influences on the sentiment allocation through a survey among participants of the module software project of the Leibniz University Hannover. The analyzed influences are the mood of the developers, the multiple participation in the survey which leads to multiple sentiment allocation for the same sentences, the project phase and the group dynamics in the development team.

The results suggest that the mood has no influence, the multiple participation leads to a changed focus and more differences in the sentiment allocation. Conflicts within a project group lead to a more negative perception and sentiment allocation. Furthermore, there could be a connection between the type of project and the sentiment allocation.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einleitung

Um Softwareprojekte durchzuführen, wird überwiegend auf Teamarbeit gesetzt, statt einzelne Entwickler mit einem Projekt zu beauftragen [21]. Ein Grund dafür ist, dass Softwareprojekte bzw. die zu entwickelnde Software immer komplexer werden [21]. Die eingesetzten Teams sind oftmals global verteilt, wodurch eine Kommunikation im Team in Teilen nur in digitaler schriftlicher Form möglich ist [23]. Eine Studie aus dem Jahr 2017 stellte fest, dass ca. 60% der Softwareteams global verteilt sind, was die Notwendigkeit digitaler Kommunikation weiter unterstreicht [23].

Im *Chaos Report 2015* [55] der *Standish Group*[1] wird dargestellt wie erfolgreiche Softwareprojekte durchgeführt wurden [55]. Dabei werden verschiedene Kategorien betrachtet und dann analysiert, ob diese erfolgreich durchgeführt wurden [55]. Aus dem Report wird deutlich, dass mit steigender Größe und Komplexität die Erfolgsquote sinkt und der Misserfolg eines Projektes wahrscheinlicher wird [55]. Eine Umfrage unter Entwicklern ergab, dass bereits ab einer Teamgröße von 4 Personen der Erfolg deutlich unwahrscheinlicher wird [3]. Große, komplexe und globale Projekte sind demnach zum einen notwendig [23], aber zum anderen auch risikoreicher [55] [3].

Jedoch lässt sich die Erfolgsquote eines Softwareprojektes durch eine gute Stimmung unter den Entwicklern heben [12]. Eine Studie unter Softwareentwicklern hat die positiven Wirkungen von glücklichen Entwicklern aufgezeigt [12]. Diese arbeiten produktiver und haben eine bessere Problemlösekompetenz als negativ gestimmte Entwickler [12]. Negative Stimmung kann sich innerhalb eines Teams von einzelnen Mitgliedern auf das gesamte Team ausbreiten, wodurch im Extremfall das gesamte Team mit verminderter Produktivität und Problemlösekompetenz arbeitet [47]. Diese Entwicklungen innerhalb eines Teams stellen bei Nichtbeachtung oder Vernachlässigung demnach ein weiteres Risiko da.

[1] https://standishgroup.com

Mittels Stimmungsanalyse können Stimmungen ermittelt werden und zu Gesamtstimmungsbildern einer Einzelperson oder eines gesamten Teams zusammengesetzt werden [25]. Für die Analyse schriftlicher Kommunikation können Stimmungsanalyse- oder Sentiment-Analyse Tools verwendet werden [27] [14]. Diese werten die Kommunikation zwischen Teammitglieder über verschiedene Plattformen (z.B. *E-Mails, GitHub* Einträge etc.) aus und weisen dieser dann ein Sentiment bzw. ein Label zu, welches positiv, negativ oder neutral sein kann [27]. Mittels dieser Tools ist es möglich, aus der textbasierter Kommunikation in einem Softwareprojekt Rückschlüsse auf die Stimmung der Entwickler zu ziehen [21] [54].

1.1 Motivation

Sentimentanalyse-Tools analysieren Texte automatisiert und liefern als Ergebnis ein Label für diesen Text. Dabei können jedoch Probleme auftreten, wodurch ein Sentiment nicht passend zugeordnet wird [16] [19]. Verursacht werden die Probleme durch z.B. die unterschiedliche subjektive Wahrnehmungsweise der Leser [50] [66]. Rater können durch Subjektivität verschiedene Label für den gleichen Satz vergeben [66]. Beispielsweise kann der Satz *“It’s always sad to see a reference like that go, but it was probably a good move.”*[2] betrachtet werden [66]. Dieser kann negativ bewertet werden aufgrund des Satzanfangs [66]. Der Teil *„but it was probably a good move“*[3] sollte jedoch positiv bewertet sein. Der Teil des Satzes, der vom Leser auf subjektive Weise stärker wahrgenommen wird, entscheidet in diesem Fall über das vergebene Label [66].

Für den Projekterfolg kann es jedoch von Vorteil sein, zu wissen, wie eine Aussage wahrgenommen wird, um die Stimmung dadurch richtig zu deuten [12]. Von daher wird in dieser Arbeit analysiert, welche Einflussfaktoren es auf die Sentimentvergabe bei Entwicklern geben kann.

[2] 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?," *2020 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, 2020, pp. 70-80, doi: 10.1109/ICSME46990.2020.00017

[3] 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?," *2020 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, 2020, pp. 70-80, doi: 10.1109/ICSME46990.2020.00017

1.2 Herangehensweise

Zur Untersuchung der Einflussfaktoren auf die Sentimentvergabe wird eine Umfrage erstellt. Die Umfrage zielt darauf ab, verschiedene Faktoren wie z.B. die Stimmung zu erfragen und anschließend auszuwerten, ob diese einen Einfluss auf die Sentimentvergabe haben. Um die Einflüsse ermitteln zu können, werden in der Umfrage Sätze der Teilnehmenden mit Labeln versehen. Dabei sollen die folgenden Forschungsfragen beantwortet werden:

1. Welchen Einfluss hat die Stimmung auf die Sentimentvergabe?
2. Wie entwickelt sich die Sentimentvergabe bei mehrfacher Umfrageteilnahme?
3. Welchen Einfluss haben verschiedene Phasen des aktuellen Softwareprojekts und die Gruppendynamik in einem Softwareprojekt auf die Sentimentvergabe?

Die Umfrage wird zu mehreren Zeitpunkten eines Softwareprojektes durchgeführt. Dabei werden die verschiedenen Zeitpunkte *Ende Iteration 1*, *Mitte Iteration 2*, *Ruhephase nach einer Weihnachtspause* und *Abnahme* abgefragt.

1.3 Struktur der Arbeit

In Kapitel 2 der Arbeit werden die Grundlagen der Sentimentanalyse und die Grundbegriffe Emotion, Stimmung und Affekt sowie ein Emotionsmodell und Skalen zur Messung von Emotionen vorgestellt.

Kapitel 3 verwandte Arbeiten vor.

Kapitel 4 beschreibt den Aufbau der Umfrage und die gewählten statistischen Maße zur Analyse.

Kapitel 5 präsentiert die Ergebnisse aus den Umfragen. Dazu werden zunächst Einzelergebnisse aus den jeweiligen Umfrageteilen vorgestellt und diese dann zur Beantwortung der Forschungsfragen auf Zusammenhänge untersucht. Danach werden die *Threats of Validity* aufgezeigt

Kapitel 6 gibt eine kurze Zusammenfassung der Erkenntnisse aus den Ergebnissen und diskutiert diese. Abschließend wird ein Ausblick gegeben.

Kapitel 2

Grundlagen

2.1 Sentimentanalyse

Die Sentimentanalyse ist eine Art der Stimmungsanalyse, die häufig zur Textanalyse genutzt wird [20]. Texte können auf emotionale Informationen untersucht werden, wobei diese auch einzelne Sätze oder Aussagen sein können [24] [59]. Ziel der Analyse ist die Ermittlung der Stimmung, die dem Text zugrunde liegt. Ausgedrückt werden kann die Stimmung durch die Zuordnung einer Polarität bzw. eines Labels [57]. Unterschieden wird zwischen *positiver*, *negativer* und *neutraler* Polarität, wobei eine neutrale Polarität dann vorliegt, wenn weder eine positive noch eine negative Stimmung ermittelt werden konnte [57]. Den Stimmungen können zur besseren Differenzierung Emotionen zugeordnet werden [24]. Beispielsweise können die Emotionen fröhlich oder optimistisch der positiven Polarität zugeordnet werden [49]. Emotionen der negativen Polarität sind z.B. nervös oder enttäuscht [49]. Aus der Betrachtung der Emotionen lassen sich durch Zuordnung zu den Stimmungen auch wieder die übergeordneten Polaritäten erzeugen.
Neben der Zuordnung einer Polarität zu einer einzelnen Aussage ist eine der Hauptaufgaben der Sentimentanalyse eine zusammenfassende Stimmung für mehrere zusammenhängende Aussagen oder einen ganzen Text zu ermitteln [49]. Dies kann zum Beispiel im Bereich von *Social Media* Plattformen genutzt werden, um Kommentare unter Beiträgen zu filtern [29]. Durch die Klassifizierung können Kommentare, die eine negative Stimmung erzeugen, identifiziert und entfernt werden [29]. Ein weiteres mögliches Anwendungsgebiet sind Bewertungen von Apps [27], die ebenfalls gefiltert werden können. Dadurch lassen sich, aus der Betrachtung der negativen Bewertungen, die eigenen Schwächen herausfinden und Verbesserungspotential ableiten [29].

2.1.1 Software Engineering spezifische Sentimentanalyse

Sentimentanalyse wird auch im Bereich des *Software Engineerings* benutzt. Es gibt zahlreiche Datensätze von Aussagen mit zugeordneter Polarität aus diesem Bereich [33].
Ein Anwendungsgebiet im Software Engineering ist beispielsweise die Analyse von Aussagen auf Plattformen wie *GitHub* [33]. Jeweils eine positive, negative und neutrale Aussage aus einer solchen Betrachtung sind in der folgenden Tabelle 1 aufgeführt.

Aussage	Label
Yay for improving consistency, +1	positiv
OMG stupid me	negativ
final class?	neutral

Tabelle 1 Aussagen aus einem GitHub Datensatz mit dem zugehörigen Label [32]

In dem Datensatz wurden die Aussagen auf Grundlage von den festgestellten Emotionen gelabelt [33]. Ein neutrales Label wurde vergeben, wenn keine Emotion erkannt wurde. Durch die Wörter „*Yay*“ und „*improving*“ oder auch das „*+1*“ am Ende der ersten Aussage werden positive Emotionen wie freudig ausgelöst und führen dadurch zu einer positiven Stimmung[33].
In der zweiten Aussage werden durch „*OMG*“ als Abkürzung für „*Oh my God*“ und „*stupid*“ negative Emotionen (z.B. frustriert) ausgelöst und diese insgesamt negativ bewertet [33].
Die letzte Aussage wurde neutral bewertet, da es sich um eine Frage ohne erkennbare Wertung handelt [33].
Für den Bereich des Software Engineering gibt es jedoch einige Besonderheiten zu beachten. Aussagen können bei Verwendung normaler Sentimentanalyse-Tools zu fehlerhaften Analysen führen [17]. Für korrekte Analysen sollte die Domäne, in der die Daten entstanden sind, bei der Wahl des Analysetools beachtet werden [17]. In der Domäne Software Engineering gibt es einige Beispiele für Wörter, die in einer allgemeinen Betrachtung als negativ klassifiziert werden würden, in dem Kontext Software Engineering jedoch nicht zwingend negativ gemeint sein müssen [17]. „*Default*“[4] oder „*Error*“[5], welche ohne Domänenbetrachtung negativ bewertet werden würde, können durchaus auch positiv sein, da sie bei der Fehlerbehandlung helfen können und das Programm letztendlich sogar verbessern. Weitere solcher domänenspezifischer Beispiele sind „*Dead*“[6], „*Support*“[7] oder „*Resolve*“[8] [17].
Aus diesem Grund gibt es eigens für den Bereich des Software Engineerings entwickelte Sentimentanalyse-Tools. Eines dieser Tools ist *SentiStrengthSE* [17], welches wiederum eine Erweiterung des Tools *SentiStrength* [10] ist. Für die Bewertung von Aussagen gibt es bei diesem Tool ein Lexikon, das Wörtern eine Bewertung zuordnet, wie die Beispiele in Tabelle 2 zeigen.

Wort	Bewertung
fürchten	-4
genießen	+3
meckern	-3

Tabelle 2 Beispiele für die Bewertung von Wörtern in der deutschen Version von SentiStrength [10]

Die Bewertung kann dabei von -5 bis +5 erfolgen [56]. Die negativen Werte stellen eine negative Stimmung da, mit dem Höchstwert von -5 als Ausdruck der maximal möglichen Negativität eines Wortes.

[4-8] Md Rakibul Islam, Minhaz F. Zibran**.** *SentiStrength-SE: Exploiting Domain Specificityfor Improved Sentiment Analysis in Software Engineering Text.* s.l. : The Journal of Systems & Software, 2018. doi: https://doi.org/10.1016/j.jss.2018.08.030, S. 12

Die positiven Werte stellen in gleicher Art die positive Stimmung dar. Auch hier ist +5 die maximal mögliche Positivität.
Für die neutrale Aussage ist die 0 vorgesehen [56] Jede Aussage startet mit dem positiven Wert +1 und dem negativen Wert -1 [56]. Anhand der im Lexikon aufgelisteten Werte werden die einzelnen Wörter des Satzes dann analysiert und das jeweilige Maximum (positiv wie negativ) bestimmt [56]. Am Ende der Analyse wird dann die Summe aus beiden Maxima errechnet, welche die Stimmung des Satzes repräsentiert. Ein positiver Endwert bedeutet, dass der Satz insgesamt eine positive Stimmung erzeugt, bei einem negativen Wert liegt entsprechend eine negative Stimmung vor und bei einer 0 eine neutrale Stimmung [56]. Bei diesem Verfahren kann auch bei gefundenen Emotionen ein neutrales Label vergeben werden, da die Summe entscheidend ist [56]. In der Weiterentwicklung *SentiStrengthSE* wird diese Funktionsweise nicht verändert, sondern ein modifiziertes Lexikon verwendet, in dem die Werte für spezielle Wörter angepasst wurden [17].

2.1.2 Sentimentanalyseverfahren

Sentimentanalyse-Tools können unterschiedliche Funktionsweisen aufweisen. Eine Funktionsweise ist die Analyse einzelner Wörter einer Aussage hinsichtlich ihrer Polarität, was auch als linguistische Analyse bezeichnet wird [54]. Dabei kann dieses Verfahren nochmals in zwei unterschiedliche Ansätze zur Ermittlung der Polaritäten unterteilt werden [54].
Einerseits gibt es das lexikonbasierte Verfahren [54], welches auch im bereits beschriebenen *SentiStrength* Tool Anwendung findet [17]. Es existiert bei diesem Verfahren ein Lexikon, das als Grundlage für die spätere Analyse dient. Im Lexikon werden Polaritäten für einzelne Wörter definiert [54]. Die Wörter werden gemäß der im Lexikon stehenden Polarität bewertet und das Ergebnis ergibt sich aus der Summe der Einzelpolaritäten. Neben der bereits beschriebenen Skala (-5 bis +5) sind auch andere Wertebereich möglich, beispielsweise kann auf die Abstufung bei den positiven und negativen Werten verzichtet werden und lediglich mit +1 bzw. -1 gewertet werden [54].
Andererseits gibt es noch das keywordbasierte Verfahren [30]. Bei diesem Verfahren wird der Text oder die Aussage nach bestimmten Wörtern, den Keywords, durchsucht, die dann für die Polarität ausschlaggebend sind [30]. Ein Keyword muss sich nicht zwingend auf ein Wort beschränken, es kann auch eine ganze Aussage oder ein Emoticon sein [30]. Im Tool *Senti4SD* sind definierte Keywords beispielsweise „*excellent*“, „*:)*“ oder „*hate*“ [30].
Wörter wie „sehr“ oder „nicht“, sogenannte Verstärkungen und Negationen, können Aussagen verfälschen [17]. Bei linguistischen Verfahren kann die Polarität nachfolgender Wörter geändert werden, wenn eines dieser Wörter erkannt wird [17]. Als Beispielkonstrukt kann die Aussage „[…] sehr schlecht“ betrachtet werden mit den Polaritäten +1 für „sehr“ und -1 für „schlecht“. Betrachtet man rein die einzelnen Wörter ohne einen Zusammenhang zwischen den beiden, könnte kein klares Label vergeben werden, da sowohl positive als auch negative Emotionen gefunden worden sind. Eine Aussage über etwas, dass sehr schlecht ist, vermittelt jedoch eher eine klare negative Stimmung. Deswegen kann bei Erkennen des Wortes „sehr“ automatisch das nächste Wort mitbetrachtet werden und dieses als eine Verstärkung der Polarität wirken, von z.B. -1 auf -2 [17].

Ähnlich kann bei einer Negation vorgegangen werden, indem die Polarität des folgenden Wortes umgedreht wird [17].
Neben der linguistischen Analyse gibt es die Möglichkeit maschinelles Lernen einzusetzen [1]. Ein Beispiel für dieses Verfahren ist das Tool *SentiCR* [1], welches speziell dafür entwickelt wurde, Code Review Comments zu analysieren. Generell werden bei solchen Verfahren zunächst die Datensätze manuell analysiert und das Programm dann mithilfe dieses Datensatzes trainiert [1]. Im Falle von *SentiCR* wurden als Datenreview *comments* manuell gelabelt und dann als Datensatz genutzt. Die Qualität der Ergebnisse einer solchen Analyse hängen maßgeblich vom verwendeten Datensatz ab [1]. Die beiden Verfahren können auch kombiniert werden wie z.B. im bereits erwähnten Tool *Senti4SD*, in dem Sätze aus *Stack Overflow* als Datensatz verwendet wurden [30].
Die bis jetzt betrachteten Methoden sind alles automatische Analyseverfahren. Es ist jedoch auch möglich, eine solche Analyse manuell vorzunehmen, indem z.B. Entwickler gebeten werden, Aussagen ein Label zuzuordnen. Dadurch ist z.B. eine Kontrolle der automatisch ermittelten Ergebnisse möglich.

2.2 Stimmung

Um zu untersuchen, wodurch die wahrgenommene Stimmung eines Entwicklers beeinflusst werden kann, ist es wichtig zu verstehen, wie eine Stimmung überhaupt entsteht und wodurch diese beeinflusst werden kann.

Die Stimmung eines Individuums lässt sich grundsätzlich als ein Zustand definieren [64]. Diesen Zustand kennzeichnen Gefühle, die „als entweder angenehm oder unangenehm erlebt werden“[5] [64]. Die erlebten Gefühle sind bei Stimmungen von längerer Dauer und nur schwachem Ausmaß [64]. Eine Stimmung ist demnach ein längerfristiger Zustand mit schwach ausgeprägten Gefühlen. Weiterhin sind Stimmungen nicht gegen ein bestimmtes Objekt oder gegen eine konkrete Person gerichtet [64]. Eng verknüpft mit der aktuellen Stimmung sind die Bedürfnisse und die Verfassung des Individuums, weshalb die Erfüllung (oder die Nichterfüllung) von Bedürfnissen und die aktuelle Verfassung ausschlaggebende Faktoren für die Stimmungsbildung sind [64] [63].
Das *Core Affect Modell* von Russell beschreibt Stimmung als „*core affect with no Object*“[6] [42]. Der *Core Affect* ist ein Gefühl, welches durch Mischung von zwei Merkmalsausprägungen, Lust oder Unlust und schläfrig oder aktiv, entsteht [42]. Welches Gefühl konkret durch diese Mischung entsteht, lässt sich beispielsweise mittels eines *Circumplexmodells* ermitteln [43]. Ein solches Modell hat Russell für den *Core Affect* ebenfalls entwickelt [43]. Im *Circumplexmodell* werden zwei Dimensionen betrachtet aus deren Kombination sich das konkrete Gefühl ergibt [43].

[5] Markus Antonius Wirtz (Hrsg.): Dorsch – Lexikon der Psychologie, https://dorsch.hogrefe.com/stichwort/stimmung#search=38f9a176c65e63b7a752a83c7d01ee44&offset=0 vom 24.01.2022
[6] James A. Russell (2003): Core Affect and the Psychological Construction of Emotion, S. 147

Abbildung 1 Core Affect Modell nach Russell [7]

Abbildung 1 zeigt das Modell mit den Achsenausprägungen (*Activation und Deactivation, Pleasure und Displeasure*), die die Dimensionen darstellen. Zwischen den Achsen befinden sich Beispielemotionen, die sich durch Kombination ergeben (z.B. *Excited*) [43].

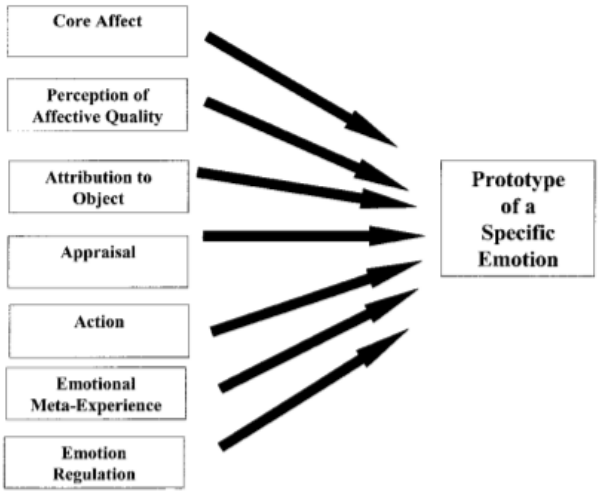


Abbildung 2 Entstehung einer Emotion nach Russell[8]

In Abbildung 2 wird gezeigt, welche Bausteine bei der Entstehung einer Emotion nach dem *Core Affect Modell* relevant sind [42]. Gemäß dem Modell ist der *Core Affect* ebenfalls ein Baustein der Emotionsentstehung. Stimmungen und Emotionen haben nach dem Modell somit (zumindest zum Teil) einen gemeinsamen Ursprung [42]. Weiterhin impliziert das Modell eine wechselseitige Beziehung zwischen Emotionen und Stimmung [63]. Emotionen können demnach die Stimmung beeinflussen [42] [63].

[7] James A. Russell (2003): Core Affect and the Psychological Construction of Emotion, S. 148
[8] James A. Russell (2003): Core Affect and the Psychological Construction of Emotion, S. 152

2.2.1 Emotionen und Gefühle

Oftmals werden Emotionen und Gefühle synonym verwendet [53]. Die beiden Begriffe lassen sich jedoch inhaltlich voneinander unterschieden [63]. Ein Gefühl kann dabei als der subjektive Teil von Emotionen angesehen werden und ist damit ein Teil der Emotion [63] [64]. Für Emotionen gibt es in der Literatur keine universell gültige Definition oder ein allgemeingültiges Verständnis. Aus einer umfangreichen Literaturrecherche [20] ging eine sehr umfassende Definition für den Emotionsbegriff hervor, welcher von Otto et al. [38] ins Deutsche übersetzt wurde:

> „Emotion ist ein komplexes Interaktionsgefüge subjektiver und objektiver Faktoren, das von neuronal/hormonalen Systemen vermittelt wird, die (a) affektive Erfahrungen, wie Gefühle der Erregung oder Lust/Unlust, bewirken können; (b) kognitive Prozesse, wie emotional relevante Wahrnehmungseffekte, Bewertungen, Klassifikationsprozesse, hervorrufen können; (c) ausgedehnte physiologische Anpassungen an die erregungsauslösenden Bedingungen in Gang setzen können; (d) zu Verhalten führen können, welches oft expressiv, zielgerichtet und adaptiv ist.“[9] [38]

Besonders interessant für diese Arbeit sind die Teile a und b. Teile c und d beschäftigen sich mit dem gezeigten Verhalten und physiologischen Prozessen, welche für dieses Thema keinen entscheidenden Einfluss haben, für den vollständigen Emotionsbegriff aber notwendig sind [38].

