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

Evaluation der wahrnehmungsbasierten Zuordnung von Entwicklern zur Individualisierung der Stimmungsanalyse in Softwareprojekten

Evaluating Perception-Based Assignment of Developers to Individualize Sentiment Analysis in Software Projects

Bachelorarbeit

im Studiengang Informatik

von

Laura Elgert

Prüfer: Prof. Dr. rer. nat. Kurt Schneider
Zweitprüfer: Dr. rer. nat. Jil Ann-Christin Klünder
Betreuer: Marc Herrmann, Alexander Specht

Hannover, 28.12.2023

Erklärung der Selbstständigkeit

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

Hannover, den 28.12.2023

Laura Elgert

Zusammenfassung

Im Gebiet der Softwareentwicklung werden die Anforderungen an die zu erstellende Software immer größer. Infolgedessen ist die Zusammenarbeit in einem Team durch die zuständigen Software-Entwickler umso wichtiger, um gemeinsam im jeweiligen Softwareprojekt bessere Ergebnisse zu erzielen. Neben der Implementierung von Funktionen ist die Kommunikation der Entwickler untereinander von großer Bedeutung, die nicht nur direkt, sondern oft auch in der Form von E-Mails oder Kommentaren in Verwaltungssystemen der Software stattfindet. Betrachtet man die Zusammenarbeit von Entwicklern, so stellt man nicht selten auch Schwierigkeiten fest. So kann eine Misskommunikation sich negativ auf die Stimmung auswirken und dafür sorgen, dass der Projekterfolg ebenfalls beeinflusst wird. Schlimmstenfalls führen die Diskrepanzen sogar zu einem Burn-Out oder dem Ausstieg einiger Software-Entwickler aus dem Softwareprojekt. Besonders für Projektleiter bieten sogenannte Stimmungsanalysetools daher den Vorteil, die Stimmungen der Entwickler während des Projektes zu analysieren und rechtzeitig zu invertieren. Die Stimmungsanalyse wird bisher nur in Gebieten ausserhalb der Softwareentwicklung angewandt und ist nicht angepasst auf die Subjektivität individueller Software-Entwickler, welche unterschiedliche Wahrnehmungen hervorruft. Im Rahmen dieser Arbeit wird untersucht, ob eine Kalibrierung von Stimmungsanalysetools sinnvoll ist, indem 22 Informatiker einige Aussagen manuell nach ihrer Wahrnehmung annotieren. Basierend auf einer wahrnehmungsbasierten Zuordnung der Entwickler werden diesen neue Aussagen mit berechneten Wahrnehmungen präsentiert. Indem die Aussagen mit den berechneten Wahrnehmungen nun nach ihrer Zustimmung bewertet werden, wird so überprüft, ob solch eine Kalibrierung möglich ist. Die Erkenntnisse dieser Arbeit sollen also den Einsatz der Stimmungsanalyse in der Softwareentwicklung unterstützen.

Abstract

In software development, the requirements for the software to be created are constantly growing. Consequently, collaboration within a team of the responsible software developers becomes increasingly important in order to achieve better results in the respective software project. In addition to the implementation of functions, communication among developers is of great importance, taking place not only directly but also in the form of emails or comments within software management systems. Considering the collaboration of developers, some difficulties are not uncommon. Miscommunication can negatively impact the atmosphere and, thus, the success of the project. In the worst case, these discrepancies may even lead to burnout or the departure of some software developers from the project. Especially for project managers, so-called sentiment analysis tools offer the advantage of analyzing developers' moods during the project and reversing them in a timely manner. Sentiment analysis is currently only applied in areas outside of software development and is not adapted to the subjectivity of individual software developers which can lead to different perceptions. This study examines whether the calibration of sentiment analysis tools is meaningful by having 22 computer scientists manually annotate statements based on their perception. New statements with calculated perceptions are then presented to these developers based on a perception-based assignment. By evaluating these statements in terms of their agreement with the calculated perceptions, it is checked whether such calibration is possible. The findings of this study are intended to support the use of sentiment analysis in software development.