Aus Teil a lässt sich der Einfluss von Emotionen auf die Stimmung ablesen. „Gefühle wie Erregung oder Lust/ Unlust“ entsprechen dem *Core Affect* aus dem Modell von Russell und sind damit Bestandteil der Stimmung [42] [38]. Teil b beschreibt den Zusammenhang zwischen Emotionen und verschiedenen kognitiven Prozessen [38]. Teil davon sind auch Bewertungen und Klassifizierungsprozesse, die durch Emotionen verursacht werden. Da Gefühle ein Teil der Emotionen sind, haben auch diese einen Einfluss auf die Prozesse [38]. Darüber hinaus können wir aus dem *Core Affect* Modell schließen, dass die Stimmung ebenfalls einen Einfluss auf die Emotionen hat und damit auch auf die genannten Prozesse [42] [38].

2.2.2 Affekt

Als letzte Begriffsabgrenzung wird der Affekt betrachtet [64]. Ein Affekt ist ein sehr kurzer Zustand, der dafür jedoch besonders intensiv ist [64]. Als Beispiel hierfür dient der Schreck, der ein sehr intensives Gefühl auslöst, aber nur von kurzer Dauer ist [64]. Im Allgemeinen wird unter einem Affekt deswegen auch eine sehr starke Gemütsbewegung verstanden [64].

---

[9] Otto, J., Euler, H. & Mandl, H. (2000). Begriffsbestimmungen. In J. Otto, H. Euler & H. Mandl (Hrsg.), Emotionspsychologie. Ein Handbuch, S. 15

Da ein Affekt nur von sehr kurzer Dauer ist, ist er für die Betrachtung von Emotionen und Stimmungen nicht relevant und sollte deshalb von diesen losgelöst betrachtet bzw. bei der Analyse von Stimmungen und Emotionen nicht ausschlaggebend betrachtet werden [64].

2.2.3 Emotionsmodell

Um dem theoretischen Konstrukt einer Emotion nun konkrete einzelne Emotionen zuordnen zu können, werden z.B. Emotionsmodelle verwendet [60].

Es gibt verschiedene Ansätze für Emotionsmodelle. Unterschieden werden können Basisemotions-, Klassifikations- und dimensionale Modelle [60].

In einem Basisemotionsmodell wird davon ausgegangen, dass sogenannte Basisemotionen existieren [18]. Auszeichnen tun sich diese durch ihre bereits angeborene Existenz, wodurch sie nicht von externen Einflüssen beeinflussbar sind und bei jedem in gleicher Art und Weise vorhanden sind [18]. Die genaue Anzahl an Basisemotionen und welche es genau gibt kann von Modell zu Modell teils stark variieren [18] [40]. Exemplarisch dafür kann man die Modelle von Izard und Plutchik vergleichen [18] [40]. In den Modellen gibt es einmal acht und einmal zehn Basisemotionen, von denen jedoch gerade einmal drei und damit weniger als die Hälfte miteinander übereinstimmen [18] [40].

Klassifikationsmodelle sortieren Emotionen und bilden anhand dieser Sortierung übergeordnete Kategorien, denen diese Emotionen dann zugeordnet werden [60]. Ein Beispiel für ein solches Emotionsmodell ist das Emotionsmodell von Shaver et al. [49]. Für dieses haben Psychologiestudierende die Sortierung der Emotionen vorgenommen [49]. Daraus ergaben sich dann die Kategorien bzw. die übergeordneten Gefühle aus Abbildung 3.

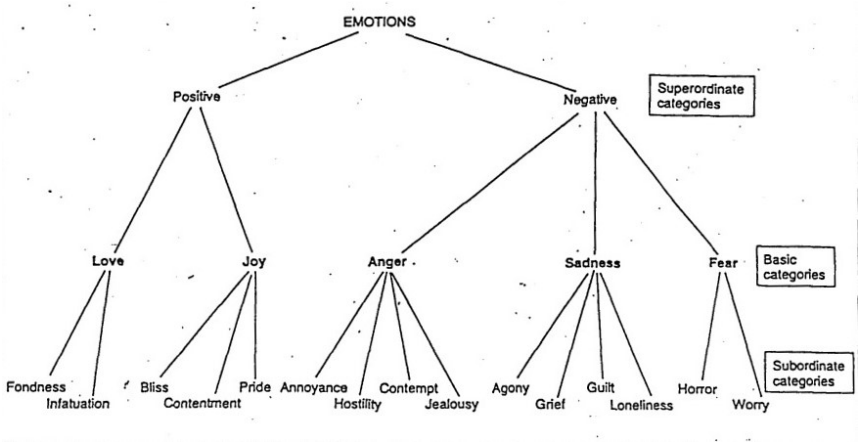


Abbildung 3 Emotionshierarchie nach Shaver et al. (vereinfachte Darstellung)[10]

[10] Fischer, Kurt W., Shaver, Philip R., Carnochan, Peter (1990): How Emotions Develop and How they Organise Development in Cognition and Emotion, S. 90

Auf der obersten Ebene wird nur zwischen positiven und negativen Emotionen unterschieden [49]. Ebene 2 unterscheidet zwischen den Oberkategorien der konkreten Emotionen (*Love, Joy, Anger, Sadness, Fear*) [49]. Konkrete Emotionen zu den jeweiligen Oberkategorien werden in der untersten Ebene 3 dargestellt [49]. Es ergibt sich durch diesen Aufbau ein hierarchischer Aufbau, der zunehmend konkreter wird und an dessen Ende konkrete Emotionen stehen [49]. Um dem Modell von Shaver et al. weitere Details hinzuzufügen, kann das Modell von Parrott [39] herangezogen werden. Dieses ist vom Aufbau der Hierarchien identisch mit dem Shaver Modell [39]. Parrott unterscheidet jedoch noch zwischen Primär-, Sekundär- und Tertiäremotionen, wobei die Primäremotionen mit denen von Ebene 2 des Shaver Modells übereinstimmen [2]. Anschließend werden diese Emotionen in der sekundären und tertiären Ebene weiter konkretisiert und sind somit noch einmal eine Ebene feiner als das Shaver Modell[11] [2] [39].

Ein dimensionales Emotionsmodell und deren Funktionsweise wurde im Abschnitt 2.2 bereits mit dem *Circumplexmodell* von Russell vorgestellt [43]. Generell sind dimensionale Modelle jedoch nicht auf zwei Dimensionen, wie im *Circumplexmodell* begrenzt, es können auch mehr Dimensionen verwendet werden [28]. Ein Modell, welches ebenfalls von Russell in Zusammenarbeit mit Mehrabian [28] entwickelt wurde, verwendet drei Dimensionen. Neben den beiden bereits bekannten Dimensionen aus dem ersten Modell, wird hier zusätzlich die Dimension „Dominanz“ eingeführt [28]. Generell lassen sich bei einigen Modellen bei der Auswahl der Dimensionen „Vergnügung“ und „Erregung“ Übereinstimmungen finden [28] [37] [62]. In Teilen werden diese zwar anders benannt, sind jedoch inhaltlich übereinstimmend. Weiteres Beispiel sind das Model von Osgood, Suci und Tannenbaum [37] oder das Model von Watson und Tellegen, die ebenfalls die beiden Dimensionen verwenden [62].

Im Folgenden wurde sich auf das Modell von Shaver et al. bezogen mit der Ergänzung des Modells von Parrott, da dieses sehr konkrete Emotionen liefert, welche genau definiert sind, wodurch sich bei einer späteren Abfrage mittels einer Skala eindeutig eine Schnittmenge bestimmen lässt. Darüber hinaus wurde das Modell auch in einem Datensatz als Grundlage für die Vergabe von Labels benutzt wurde (siehe 4.5).

[11] Siehe Anhang 1 Emotionsmodell nach Parrott

2.3 Stimmungsskalen

Um die Stimmung bestimmen und messen zu können werden unter anderem Stimmungsskalen verwendet [5].

Zur Messung eines allgemeinen Stimmungsbildes, im Sinne einer „überdauernde Stimmungslage“[12], kann die Stimmungsskala von Bohner & Schwarz [5] verwendet werden. Sie stellt die deutsche Version der „*mood survey*“ von Underwood und Froming [61] dar. Die Skala lässt sich in zwei Teilskalen aufteilen, eine die wie beschrieben die überdauernde Stimmungslage erfasst und eine, die die Reaktivität einer Person erfragt [5]. Reaktivität bezeichnet „die Intensität der Stimmungsveränderungen“[13]. Insgesamt hat die Skala 15 Aussagen, zu denen auf einer 7-stufigen Skala eine Bewertung abgegeben wird [5].

Nr.	Item	Polung	Sub-skala
1	Manchmal pendelt meine Stimmung mehrmals zwischen glücklich und traurig in einer einzigen Woche.	+	R
2	Ich fühle mich meist ziemlich fröhlich.	+	S
3	Meine Stimmung ist oft bedrückt.	-	S
4	Ich sehe im Allgemeinen mehr die Sonnenseiten des Lebens.	+	S
5	Verglichen mit meinen Freunden gehen meine Stimmungen weniger rauf und runter.	-	R
6	Ich bin selten in wirklicher Hochstimmung.	-	S
7	Manchmal schwankt meine Stimmung sehr schnell hin und her.	+	R
8	Ich fühle mich meist so, als ob ich vor Freude übersprudeln würde.	+	S
9	Meine Stimmungen sind sehr konsistent; sie ändern sich fast nie.	-	R
10	Ich halte mich für eine glückliche Person.	+	S
11	Verglichen mit meinen Freunden denke ich weniger optimistisch über das Leben.	-	S
12	Ich bin eine Person, die sich oft ändert.	+	R
13	Ich bin nicht so fröhlich wie die meisten Leute.	-	S
14	Ich bin weniger von meinen Stimmungen abhängig als die meisten Leute, die ich kenne.	-	R
15	Meine Freunde scheinen oft zu glauben, dass ich unglücklich bin.	-	S

Tabelle 3 Items der Stimmungsskala mit Polung und Sub-Skala[14]

[12] Bohner, G. & Schwarz, N. (2014). Die Stimmungsskala. Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS). https://doi.org10.6102/zis53, S. 3
[13] Bohner, G. & Schwarz, N. (2014). Die Stimmungsskala. Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS). https://doi.org10.6102/zis53, S. 3
[14] Bohner, G. & Schwarz, N. (2014). Die Stimmungsskala. Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS). https://doi.org10.6102/zis53, S. 2

Tabelle 3 zeigt die einzelnen Aussagen mit ihrer jeweiligen Polung und der zugehörigen Sub-skala, wobei „R“ für Reaktivität und „S“ für Stimmung steht. Items mit einer negativen Polung müssen bei der Auswertung der Skala umgepolt werden, damit ein richtiges Ergebnis berechnet werden kann [5]. Nach der Umpolung stehen höhere Werte für eine besser Stimmung bzw. stärkere Reaktivität [5].

Für die Erfassung der kurzfristigen Stimmung, z.B. der letzten 7 Tage, kann *PANAS* [7] (*Positive and Negative Affect Schedule*) bzw. die dazugehörige deutsche Version von Breyer und Bluemke [7] verwendet werden. *PANAS* umfasst 20 Adjektive, die sich in positive und negative Affekte sortieren lassen [7]. Auch hier existieren somit zwei Unterskalen. Bewertet werden die einzelnen Adjektive aus einer 5-Punkt-Skala [7]. Tabelle 4 zeigt die einzelnen Adjektive von *PANAS* und deren Zuordnung zu negativem (NA) und positivem (PA) Affekt.

Nr.	**Deutsch**	**Englisch**	**Dimension**
1	aktiv	active	PA
2	bekümmert	distressed	NA
3	interessiert	interested	PA
4	freudig erregt	excited	PA
5	verärgert	upset	NA
6	stark	strong	PA
7	schuldig	guilty	NA
8	erschrocken	scared	NA
9	feindselig	hostile	NA
10	angeregt	inspired	PA
11	stolz	proud	PA
12	gereizt	irritable	NA
13	begeistert	enthusiastic	PA
14	beschämt	ashamed	NA
15	wach	alert	PA
16	nervös	nervous	NA
17	entschlossen	determined	PA
18	aufmerksam	attentive	PA
19	durcheinander	jittery	NA
20	ängstlich	afraid	NA

Tabelle 4 Items von PANAS und die dazugehörige Dimension[15]

Für die Auswertung der Skala werden die einzelnen Adjektive entsprechend ihrer Dimension geordnet und die abgegebenen Bewertungen addiert [7].

[15] Breyer, B. & Bluemke, M. (2016). Deutsche Version der Positive and Negative Affect Schedule PANAS (GESIS Panel). Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS). https://doi.org/10.6102/zis242

Kapitel 3

Verwandte Arbeiten

Mit dem Einsatz von Sentimentanalyse im Bereich des Software Engineerings haben sich bereits viele Wissenschaftler beschäftigt. In einer umfangreichen Literaturanalyse wurden Probleme der Stimmungsanalyse im Bereich des Software Engineerings aufgezeigt [21]. Ein häufig gefundenes Problem entsteht aus der Subjektivität der Labelvergabe, die dazu führen kann, dass eine Aussage unterschiedlich gelabelt werden kann [21].

Mit dem Einfluss von Emotionen im Software Engineering Bereich beschäftigten sich bereits Novielli et al. [34]. Diese untersuchten den Zusammenhang zwischen Fragen im Bereich der *Q&A-Seiten* von *Stack Overflow* und Emotionen. Nach ihren Ergebnissen haben die Emotionen, die in einer Frage erkennbar sind, einen Einfluss auf deren Beantwortung [34].

Eine weitere Analyse zum Einfluss von Emotionen im Bereich des Software Engineerings untersucht den Einfluss auf *Commit* Nachrichten [36] [14]. Dabei wurden positive Zusammenhänge zwischen der Problembehebung und positiven Emotionen gefunden [14]. Weiterhin konnte festgestellt werden, dass *Commit* Nachrichten an Montagen und in *Java* mehr negative Emotionen enthalten, während bei globalen Teams mehr positive Emotionen in *Commit* Nachrichten existieren [36].

Graziotin [13] beschäftigte sich mit den Auswirkungen sozialer Aspekte von Entwicklern auf die Produktivität. Dabei wurden auch die Gründe für z.B. unglückliche Entwickler untersucht.

Im Bereich Sentimentanalyse in Entwicklerteams haben Tourani et al. [58] untersucht, welche Art von Sentiments in Maillisten gefunden werden können [58]. Die Maillisten stammen aus *Open-Source* Softwareprojekten. Tourani et al. fanden dabei heraus, dass positive und negative Sentiments in spezifischen Themen auftreten, jedoch nur sehr selten im gleichen Thema [58].

Auch im Bereich der Stimmung in studentischen Softwareprojekten wurde bereits geforscht [46]. Schneider et al. [46] untersuchten, wie sich Konflikte im Laufe eines Softwareprojekts entwickeln und stellten fest, dass soziale Konflikte stetig anstiegen, während arbeitsbezogene Konflikte anfangs zunahmen und zum Ende wieder abnahmen.

Schiller [45] hat ebenfalls im Rahmen von Softwareprojekten Zusammenhänge zwischen Stimmung und Interaktionen in Meetings untersucht. Interaktionen wurden Polaritätsklassen zugeordnet und mittels dieser Information auf Zusammenhänge untersucht [45].

Neben den in dieser Arbeit verwendeten Datensätzen existieren noch weitere [31]. Novielli et al. [31] haben für *Stack Overflow* einen weiteren Datensatz erstellt, der 4.800 Sätze mit zugehörigem Label enthält [31]. Neben *GitHub*- und *Stack Overflow* Datensätzen können auch solche von der Plattform *Jira* verwendet werden [35]. Ortu et al. [35] haben aus einem *Jira* Projekt einen Datensatz mit circa 6.000 Kommentaren erstellt.

Kapitel 4

Umfrage

In Kapitel 4 werden die Elemente der Umfrage erklärt und anschließend die statistischen Methoden zur Auswertung erläutert.

4.1 Teilnehmende der Umfrage

Die Umfrage wurde im Rahmen der Universitätsveranstaltung *Software-Projekt* des Fachgebiets Software Engineering der Leibniz Universität Hannover im Wintersemester 2021/2022 durchgeführt [11]. Im Regelstudienplan ist diese Veranstaltung für das 5. Semester im Bachelor Informatik vorgesehen [11]. Ziel der Veranstaltung ist es, für externe oder interne Kunden ein Produkt in Form einer Software zu erstellen. Externe Kunden sind Vertreter realer Unternehmen. Interne Kunden repräsentieren universitätsinterne Projekte, verhalten sich in der Kommunikation mit den Studierenden jedoch wie externe Kunden.

Die Studierenden arbeiten in Projektgruppen und setzen Prinzipien des Software-Engineerings (z.B. regelmäßige *SCRUM-Meetings*) um [11]. Zur Teilnahme am Projekt müssen die Studierenden bestimmte Module aus dem Studium bereits bestanden haben. Sie müssen zum einen Softwaretechnik oder Softwarequalität, zum anderen Programmieren 1, Programmieren 2 oder Programmierprojekt bestanden haben [11]. Jeder Teilnehmende hat somit bereits Programmiererfahrung und kennt Prinzipien des Software Engineerings.

Befragt wurden die Studierenden zu verschiedenen Zeitpunkten im Verlauf der Veranstaltung. Insgesamt wurde die Umfrage viermal wiederholt, um unterschiedliche Phasen eines Projektes abzufragen. Zunächst wurde eine eher stressige Phase mit dem Ende der ersten Iteration abgefragt. Die zweite Umfrage wurde zur Mitte der zweiten Iteration durchgeführt, welches eine etwas ruhigere Phase darstellt. In einer noch ruhigeren Projektphase befanden sich die Studierenden zum Zeitpunkt der dritten Umfrage nach der Weihnachtsunterbrechung. Nach den Abnahmetestfällen wurde die vierte Umfrage durchgeführt. Diese Phase kann entweder sehr stressig durch gescheiterte Abnahmetestfälle oder eher ruhig und erleichternd nach einer erfolgreichen Abnahme sein. Da in der Umfrage immer die letzten sieben Tage berücksichtigt wurden, kann der genaue Stichtag individuell abweichen.

Durch eine gesetzte Frist von sieben Tagen zum Ausfüllen der Umfrage ist jedoch die avisierte Phase immer zum Teil in den Umfragewerten mit vertreten.

4.2 Aufbau

Die Umfrage wurde mithilfe von *LimeSurvey*[16] erstellt. Neben vorgefertigten Tools bietet *LimeSurvey* auch die Möglichkeit, die Resultate aus der Umfrage herunterzuladen.

Strukturiert wird die Umfrage in *LimeSurvey* in verschiedene Fragegruppen, denen dann einzelne Fragen zugeordnet werden. Der Umfrage vorgeschaltet ist ein Willkommenstext[17], in dem erklärt wird, warum die Umfrage durchgeführt wird und welche Fragen in der Umfrage zu erwarten sind.

Die einzelnen Fragegruppen lauten wie folgt:

1. Allgemeine Daten
2. Stimmungsanalyse (allgemein)
3. Stimmungsanalyse (letzte 7 Tage)
4. Lebenssituation
5. Labelvergabe (Labeln von 10 Sätzen)
6. Labelvergabe (Labeln von 10 Sätzen)
7. Labelvergabe (Labeln von 10 Sätzen)
8. Labelvergabe (Gründe für die Labelvergabe)
9. Gruppendynamik

Punkt 2 wurde in der zweiten Umfrage entfernt, da sich die allgemeine Stimmung in der Regel nicht ändern wird. In späteren Umfragen wurde abgefragt, ob bereits an einer Umfrage teilgenommen wurde. Falls ja, wurde dieser Teil übersprungen.

In der Umfrage wurde bei allen Skalen auf Konsistenz geachtet. Ein niedriger Wert entsprach „total unzufrieden“, „sehr unsicher“ usw., während die höchsten Werte „sehr zufrieden“, „sehr sicher“ usw. entsprachen.

4.3 Auswahl der Fragen

Bei den allgemeinen Daten wurde erfragt, ob der Umfrageteilnehmer Teil eines externen oder internen Projektes ist. Da ein externes Projekt näher an einem richtigen Softwareprojekt ist als ein internes Projekt, wurde dieses Merkmal zur späteren Unterscheidung abgefragt.

Um die Daten bei mehrfacher Umfrageteilnahme einer Person zuzuordnen, wurden die Teilnehmer gebeten, einen Code zu erzeugen.

[16] https://www.limesurvey.org/de
[17] Siehe Anhang 2

Damit keine Rückschlüsse auf persönliche Daten möglich sind, wird der Code aus:

a) dem Tag des Geburtstages,
b) dem ersten Buchstaben des Vornamens,
c) dem zweiten Buchstaben des Nachnamens und
d) dem ersten Buchstaben der Straße der Wohnadresse

erzeugt. Für Max Mustermann, geboren am 01.02.1995, aus der Straße „Am Musterweg“, wäre der Code *01MuA*. Da die zu bewertenden Sätze auf Englisch sind, wurde im ersten Punkt der Umfrage auch das Englischniveau über eine Skala von 1-5 abgefragt. Diese wird angezeigt, wenn Englisch keine Muttersprache des Teilnehmers ist. Ab der dritten Umfrage wurde dann noch zusätzlich gefragt, ob bereits eine Umfrage zuvor ausgefüllt worden ist. Wurde bereits eine Umfrage ausgefüllt, wurden die allgemeinen Stimmungsdaten den Teilnehmern zugeordnet.

Punkt 8 fragt die Teilnehmer nach dem ausschlaggebenden Grund für die Labelvergabe, ob Unsicherheiten bei der Vergabe aufkamen und aus welchem Grund.[18]

4.4 Auswahl der Skalen

In den Fragegruppen 2, 3, 4 und 9 wurden Skalen zur Messung der jeweiligen Größen verwendet.

In Fragegruppe 2 soll nach der allgemeinen Stimmung gefragt werden. Um diese zu messen, wurde die Stimmungsskala von Bohner & Schwarz [5] verwendet. Da die Stimmungsskala bereits eine validierte Skala ist, die gut die gewünschte Stimmung und darüber hinaus noch die Reaktivität erfasst, wurde diese gewählt [5].

Für die Erfassung der Stimmung der letzten 7 Tage wurde *PANAS* verwendet bzw. die deutsche Version von *PANAS* [7]. Um die Stimmung zu erfassen, und nicht den Affekt im Moment der Umfrage, wurden die Teilnehmer gebeten, die letzten 7 Tage bei ihrer Bewertung zu betrachten. Es wurde zusätzlich die Hilfe gegeben, die gesamte Lebenssituation (z.B. Hobbys oder Familie) in die Bewertung einzubeziehen, da einige Adjektive recht abstrakt sind und eventuell zu Problemen bei der erstmaligen Bewertung führen könnten. Aufgrund seiner vielfachen Verwendung bei der Erfassung von Emotionen wurde *PANAS* gewählt. Auch *PANAS* ist bereits validiert [7].Bei der Auswahl von *PANAS* wurde das Emotionsmodell von Shaver et al. [49] mit der Ergänzung des Modells von Parrott [39] berücksichtigt.

[18] Siehe Anhang 3

Ein Abgleich mit den konkreten tertiären Emotionen ergab, dass die Oberkategorien des vorgestellten Emotionsmodell mindestens einmal in *PANAS* vertreten sind und diese alle abgedeckt sind.

Punkt 4 erfragt die Zufriedenheit mit der aktuellen Lebenssituation. Dafür wurden die Aspekte:

- Studium
- Arbeit
- Sicherheit
- Wohnsituation
- Selbstbestimmtheit
- Gesundheit
- Finanzen
- Selbstverwirklichung und
- Soziales Umfeld

mit einer 5-Punkt-Skala abgefragt. Zu jedem der Aspekte wurden Beispiele genannt, um Unsicherheiten vorzubeugen. Für z.B. Sicherheit waren diese „Sicherheitsgefühl beim Verlassen des Hauses, soziale Sicherheit“ oder für Finanzen „Finanzielle Unabhängigkeit, monatliche Einnahmen verglichen mit den Ausgaben“. Diese Kriterien wurden aus einer Umfrage des Statistischen Bundesamtes zur Lebensqualität[19], ein Bericht der Bundesregierung zur Lebensqualität[20] und einem Beitrag zur Gesundheit der Bundeszentrale für gesundheitliche Aufklärung[21], zusammengestellt [15] [8] [4]. Da diese Skala jedoch noch nicht validiert ist, wurde aus den Gesamtdaten das *Cronbachs Alpha* zur Validierung berechnet [22]. Mit Cronbachs Alpha wird die Korrelation zwischen den Items einer Skala berechnet und dadurch ermittelt, wie gut diese Items zusammen ein Merkmal messen [22]. Zur Berechnung wird die Summe der Varianzen der einzelnen Items, die Varianz der Summe aller Itemsummen jedes Teilnehmers und die Anzahl der Fragen benötigt [22]. Werte größer als 0,7 sind als akzeptabel zu bewerten [22].