Inhaltsverzeichnis

Kapitel 1

Einleitung

Die Entwicklung von Software bietet stets neue Herausforderungen und Hindernisse [22]. Neben der Implementierung der Funktionalität spielt besonders die Kommunikation unter Entwicklern eine große Rolle in der Zukunft eines Projektes [27]. Um den Erfolg des Projektes zu gewährleisten, muss daher auch der soziale Aspekt betrachtet werden [31]. Die Zufriedenheit in einem Projekt führt zu einer guten Mitarbeit oder zu innovativen Ideen, die vor allem im Design hilfreich sind [12]. Unzufriedene Entwickler hingegen tendieren dazu, Aufgaben zu verschieben oder sogar das Projekt abzubrechen [23]. Diese Ansichten spiegeln sich nicht nur in den Commit-Nachrichten bei der Softwareerstellung wider, sondern auch in der direkten Kommunikation der Entwickler untereinander. Negative Nachrichten begünstigen außerdem Missverständnisse und Diskrepanzen, die sich auf den Erfolg des Projektes auswirken [13]. In der Forschung stellt die Stimmungsanalyse daher eine große Relevanz für die Kommunikation unter Entwicklern in der Softwareentwicklung dar [23].

1.1 Problemstellung

Um das Problem der Missverständnisse und Diskrepanzen unter Entwicklern zu analysieren, wird die Stimmungsanalyse verwendet. Mit Hilfe von Stimmungsanalysetools werden Texte im Hinblick auf ihre Sentiment-Polarität (positiv, neutral oder negativ) überprüft [11]. Dieses Verfahren hilft unter anderem Projektleitern dabei, sich einen Überblick über die aktuelle Lage im Projekt zu verschaffen [34]: Wie gut schreitet das Projekt voran? Gibt es Probleme zwischen den Entwicklern? Welchen Einfluss hat die Mitarbeit auf den Erfolg des Projektes? Solche Informationen sind von wichtiger Bedeutung für den Projektleiter, der dann im Falle von möglichen Problemen intervenieren kann [11].

In einer Masterarbeit von Herrmann [15] wurden Umfrageergebnisse aus einer früheren Arbeit [16] verwendet und auf verschiedene Merkmale

in der Wahrnehmung der Umfrageteilnehmer untersucht. Die Umfrageteilnehmer sind hauptsächlich Entwickler oder angehende Entwickler mit Programmiererfahrung. In der Umfrage wurden den Entwicklern mehrere Aussagen präsentiert, welche mit den Sentiment-Polaritäten entsprechend der Wahrnehmung der Teilnehmer annotiert wurden. Die Erkenntnisse aus der Masterarbeit zeigen, dass eine reine Analyse eines Textes durch Stimmungsanalysetools nicht ausreicht, um die Wahrnehmung individueller Entwickler zu reflektieren. Aufgrund der unterschiedlichen Wahrnehmung zwischen den Umfrageteilnehmern stellt sich die Hypothese auf, ob Stimmungsanalysetools auf die Kalibrierung angewiesen sind. Eine solche Kalibrierung meint die Einbeziehung von Aussagen, die vorher individuell annotiert werden. Das Stimmungsanalysetool soll dadurch in der Lage sein, eine bessere Einschätzung der Wahrnehmung von Aussagen zu treffen.

1.2 Lösungsansatz

Im Rahmen dieser Arbeit sollen die Ergebnisse der Masterarbeit von Herrmann [15] validiert werden. Dies beinhaltet unter anderem die Hypothese, dass die Kalibrierung des verwendeten Stimmungsanalysetools dabei hilft, die Wahrnehmungen der Entwickler zu treffen. Mit Hilfe einer Umfrage soll also geprüft werden, ob die Studienteilnehmer mit den vorhergesagten Sentiment-Polaritäten einverstanden sind: Dazu sollen die Studienteilnehmer zu Beginn Aussagen annotieren, wodurch sie einer der zwei Gruppen aus Herrmanns Arbeit zugeordnet werden [15]. Jede Gruppe besitzt eine unterschiedliche Wahrnehmung. Die Zuordnung geschieht mittels der logistischen Regressionsanalyse, dessen Modell bereits existiert. Auf Basis der jeweiligen Wahrnehmung werden den Teilnehmern danach die neuen vorhergesagten Aussagen präsentiert. Anschließend wird die Zuordnung der Umfrageteilnehmer durch das logistische Regressionsanalyse evaluiert. Dabei wird überprüft, ob die Kalibrierung dem Stimmungsanalysetool tatsächlich dazu verhilft, die verschiedenen Wahrnehmungen der Entwickler zu treffen.