[19] Bertelsmann Stiftung (2010), Statista GmbH (Hrsg.): Was ist Ihnen für Ihre Lebensqualität wichtig?, URL: https://de.statista.com/statistik/daten/studie/163877/umfrage/umfrage-wichtige-faktoren-fuer-die-lebensqualitaet/#professional vom 24.01.2022
[20] Presse- und Informationsamt der Bundesregierung (Hrsg.): Regierungsbericht zur Lebensqualität, URL: https://www.gut-leben-in-deutschland.de vom 24.01.2022
[21] Klaus Hurrelmann , Matthias Richter (2018): Determinanten von Gesundheit, doi: 10.17623/BZGA:224-i008-1.0, URL: https://leitbegriffe.bzga.de/alphabetisches-verzeichnis/determinanten-von-gesundheit/ vom 24.01.2022

$$\alpha = \frac{N}{N-1} * \left(1 - \frac{\sum_{i=1}^{N} {S^2}_{Yi}}{{S^2}_X}\right)$$

$$N = Anzahl\ der\ Fragen$$
$${S^2}_{Yi} = Varianz\ der\ Frage\ i$$
$${S^2}_X = Varianz\ der\ Summe\ aller\ Itemsummen$$

Aus den Werten der Umfragen ergibt sich ein Wert von Cronbachs Alpha von 0,78, welcher als akzeptabel anzunehmen ist. Die Skala kann zur Messung genutzt werden[22].

In der letzten Fragengruppe der Umfrage wurde nach möglichen Konflikten innerhalb der Softwareteams gefragt. Dafür wurde die deutsche Übersetzung der *Intragroup Conflict Scale* [26] verwendet.

Item Nr.	Englisch	Deutsch
rc1	How much friction is there among members in your work unit?	Wie viele Reibereien gibt es zwischen den Gruppenmitgliedern?
rc2	How much are personality conflicts evident in your work unit?	Wie offensichtlich sind persönliche Konflikte in der Gruppe?
rc3	How much tension is there among members in your work unit?	Wie viele Spannungen gibt es zwischen den Gruppenmitgliedern?
rc4	How much emotional conflict is there among members in your work unit?	Wie viele emotionale Konflikte gibt es zwischen den Gruppenmitgliedern?
tc5	How often do people in your work unit disagree about opinions regarding the work being done?	Wie oft sind sich die Gruppenmitglieder uneinig, wie die Arbeit zu erledigen ist?
tc6	How frequently are there conflicts about ideas in your work unit?	Wie häufig gibt es Ideenkonflikte in der Gruppe?
tc7	How much conflict about the work you do is there in your work unit?	Wie viele die Arbeit betreffende Konflikte gibt es in der Gruppe?
tc8	To what extent are there differences of opinion in your work unit?	In welchem Ausmaß gibt es Meinungsverschiedenheiten in der Gruppe?

Tabelle 5 Intragroup Conflict Scale mit deutscher Übersetzung [26]

Tabelle 5 zeigt die einzelnen Fragen der Skala inklusive Zuordnung zu zwei Subskalen. Dabei steht *rc* für *Relationship Conflict* und *tc* für *Task Conflict* [26].

[22] Siehe Anhang 4

Es kann dadurch nach dem Grund für die Konflikte unterschieden werden. Bewertet werden die einzelnen Fragen auf einer 6-Punkt-Skala [26].

Tabelle 6 zeigt die Beschreibungstexte der einzelnen Skalen aus der Umfrage und die jeweiligen Pole der Skala im Überblick.

Skala	Beschreibung	Pole
Stimmungsskala	Bitte gib an, in welchem Maße die folgenden Sätze im Allgemeinen auf dich zutreffen.	1 = gar nicht zutreffend 7 = äußerst zutreffend
PANAS	Bitte gib hier an, in welchem Maße die aufgelisteten Adjektive auf dich innerhalb der letzten Woche (7 Tage) zutreffend waren. Bezieh dabei deine gesamte aktuelle Lebenssituation mit ein (z.B. Freunde, Familie, Hobbys, Arbeit, Studium...).	1 = gar nicht zutreffend 5 = äußerst zutreffend
Bewertung der Lebensumstände	Bitte gib an, wie zufrieden du aktuell mit deiner Lebenssituation in den aufgelisteten Bereichen bist.	1 = total unzurieden 5 = sehr zufrieden
Intragroup Conflict Scale	Bitte schätze ein, wie oft folgende Situationen in euer Gruppe auftreten.	1 = nie 6 = sehr oft

Tabelle 6 Skalenbeschreibungen und Pole im Überblick

4.5 Auswahl der Sätze zum Labeln

Für die Sätze, die in der Umfrage mit Labeln zu versehen sind, wurden Datensätze von *GitHub*[23] und *Stack Overflow*[24] benutzt. In der *SentiSurvey* des Fachgebiets Software Engineering wurden insgesamt 100 Sätze verwendet, für diese Umfrage wurde eine Teilmenge von 30 Sätzen verwendet, welche je zur Hälfte aus *GitHub* und *Stack Overflow* stammen. Für jeden der Sätze aus diesen Daten existiert bereits ein zugeordnetes Label. Diese wurden entweder durch eine ad hoc (*Stack Overflow*) Bewertung [27] oder unter Verwendung von Guidelines (*GitHub*) [33] ermittelt. Bei einer *ad hoc* Bewertung erfolgt die Bewertung nur anhand der wahrgenommenen Stimmung beim Lesen der Sätze [27]. Werden *Guidelines* verwendet, erfolgt die Bewertung unter Berücksichtigung eines Emotionsmodells [33].

[23] https://figshare.com/articles/dataset/A_gold_standard_for_polarity_of_emotions_of_software_developers_in_GitHub/11604597

[24] https://sentiment-se.github.io/replication.zip

Die zugeordneten Emotionen können dann den Labeln positiv, negativ oder neutral zugeordnet werden. Das zugrunde liegende Emotionsmodell war das Modell von Shaver et al. [33].

Bei der Auswahl der Sätze wurde auf eine gleichmäßige Verteilung zwischen den Labeln geachtet, sodass aus jedem Datensatz 5 positive, 5 neutrale und 5 negative ausgewählt wurden. Die Auswahl geschah zufällig über eine Excel-Tabelle, welche Zufallszahlen erzeugt, anhand derer die Sätze ausgewählt worden sind. Tabelle 7 und Tabelle 8 zeigen die ausgewählten Sätze für diese Umfrage. Dazu sind diese in positiv, negativ und neutral unterteilt. Zuerst wird die erzeugte Zufallszahl ausgegeben, anschließend die Orginal-ID im Datensatz und zum Schluss der Satz.

Positiv			
	Zufalls ID		
	14	6184	Very good example of steady pooling readHere.
	5	2231	It does its job, but was built for our use case. The following code therefore, might be (as I have not attempted using it) a better one, if you are willing to use
	2	5626	Mojarra specific classes.
	12	5924	(Hopefully with a good example.)
	1	6238	Now we're getting to the good part.
Negativ			
	Zufalls ID		
	7	3382	The following is the error I keep getting.
	15	3853	I don't know what else to do to make things to work. We ran into the same sort of problem with Flex and
	5	5310	JPA/Hibernate. The data structure you are saving your data is not very
	4	4596	optimal for the days with daylight saving time. I have looked and found that there are some packages that would automatically enter my keyring on login but
	6	5655	that isn't really an option.
Neutral			
	Zufalls ID		
	7	238	Hope this helps.
	2	5078	Here's an example of a JSON structure.
	8	3856	It dependes where it is exactly located. Let me know if there are any other details that might
	12	3368	help!
	15	1284	which someone converted to a JSON string as follows.

Tabelle 7 Stack Overflow Sätze mit der erzeugten Zufallszahl, Orginal ID und dem Satz, geordnet nach Polaritäten

Positiv		
Zufalls ID	Orginal ID	Satz
16	1608237	If I still wasn't explicit enough: 9cc49763934a2ec3016b5bb882c3b8efd2265376 I can either submit a pull request or just forget it :)"
4	364658	@timmywil Sounds good!"
15	763921	Most awesome! :+1:"
14	3573171	Yay for improving consistency, +1"
11	148730	lol :)"
Negativ		
Zufalls ID		
12	251282	i was afraid it'd do unimaginable things!"
14	261781	OMG stupid me"
3	4089417	oh nice find, that's been bugging the crap out of me" This is the only one that bothers me. If an old compilers fails to optimize a static `strlen`, that's ok, the code will run a bit slower... But if an old compiler fails to optimize this one, this will just not compile. We need to change
1	103030	this to a numeric "
11	5491	Why was this reverted? :("
Neutral		
Zufalls ID		
16	284725	final class?" That's what obfuscation does. We do our best to unobfuscate but only for class names. This is a mojang
9	250794	thing, not us."
1	287899	md5 not good enough for you?" I ended up using `strtotime()` as the various formats that I saw in the wild and in the specs resulted in icky code. However, strtotime() seems to be able to sort
5	2676081	everything out."
3	5700222	Is this good to go then?"

Tabelle 8 GitHub Sätze mit der erzeugten Zufallszahl, Orginal ID und dem Satz, geordnet nach Polaritäten

4.6 Analyse der Umfrage

4.6.1 Skalenniveaus

Zur Auswahl der statistischen Methoden müssen zunächst die Skalenniveaus bestimmt werden, da nicht alle Methoden für alle Skalenniveaus anwendbar sind [48].

Unterschieden werden Nominal-, Ordinal- und Kardinalskala [48].

Bei einer Nominalskala kann zwischen den Daten unterschieden werden. Alle Werte, die angenommen werden können, sind gleichberechtigt und dienen nur der Unterscheidung von Werten [48].

Ordinalskalen werden auch als Rangskalen bezeichnet, woraus sich auch der wesentliche Unterscheid zur Nominalskala ergibt [48]. Den Daten können Ränge zugeordnet werden, wodurch sich eine Reihenfolge ergibt. Abstände zwischen zwei Datenpunkten können jedoch nicht verglichen werden.

Daten auf einer Kardinalskala, auch Intervallskala genannt, können darüber hinaus noch über Intervalle oder Verhältnisse verglichen werden [48].

Charakteristisch für diese Skala ist darüber hinaus die Existenz eines Nullpunktes [48]. Daten die aus solchen Skalen stammen werden metrische Daten genannt [48].

Bei den vorliegenden Skalen handelt es sich fast ausschließlich um Ordinalskalen. Die einzelnen Items der Skalen oder auch deren Gesamtergebnis können in eine Reihenfolge gebracht werden, es ist jedoch nicht möglich zu sagen, dass eine Bewertung mit 4 doppelt so gut ist wie eine Bewertung mit 2. Es kann auch nicht davon ausgegangen werden, dass alle Studierende bei der Bewertung die Unterschiede zwischen den Bewertungsschritten als gleich bewertet haben. Bei den verwendeten Skalen gibt es ebenfalls keinen Nullpunkt, da die Bewertung stets bei 1 startet.

Aus den Daten können jedoch beispielsweise eine Trefferquote oder die Anzahl der Treffer innerhalb der positiven Aussagen errechnet werden. Diese neu errechneten Daten sind metrisch.

Für eine allgemeine Übersicht der Daten wurden zunächst für die gesamten Daten Lageparameter und Streuungsmaße angegeben. Zur Übersicht über die einzelnen Umfragen wurden die Gesamtdaten anschließend noch einmal auf die einzelnen Umfragen aufgeteilt.

Nach dieser allgemeinen Übersicht über die gesammelten Daten wird dann spezifisch für die einzelnen Forschungsfragen mit Zusammenhangsmaßen auf diese eingegangen.

4.6.2 Lageparameter

Lageparameter werden verwendet „um die zentrale Lage bzw. den Mittelpunkt einer Verteilung näher zu beschreiben“[25]. Zu den Lageparametern gehören das arithmetische Mittel, der Median und der Modus [52].

Das arithmetische Mittel wird oft auch als Mittelwert bezeichnet [48]. Berechnet wird das arithmetische Mittel durch Bilden der Summe über alle Datenpunkte und anschließender Division der Summe durch die Anzahl an Datenpunkten [48]. Das arithmetische Mittel kann nur für mindestens metrisch skalierte Daten berechnet werden [48].

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$\bar{x}$ = arithmetisches Mittel
x_i=Datenpunkte
N=Anzahl an Werte

[25] Statista GmbH; Lexikon – Definition Lageparameter; URL: https://de.statista.com/statistik/lexikon/definition/80/lageparameter/ vom 30.01.2022

Zum arithmetischen Mittel lässt sich zusätzlich der Standardfehler berechnen. Dieser drückt den Unterscheid zwischen dem arithmetischen Mittel, der Stichprobe und der Grundgesamtheit aus [48]. Zur Berechnung wird die Standardabweichung durch die Wurzel der Stichprobengröße dividiert [48]. Mittels dieser Berechnung ergibt sich der geschätzte Standardfehler. Da Mittelwert und Standardabweichung der Gesamtheit nicht bekannt sind, muss dieser geschätzt werden [48].

$$\hat{\sigma}_{\bar{x}} = \frac{s_x}{\sqrt{n}}$$

$\hat{\sigma}_{\bar{x}}$=geschätzter Standardfehler
s_x=Standardabweichung
n=Anzahl Werte

Der Median wird über eine nach Größe geordnete Datenreihe gebildet [48]. Ausgewählt wird der Mittelpunkt der Reihe, welcher der Median ist. Für eine Berechnung müssen mindestens ordinalskalierte Daten vorliegen [48].

Als Modus wird der Wert bezeichnet, der innerhalb einer Datenreihe am häufigsten vertreten ist [48]. Der Modus kann für mindestens nominalskalierte Daten berechnet werden [48].

4.6.3 Streumaße

Streumaße geben „Auskunft über den Verlauf der Daten […] rechts und links des Mittelpunkts“[26]. Unterschieden werden Varianz, Standardabweichung und Spannweite [52].

Die Varianz zeigt, wie sehr die Werte einer Datenreihe im Mittel quadratisch um das arithmetische Mittel gestreut sind [48]. Da für die Berechnung das arithmetische Mittel notwendig ist, müssen metrisch skalierte Daten vorliegen [48].

$$S^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{X})^2$$

S^2=Varianz
N=Anzahl Werte
x_i=Wert
$\bar{X}$=arithmetisches Mittel

Von jedem Beobachtungswert wird das arithmetische Mittel abgezogen und anschließend das Ergebnis quadriert [48].

[26] Statista GmbH; Lexikon – Definition Lageparameter; URL: https://de.statista.com/statistik/lexikon/definition/80/lageparameter/ vom 30.01.2022

Dies wird für alle Beobachtungswerte durchgeführt und anschließend die Summe aller Ergebnisse gebildet. Wurde die Summe berechnet, wird diese durch die Stichprobengröße geteilt [48].

Die Standardabweichung beschreibt die durchschnittliche Streuung der Werte um das arithmetische Mittel [48]. Wie bei der Varianz müssen metrische Daten vorliegen [48].

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2} \quad oder\ s = \sqrt[+]{s^2}$$

s=Standardabweichung
N=Anzahl Werte
x_i=Wert
$\bar{X}$=arithemtisches Mittel
s^2=Varianz

Die Spannweite ist der Unterschied zwischen dem größten und dem kleinsten Wert einer Datenreihe [6]. Die Datenreihe muss mindestens ordinalskalierte Daten enthalten [48]. Die Berechnung ergibt sich aus der Definition, indem vom Maximum das Minimum abgezogen wird.

$$R = x_{max} - x_{min}$$

R=Spannweite
x_{max}=Maximum
x_{min}=Minimum

4.6.4 Zusammenhangsmaße

Korrelationen werden berechnet, um Zusammenhänge zwischen Daten zu ermitteln [51]. Dabei wird ein Korrelationskoeffizient ermittelt, der die Stärke des Zusammenhangs angibt [51]. Der Koeffizient kann zwischen -1 und +1 liegen. Bei einem Korrelationskoeffizient von null liegt kein Zusammenhang vor. Werte größer als null zeigen positive Korrelation an, woraus sich schließen lässt, dass wenn eine Variable steigt, die andere Variable ebenfalls steigt. Werte kleiner als null zeigen negative Korrelation an, wenn eine Variable steigt, sinkt die andere Variable [51].

Da ordinalskalierte Daten verglichen werden, muss ein Rangkorrelationskoeffizient berechnet werden, bei dem nicht die konkreten Werte verglichen werden, sondern deren Rang [51]. Dieser spezielle Koeffizient wird Spreaman´sche Rangkorrelation genannt [51].

$$r_s = 1 - \frac{6 \sum D^2}{N(N^2 - 1)}$$

r_s=Korrelationskoeffizient
D=Differenz der Ränge der Werte
N=Anzahl an Werten

Für die Interpretation des Korrelationskoeffizienten wird folgende Skala nach Cohen [9] genutzt.

0,5 < r < 1	starker Zusammenhang
0,3 < r < 0,49	mittlerer Zusammenhang
0,1 < r < 0,29	schwacher Zusammenhang
r <=0,09	kein Zusammenhang

Tabelle 9 Korrelationsinterpretation [9]

Um zu entscheiden, ob eine Korrelation signifikant ist, werden zu den Korrelationskoeffizienten die p-Werte berechnet [41]. Liegen diese unter einem festgelegten Signifikanzniveau, ist die Korrelation statistisch signifikant [51]. Das Signifikanzniveau gibt die höchste Wahrscheinlichkeit an, mit der eine eigentlich korrekte These abgelehnt wird [51]. In diesem Fall ist die These die Korrelation, die abgelehnt werden würde, obwohl sie korrekt ist. Typische Signifikanzniveaus sind 0,05 oder 0,01 [51]. Für diese Arbeit wurde ein Signifikanzniveau von 0,05 festgelegt. Zur Berechnung der p-Wert müssen zunächst t-Werte berechnet werden [41].

$$t = r_s \frac{\sqrt{N-1}}{\sqrt{1 - {r_s}^2}}$$

t=t-Wert
r_s=Korrelationskoeffizient
N=Anzahl der Werte

Die resultierenden t-Werte können mit der Microsoft Excel Funktion T.VERT.2S[27] in p-Werte umgerechnet werden. Dies erspart das Nachschlagen in Tabellen für kritische Werte. Liegt der p-Wert unter dem Signifikanzniveau von 0,05, gilt die Korrelation als statistisch signifikant [51].

[27] https://support.microsoft.com/de-de/office/t-vert-2s-funktion-198e9340-e360-4230-bd21-f52f22ff5c28

Neben der Korrelation wird noch der Fleiss´ Kappa Wert betrachtet [44]. Dieser wird verwendet, um die Übereinstimmung zwischen Beurteilern zu messen. Berechnet werden kann Fleiss´ Kappa ab zwei oder mehr Beurteilern [44].

$$k = \frac{\bar{P} - \bar{P_e}}{1 - \bar{P_e}}$$

$$mit$$

$$\bar{P} = \frac{1}{Nn(n-1)} \sum_{i=1}^{N} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1)$$

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

N=Anzahl der Daten
n=Anzahl Bewertungen pro Bewerter
k=Anzahl Antwortmöglichkeiten

Zur Beurteilung des Kappa Wertes wird die folgende Skala verwendet [44].

k<0	Keine Übereinstimmung
0≤k<0,2	Leichte Übereinstimmung
0,2≤k<0,4	Ausreichende Übereinstimung
0,4≤k<0,6	Moderate Übereinstimmung
0,6≤k<0,8	Beachtliche Übereinstimmung
0,8≤k<1,0	(fast) perfekte Übereinstimmung

Tabelle 10 Interpretation des Fleiss´ Kappa Wertes [44]

Kapitel 5

Ergebnisse

In diesem Kapitel werden die Ergebnisse der Umfrage mit den in Punkt 4.6 beschriebenen Methoden vorgestellt und Zusammenhänge analysiert. Zunächst werden die Daten einzeln ausgewertet und anschließend Zusammenhänge zwischen Daten betrachtet.

Insgesamt wurde die Umfrage 147-mal vollständig ausgefüllt, die Aufteilung zwischen den Umfragen zeigt

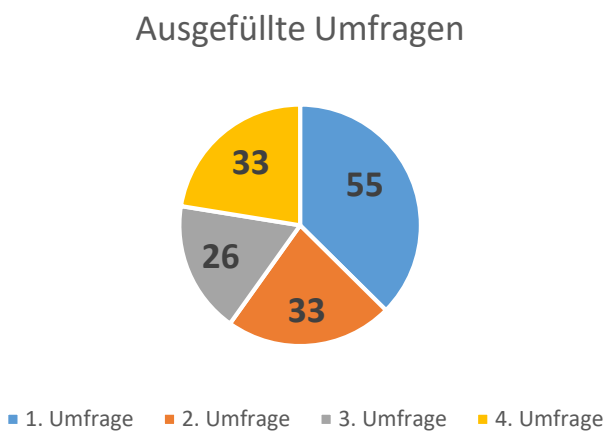


Abbildung 4 Aufteilung der Anzahl an vollständig ausgefüllten Umfragen

Unter allen Teilnehmenden gab es insgesamt 8 englische Muttersprachler, bei einem ansonsten überdurchschnittlichem Englisch Niveau (Median 4 von 5).

5.1 Stimmungsskala

Zur Auswertung der Stimmungsskala gehören die *Stimmung* und die *Reaktivität*. Gemäß dem Schema in Tabelle 3 müssen die zu bewertenden Aussagen der Skala in der Auswertung für eine korrekte Analyse umgepolt werden. Konkret betrifft dies die Items 3,5,6,9,11,13,14 und 15. Hohe Werte auf dieser Skala repräsentieren eine gute allgemeine Stimmung bzw. eine sehr reaktive Person, die ihre Meinung häufig ändert. Maximal sind 63 Punkte auf der Sub-Skala Stimmung und 42 Punkte für die Reaktivität möglich.

Insgesamt hat die Sub-Skala Stimmung einen Median von 39 (n=68). Damit liegt der Median über der Hälfte der möglichen Punkte und lässt auf eine tendenziell positivere Stimmung der Studierenden schließen. Mit einem Maximum von 59 und einem Minimum von 11 gibt es starke Schwankungen zwischen den einzelnen Studierenden.

Die Sub-Skala Reaktivität hat einen Median von 19,5 und liegt damit knapp unter der Hälfte der maximal möglichen Punkte. Die einzelnen Werte schwanken wie bei der Stimmung stark mit einer Spannweite von 31 (Maximum 41, Minimum 10).

5.2 PANAS

Bei der Auswertung von *PANAS* müssen die Sub-Skalen *Positiver-* und *Negativer Affekt* beachtet werden.[28] Die einzelnen Items können gemäß Tabelle 4 den einzelnen Dimensionen zugeordnet werden. Höhere Werte bedeuten auf beiden Sub-Skalen einen stärker ausgeprägten Affekt. Die maximal mögliche Punktzahl beträgt jeweils 50 Punkte.

Abbildung 5 zeigt den Median für den positiven- und negativen Affekt.