1.3 Struktur der Arbeit

Diese Arbeit ist wie folgt strukturiert. Um dem Leser ein besseres Verständnis über das Forschungsdesign und in den restlichen Kapiteln zu geben, werden in Kapitel 2 die in dieser Arbeit verwendeten grundlegenden Verfahren erklärt. In Kapitel 3 wird die Arbeit von verwandten Arbeiten abgegrenzt. Kapitel 4 gibt Aufschluss über das Forschungsdesign und in Kapitel 5 folgen die Ergebnisse der Studie und deren Evaluierung. Nach einer Diskussion in Kapitel 6 endet die Arbeit mit einer Zusammenfassung der in dieser Arbeit gewonnenen Erkenntnisse in Kapitel 7.

Kapitel 2

Grundlagen

Um das Verständnis der nachfolgenden Kapitel zu erleichtern, werden in diesem Kapitel einige Konzepte und Verfahren erläutert. Dazu wird auf den Begriff der Stimmungsanalyse sowie auf die logistische Regressionsanalyse eingegangen.

2.1 Stimmungsanalyse

Natural Language Processing (NLP) nimmt die menschliche Sprache in der textuellen oder verbalen Kommunikation entgegen und befasst sich damit, diese maschinell zu verarbeiten [32]. Die Stimmungsanalyse als Teilgebiet, ebenfalls unter dem Namen Meinungsforschung bekannt, analysiert eben solchen Text speziell auf seine Sentiment-Polarität. Die Sentiment-Polarität stellt die Stimmung des Textes dar. Die Stimmungsanalyse kann jedoch nicht allein auf NLP, sondern auch auf andere Verfahren wie das KI- oder Lexikon-Verfahren basieren [43]. Um den Text zu analysieren, wird die benötigte Information herausgefiltert und danach die Stimmung des Textes einem Wert zugeordnet [11]. In der Regel liegt die häufigste Zuordnung bei einem Wert von -1, 0 und 1, wobei -1 einer negativen und 1 einer positiven Wahrnehmung entspricht [11].

Andere Stimmungsanalysetools wie das SentiStrength können auch größere Skalen verwenden. Bei dem Lexikon-basierten Verfahren wird jedes Wort eines zu analysierenden Satzes mit einem Wert zwischen -5 (sehr negativ) und 5 (sehr positiv) für die jeweilige Sentiment-Polarität bewertet. Im Anschluss werden der größte (Maximum) und der kleinste vorkommende Wert (Minimum) miteinander addiert. Bei einem Maximum von 2 und einem Minimum von -2 würde die Addition eine 0 ergeben und damit für eine neutrale Bewertung sprechen [39]. Während die Sprache eindeutige Worte enthalten kann wie beispielsweise „wundervoll“ oder „unglücklich“, die jeweils für eine positive und negative Stimmung stehen und korrekt eingeordnet werden, so gibt es doch einige erschwerliche Begriffe [43]. Als

besonders schwierig stellt sich die Analyse bei einer Individualität in der Ausdrucksweise der Verfasser und bei Referenzen heraus [38]. So kann der Text auch subjektive sowie objektive Äußerungen enthalten, welche schwer einzuordnen sind [43].

2.1.1 Stimmungsanalyse in Social Media

Die Stimmungsanalyse hat ein weitreichendes Spektrum an Anwendungsgebieten. Neben dem Einsatz auf Social Media Plattformen wird die Stimmungsanalyse ebenfalls für die Analyse von Kundenzufriedenheit bei dem Kauf und der Bewertung von Produkten verwendet [38, 24].