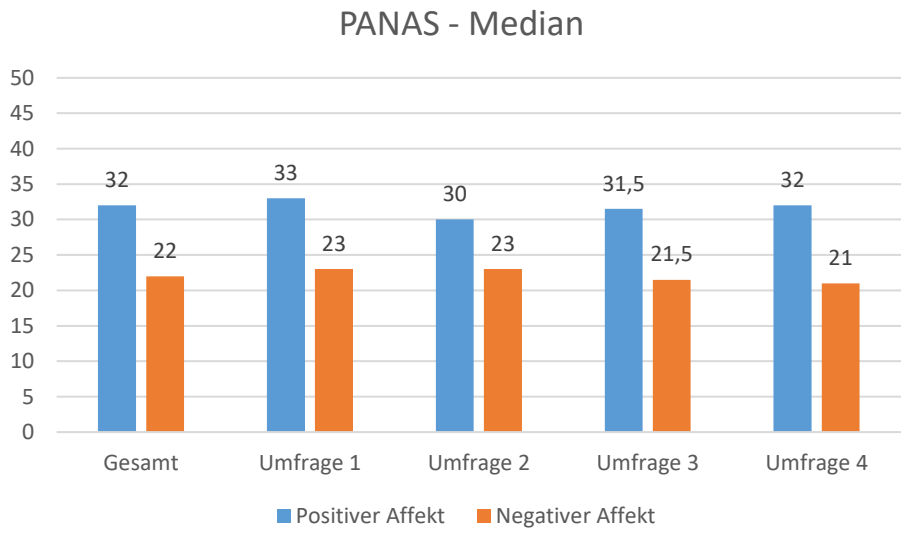


Abbildung 5 Median der Dimensionen von PANAS aufgeteilt in positiver- und negativer Affekt

Trotz der vielen verschiedenen Datenpunkte ergibt sich ein sehr homogener Median über alle Umfragen. Die Spannweite beträgt 3 im positiven Affekt (30-33) und 2 im negativen Affekt (21-23). Während der positive Affekt etwas über der Mitte liegt und damit eher zu einer stärkeren Ausprägung der positiven Stimmung tendiert, liegt der negative Affekt knapp unter der Mitte, wodurch leicht zu einer weniger stark ausgeprägten negativen Stimmung tendiert wird.

[28] Siehe 2.3

Auffällig ist, dass die Spannweite beim negativen Affekt stets höher ist als beim positiven Affekt (Graue Balken in Abbildung 6 und Abbildung 7). Die Spannweite im negativen Affekt ist demnach höher als beim positiven. Dies entsteht hauptsächlich durch die allgemein stärkere minimale Ausprägung des positiven Affekts von 15 Punkten im Vergleich zur maximal minimalen Ausprägung des negativen Affekts von 13 (Gelbe Balken). Der positive Affekt ist im Minimum stärker ausgeprägt als der Negative (15 im Vergleich zu 10), was die Schlussfolgerung der Tendenz zur stärkeren Ausprägung des positiven Affektes stützt. Abbildung 6 und Abbildung 7 zeigen grafisch die Verteilung der Spannweiten, Max- und Minima für die einzelnen Umfragen.

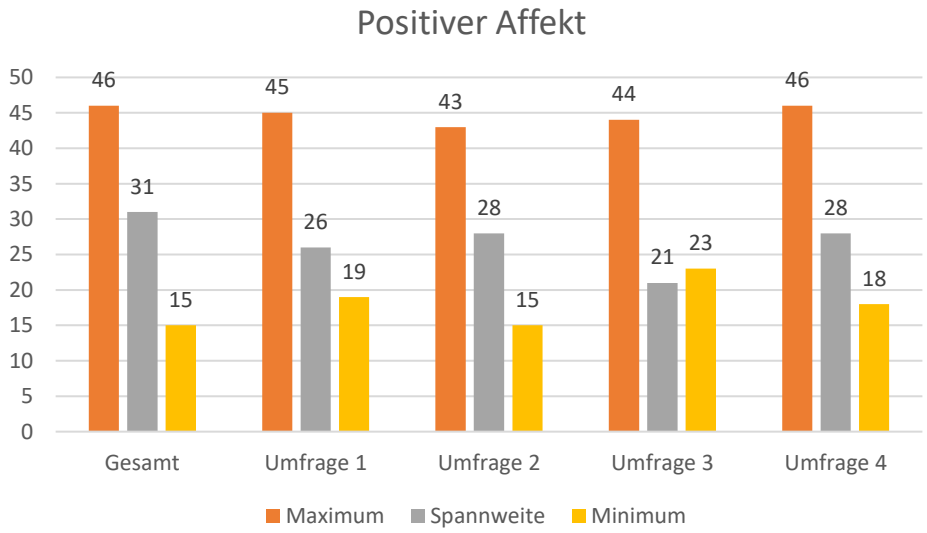


Abbildung 6 Spannweite, Max- und Minimum des positiver Affekts

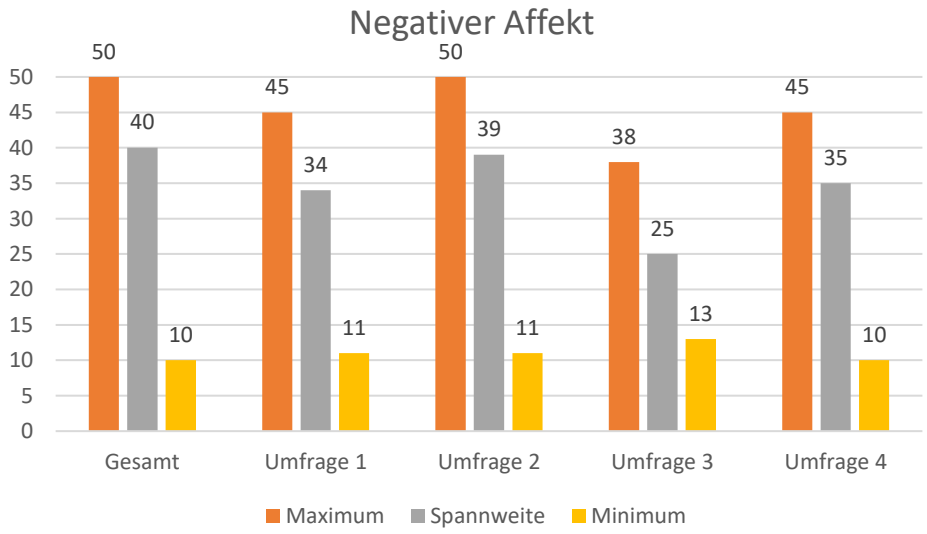


Abbildung 7 Spannweite, Max- und Minimum des negativer Affekts

5.3 Lebensumstände

Der Median der einzelnen Umfragen schwankt noch weniger als bei den bisher vorgestellten Ergebnissen. Bei einer Spannweite von gerade einmal eins (30-31) liegt der Median 7,5 bzw. 8,5 Punkte über der Hälfte der maximal möglichen Punkte (22,5).

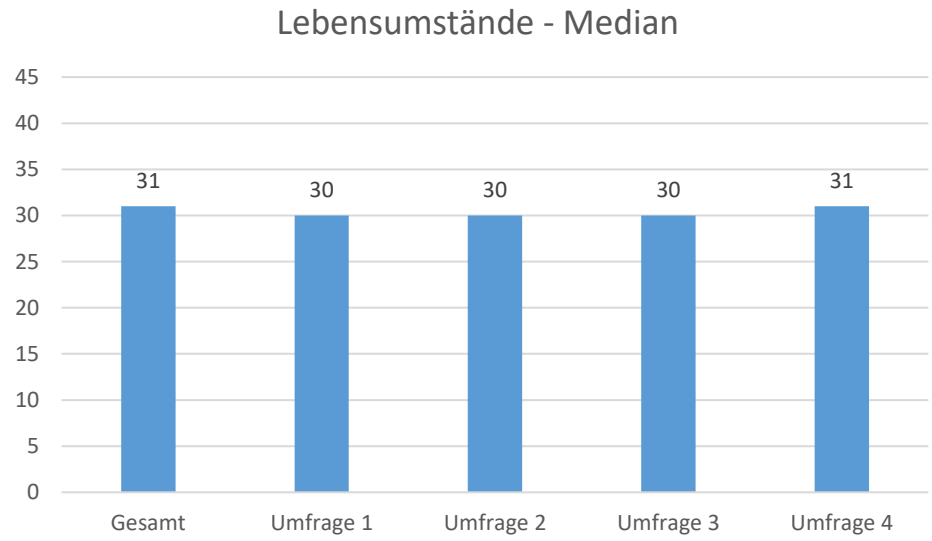


Abbildung 8 Median der Lebensumstände

Abbildung 8 zeigt deutlich den sehr homogenen Median der Einzelumfragen. Bei einem gegebenen Maximalwert von 41 Punkten und einem Minimalwert von 12 Punkten schwanken die einzelnen Werte wieder stark. Begünstigt werden die starken Schwankungen durch die Möglichkeit statt einer Bewertung „Nicht zutreffend“ anzugeben. Eine solche Bewertung wird mit 0 Punkten gewertet.

5.4 Labelvergabe

Bei der Labelvergabe werden die von den Studierenden vergebenen Label in Bezug zu den Labeln der zugrunde liegenden Datensätze betrachtet.

Durch diesen Abgleich lässt sich Trefferquote der Studierenden bestimmen. In Abbildung 9 sind die Trefferquote und die Standardabweichung dargestellt.

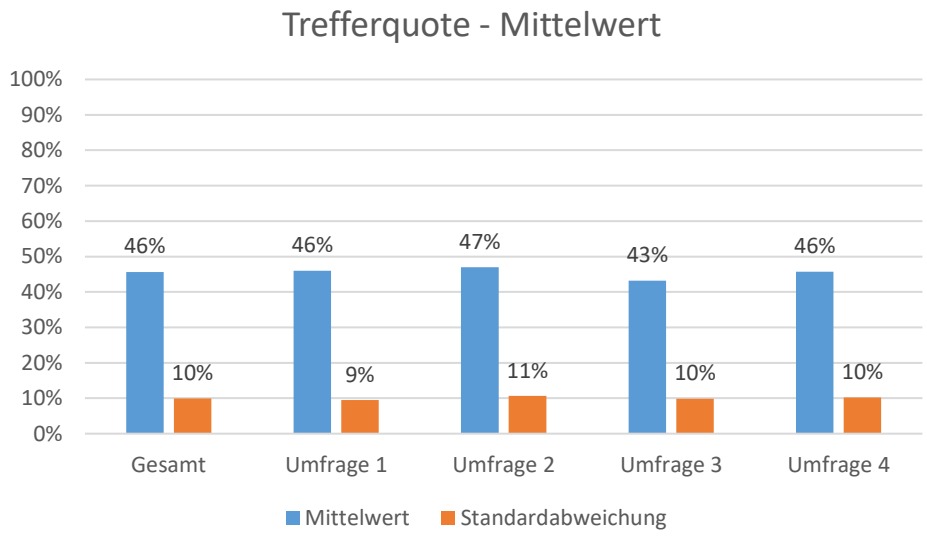


Abbildung 9 Mittelwert und Standardabweichung der Trefferquote

Über alle Datensätze liegt die Trefferquote knapp unter der Hälfte. In absoluten Zahlen wurden im Schnitt 13,7 von 30 Labeln korrekt vergeben. Lediglich in Umfrage 3 weicht die Trefferquote mehr als 1% vom Mittelwert über alle Datenpunkte ab. Für den Mittelwert der Trefferquote ergibt sich ein geschätzter Standardfehler von 0,82%. Um diesen Mittelwert streuen die Einzelergebnisse im Durchschnitt um 10% (Standardabweichung) bzw. um 2,99 korrekt vergebenen Labeln. Die Betrachtung der einzelnen Trefferquoten für positive, neutrale und negative Sätze zeigt, dass neutralen Aussagen am häufigsten das richtige Label bei einem geschätzten Standardfehler von 1,45% zugeordnet wurde (siehe Abbildung 10). Am schlechtesten wurden über alle Umfragen hinweg negative Aussagen erkannt (maximal 31%, geschätzter Standardfehler 1,64%).

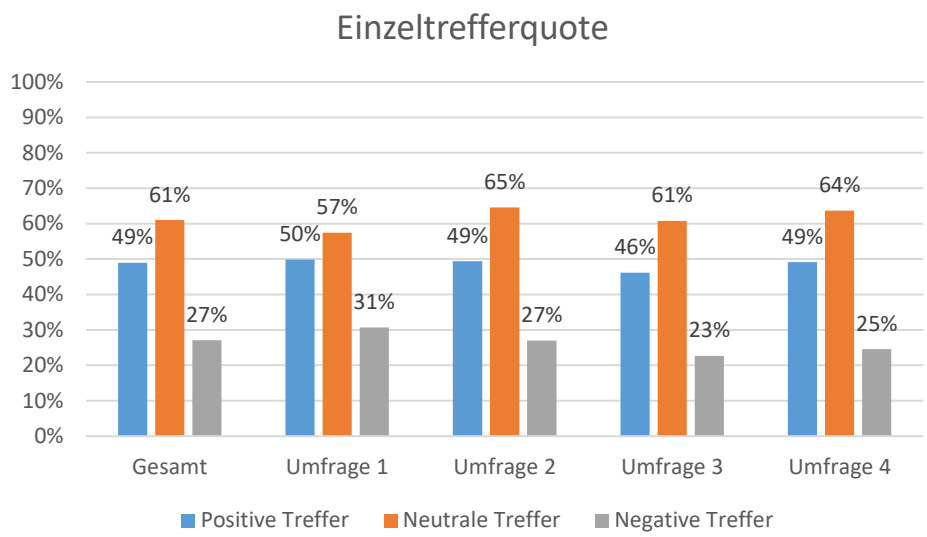


Abbildung 10 Einzeltrefferquoten differenziert nach den Polaritäten

Die Standardabweichung der einzelnen Trefferquoten ist bei den positiven Aussagen leicht höher als die der negativen Treffer. Am niedrigsten ist die Standardabweichung der neutralen Treffer[29].

Zusätzlich zu den einzelnen Trefferquoten zeigt Abbildung 11 die absolute Anzahl der vergebenen Label aufgeteilt nach Polaritäten.

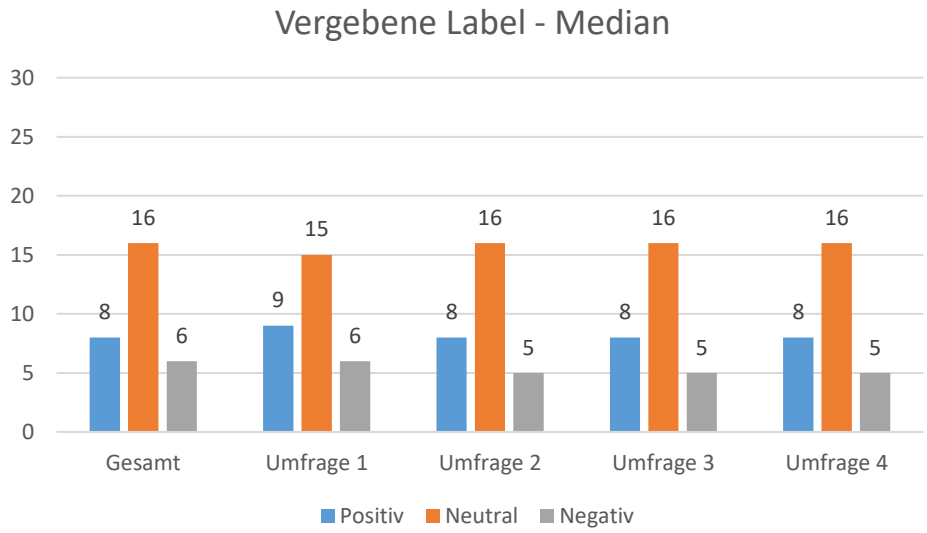


Abbildung 11 Median der vergebenen Label nach Polaritäten

Deutlich zu sehen ist, dass neutrale Label mit Abstand am häufigsten vergeben worden sind.

[29] Siehe Anhang 5, Anhang 6 und Anhang 7

Bei insgesamt 30 Labeln die zu vergeben waren, war im Schnitt jedes zweite Label neutral. Positive Label wurden nur halb so oft vergeben, wie neutrale Label und negative Label wurden am seltensten vergeben. Auffällig ist, dass fast keine Schwankungen in den Werten sind. In Umfrage zwei bis vier waren die Werte komplett identisch und unterscheiden sich von der ersten Umfrage nur um eins in den einzelnen Medianen. Dieselbe Verteilung ergibt sich aus der Betrachtung der Modi, bei denen leicht mehr positive und bis auf eine Ausnahme weniger negative Label vergeben wurden[30]. Aus dem Vergleich der Werte aus Abbildung 11 und Abbildung 10 lässt sich schließen, dass neutrale Label am häufigsten richtig erkannt wurden, diese aber auch am häufigsten vergeben worden sind. Positive Label wurden zwar deutlich weniger vergeben, sind dafür in der Trefferquote jedoch nur 12% schlechter.

Die Trefferquote lässt sich auch auf die zwei Datensätze aufteilen, um diese miteinander zu vergleichen.

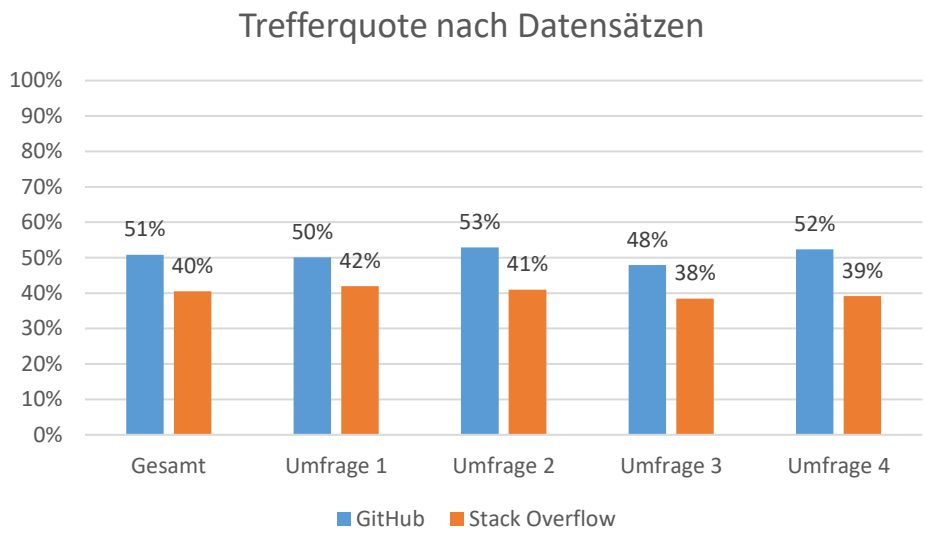


Abbildung 12 Trefferquote verteilt auf die Datensätze von GitHub und Stack Overflow

Der Vergleich in Abbildung 12 zeigt, dass die Aussagen aus dem *GitHub* Datensatz in allen Umfragen besser richtig gelabelt worden sind, wobei ein Unterschied von 11% beim gesamten Datensatz einen Unterschied von ca. 1,6 Aussagen richtiger gelabelter Sätze ausmacht. Darüber hinaus sind beide Quoten mit einer Spannweite von 5% (*GitHub,* 48%-53%) bzw. 4% (*Stack Overflow,* 42%-38%) nahezu gleich konstant. Die Quote für den *GitHub* Datensatz liegt jedoch in 4 von 5 Betrachtungen über 50%, während beim *Stack Overflow* Datensatz maximal 42% im Mittelwert richtig gelabelt wurden. Zu beachten ist der geschätzte Standardfehler von 1,08% bzw. 0,95%.

[30] Siehe Anhang 8

Ein Datensatz, der mit zugrundeliegenden *Guidelines* gelabelt worden ist, schneidet im Schnitt beim manuellen *ad hoc* Labeln bezogen auf die Trefferquote besser ab als ein *ad hoc* vorgelabelter Datensatz.

5.5 Intragroup Conflict Scale

Aus den Ergebnissen der *Intragroup Coflict Scale* lassen sich *Task*- und *Relationship-Konflikte* ableiten.

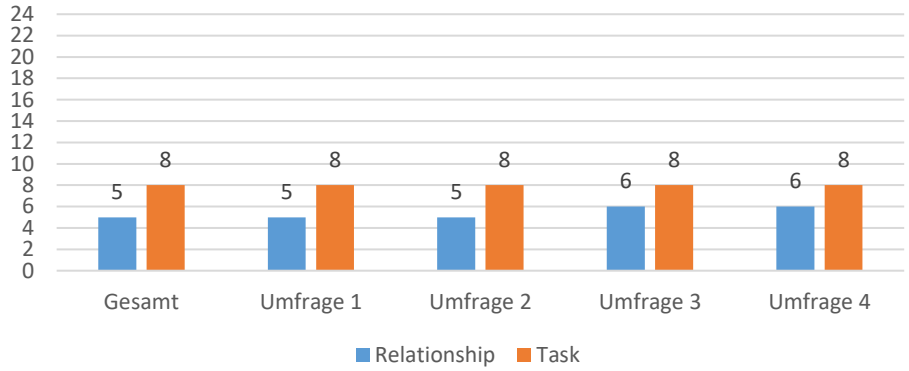


Abbildung 13 Median der Relationship- und Task-Konflikte

Abbildung 13 zeigt den Vergleich zwischen dem Median der *Relationship-Konflikte* und dem Median der *Task-Konflikte*. Klar erkennbar ist, dass *Task-Konflikte* häufiger auftreten als *Relationship-Konflikte*. Dabei sind jedoch beide Werte immer unterhalb der Hälfte der möglichen 24 Punkte. Tendenziell gab es innerhalb der Softwareprojekte demnach wenige Konflikte. Bemerkenswert ist die Konstanz der *Relationship-* und *Task-Konflikte* über alle betrachteten Datenpunkte. Der Modus der *Relationship-Konflikte* ist noch einen Punkt geringer als der Median, während der Modus der *Task-Konflikte* gleich dem Median ist. Auch die Modi zeigen, dass es mehr *Task-Konflikte* gab[31].

Die Werte schwanken dabei sehr stark von der niedrigsten möglichen Punktzahl vier bis zu 22 Punkten bei den *Relationship-Konflikten* und 19 Punkten bei den *Task-Konflikten*. Erstaunlich ist das höhere Maximum bei den *Relationship-Konflikten*, da der Median im Vergleich deutlich geringer war.

5.6 Gründe für die Labelvergabe

Der überwiegende Teil der Studierenden hat die Label nach Inhalt oder herausgelesenem Klang bzw. Ton vergeben. Lediglich knapp unter der Hälfte aller Studierenden haben Smileys oder Emoticons als ausschlaggebend angegeben (5 von 30 Aussagen enthielten Emoticons[32]).

[31] Siehe Anhang 9
[32] Siehe Tabelle 7 und Tabelle 8

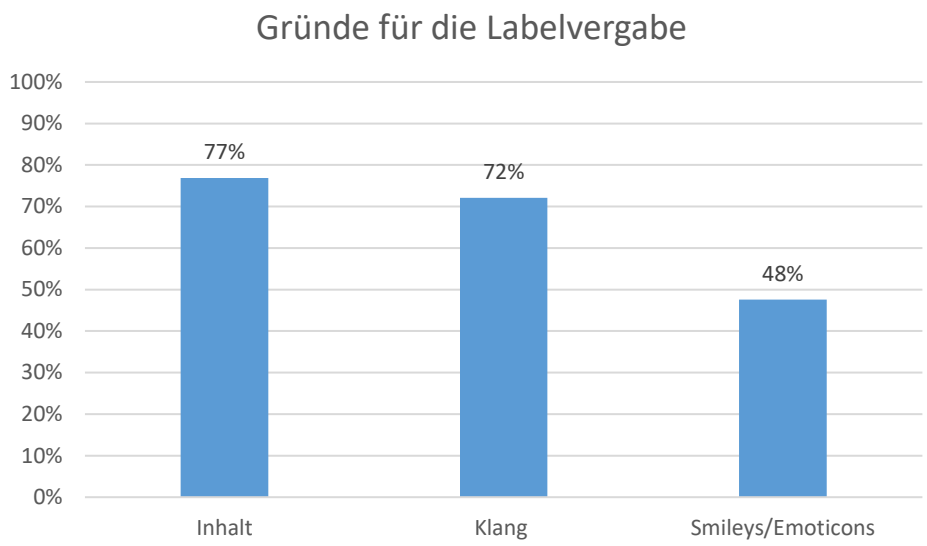


Abbildung 14 Überblick über die Gründe für die Labelvergabe

Die Studierenden hatten auch die Möglichkeit, weitere Gründe, falls vorhanden, zu nennen.[33] Dabei gab es keine Häufungen von Gründen, die weitere Rückschlüsse zulassen.

Abbildung 15 zeigt, dass etwas mehr als die Hälfte aller Teilnehmenden bei der Labelvergabe zu viele Ansatzpunkte für die Auswahl eines Labels gefunden haben und dadurch verunsichert waren. Etwa ein Drittel haben widersprüchliche Emotionen erkannt oder konnten kein klares Label bestimmen. Nur 21% waren sich bei der Labelvergabe nicht unsicher.

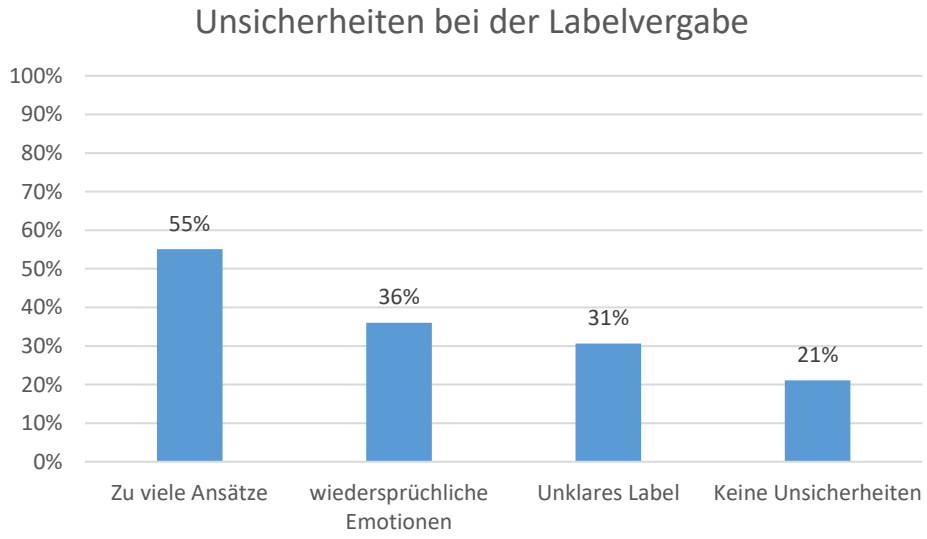


Abbildung 15 Gründe der Teilnehmenden für Unsicherheiten bei der Labelvergabe

[33] Siehe Anhang 10 Sonstige Gründe der Studierenden bei der Labelvergabe

Die Studierenden hatten hier ebenfalls die Möglichkeit, weitere Gründe für Unsicherheiten, falls vorhanden, anzugeben.[34] Sehr übereinstimmend wurde der mangelnde Kontext als Grund angegeben. Nur einer der Teilnehmenden empfand die Möglichkeit zwischen drei Labeln zu wählen als zu generell.

5.7 Beantwortung der Forschungsfragen

Die zuvor vorgestellten Daten werden nachfolgend mithilfe von Zusammenhangsmaßen[35] analysiert, um die Forschungsfragen zu beantworten. Für die Zusammenhangmaße wurden aus den Daten weitere Informationen abgeleitet. Es wurden die Größen *Positive Label*, *Neutrale Label* und *Negative Label* errechnet. Diese drücken für das jeweilige Label aus, wie oft dieses insgesamt vergeben wurde. Darüber hinaus wurde mittels Abgleich der vergebenen und Musterlabel ermittelt, wie oft Studierende eine Aussage positiver oder negativer als das Musterlabel bewertet haben. Die Anzahl der Sätze, bei denen dies der Fall war, drücken die Werte Positiver- und Negativer bewertet aus. Eine stark positivere Bewertung lag vor, wenn für einen negativen Satz das Label positiv vergeben wurde. Eine stark negative Bewertung bezeichnet das Gegenteil. Alle Zahlen wurden auf 2 Nachkommastellen gerundet.

5.7.1 Auswirkungen der Stimmung auf die Sentimentvergabe

Für die Analyse des Zusammenhangs zwischen Stimmung und Sentimentvergabe werden drei Teilbereiche betrachtet: Stimmung aus der Stimmungsskala, positiver- und negativer Affekt aus *PANAS* und die Bewertung der Lebensumstände. Zu jedem der einzelnen Teilbereiche können verschiedene Korrelationskoeffizienten berechnet werden.

Im Teilbereich Stimmung wurden für die Analyse zunächst alle Datenpunkte ermittelt, bei denen eine Stimmung vorliegt. Da die Stimmung in der zweiten Umfrage nicht abgefragt wurde, ist die Stichprobe mit einer Größe von 128 Datensätzen kleiner als die Gesamtzahl der Datensätze.

[34] Siehe Anhang 11
[35] Siehe Kapitel 4.6.4

Korrelation Stimmung zu	Korrelationskoeffizient	p-Wert	Signifikant
Treffer	0,02	0,79	Nein
Trefferquote	0,02	0,79	Nein
Positive Treffer	0,16	0,07	Nein
Neutrale Treffer	-0,07	0,41	Nein
Negative Treffer	-0,07	0,41	Nein
Positive Label	0,21	0,02	Ja
Neutrale Label	-0,13	0,15	Nein
Negative Label	-0,08	0,38	Nein
Positiver bewertet	0,13	0,14	Nein
Negativer bewertet	-0,14	0,11	Nein
Stark positive Abweichung	0,16	0,08	Nein
Stark negative Abweichung	-0,09	0,34	Nein

Tabelle 11 Korrelationskoeffizienten zur Stimmung mit zugehörigem p-Wert

Tabelle 11 zeigt die verschiedenen Korrelationskoeffizienten der Stimmung. Aus den Werten wird deutlich, dass nur der Zusammenhang zwischen Stimmung und Positive Label in Zeile 7 der Tabelle statistisch signifikant ist. Mit einem Korrelationskoeffizienten von 0,21 ist der Zusammenhang trotzdem nur schwach. Es lässt sich daraus ableiten, dass die Stimmung einen schwach positiven Einfluss auf die Anzahl an vergebenen positiven Labeln hat. Nimmt die Stimmung eines Entwicklers zu, werden mehr positive Label vergeben.

Im zweiten und dritten Teilbereich musste die Stichprobe nicht eingeschränkt werden, da die nötigen Daten bei jeder Umfrage abgefragt worden sind. Die Stichprobengröße beträgt für diese beiden Bereiche 147.

Keine der Korrelationen des positiven Affekts aus *PANAS* sind statistisch signifikant ist.[36] Der positive Affekt hatte keine signifikanten Auswirkungen auf die Bewertungen der Studierenden. Werden nur die Korrelationen betrachtet, fällt zusätzlich auf, dass keine hoch genug ist, um einen Zusammenhang erkennen zu können.

[36] Siehe Anhang 12

Korrelation negativer Affekt zu	Korrelationskoeffizient	p-Wert	Signifikant
Treffer	0,09	0,29	Nein
Trefferquote	0,09	0,29	Nein
Positive Treffer	-0,04	0,67	Nein
Neutrale Treffer	0,02	0,84	Nein
Negative Treffer	0,15	0,08	Nein
Positive Label	-0,07	0,43	Nein
Neutrale Label	-0,05	0,54	Nein
Negative Label	0,16	0,05	Ja
Positiver bewertet	-0,19	0,02	Ja
Negativer bewertet	0,09	0,30	Nein
Stark positive Abweichung	0,03	0,70	Nein
Stark negative Abweichung	0,11	0,19	Nein

Tabelle 12 Korrelationskoeffizienten zum negativen Affekt mit zugehörigem p-Wert

Der zweite Teil der *PANAS* Analyse ist in Tabelle 12 dargestellt. Im Gegensatz zum positiven Affekt hat der negative Affekt zwei signifikante Auswirkungen auf die Bewertung. Sowohl die Anzahl negativer Label, als auch die positivere Bewertung der Aussagen weisen einen statistisch signifikanten Zusammenhang zum negativen Affekt auf. Zum einen lässt sich ableiten, dass je höher der negative Affekt der letzten 7 Tage war, desto höher ist die Anzahl vergebener negativer Label. Zum anderen werden mit steigendem negativem Affekt Aussagen weniger positiv bewertet. Die Korrelationen könnten statistisch voneinander abhängig sein, da bei mehr negativen Labeln weniger Aussagen positiver bewertet werden können. Bei Betrachtung der Korrelationskoeffizienten zeigt sich, dass diese nur einen schwachen Zusammenhang anzeigen.

Korrelation Lebensumstände zu	Korrelationskoeffizient	p-Wert	Signifikant
Treffer	0,06	0,45	Nein
Trefferquote	0,06	0,45	Nein
Positive Treffer	0,09	0,30	Nein
Neutrale Treffer	0,11	0,18	Nein
Negative Treffer	-0,08	0,32	Nein
Positive Label	0,03	0,74	Nein
Neutrale Label	0,09	0,27	Nein
Negative Label	-0,19	0,02	Ja
Positiver bewertet	0,04	0,64	Nein
Negativer bewertet	-0,13	0,12	Nein
Stark positive Abweichung	-0,01	0,93	Nein
Stark negative Abweichung	-0,28	0,00	Ja

Tabelle 13 Korrelationskoeffizienten zu den Lebensumständen mit zugehörigem p-Wert

Die Betrachtung der Lebensumstände aus Tabelle 13 ist die letzte Teilbetrachtung der Forschungsfrage. Signifikant sind die Zusammenhänge zwischen Lebensumständen und negativen Labeln und stark negativen Abweichungen. Eine höhere Zufriedenheit mit den Lebensumständen sorgt für weniger negative Label und weniger stark negative Abweichungen bei der Labelvergabe. Auffällig ist die höhere Korrelation zu den negativen Labeln im Vergleich zu den anderen Polaritäten. Werden weniger negative Label vergeben, müssen diese Label auf die anderen Polaritäten verteilt werden. Dabei scheint keins der anderen beiden Label zu überwiegen und diese zusätzlichen Label werden aufgeteilt. Zusammenhängen können auch hier die beiden signifikanten Korrelationen, da mehr negative Label zu mehr stark negativen Abweichungen führen können. Auffällig ist, dass für negativer bewertet der Koeffizient nicht signifikant und kleiner ist, als bei stark negativen Abweichungen. Die Korrelation zwischen Lebensumständen und der Vergabe negativer Label ist erneut ein schwacher Zusammenhang. Für die stark negativen Abweichungen liegt die Korrelation ebenfalls im Bereich des schwachen Zusammenhangs. Gleichzeitig ist dies die stärkste Korrelation aller Zusammenhänge für diese Forschungsfrage. Andere Zusammenhänge sind statistisch nicht signifikant.

Für die Forschungsfrage 1 können zusammenfassend folgende Schlussfolgerungen festgehalten werden:

1. Eine Erhöhung der allgemeinen Stimmung führt zu einer schwachen Erhöhung der Anzahl an vergebenen positiven Label.
2. Der positive Affekt hat keinen signifikanten Einfluss auf die Sentimentvergabe.
3. Der negative Affekt hat einen schwachen signifikanten positiven Einfluss auf die Anzahl an vergebenen negativen Labeln und einen schwachen negativen Einfluss auf die positivere Bewertung der Aussagen.
4. Die Lebensumstände haben schwach negative signifikante Einflüsse auf die Anzahl negativer Label und die stark negativen Abweichungen.

5.7.2 Auswirkungen mehrfacher Teilnahme auf die Sentiment Vergabe

Für die Beantwortung der zweiten Forschungsfrage wurden die vorhandenen Daten nach der Anzahl der Teilnahmen geordnet. Daraus ergaben sich die Teilnahmezahlen aus Abbildung 16.

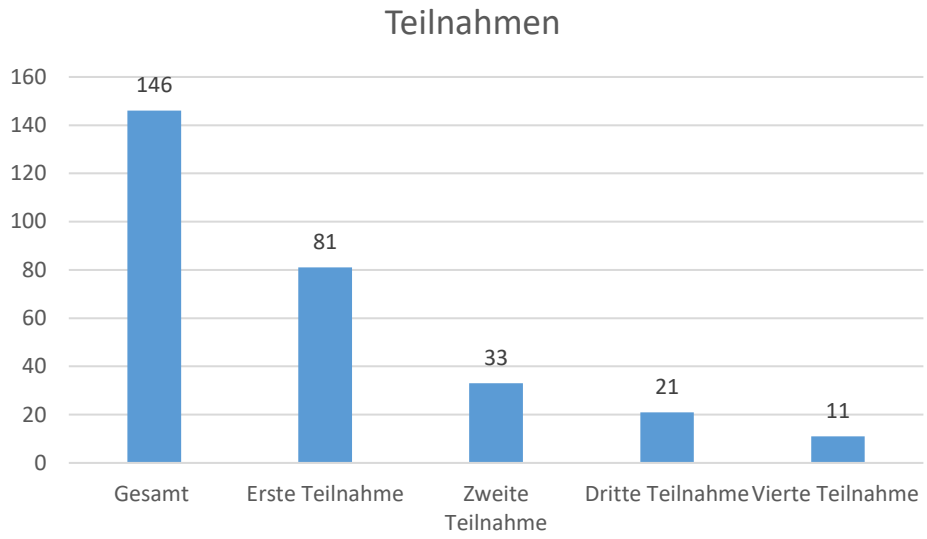


Abbildung 16 Anzahl an Teilnahmen geordnet nach der Häufigkeit der Teilnahme

Zunächst wird analysiert, inwieweit sich eine mehrfache Teilnahme auf die Trefferquote auswirkt, um zu prüfen, ob sich die Sentimentvergabe der Studierenden durch mehrfache Teilnahme an die Label der Datensätze anpasst. Danach werden Änderungen in den Gründen für die Sentimentvergabe und die Konsistenz in der Bewertung untersucht.

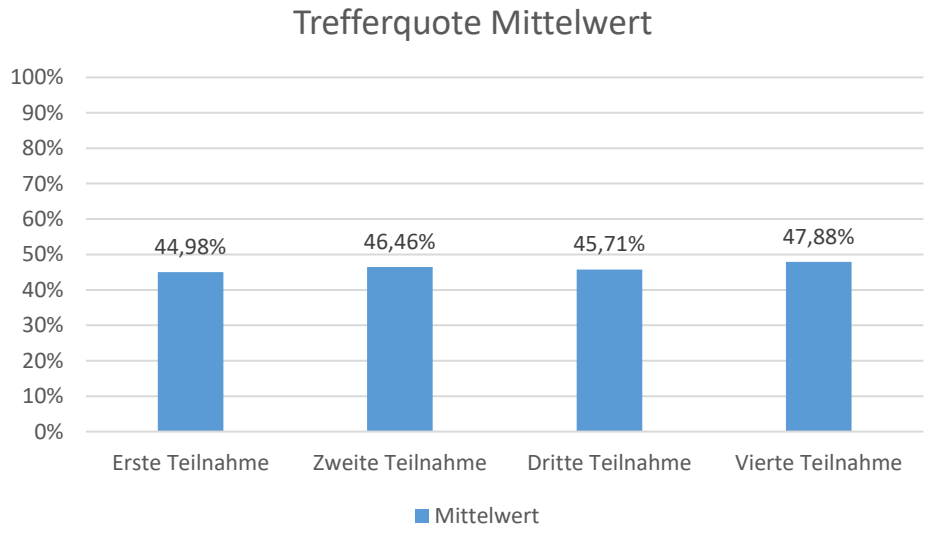


Abbildung 17 Mittelwerte der Trefferquote

Die Betrachtung der Trefferquote im Verlauf der Teilnahmen aus Abbildung 17 zeigt, dass diese sich von der ersten zur vierten Teilnahme leicht verbessert hat, zwischendurch von der zweiten auf die dritte Teilnahme jedoch auch wieder gefallen ist.

Der Verlauf ist nicht monoton steigend und könnte durch Schwankungen bedingt sein. Es ist, wenn überhaupt, nur ein sehr leichter Aufwärtstrend erkennbar.

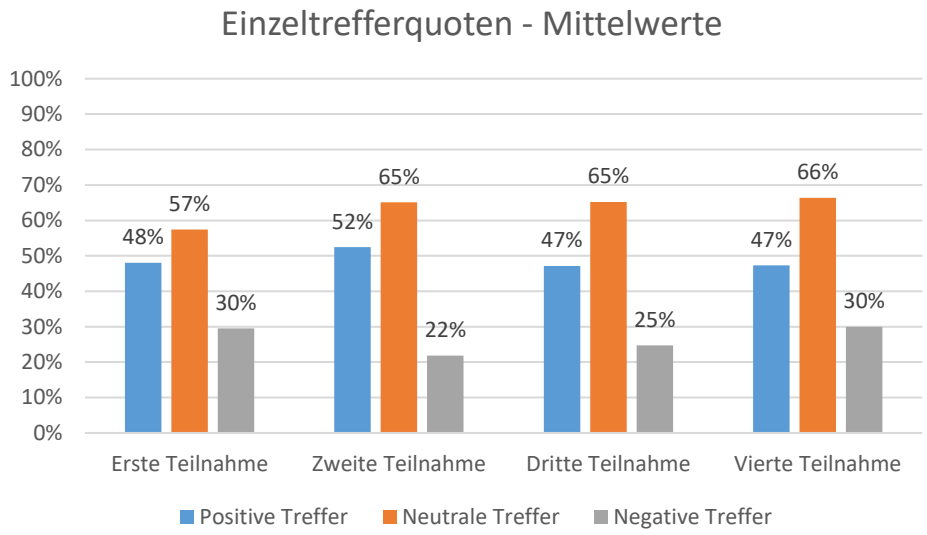


Abbildung 18 Einzeltrefferquoten gegliedert nach Polaritäten

Ergänzend zu den Daten aus Abbildung 17 kann Abbildung 18 betrachtet werden. Diese zeigt den Verlauf der einzelnen Trefferquoten der Polaritäten. Zu erkennen ist ein monoton steigender Verlauf der neutralen Treffer, die zunächst sogar sehr stark und dann nur noch minimal ansteigen. Neutrale Aussagen wurden ab der zweiten Teilnahme besser erkannt als bei nur einmaliger Teilnahme. Aus den anderen Verläufen lassen sich keine klaren Trends erkennen.

Eine Betrachtung der Anzahl der einzelnen vergebenen Labels bietet keine weiteren Erkenntnisse, da die Anzahlen bis auf kleine Schwankungen und eine kleine Abnahme der positiven Label weitestgehend konstant verlaufen[37].

Neben den reinen vergebenen Labeln und deren Trefferquoten kann auch betrachtet werden, welche Gründe für die Labelvergabe ausschlaggebend waren, um zu analysieren, ob sich Gründe für die Sentimentvergabe verschieben.

[37] Siehe Anhang 13 Vergebene Label

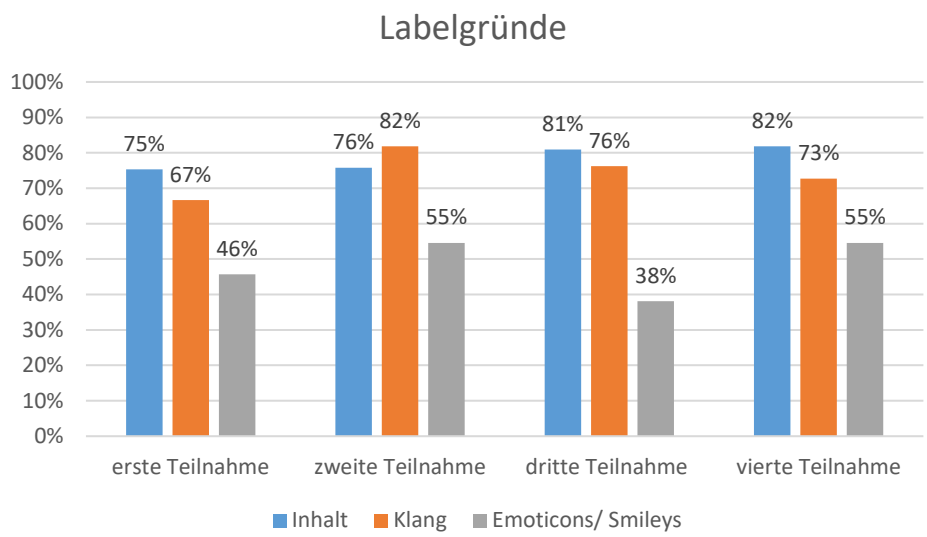


Abbildung 19 Gründe für die Labelvergabe

Abbildung 19 fasst zusammen, für wie viel Prozent der Studierenden einer der aufgelisteten Gründe ausschlaggebend für die Labelvergabe war. Dabei fällt auf, dass die Werte für herausgelesenen Klang und Emoticons/ Smileys stark schwanken und keinen klaren Trend erkennen lassen. Streng monoton steigend ist jedoch der Inhalt. Bei der Bewertung der Aussagen scheint mit häufigerer Teilnahme der Inhalt immer wichtiger zu werden. Dieser ist in drei von vier Datenpunkten zusätzlich der wichtigste Faktor und wird bei Beibehaltung des Trends zunehmend wichtiger werden. Dazu ergänzend lassen sich die Gründe für Unsicherheiten bei der Labelvergabe in Abbildung 20 betrachten.

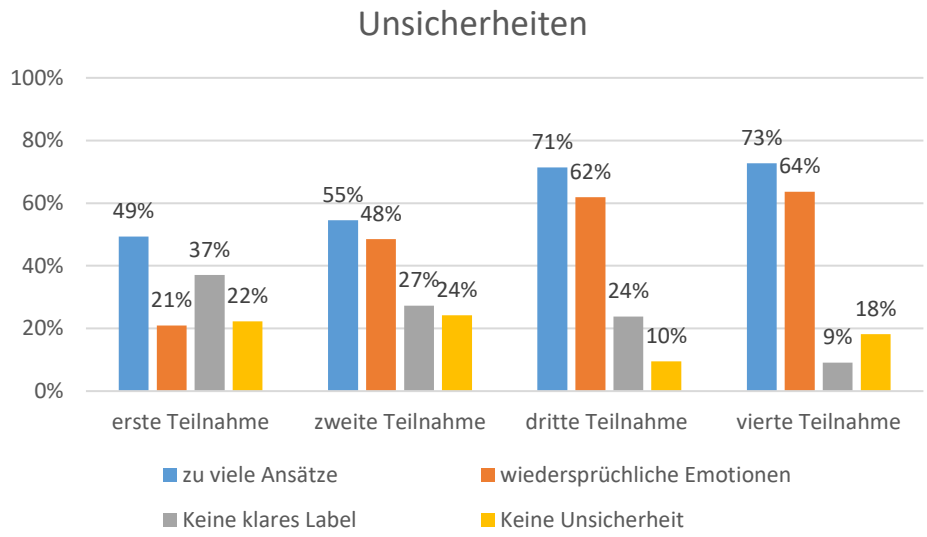


Abbildung 20 Gründe für Unsicherheiten bei der Labelvergabe

Durch die mehrfache Teilnahme nahmen die Studierenden zunehmend mehr Ansätze für eine Labelvergabe und widersprüchliche Emotionen wahr. Stark abgenommen hat die Wahrnehmung keines klaren Labels. Kein klarer Trend ist bei „Keine Unsicherheiten“ zu erkennen, obwohl diese im Vergleich von erster und vierter Teilnahme leicht abgenommen haben. Mit der Betrachtung der Erkenntnisse aus Abbildung 19 und Abbildung 20 können mögliche Zusammenhänge abgeleitet werden. Die stärkere Fokussierung auf den Inhalt und der starke Fokus auf den Klang der Aussagen hat möglicherweise die Identifizierung eines klaren Labels für die Studierenden erleichtert, kann jedoch gleichzeitig dafür gesorgt haben, dass mehr Ansätze und mehr widersprüchliche Emotionen erkannt wurden.