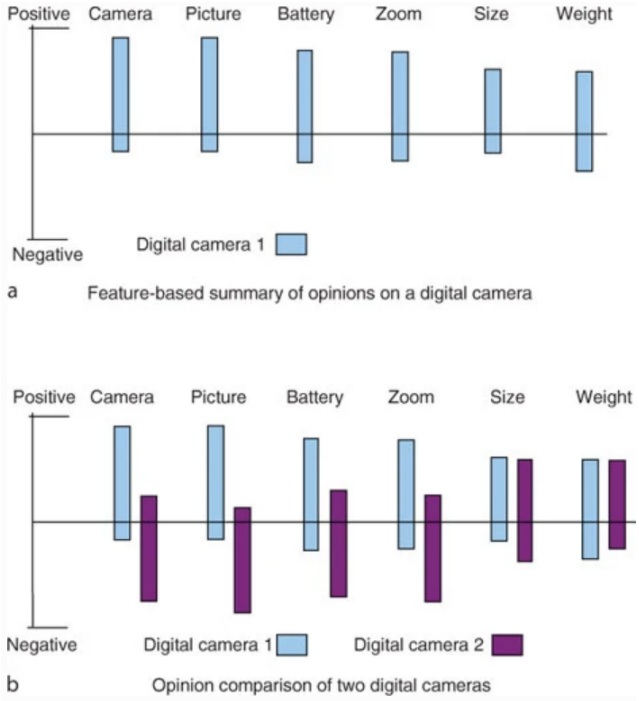

Abbildung 2.1: Visualisierung der eigenschaftsbasierten Zusammenfassung und Vergleich von Meinungen

In der obenstehenden Abbildung 2.1 sind zwei Balkendiagramme zu sehen. Beide Diagramme visualisieren die aus Rezensionen bestimmte Stimmungen zu den einzelnen Eigenschaften eines Produktes, hier einer Digitalkamera. Bei diesem Beispiel kommt die Aspekt-basierte Stimmungsanalyse zum Einsatz [24]. Dieses Verfahren unterteilt größere Abschnitte von Text in Eigenschaften und bietet sich daher gut für Produktrezensionen an. Wie man in der Abbildung 2.1 sehen kann, sind die aus mehreren Rezensionen bestimmten Eigenschaften die Kamera, Bildqualität, Batterieleistung, Zoom-Fähigkeit, Größe und Gewicht.

Der Prozess der Aspekt-basierten Stimmungsanalyse läuft wie folgt ab. Zu Beginn werden die Produkteigenschaften identifiziert, indem die Nomen jedes Satzes auf ihre Häufigkeit überprüft werden [24]. Im nächsten Schritt werden die Sentiment-Polaritäten der einzelnen Sätze zu der jeweiligen Eigenschaft bestimmt. Für diesen Schritt eignen sich Lexikon-basierte Verfahren sehr gut, es können jedoch auch andere Ansätze verfolgt werden [24]. Zuletzt werden die Eigenschaften jeder Produktrezension gruppiert, da es möglich ist, dass die Autoren der Rezensionen Synonyme verwenden. Eine Rezension kann zum Beispiel den Begriff „Fotos“ beinhalten und eine andere „Bilder“, jedoch beide dieselbe Eigenschaft ansprechen [24].

Die Balken in beiden Diagrammen der Abbildung 2.1 stellen die Häufigkeitsverteilung zwischen den positiven und negativen Wahrnehmungen dar. Diagramm a zeigt beispielsweise, dass mehr positive (insgesamt 125) als negative (insgesamt 7) Meinungen zu der Kamera in allen Rezensionen zu finden sind. Während in Diagramm a in Abbildung 2.1 die einzelnen Stimmungen zu einer Kamera zusammmengefasst zu sehen sind, zeigt Diagramm b den Vergleich zwischen den Stimmungen zweier verschiedener Digitalkameras. Dazu sind jeweils zwei Balken zu jeder Kamera in jeder Eigenschaft nebeneinander dargestellt. Die Abbildung 2.1 stellt lediglich eine mögliche Visualisierung der Ergebnisse der Stimmungsanalyse dar.