Um festzustellen, inwieweit sich die Studierenden bei der Vergabe der Label mit sich selbst einig sind, wurde für Studierende die zwei- drei- oder viermal an der Umfrage teilgenommen haben der Fleiss´ Kappa Wert berechnet. Mit mittleren Fleiss´ Kappa Werten von 0,47 (zweifache Teilnahme), 0,53 (dreifache Teilnahme) und 0,51 (vierfache Teilnahme) haben die Studierende immer eine moderate Übereinstimmung mit sich selbst[38]. Im ersten Fleiss´ Kappa Wert sind durch negative Werte im Mittelwert stark abweichende Werte enthalten. Wären diese negativen Werte nicht enthalten ergäbe sich ein Fleiss´ Kappa Wert von 0,55. Generell schwanken die Fleiss´ Kappa Werte sehr stark. Über alle Teilnahmen ist das Maximum 0,85 und das Minimum -0,10 bzw. der niedrigste Wert größer als null ist 0,08. In jeder Teilnahme schwanken die Fleiss´ Kappa Werte mindestens um 0,69.[39]

Zusätzlich kann die Anzahl konsistenter Antworten errechnet werden.

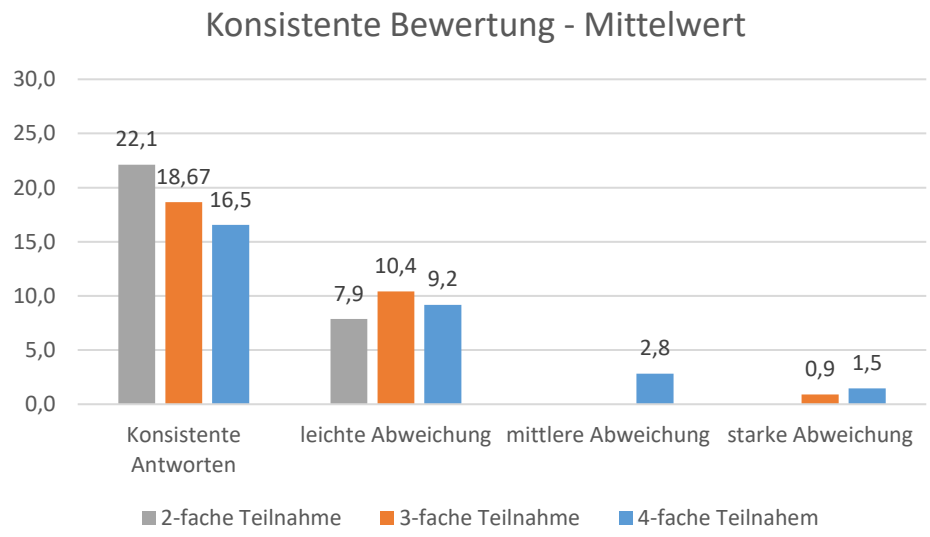


Abbildung 21 Konsistenz in der Bewertung

[38] Siehe Anhang 14 2-fache Telnahme Fleiss´ Kappa Werte
[39] Siehe Anhang 14, Anhang 15 und Anhang 16

Konsistente Antworten in Abbildung 21 liegen vor, wenn stets dasselbe Label vergeben wurde, eine leichte Abweichung bei einer abweichenden Antwort und eine starke Abweichung bei zwei abweichenden Antworten. Eine mittlere Abweichung existiert nur bei 4-facher Teilnahme, wenn zwei Mal zwei konsistente Label vergeben wurden. Da in den Daten der Fall von zwei positiven und zwei negativen Labeln nicht vorkommt, wurde dieser Fall nicht gesondert betrachtet[40].

Bei der zweiten Teilnahme gab es mit Abstand am meisten konsistente Antworten. Bei häufigerer Teilnahme wichen die Studierenden stärker von ihrer vorherigen Antwort ab. Da bereits gezeigt wurde, dass die Stimmung keinen starken Einfluss auf die Labelvergabe hat, wird analysiert, ob die Reaktivität einen Einfluss auf die Abweichung der Label haben kann.

Für eine vierfache Teilnahme konnte keine signifikante Korrelation gefunden werden.[41]

Korrelation	Korrelationskoeffizient	p-Wert	Signifikant
Fleiss´ Kappa zu Reaktivität	-0,19	0,42	Nein
Fleiss´ Kappa zu konsistenten Antworten	0,72	0,00	Ja
Reaktivität zu Konsistente Antworten	-0,14	0,57	Nein
Reaktivität zu leichte Abweichung	0,10	0,66	Nein
Reaktivität zu starke Abweichung	0,04	0,87	Nein

Tabelle 14 Korelation Fleiss´ Kappa 3-fache Teilnahme

Aus Tabelle 14 geht hervor, dass bei dreifacher Teilnahme nur die Korrelationen zwischen Fleiss´ Kappa und konsistenten Antworten statistisch signifikant ist. Da per Definition von Fleiss´ Kappa dies die Übereinstimmung mit sich selbst ist, ist der starke Zusammenhang vorhersehbar gewesen. Mehr konsistente Antworten resultieren in einer stärkeren Übereinstimmung. Für eine zweifache Teilnahme kann der gleiche Zusammenhang gezeigt werden, auch wenn dieser schwächer ist als bei dreifacher Teilnahme.[42] Die Reaktivität hat aber bei keiner der Teilnahmen einen signifikanten Einfluss auf die Abweichungen in den Antworten. Eine erhöhte Reaktivität steht nicht im Zusammenhang mit veränderter Sentimentvergabe bei mehrfacher Sentimentvergabe.

Zusammenfassend lassen sich folgende Aussagen für die Beantwortung der Forschungsfrage 2 formulieren:

1. Ab der zweiten Teilnahme werden neutrale Aussagen besser erkannt.

[40] Siehe Anhang 17
[41] Siehe Anhang 18
[42] Siehe Anhang 19

2. Mehrmaliges Teilnehmen erhöht den Fokus bei der Sentimentvergabe auf den Inhalt und den Klang.
3. Es werden zunehmend mehr klare Label zugeordnet und gleichzeitig mehr Ansätze und widersprüchliche Emotionen erkannt.
4. Bei mehrfacher Teilnahme gibt es mehr Abweichungen zu den vorherigen Antworten.
5. Eine höhere Reaktivität verändert nicht die Sentimentvergabe.

5.7.3 Auswirkungen der Projektphase und der Gruppendynamik auf die Sentimentvergabe

Zur Beantwortung der dritten Forschungsfrage wurden die Daten nach Zeitpunkten und Projektart gefiltert, woraus sich die Verteilung aus Abbildung 22 ergab.

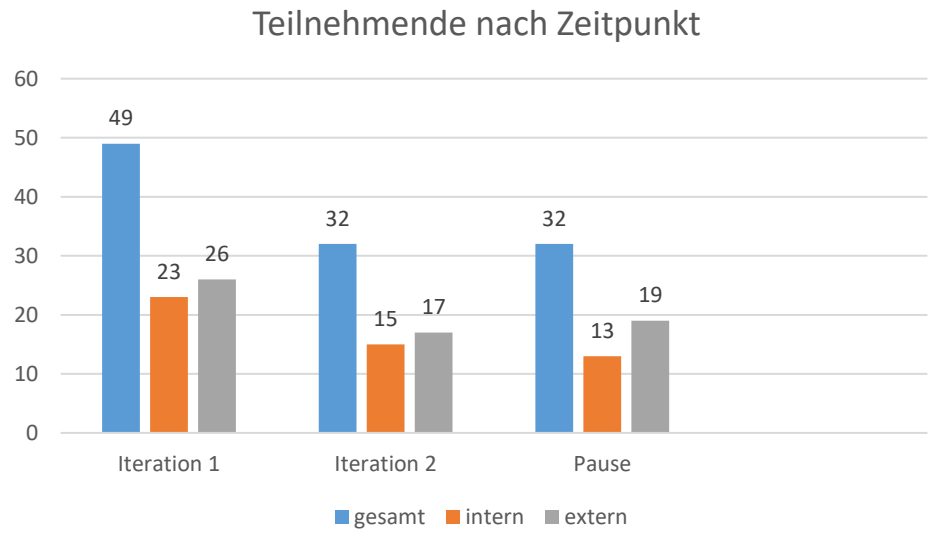


Abbildung 22 Teilnehmende nach Zeitpunkt und Projektart sortiert

Zunächst wird betrachtet, welchen Einfluss die Phase eines Projektes und der damit verbundene Stresslevel auf die Sentimentvergabe haben. Dazu kann analysiert werden, wie sich die vergebenen Label über die Phasen entwickeln.

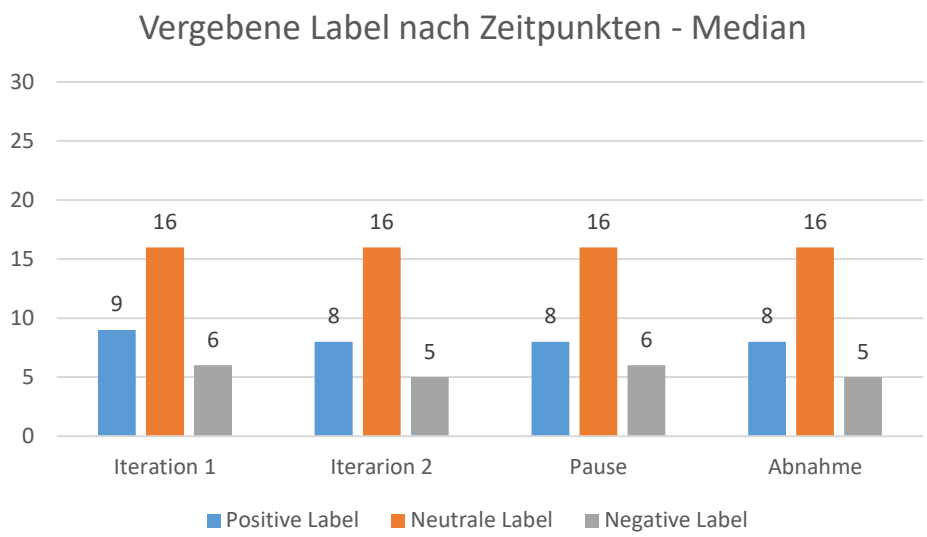


Abbildung 23 Vergebene Label nach Zeitpunkten

Aus Abbildung 23 wird deutlich, dass die Anzahl der vergebenen Label über die Projektphasen nahezu unverändert ist. Lediglich kleine Schwankungen um ein Label sind bei den positiven und negativen Labeln zu beobachten. Aus der Unterteilung nach internen und externen Projekten lassen sich keine weiteren Erkenntnisse ziehen. Die Werte sind weitestgehend übereinstimmend mit denen aus der Gesamtbetrachtung[43]. Sowohl die Projektart als auch die Projektphase scheint daher keinen Einfluss auf die Anzahl vergebener Label zu haben. Eine weitere mögliche Betrachtung ist die Anzahl der Treffer aus Abbildung 24.

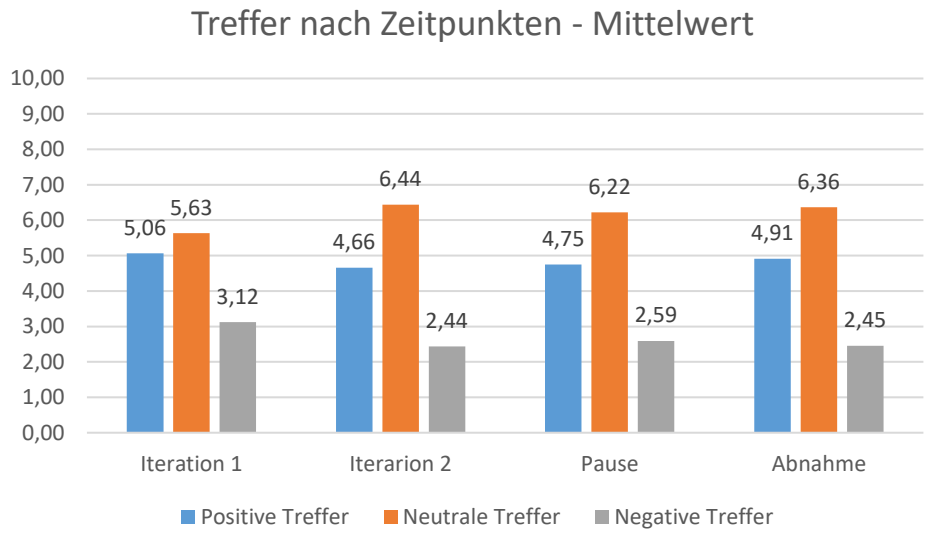


Abbildung 24 Treffer nach Zeitpunkten und Polaritäten sortiert

[43] Siehe Anhang 20 und Anhang 21

Die positiven Treffer sinken zunächst mit abnehmenden Stresslevel, steigen dann zur Pause jedoch wieder leicht an und zur Abnahme gibt es erneut eine Steigerung. Ein klarer Zusammenhang zum Stresslevel und der Projektphase ist nicht zu erkennen. Neutrale Treffer steigen zunächst an, nehmen danach ab und nehmen zur Abnahme wieder zu, woraus ebenfalls kein klarer Zusammenhang zum Projekt zu erkennen ist. Die negativen Treffer entwickeln sich in der Tendenz gleich der Anzahl negativer Label, woraus sich die gleiche Schlussfolgerung ableitet wie zuvor.

Zusätzlich zu der Analyse der Zusammenhänge zwischen Projektphase und Sentimentvergabe, wurden auch Zusammenhänge zwischen Konflikten innerhalb der Projektgruppe und der Sentimentvergabe analysiert.

Gesamt	**Korrelationskoeffizient**	**p-Wert**	**Signifikant**
Relationship - negative Label	0,07	0,39	Nein
Relationship - positive Label	-0,05	0,51	Nein
Relationship -neutrale Label	-0,05	0,58	Nein
Task - negative Label	0,17	0,03	Ja
Task - positive Label	0,00	1,00	Nein
Task - neutrale Label	-0,11	0,18	Nein

Tabelle 15 Korrelationen Konflikte zur Labelvergabe

Intern	**Korrelationskoeffizient**	**p-Wert**	**Signifikant**
Relationship - negative Label	-0,00	0,99	Nein
Relationship - positive Label	-0,14	0,25	Nein
Relationship -neutrale Label	0,09	0,44	Nein
Task - negative Label	0,10	0,43	Nein
Task - positive Label	-0,08	0,54	Nein
Task - neutrale Label	0,02	0,86	Nein

Tabelle 16 Korrelationen Konflikte zur Labelvergabe interner Projekte

Extern	**Korrelationskoeffizient**	**p-Wert**	**Signifikant**
Relationship - negative Label	0,13	0,27	Nein
Relationship - positive Label	0,04	0,766	Nein
Relationship -neutrale Label	-0,17	0,13	Nein
Task - negative Label	0,24	0,03	Ja
Task - positive Label	0,10	0,41	Nein
Task - neutrale Label	-0,26	0,02	Ja

Tabelle 17 Korrelationen Konflikte zur Labelvergabe externer Projekte

In Tabelle 15 werden die Korrelationen zwischen der Labelvergabe und den Gruppenkonflikten berechnet und auf Signifikanz untersucht. Darüber hinaus sind die Korrelationen für den Gesamtdatensatz und getrennt für interne in Tabelle 16 und externe Projekte in Tabelle 17 aufgelistet.

Zunächst fällt auf, dass *Relationship-Konflikte* keinen signifikanten Einfluss auf die Sentimentvergabe haben. Bei internen Projekten gibt es sowohl bei *Relationship-* als auch bei *Task-Konflikten* keinen signifikanten Einfluss.

Signifikant ist der schwache Einfluss der *Task-Konflikte* auf die Vergabe negativer Label im Bereich der Gesamtdatensätze. Bei mehr *Task-Konflikten* steigen die vergebenen negativen Label an

Unter Teilnehmenden an externen Projekten gibt es signifikante Einflüsse der *Task-Konflikte* auf die Vergabe negativer und neutraler Label. Mit zunehmender Anzahl an *Task-Konflikten* werden mehr negative und weniger neutrale Label vergeben. Beide Korrelationen lassen sich als schwacher Zusammenhang interpretieren.

Die Korrelationen lassen sich ebenfalls für die einzelnen Projektphasen errechnen, um zusätzlich zu analysieren, wie sich die Korrelationen bei unterschiedlichen Stressniveaus entwickeln. Der Großteil der Korrelationen ist statistisch nicht signifikant. In Tabelle 18 und Tabelle 19 sind deshalb nur die signifikanten Korrelationen aufgeführt.

Iteration 1 intern	**Korrelationskoeffi-zient**	**p-Wert**	**Signifikant**
Relationship - negative Label	0,29	0,18	Nein
Relationship - positive Label	-0,57	0,00	Ja
Relationship -neutrale Label	0,18	0,40	Nein
Task - negative Label	0,31	0,165	Nein
Task - positive Label	-0,49	0,01	Ja
Task - neutrale Label	0,17	0,43	Nein

Tabelle 18 Korrelationen Iteration 1 in internen Projekten

Signifikant sind die Korrelationen zwischen *Relationship-* und *Task-Konflikten*, und der Vergabe positiver Label interner Projekte. Beide sind negativ korreliert. *Relationship-Konflikte* und positive Label sind mit einem Korrelationskoeffizienten von -0,57 stark negativ korreliert. Mittelstark negativ korreliert sind mit -0,49 die *Task-Konflikte* und positive Label. Unter einem erhöhten Stressniveau in einem internen Projekt nehmen die positiven Label bei Erhöhung der *Relationship-* oder *Task-Konflikten* stark bzw. mittelstark ab.

Pause Extern	Korrelationskoeffizient	p-Wert	Signifikant
Relationship - negative Label	0,49	0,03	Ja
Relationship - positive Label	-0,04	0,86	Nein
Relationship -neutrale Label	-0,42	0,07	Nein
Task - negative Label	0,50	0,02	Ja
Task - positive Label	-0,14	0,56	Nein
Task - neutrale Label	-0,38	0,10	Nein

Tabelle 19 Korrelationen Pause in externen Projekten

Darüber hinaus gibt es in der Ruhephase in externen Projekten signifikante Korrelationen. Die Vergabe negativer Label korreliert mittelstark mit den *Relationship-* und den *Task-Konflikten*. Mit einem Korrelationskoeffizient von 0,49 nehmen die negativen Label mittelstark zu bei Erhöhung der *Relationship-Konflikte*. Einen minimal höheren Korrelationskoeffizienten, und damit einen starken Zusammenhang, weisen die negativen Label und die *Task-Konflikte* auf. Bei einem sehr niedrigen Stressniveau nehmen negative Label zu, wenn *Relationship-* oder *Task-Konflikte* zunehmen.

Zusammenfassend lassen sich folgende Aussagen für die Beantwortung der Forschungsfrage 3 formulieren:

1. Unterschiedliche Stressniveaus haben keinen nachweisbaren direkten Einfluss auf die Sentimentvergabe.
2. Im Allgemeinen haben *Relationship-Konflikte* keinen Einfluss auf die Labelvergabe.
3. In der Gesamtbetrachtung und in externen Projekten nimmt die Anzahl negativer Label schwach zu, wenn in der Gruppe *Task-Konflikte* zunehmen. Neutrale Label in externen Projekten hingegen nehmen schwach ab bei Zunahme von *Task-Konflikten*.
4. Bei erhöhtem Stressniveau in einem internen Projekt nehmen die positiven Label bei Erhöhung der *Relationship-* oder *Task-Konflikte* stark bzw. mittelstark ab.
5. Bei einem sehr niedrigen Stressniveau in einem externen Projekt nehmen negative Label mittelstark bzw. stark zu, wenn *Relationship-* und *Task-Konflikte* zunehmen.

5.8 Validity Threats

Für die Betrachtung der *Validity Threats* werden die *Threats* nach Wohlin [65] genutzt.

Teilnehmende an der Umfrage können Antworten zufällig gegeben haben, die von den wirklichen Antworten abweichen oder Ergebnisse verfälschen. Aus den Datensätzen ist nicht erkennbar, ob Antworten zufällig gegeben worden sind, da nicht mit absoluter Sicherheit festgestellt werden kann, dass die Antworten wirklich zufällig sind. Diese Möglichkeit ist ein *Threat to Conclusion Validity*. Es wurde in der Arbeit deswegen Durchschnittswerte berechnet und nicht mit Einzelwerten gerechnet.

Die Stichprobe ist in Teilen sehr klein (11 Teilnehmende, die an allen 4 Umfragen teilgenommen haben), wodurch Rückschlüsse auf die Grundgesamtheit nur bedingt bzw. nicht zuverlässig möglich sind. Es gibt jedoch auch große Stichproben (n=147). Ein Transfer auf die Grundgesamtheit ist nicht aussagekräftig möglich und es besteht ein *Threat to Conclusion Validity*.

Durch die Verwendung von Teilmengen aus zwei Datensätze ist es möglich, dass die beobachteten Ergebnisse und Zusammenhänge nur für die ausgewählten Sätze gelten. Die Ergebnisse könnten nicht allgemeingültig sein und bei Verwendung anderer Sätze oder anderer Datensätze nicht gelten. Aufgrund dieser Möglichkeit existiert ein *Threat to External Validity.*

Die Gruppe der Befragten stellt einen weiteren *Threat to External Validity* dar. Bei der Befragung von echten Softwareentwicklern können andere Ergebnisse zustande kommen. Studierende die an einem Softwareprojekt arbeiten, entwickeln zwar Software, unterscheiden sich jedoch in vielen Punkten (z.B. Alter und Arbeitsumgebung) von richtigen Softwareentwicklern.

Auch ist es möglich, dass bei der Auswertung der Daten Fehler aufgetreten sind, wodurch ein *Threat to Internal Validity* besteht. Es wurde beim Auswerten darauf geachtet Kontrollgrößen zu berechnen, um Fehler zu minimieren.

Durch die Verwendung bereits existierender Skalen [26] [7] [5] werden die *Threats to Validity* dieser Skalen übernommen. Da eine Skala selbst entwickelt wurde, ist es möglich, dass trotz Prüfung der Validität die Skala nicht akkurat die gewünschte Zielgröße abfragt. Es besteht ein *Threat to Construct Validity.*

Kapitel 6

Diskussion

In diesem Kapitel werden zunächst die zentralen Aussagen aus den Ergebnissen zu den Forschungsfragen wiederholt, um diese dann anschließend zu diskutieren und einen Ausblick zu geben.

6.1 Zusammenfassung der Ergebnisse

6.1.1 Forschungsfrage 1

Welchen Einfluss hat die Stimmung auf die Sentimentvergabe?

1. Eine Erhöhung der allgemeinen Stimmung führt zu einer schwachen Erhöhung der Anzahl an vergebenen positiven Label.
2. Der positive Affekt hat keinen signifikanten Einfluss auf die Sentimentvergabe.
3. Der negative Affekt hat einen schwachen signifikanten positiven Einfluss auf die Anzahl an vergebenen negativen Labeln und einen schwachen negativen Einfluss auf die positivere Bewertung der Aussagen.
4. Die Lebensumstände haben schwach negative signifikante Einflüsse auf die Anzahl negativer Label und die stark negativen Abweichungen.

6.1.2 Forschungsfrage 2

Wie entwickelt sich die Sentimentvergabe bei mehrfacher

Umfrageteilnahme?

1. Ab der zweiten Teilnahme werden neutrale Aussagen besser erkannt.
2. Mehrmaliges Teilnehmen erhöht den Fokus bei der Sentimentvergabe auf den Inhalt und den Klang.
3. Es werden zunehmend mehr klare Label zugeordnet und gleichzeitig mehr Ansätze und widersprüchliche Emotionen erkannt.
4. Bei mehrfacher Teilnahme gibt es mehr Abweichungen zu den vorherigen Antworten.
5. Eine höhere Reaktivität verändert nicht die Sentimentvergabe.

6.1.3 Forschungsfrage 3

Welchen Einfluss haben verschiedene Phasen des aktuellen Softwareprojekts und die Gruppendynamik in einem Softwareprojekt auf die Sentimentvergabe?

1. Unterschiedliche Stressniveaus haben keinen nachweisbaren direkten Einfluss auf die Sentimentvergabe.
2. Im Allgemeinen haben *Relationship-Konflikte* keinen Einfluss auf die Labelvergabe
3. In der Gesamtbetrachtung und in externen Projekten nimmt die Anzahl negativer Label schwach zu, wenn in der Gruppe *Task-Konflikte* zunehmen. Neutrale Label in externen Projekten hingegen nehmen schwach ab bei Zunahme von *Task-Konflikten*.
4. Bei erhöhtem Stressniveau in einem internen Projekt nehmen die positiven Label bei Erhöhung der *Relationship-* oder *Task-Konflikte* stark bzw. mittelstark ab.
5. Bei einem sehr niedrigen Stressniveau in einem externen Projekt nehmen negative Label mittelstark bzw. stark zu, wenn *Relationship-* und *Task-Konflikte* zunehmen.

6.2 Interpretation

Aus den Ergebnissen der ersten Forschungsfrage lässt sich ableiten, dass Emotionen bei der Sentimentvergabe maximal einen schwachen signifikanten Einfluss haben. Im Hinblick auf die subjektive Wahrnehmung von Aussagen und den daraus resultierenden Uneinigkeiten bei der Sentimentvergabe, können Emotionen als Auslöser nahezu ausgeschlossen werden. Subjektivität ist zwar ein Grund für abweichende Labelvergabe, sie wird jedoch nicht beeinflusst von Emotionen. Eine Erklärung für diese Abweichungen lässt sich nicht mithilfe der Betrachtung von Emotionen finden. Die subjektive Wahrnehmung ist demnach unberührt von Emotionen und Aussagen werden losgelöst von diesen betrachtet. Der schwache Zusammenhang zwischen Zufriedenheit mit den Lebensumständen und den stark negativen Abweichungen könnte bei starken Veränderungen in der Zufriedenheit die Sentimentvergabe beeinflussen. Tendenziell werden Aussagen negativer aufgefasst bei stark sinkender Zufriedenheit. Dies könnte eine Erklärung für schwache Abweichungen zwischen mehreren Entwicklern in der Sentimentvergabe sein.

Auffällig ist der Widerspruch, dass mehr klare Label vergeben werden, bei mehr widersprüchlichen Emotionen und mehr Ansatzpunkten zur Sentimentvergabe. Für die Unstimmigkeiten bei der Sentimentvergabe sind widersprüchliche Emotionen oder viele Ansatzpunkte in den Aussagen anscheinend nicht ausschlaggebend. Sie sorgen für Unsicherheit, aber nicht dafür, dass das Label nicht klar ist.

Der starke Fokus auf Ton und Inhalt der Aussagen könnte darauf hindeuten, dass durch unterschiedliche Wahrnehmung dieser Punkte die Sentimentvergabe stärker beeinflusst wird. Die unterschiedliche Wahrnehmung dieser Punkte könnte bedingt sein durch widersprüchliche Emotionen oder zu viele Ansatzpunkte in einer Aussage. Wahrgenommene Emotionen könnten einen stärkeren Einfluss haben als die eigenen Emotionen bei der Vergabe von Sentimenten. Dies würde auch dafür sprechen, dass die Bewertung von Aussagen losgelöst von den eigenen Emotionen erfolgt.

Darüber hinaus scheinen Label eher abzuweichen, wenn häufiger über die gleiche Aussage entschieden wird.

Im Zusammenhang zu den zunehmend erkannten Emotionen und Ansatzpunkten, könnten diese bei mehrfacher Bewertung zum Überdenken der ersten Meinung führen. Werden Aussagen mehrfach getätigt, könnte sich die Wahrnehmung verändern bzw. die vorherige Wahrnehmung überdacht werden. Stützen tun dies die Fleiss´ Kappa Werte, die zwar eine moderate Übereinstimmung zeigen, jedoch stark schwanken und oft auch sehr niedrige Werte annehmen.

Konflikte haben den größten gemessenen Einfluss auf die Sentimentvergabe. Dabei haben Konflikte über die Aufgabe einen höheren Einfluss auf die Sentimentvergabe. Sollen Aussagen möglichst positiv bewertet werden, sollten Task Konflikte vermieden werden. Interessant sind auch die Unterschiede zwischen internen und externen Projekten mit in Teilen unterschiedlich gepolten Korrelationen. Es scheint einen Unterschied zu machen, ob ein Projekt von einem externen oder internen Kunden betreut wird. Eventuell könnte dies an der Größe der Stichproben liegen (69 interne zu 77 externe). Externe Projekte sind näher an echten Softwareprojekten, Zusammenhänge aus externen Projekten könnten deshalb eher auf echte Softwareprojekte übertragbar sein. Durch das Arbeiten für einen Externen können Konflikte auch anders wahrgenommenen werden, als bei internen Projekten, welche eher als eine Universitätsveranstaltung gesehen werden könnten. Auch könnte Stress unterschiedlich wahrgenommen werden zwischen den Projektgruppen. Bei der Sentimentvergabe macht die Projektart einen messbaren Unterschied aus.

6.3 Ausblick

Die genauen Einflussfaktoren, die bei der Vergabe eines Sentiments wichtig sind, sind durch die vielen möglichen Faktoren nur schwer exakt zu bestimmen. Einen ersten Ansatz liefert diese Arbeit. Die Ergebnisse müssen jedoch weiter verifiziert und die gefundenen Zusammenhänge unter anderen Umständen (z.B. andere Sätze oder hauptberufliche Entwickler als Zielgruppe) geprüft werden. Es kann nicht ausgeschlossen werden, dass die gefundenen Zusammenhänge nur in dem beobachteten Rahmen gelten und unter anderen Umständen andere Zusammenhänge gelten würden. Auch sollten neben den in dieser Arbeit abgefragten Faktoren weitere Faktoren, wie z.B. kulturelle und soziale Einflüsse oder das Bildungsniveau, untersucht werden, um die wesentlichen Einflussfaktoren auf die Sentimentvergabe zu finden und mit einer größeren Stichprobe zu verifizieren.

Besonders Zusammenhänge zwischen Projektart und Sentimentvergabe sollten weiter untersucht werden. Dabei sollten eventuell auch weitere Konflikte unterschieden werden, um die genaue Ursache für die Zusammenhänge zwischen Konflikten und Sentimentvergabe zu finden.

Weitere Forschung nach den Einflüssen auf die Sentimentvergabe könnte weitere spannende Erkenntnisse liefern und Unterschiede in der Sentimentvergabe weiter erklären.

Anhangverzeichnis

Anhang

Primäre Emotionen	Sekundäre Emotionen	Tertiäre Emotionen
Liebe („love“)	Zuneigung („affection“)	Verehrung („adoration“), Zuneigung („affection“), Liebe („love“), Vorliebe („fondness“), Mögen („liking“), Anziehung („attraction“), Fürsorge („caring“), Zärtlichkeit („tenderness“), Mitgefühl („compassion“), Empfindsamkeit („sentimentality“)
	Begierde („lust“)	Erregung („arousal“), Verlangen („desire“), Begierde („lust“), Leidenschaft („passion“), Vernarrtheit („infatuation“)
	Sehnsucht („longing“)	Sehnsucht („longing“)
Freude („joy“)	Fröhlichkeit („cheerfulness“)	Amüsement („amusement“), Glückseligkeit („bliss“), Frohsinn („cheerfulness“), Fröhlichkeit („gaiety“), Freude („glee“), Lustigkeit („jolliness“), Heiterkeit („joviality“), Freude („joy“), Entzückung („delight“), Genuss („enjoyment“), Fröhlichkeit („gladness“), Glück („happiness“), Jubel („jubilation“), Hochstimmung („elation“), Zufriedenheit („satisfaction“), Ekstase („ecstasy“), Euphorie („euphoria“)
	Begeisterung („zest“)	Enthusiasmus („enthusiasm“), Eifer („zeal“), Begeisterung („zest“), Aufregung („excitement“), Nervenkitzel („thrill“), Hochgefühl („exhilaration“)
	Zufriedenheit („contentment“)	Zufriedenheit („contentment“), Vergnügen („pleasure“)
	Stolz („pride“)	Stolz („pride“), Triumphgefühl („triumph“)
	Optimismus („optimism“)	Begierde („eagerness“), Hoffnung („hope“), Optimismus („optimism“)
	Intensives Entzücken („enthrallment“)	intensives Entzücken („enthrallment“), Entzücken („rapture“)
	Relief („relief“)	Erleichterung („relief“)
Überraschung („surprise“)	Überraschung („surprise“)	Verblüffung („amazement“), Überraschung („surprise“), Staunen („astonishment“)

Ärger („anger“)	Irritationen („Irritation“)	Verschlimmerung („aggravation“), Reizung („irritation“), Aufregung („agitation“), Ärger („annoyance“), Gereiztheit („grouchiness“), Grantigkeit („grumpiness“)
	Verzweiflung („exasperation“)	Verzweiflung („exasperation“), Frustration („frustration“)
	Wut („rage“)	Ärger („anger“), Wut („rage“), Empörung („outrage“), Rage („fury“), Zorn („wrath“), Feindseligkeit („hostility“), Wildheit („ferocity“), Bitterkeit („bitterness“), Hass („hate“), Verabscheuung („loathing“), Verachtung („scorn“), Bosheit („spite“), Rachsucht („vengefulness“), Abneigung („dislike“), Verbitterung („resentment“)
	Ekel („disgust“)	Ekel („disgust“), Abscheu („revulsion“), Verachtung („contempt“)
	Neid („envy“)	Neid („envy“), Eifersucht („jealousy“)
	Qual / Schmerz („torrment“)	Qual / Schmerz („torrment“)
Trauer („Sadness“)	Leid („suffering“)	Heftiger Schmerz („agony“), Leid („suffering“), Verletzung („hurt“), Schmerz („anguish“)
	Trauer („sadness“)	Niedergeschlagenheit („depression“), Verzweiflung („despair“), Hoffnungslosigkeit („hopelessness“), Düsterkeit („gloom“), Trostlosigkeit („glumness“), Trauer („sadness“), Unglücklichkeit („unhappiness“), Kummer („grief“), Trauer („sorrow“), Not („woe“), Elend („misery“), Melancholie („melancholy“)
	Enttäuschung („disappointment“)	Bestürzung („dismay“), Enttäuschung („disappointment“), Missfallen („displeasure“)
	Scham („shame“)	Schuld („guilt“), Scham („shame“), Bedauern („regret“), Reue („remorse“)
	Vernachlässigung („neglect“)	Entfremdung („alienation“), Isolation („isolation“), Vernachlässigung („neglect“), Einsamkeit („loneliness“), Ablehnung („rejection“), Heimweh („homesickness“), Niederlage („defeat“), Niedergeschlagenheit („dejection“), Unsicherheit („insecurity“), Verlegenheit („embarrassment“), Demütigung („humiliation“), Beleidigung („insult“)
	Sympathie („sympathy“)	Mitleid („pity“), Verständnis („sympathy“)
Angst („Fear“)	Schrecken („horror“)	Alarmiertheit („alarm“), Schock („shock“), Angst („fear“), Schreck („fright“), Horror („horror“), schreckliche Angst („terror“), Panik („panic“), Hysterie („hysteria“), Kränkung („mortification“)

Nervosität („nervousness“)	Angst („anxiety“), Nervosität („nervousness“), Angespanntheit („tenseness“), Unbehaglichkeit („uneasiness“), Verständnis („apprehension“), Sorge („worry“), Bedrängnis („distress“), große Angst („dread“)

Anhang 1 Emotionsmodell nach Parrott[44]

Danke, dass du dir die Zeit nimmst an meiner Umfrage teilzunehmen. Ich schreibe gerade meine Bachelor Arbeit am Fachgebiet Software Engineering zu dem Thema "Analyse von Einflüssen in der Sentimenterkennung von Entwicklern".
Alle Daten und Antworten in dieser Umfrage werden anonymisiert und es werden keine Rückschlüsse auf deine Personalien möglich sein.
In der Umfrage geht es zuerst darum deine allgemeine und deine aktuelle Stimmung zu erfassen. Anschließend wirst du deine momentane Lebenssituation bewerten. Im zweiten Teil werden dir Sätze auf Englisch präsentiert, denen du spontan ein Label (positiv, neutral, negativ) zuordnen sollst. Zuletzt wird noch gefragt wonach du gelabelt hast. In meiner Arbeit werde ich dann analysieren ob die analysierten Daten (Stimmung und Zufriedenheit mit den Lebensumständen) einen Einfluss auf die Labelzuordnung haben.
Im Laufe des Softwareprojektes werde ich die Umfrage wiederholen, damit ich verschiedene Zeitpunkte eines Projektes betrachten kann. Dafür ist es wichtig zu wissen ob zwei Datenpunkte von derselben Person stammen. Dies geschieht über einen Code, den du vor Beginn der eigentlichen Umfrage erzeugst. Die aus den anonymisierten Daten folgende Analyse wird im Rahmen meiner Bachelor Arbeit und weiteren Publikationen verwendet.
Wenn du noch Fragen zu der Umfrage hast oder während der Umfrage Probleme auftreten, kannst du mir gerne eine E-Mail schreiben:
j.martensen@stud.uni-hannover.de
Die Umfrage dauert ca. 10 Minuten.
Vielen Dank fürs Teilnehmen!!!

Anhang 2 Willkommenstext zur Umfrage

[44] Andreas Aschenbrenner (2019): Emotionserkennung bei Nachrichtenkommentaren mittels Convolutional Neural Networks und Label Propagationsverfahren, Dissertation, S. 63-64

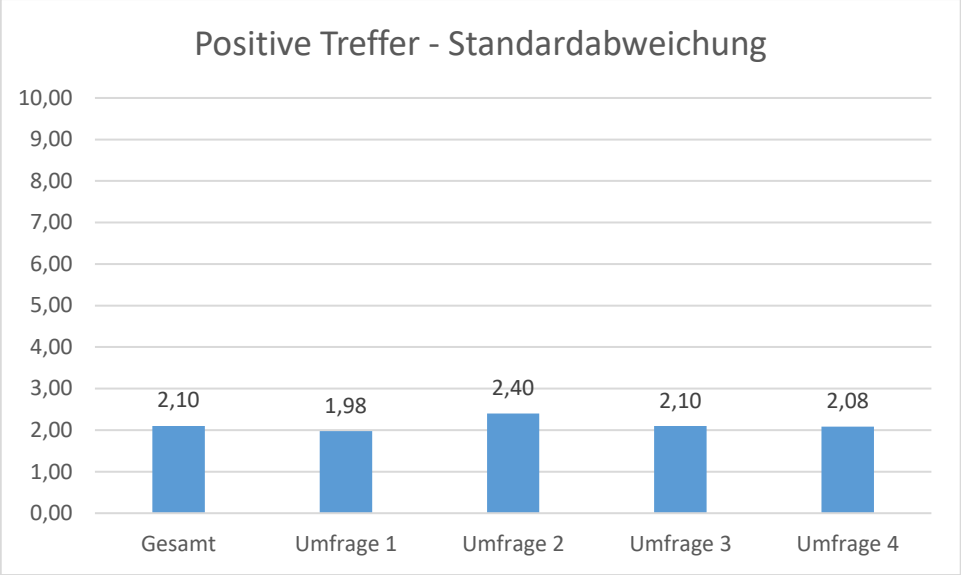
Frage	Auswahlmöglichkeiten
Was war für dich ausschlaggebend für die Verteilung der Labels?	• negativer/ neutraler/ positiver Inhalt (z.B. "mir gefällt dein Code" - positiv) • negativer/ neutraler/ positiver Klang (z.B. "Bitte nicht schon wieder..." - negativ) • negative/ neutrale/ positive Smileys bzw. Emoticons (z.B. :) - positiv) • Sonstiges: Freitextfeld
Wie sicher warst du dir bei der Vergabe der einzelnen Label?	5-Punkt Skala
Warum warst du dir unsicher bei der Vergabe der Label?	• Es gab zu viele mögliche Ansatzpunkte eine Auswahl zu treffen (Ton, Inhalt, Emoticons, etc.) • Es wurden widersprüchliche Emotionen erkannt (z.B. negativer Ton mit positiven Emoticons) • Label war nicht klar zu bestimmen, da keine der Optionen für die Aussage richtig zutreffen war • Keine Unsicherheiten • Sonstiges: Freitextfeld

Anhang 3 Frageblock 8 mit Auswahlmöglichkeiten aus der Umfrage

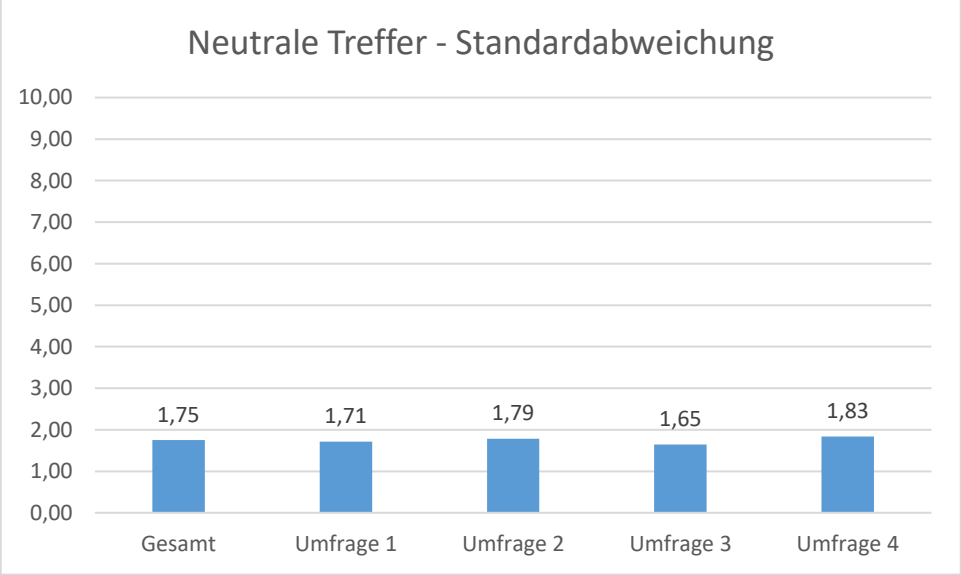
Item	Varianz
Soziales Umfeld	1,10008387
Gesundheit	1,21917808
Studium	1,00680272
Arbeit	1,34792756
Sicherheit	1,4024864
Wohnsituation	1,07743919
Selbstbestimmtheit	1,09766098
Selbstverwirklichung	0,94662258
Finanzen	1,68501921
Summe (a)	10,8832206

Cronbachs Alpha	(N/(N-1)) x (1-(a/b))
Varianz der Summen (b)	35,93570031
Fragen (N)	9
Cronbachs Alpha	0,78

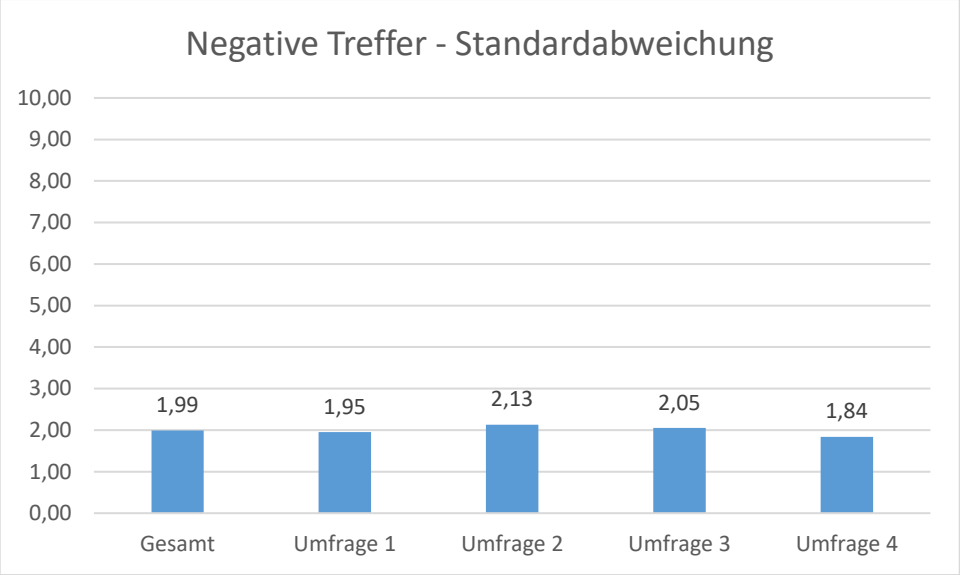
Anhang 4 Cronbachs Alpha Berechnung



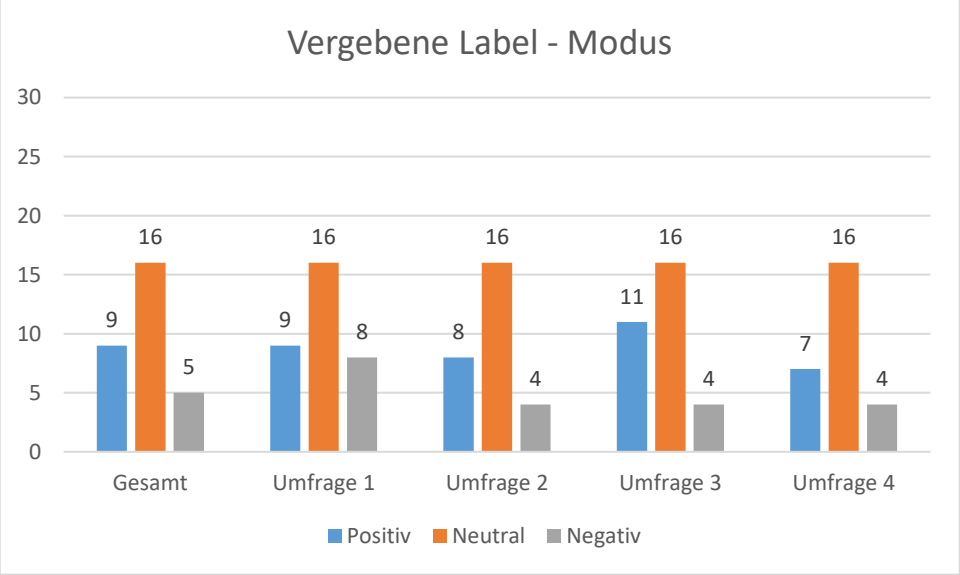
Anhang 5 Positive Treffer Standardabweichung



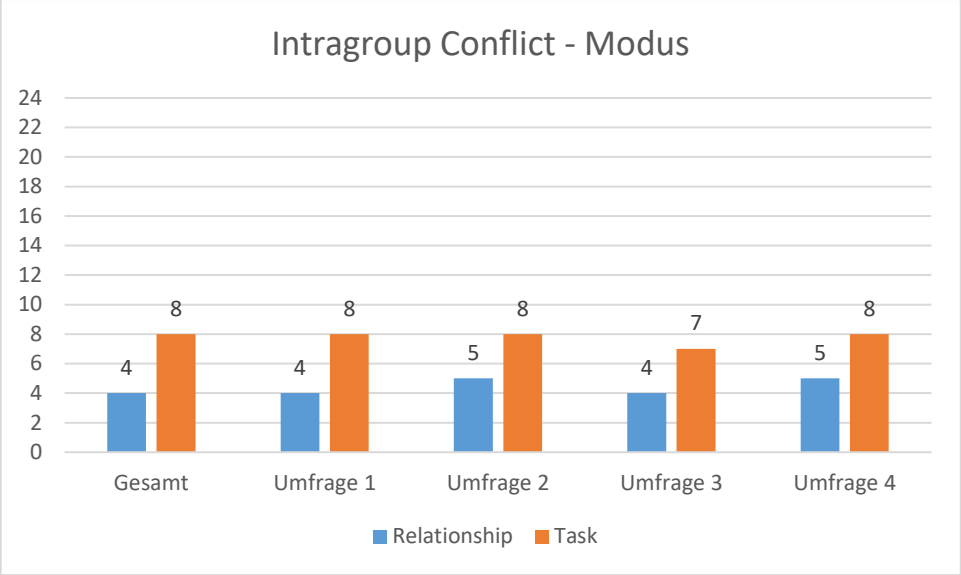
Anhang 6 Neutrale Treffer Standardabweichung