2.1.2 Stimmungsanalyse in Softwareunternehmen

Die Anwendung von Stimmungsanalysetools ist nicht nur wichtig bei der Meinungsforschung zu Produkten. Stimmungsanalysetools sollen in Zukunft auch in Softwareunternehmen eingesetzt werden [23]. In der Softwareerstellung ist die Arbeit im Team besonders wichtig. Neben der direkten Kommunikation findet auch eine textuelle Kommunikation unter Entwicklern statt, um unter anderem wichtige Aufgabenverteilungen, Termine sowie die Implementierung von Funktionen abzuklären [27]. Die für die Softwareerstellung zuständigen Entwickler kommunizieren über verschiedene Wege wie beispielsweise E-Mails mit anderen beteiligten Entwicklern im Unternehmen selbst. Eine andere Kommunikationsmöglichkeit bieten Online-Hilfeforen wie *Stack Overflow* oder auch *Reddit* [8, 29]. Auf diese Foren kann zurückgegriffen werden, wenn bei dem Programmieren Schwierigkeiten auftreten und die Möglichkeit nicht besteht, sich mit einem anderen Entwickler zu beraten [29]. Über Online-Dienste wie *GitHub* [9, 14] oder *Jira* [19, 31] finden die Versions- und Fehlerverwaltung der zu erstellenden Software im Projekt statt. In jeder Kommunikationsmöglichkeit können Schwierigkeiten wie Missverständnisse zwischen den kommunizierenden Entwicklern entstehen. Da solch eine Misskommunikation nicht selten schlechte Auswirkungen auf den Projekterfolg birgt [13], ist es sinnvoll, diese im Voraus zu erkennen und als Projektleiter zu intervenieren [34]. Das im Rahmen dieser Arbeit thematisierte Anwendungsgebiet schränkt sich auf die textuelle Analyse in

individuellen Softwareunternehmen ein. Dafür sollen Beispieldatensätze aus *GitHub* und *Stack Overflow* und herangezogen werden.

Es gibt in der Softwareentwicklung zwar spezifische Stimmungsanalysetools wie den SEnti-Analyzer für die textuelle wie verbale Textanalyse [17]. Bisher wurden Stimmungsanalysetools jedoch hauptsächlich in Social Media verwendet und sind aus diesem Grund meist nicht auf technischen Jargon oder Problemberichte abgestimmt [8]. Um Missinterpretationen von Texten aus diesem Bereich zu verhindern, müssen Stimmungsanalysetools also trainiert werden [8]. Daher werden in dieser Arbeit durch Entwickler annotierte Aussagen hinsichtlich der Wahrnehmung untersucht. Dabei wird ebenfalls zwischen positiv, neutral und negativ unterschieden.

2.2 Logistische Regressionsanalyse

Die logistische Regressionsanalyse macht Gebrauch von dem linearen Regressionsmodell. In der linearen Regression werden intervallskalierte Daten verwendet und eine Variable $y = \alpha x + \beta$ vorhergesagt, die von einer unabhängigen Variable x abhängt und bei Änderung von x ihren Wert ändert. Ein Beispiel dafür ist die Vorhersage des Blutzuckerwertes anhand des Geschlechts und Gewichtes einer Person [41]. Dazu werden bekannte Blutzuckerwerte samt Geschlecht und Gewicht der Patienten betrachtet und in einem Graphen eine Regressionslinie dargestellt. Das Geschlecht und Gewicht bilden die unabhängigen Variablen und müssen für den neuen Blutzuckerwert nicht in den Daten vertreten sein. Mit Hilfe der Regressionsgleichung und $y = \alpha x + \beta$ wird der neue Wert dann abgeschätzt, wobei α und β Eigenschaften der Regressionslinie darstellen.

Die logistische Regression wiederum findet ihre Verwendung bei nominalskalierten Daten. Dahingehend wird zwar ebenfalls eine abhängige Variable Y vorhergesagt, jedoch besitzt diese Variable Y zwei diskrete Ausprägungen. Diese Ausprägungen belaufen sich auf die Existenz einer Eigenschaft, beispielsweise ob ein Wettbewerb gewonnen wird oder nicht, auf Basis der Anzahl der vorangegangenen Siege und des Alters der teilnehmenden Person [7].

2.2.1 Bestimmung der vorhergesagten Variable

Um die Wahrscheinlichkeit für die Ausprägung der Variable Y in Abhängigkeit der Variable X und damit das Eintreffen der dazugehörigen Eigenschaft zu bestimmen, wird die folgende Formel verwendet [33]:

$$P(Y = 1|X) = \frac{e^z}{1 + e^z} \tag{2.1}$$

Die Formel 2.1 beinhaltet den normierten linearen Prädiktor z und gibt für $P(Y = 1|X)$ ein Ergebnis im Intervall [0,1] aus. Im Gegensatz stellt die

Gegenwahrscheinlichkeit $P(Y = 0|X) = 1 - P(Y = 1|X)$ dafür dar, dass Y nicht ausgeprägt ist. Folglich sagt durch eine Wahrscheinlichkeit von $P \geqq 0.5$ $Y = 1$ und durch $P < 0.5$ auf $Y = 0$ voraus [7].