Anhang 7 Negative Treffer Standardabweichung



Anhang 8 Vergebene Label Modus



Anhang 9 Intragroup Conflict Modus

1. Manche Smileys können bei bestimmten Sätzen negativ wirken. „:)“ zum Beispiel
2. Wie ich mir vorstelle, dass ich es im Moment in einem Gespräch rüberbringen würde.
3. Was könnte der Autor hier gemeint haben?
4. "stupid me" zum Beispiel klang positiv, weil es darauf schließen lässt, dass der Author eine Lösung entdeckt/ gezeigt bekommen hat, und euphorisch darüber ist. Auch, wenn es negativ erstmal aussieht, sehe ich es einfach als persönliche Rhetorik an.
5. Worte in denen ich bei der Ausprache einen Unterton vermute wie "Hoffentlich wird das beim nächsten mal besser..."
6. ob die Kommentare konstruktiv und sachlich sind anstatt emotional und persönlich zu werden
7. Klang ja, aber er kann auch eine Situation aufheitern, falls er zum Beispiel einen simplen Fehler gemacht hat ist "OMG stupid me" nicht negativ datiert sondern sorgt eventuell sogar für den einen oder anderen Lacher.
8. Produktiv oder nicht
9. Hilfreich oder nicht hilfreich
10. Möglicher Kontext
11. (Legitime) Fragen habe ich als neutral wahrgenommen..
12. Rechtschreibung, Grammatik, Satzstellung(ein Satz mit 4 Kommata sollte mehrere kurze Sätze sein). Manche Textschnipsel laßen sich passiv agressiv
13. Was mir schonmal hilfreich war - positiv

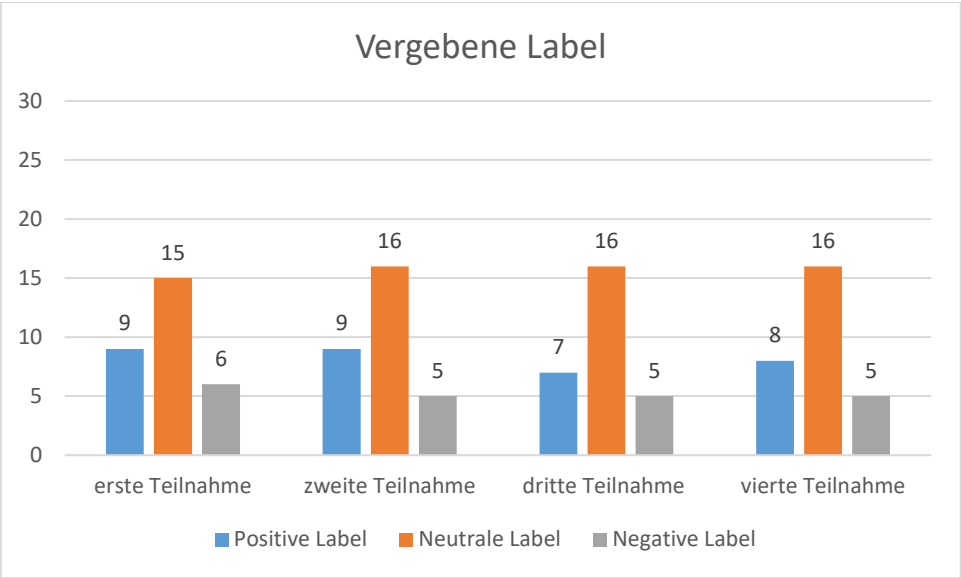
Anhang 10 Sonstige Gründe der Studierenden bei der Labelvergabe

1. Je nach Kontext kann sich die Bedeutung der Aussagen und die durch sie vermittelte Stimmung stark unterscheiden.
2. Kontext ist wichtig
3. Kontext fehlt
4. sehr generelle Möglichkeiten
5. sehr generell gehalten
6. Kontext fehlt
7. Kontext fehlte
8. unklarer Kontext der Äußerung (z.B. in welchem Verhältnis die Personen zueinander stehen)
9. Kontext ist wie bei jeder Kommunikation entscheidend. Manche der Labels können je nach Kontext auch ganz anders verteilt werden.

Anhang 11 Sonstige Gründe der Studierenden für Unsicherheiten bei der Labelvergabe

Korrelation positiver Affekt zu	Korrelationskoeffizient	p-Wert	Signifikant
Treffer	0,00	1,00	Nein
Trefferquote	0,00	1,00	Nein
Positive Treffer	0,10	0,21	Nein
Neutrale Treffer	-0,05	0,56	Nein
Negative Treffer	-0,01	0,88	Nein
Positive Label	0,14	0,09	Nein
Neutrale Label	-0,04	0,62	Nein
Negative Label	-0,02	0,77	Nein
Positiver bewertet	0,09	0,29	Nein
Negativer bewertet	-0,09	0,30	Nein
Stark positive Abweichung	0,08	0,34	Nein
Stark negative Abweichung	-0,05	0,58	Nein

Anhang 12 Korrelationen zum Positiven Affekt



Anhang 13 Vergebene Label

Rater	2-fache Teilnahme Fleiss´ Kappa
1	0,43
2	0,64
3	0,67
4	0,48
5	0,71
6	0,65
7	-0,03
8	0,69
9	0,28
10	0,62
11	0,62
12	0,34
13	0,78
14	0,42
15	0,82
16	0,44
17	0,52
18	0,67
19	0,52
20	-0,01
21	0,34
22	0,49
23	0,21
24	0,74
25	-0,10
26	0,59
27	0,64
28	-0,03
29	0,21
30	0,37
31	0,66
32	0,69
33	0,62
Mittelwert	0,47

Anhang 14 2-fache Telnahme Fleiss´ Kappa Werte

Rater	3-fache-Teilnahme Fleiss´ Kappa
1	0,40
2	0,60
3	0,54
4	0,52
5	0,65
6	0,74
7	0,39
8	0,63
9	0,40
10	0,51
11	0,85
12	0,51
13	0,67
14	0,53
15	0,08
16	0,59
17	0,65
18	0,09
19	0,53
20	0,59
21	0,68
Mittelwert	0,53

Anhang 15 3-fache Teilnahme Fleiss´ Kappa Werte

Rater	4-fache Teilnahme Fleiss´ Kappa
1	0,63
2	0,34
3	0,53
4	0,61
5	0,39
6	0,50
7	0,78
8	0,63
9	0,55
10	0,61
11	0,09
Mittelwert	0,51

Anhang 16 4-fache Teilnahme Fleiss´ Kappa Werte

	Umfrage 1	Umfrage 2	Umfrage 3	Umfrage 4
Konsistente Antworten	Positv	Positiv	Positiv	Positiv
Leichte Abweichung	Positiv	Positiv	Positiv	Neutral
Mittlere Abweichung	Positiv	Positiv	Neutral	Neutral
Starke Abweichung	Positiv	Positiv	Neutral	Negativ

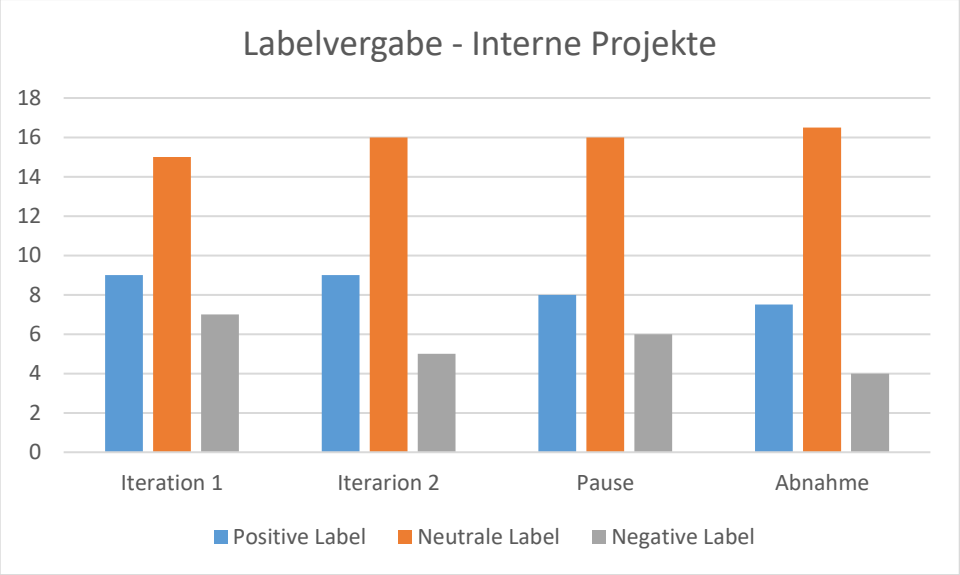
Anhang 17 Beispiel Abweichungen

Korrelation	Korrelationskoeffizient	p-Wert	Signifikant
Fleiss´ Kappa zu Reaktivität	-0,08	0,81	Nein
Fleiss´ Kappa zu konsistenten Antworten	0,58	0,06	Nein
Reaktivität zu konsistente Antworten	0,15	0,67	Nein
Reaktivität zu leichte Abweichung	-0,12	0,74	Nein
Reaktivität zu mittlere Abweichung	-0,27	0,43	Nein
Reaktivität zu starke Abweichung	0,45	0,17	Nein

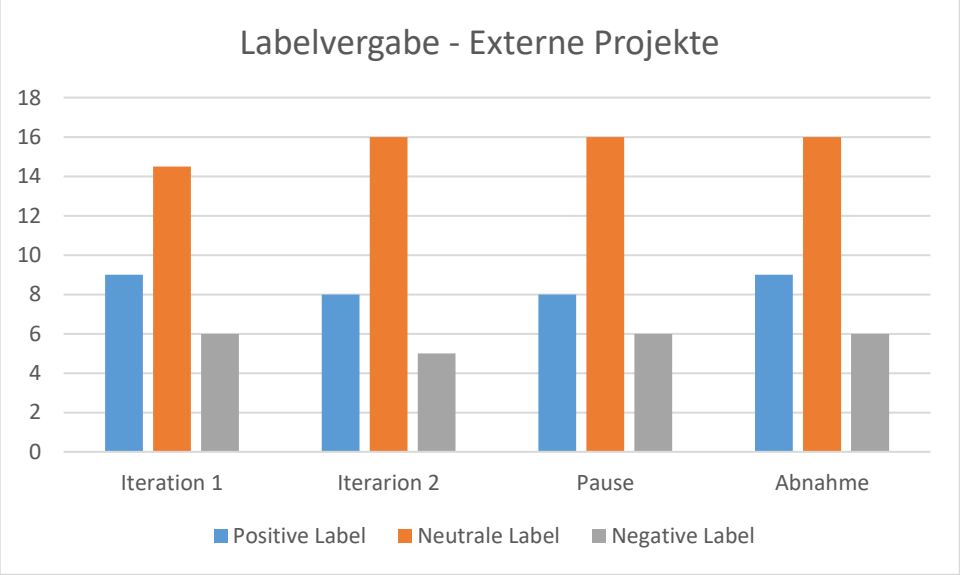
Anhang 18 Korrelationen 4-fache Teilnahme

Korrelation	Korrelationskoeffizient	p-Wert	Signifikant
Fleiss´ Kappa zu Reaktivität	-0,24	0,21	Nein
Fleiss´ Kappa zu konsistenten Antworten	0,37	0,05	Ja
Reaktivität zu konsistente Antworten	-0,24	0,20	Nein
Reaktivität zu Abweichung	0,24	0,20	Nein

Anhang 19 Korrelationen 2-fache Teilnahme



Anhang 20 Labelvergabe - Interne Projekte



Anhang 21 Labelvergabe - Externe Projekte

Literaturverzeichnis

[1] T. Ahmed, A. Bosu, A. Iqbal and S. Rahimi. "SentiCR: A customized sentiment analysis tool for code review interactions," *2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE),* 2017, pp. 106-111, doi: 10.1109/ASE.2017.8115623, 2017.

[2] A. Aschenbrenner. Emotionserkennung bei Nachrichtenkommentaren mittels Convolutional Neural Networks und Label Propagationsverfahren. Dissertation, *LMU München: Fakultät für Psychologie und Pädagogik*, 2019.

[3] C. Becker und E. Huber. Die Bilanz des (Miss)-Erfolges in IT-Projekten : harte Fakten und weiche Faktoren, Ludwigsburg: Pentaeder, 2008.

[4] Bertelsmann Stiftung. Was ist Ihnen für Ihre Lebensqualität wichtig?, 2010.

[5] G. Bohner und N. Schwarz. Die Stimmungsskala. *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*. https://doi.org10.6102/zis53, 2014.

[6] K. Bosch. Formelsammlung Statistik, 2003.

[7] B. Breyer und M. Bluemke. Deutsche Version der Positive and Negative Affect Schedule PANAS (GESIS Panel). *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS).* https://doi.org/10.6102/zis242, 2016.

[8] Bundesregierung (Hrsg.). Bericht der Bundesregierung zur Lebensqualität in Deutschland, 2020.

[9] J. Cohen. Statistical power analysis for the behavioral sciences (2nd ed.), Routledge, https://doi.org/10.4324/9780203771587, 1988.

[10] GitHub. SentiStrengthDE, URL: https://raw.githubusercontent.com/OFAI/SentiStrength_DE/main/sentistrength_de.zip, vom 14.01.2022.

[11] Gottfried Wilhelm Leibniz Universität Hannover - Fachgebiet Software Engineering. Software Projekt, URL: https://www.pi.uni-hannover.de/de/se/lehre/swp/, vom 12.02.2022.

[12] D. Graziotin, X. Wang, P. Abrahamsson. Happy software developers solve problems better: psychological measurements in empirical software engineering. PeerJ 2:e289 https://doi.org/10.7717/peerj.289, 2014.

[13] D. Graziotin and F. Fagerholm. Happiness and the productivity of software engineers, 2019.

[14] E. Guzman, D. Azócar, Y. Li. Sentiment analysis of commit comments in GitHub: An empirical study. *11th Working Conference on Mining Software Repositories, MSR 2014 - Proceedings*. 10.1145/2597073.2597118, 2014.

[15] K. Hurrelmann und M. Richter. Determinanten von Gesundheit, doi:10.17623/BZGA:224-i008-1.0, 2018.

[16] 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 *Proceedings of the 3rd International Workshop on Emotion Awareness in Software Engineering (SEmotion '18)*. Association for Computing Machinery, New York, NY, USA, 55–61. DOI:https://doi.org/10.1145/3194932.3194938, 2018.

[17] M. Islam and M. Zibran. SentiStrength-SE: Exploiting Domain Specificity for Improved Sentiment Analysis in Software Engineering Text. *Journal of Systems and Software*. 145. 10.1016/j.jss.2018.08.030, 2018.

[18] C. Izard. Die Emotionen des Menschen: eine Einführung in die Grundlagen der Emotionspsychologie, 1999.

[19] R. Jongeling, S. Datta and A. Serebrenik. "Choosing your weapons: On sentiment analysis tools for software engineering research," *2015 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pp. 531-535, doi: 10.1109/ICSM.2015.7332508, 2015.

[20] P. Kleinginna and A. Kleinginna. A categorized list of emotion definitions, with suggestions for a consensual definition. Motiv Emot 5, 345–379. https://doi.org/10.1007/BF00992553, 1981.

[21] M. Obaidi and J. Klünder. Development and Application of Sentiment Analysis Tools in Software Engineering: A Systematic Literature Review, 2021.

[22] S. Kotz, C. Read, N. Balakrishnan, B. Vidakovic, N. Johnson. Ecyclopedia of Statistical Sciences, doi: 10.1002/0471667196, 2004.

[23] M. Kuhrmann, P. Tell, J. Klünder, R. Hebig, S. Licorish, and S. MacDonell. HELENA Stage 2 Results. https://doi.org/10.13140/RG.2.2.14807.52649, 2018.

[24] R. Lazarus, C. Smith. Emotion and Adaptation. 10.2307/2075902, 1990.

[25] B. Pang and L. Lee. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval. 2. 1-135. 10.1561/1500000011, 2008.

[26] N. Lehmann-Willenbrock, A. Grohmann and S. Kauffeld. Task and Relationship Conflict at Work: Construct Validation of a German Version of Jehn's Intragroup Conflict Scale. *European Journal of Psychological Assessment - EUR J PSYCHOL ASSESS. 27*. 171-178. 10.1027/1015-5759/a000064, 2011.

[27] B. Lin, F. Zampetti, G. Bavota, M. Di Penta, M. Lanza and R. Oliveto. "Sentiment Analysis for Software Engineering: How Far Can We Go?," *2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE)*, pp. 94-104, doi: 10.1145/3180155.3180195, 2018.

[28] A. Mehrabian and J. Russell. An approach to environmental psychology, The MIT Press, 1974
.

[29] A. Moreno and C. Iglesias. Sentiment Analysis for Social Media. Applied Sciences. 9. 5037. 10.3390/app9235037, 2019.

[30] N. Novielli, F. Calefato, F. Lanubile, F. Maiorano. Sentiment Polarity Detection for Software Development. *Empirical Software Engineering*. 23. 10.1007/s10664-017-9546-9, 2018.

[31] N. Nicole, F. Calefato, F. Lanubile. A gold standard for emotion annotation in stack overflow. 14-17. 10.1145/3196398.3196453, 2018.

[32] N. Novielli, F. Calefato, D. Dongiovanni, D. Girardi, F. Lanubile. A gold standard for polarity of emotions of software developers in GitHub, 2020.

[33] N. Novielli, F. Calefato, D. Dongiovanni, D. Girardi, 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 (MSR 2020)*, 2020.

[34] Novielli, Nicole & Calefato, Fabio & Lanubile, Filippo. Towards discovering the role of emotions in stack overflow. *6th International Workshop on Social Software Engineering*, SSE 2014 - Proceedings. 33-36. 10.1145/2661685.2661689, 2014.

[35] M. Ortu, A. Murgia, G. Destefanis, P. Tourani, R. Tonelli, M. Marchesi, B. Adams. The Emotional Side of Software Developers in JIRA. 10.1145/2901739.2903505, 2016.

[36] M. Ortu, B. Adams, G. Destefanis, P. Tourani, M. Marchesi, R. Tonelli. Are Bullies More Productive? Empirical Study of Affectiveness vs. Issue Fixing Time. 10.1109/MSR.2015.35, 2015.

[37] C. Osgood, G. Suci, P. Tannenbaum. The measurement of meaning. Univer. Illinois Press, 1957.

[38] J. Otto, H. Euler, und H. Mandl. Begriffsbestimmungen. In J. Otto, H. A. Euler und H. Mandl (Hrsg.), Handbuch Emotionspsychologie (S. 11-18). Weinheim: Beltz, PsychologieVerlagsUnion, 2000.

[39] W. Parrott, (Ed.). Emotions in social psychology: Essential readings. *Psychology Press*, 2001.

[40] R. Plutchik. A psychoevolutionary theory of emotions. *Social Science Information Sur Les Sciences Sociales - SOC SCI INFORM*. 21. 529-553. 10.1177/053901882021004003, 1982.

[41] P. Ramsey. Critical Values for Spearman's Rank Order Correlation. *Journal of Educational and Behavioral Statistics* - J EDUC BEHAV STAT. 14. 245-253. 10.3102/10769986014003245, 1989.

[42] J. Russell. Core Affect and the Psychological Construction of Emotion. *Psychological review*. 110. 145-72. 10.1037//0033-295X.110.1.145, 2003.

[43] J. Russell. A Circumplex Model of Affect. *Journal of Personality and Social Psychology*. 39. 1161-1178. 10.1037/h0077714, 1980.

[44] P. Schaer, P. Mayr, P. Mutschke. Implications of Inter-Rater Agreement on a Student Information Retrieval Evaluation, 2010.

[45] D. Schiller. Untersuchung von Zusammenhängen zwischen Stimmungen und Interaktionen von Meetings in Softwareprojekten, Bachelorthesis, *Gottfried Wilhelm Leibniz Universität Hannover*, 2021.

[46] K. Schneider, O. Liskin, H. Paulsen, S. Kauffeld. Media, Mood, and Meetings. ACM Transactions on Computing Education. 15. 1-33. 10.1145/2771440, 2015.

[47] K. Schneider, J. Klünder, F. Kortum, L. Handke, J. Straube, S. Kauffeld. Positive Affect through Interactions in Meetings: The Role of Proactive and Supportive Statements. *Journal of Systems and Software*. 143. 10.1016/j.jss.2018.05.001, 2018

[48] P. Schulze. Beschreibende Statistik, Bd. 4. Auflage, 2000.

[49] P. Shaver, J. Schwartz, D. Kirson, C. O'Connor. Emotion Knowledge: Further Exploration of a Prototype Approach. *Journal of personality and social psychology*. 52. 1061-86. 10.1037//0022-3514.52.6.1061, 1987.

[50] J. Shen, O. Baysal, M. Shafiq. (2019). Evaluating the Performance of Machine Learning Sentiment Analysis Algorithms in Software Engineering. 1023-1030. 10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00185, 2019.

[51] M. Spiegel und L. Stephens, Statistik - Das Lehrbuch, 2003.

[52] Statista GmbH. Definition Lageparameter, URL: https://de.statista.com/statistik/lexikon/definition/80/lageparameter/, vom 01.02.2022.

[53] T. Städtler, Lexikon der Psychologie. Wörterbuch, Handbuch, Studienbuch, 2003.

[54] M. Taboada, Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*. 2. 10.1146/annurev-linguistics-011415-040518, 2016.

[55] The Standish Group International, Inc. Chaos Report 2015, URL: https://www.standishgroup.com/sample_research_files/CHAOSReport2015-Final.pdf, 2015.

[56] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, A. Kappas. Sentiment Strength Detection in Short Informal Text. *Journal of the American Society for Information Science and Technology.* 61. 2544-2558. 10.1002/asi.21416, 2010.

[57] M. Thelwall, K. Buckley, G. Paltoglou. Sentiment Strength Detection for the Social Web. *Journal of the American Society for Information Science and Technology*. 63. 163-173. 10.1002/asi.21662, 2012.

[58] P. Tourani, Y. Jiang, B. Adams. Monitoring sentiment in open source mailing lists: exploratory study on the apache ecosystem. In *Proceedings of 24th Annual International Conference on Computer Science and Software Engineering (CASCON '14).* IBM Corp., USA, 34–44, 2014.

[59] K. Tymann, M. Lutz, P. Palsbröker, C. Gips, Carsten. GerVADER -A German adaptation of the VADER sentiment analysis tool for social media texts, 2019.

[60] D. Ulich und P. Mayring. Psychologie der Emotionen, 1992.

[61] B. Underwood, W. Froming. The Mood Survey: A Personality Measure of Happy and Sad Moods. J*ournal of Personality Assessment - J PERSONAL ASSESS.* 44. 404-414. 10.1207/s15327752jpa4404_11, 1980
.

[62] D. Watson, A. Tellegen. Toward a Consensual Structure of Mood*. Psychological bulletin.* 98. 219-35. 10.1037/0033-2909.98.2.219, 1985.

[63] M. Wirtz. Dorsch - Lexikon der Psychologie, 2016.

[64] M. Wirtz. Dorsch - Lexikon der Psychologie, 2017.

[65] C. Wohlin, P. Runeson, M. Höst, M. Ohlsson, B. Regnell and A. Wesslén, Experimentation in Software Engineering. 10.1007/978-3-642-29044-2_10, 2012.

[66] T. Zhang, B. Xu, F. Thung, S. Haryono, D. Lo, L. Jiang. Sentiment Analysis for Software Engineering: How Far Can Pre-trained Transformer Models Go?. 70-80. 10.1109/ICSME46990.2020.00017, 2020.