Der lineare Prädiktor in Formel 2.1 besitzt Werte im Intervall $[-\infty, \infty]$ und ist normiert, um mit einem Intervall von [0,1] auf eine Wahrscheinlichkeit für die Ausprägung von Y schließen zu können [33]. Zusammensetzen lässt sich der lineare Prädiktor in der folgenden Linearkombination aus den Prädiktorvariablen $X = \{x_0, x_1, ..., x_n\}$ und den dazugehörigen Regressionskoeffizienten $\beta = \{\beta_0, \beta_1, ..., \beta_n\}$:

$$z = \beta_0 + \sum_{i=1}^{n} \beta_i x_i = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n \tag{2.2}$$

Wie in der linearen Regressionsanalyse stellt n die Teilmenge $n < p$ der Anzahl der unabhängigen Variablen p aus einer Datenbasis dar [7].

Die logistische Regressionsanalyse zielt zum Einen auf die korrekte Vorhersage von Y in Abhängigkeit der unabhängigen Variablen in X ab. Neben der Vorhersage ist es zum Anderen wichtig, nicht nur eine korrekte, sondern auch vollständige Separation zwischen den Möglichkeiten $Y = 1$ und $Y = 0$ zu erreichen [15].

2.3 Statistische Testverfahren

Im Hinblick auf die Auswahl eines geeigneten Tests für die statistische Untersuchung der von Daten wurde die nachstehende Tabelle 2.1 aus der Arbeit von Herrmann [15] erstellt.

Normalverteilt	**Homoskedastizität**	**Statistischer Test**
✓	✓	Zwei-Stichproben-t-Test
✓	✗	Welch-Test
✗	✗	Mann-Whitney-U-Test

Tabelle 2.1: Überblick über das Auswahlverfahren für ein statistisches Testverfahren in Abhängigkeit von den zu erfüllenden Vorbedingungen [15]

Im Folgenden werden die für diese Arbeit relevanten Statistikverfahren erläutert. Wie man in der Tabelle 2.1 sehen kann, wird je nach erfüllter Vorbedingung ein anderer Test verwendet.

2.3.1 Mann-Whitney-U-Test

Der Mann-Whitney-U-Test ist ein Verfahren aus der nichtparametrischen Statistik [35]. Die Art und Anzahl der Parameter ist nicht vorher festgelegt und variabel. Der Test wird auf gemessene und ordinalskalierte Daten

angewandt [21]. Ordinalskalierte Daten enthalten nominale Kategorien, die sich mit einem Rang hierarchisch vergleichen lassen [5]. Eine Kategorie kann also größer oder kleiner als die jeweils andere sein [28]. Ein Beispiel für ordinalskalierte Daten ist das Benotungssystem in der Schule. Unterscheiden gilt es zwischen den Noten „*Sehr gut*“, „*Gut*“, „*Befriedigend*“, „*Ausreichend*“, „*Mangelhaft*“ und „*Ungenügend*“, welche in der genannten Reihenfolge in der Hierarchie absteigen. Die Benotung von „*Sehr gut*“ entspricht einer 1. Die Benotung mit einer guten Note enspricht der Zahl 2 und ist bekanntlich schlechter als eine Benotung mit einer sehr guten Note.

Den Ansatz der Rangordnung macht sich der Mann-Whitney-U-Test zunutze. Der Test vergleicht die Verteilung zweier unabhängiger Teilnehmergruppen, indem er den Daten einen Rang vergibt und anschließened miteinander verrechnet [25]. Liegt der Wert der Berechnung unter dem gewählten Signifikanzniveau α, so spricht man von einem signifikantem Unterschied zwischen den beiden Gruppen. Bei einem signifikantem Unterschied bestätigt sich die Nullhypothese, dass ein zufällig ausgewählter Wert aus der einen Gruppe immer größer oder kleiner ist als ein zufällig ausgewählter Wert aus der anderen Gruppe [26].

2.3.2 Zwei-Stichproben-t-Test

Der Zwei-Stichproben-t-Test ist ein statistisches Testverfahren und wird auf normalverteilte Daten angwandt [37]. Die Normalverteilung der Merkmalswerte wird durch den Shapiro-Wilk-Test überprüft. Unterschreiten die Ergebnisse des Shapiro-Wilk-Tests das gewählte Signifikanzniveau, so spricht dies für die Verwendung des in Kapitel 2.3.1 erläuterten Mann-Whitney-U-Tests. Liegen die Ergebnisse jedoch über dem Signifikanzniveau, dann muss der Zwei-Stichproben-t-Test unter der Voraussetzung einer weiteren Bedingung verwendet werden, wie in Tabelle 2.1 zu sehen ist [36].

Die Homoskedastizität als solche Bedingung sagt aus, dass die Standardvarianzen der Daten zweier Teilnehmergruppen konstant sind. Dafür müssen die Standardvarianzen für die untersuchten Merkmalswerte ähnlich sein [20]. Man nehme also beispielsweise an, man würde das Stresslevel und das Gewicht von 20-jährigen Mädchen in einem statistischen Test untersuchen. Verhalten sich die Standardvarianzen des Stresslevels für das Gewicht von 45, 50 und 60 ähnlich, so erfüllen die untersuchten Daten das Kriterium der Homoskedastizität. Die Bedingung kann mit dem Bartlett-Test überprüft werden. Fällt dieser nicht signifikant aus, so sind die Daten konstant (Homoskedastizität) und der Zwei-Stichproben-Test muss angewandt werden [37]. Andernfalls weichen die Standardvarianzen ab und die daraus resultierende Heteroskedastizität schließt auf die Verwendung eines anderen statistischen Tests wie dem Welch-Test [42]. Fällt der Zwei-Stichproben-t-Test selbst signifikant aus, so heißt dies, dass es einen statistischen Unterschied in den Durchschnitten der Merkmalswerte der Teilnehmergruppen gibt [3].

Kapitel 3

Verwandte Arbeiten

Dieses Kapitel behandelt verwandte Arbeiten, die sich inhaltlich mit dieser Bachelorarbeit verknüpfen lassen. Zuerst wird eine Arbeit vorgestellt, deren Ergebnisse relevant für diese Arbeit sind und welche im Forschungsaufbau in dieser Arbeit weiterverwendet werden. Danach werden Arbeiten thematisiert, die die Stimmungsanalyse im Rahmen von Softwareunternehmen gebrauchen. Das Ziel dieses Kapitels ist es letztlich, die Bachelorarbeit von verwandten Arbeiten abzugrenzen.

3.1 Stimmungsanalyse in Softwareprojekten

Unterschiedliche Wahrnehmungen in der Kommunikation unter Softwareentwicklern können Missverständnisse hervorrufen. Neben einer Unzufriedenheit kann solche Misskommunikation im schlimmsten Fall einen Burn-Out als Folge haben [40]. In einer Masterarbeit „*Analyse der Wahrnehmung von Stimmung in Softwareprojekten durch explorative Datenanalyse*“ untersucht Herrmann [15] unter Verwendung der Stimmungsanalyse daher anhand verschiedener Teilnehmer in einer Umfrage, wie viele unterschiedlche Wahrnehmungen verschiedene Entwickler haben und wie sich diese einordnen lassen.

Die durchgeführte Umfrage umfasst 100 Aussagen und addressiert ausschließlich Teilnehmer, die Programmiererfahrung haben. Unter den Teilnehmern befinden sich aus diesem Grund Informatikstudenten sowie Beschäftigte in der Informatikbranche mit unterschiedlicher Erfahrung in Entwicklungsteams [15]. Den Aussagen wurden durch die Teilnehmer die Sentiment-Polaritäten positiv, neutral oder negativ zugewiesen und im Anschluss durch Herrmann [15] interpretiert. Außerdem bestand die Möglichkeit, anzugeben, anhand welches Kriteriums die Bewertung der Aussagen erfolgt ist.

Auffällig ist, dass sich unter der Verwendung von Clusteranalyse zwei Teilnehmergruppen bilden, welche signifikante Unterschiede in der Wahrnehmung aufweisen. Diese Wahrnehmungen sind in der eigenen Teilneh-

mergruppe jedoch ähnlich [15]. Des Weiteren lassen sich fünf Aussagen bestimmen, durch welche eine korrekte Vorhersage der Wahrnehmung beider Teilnehmergruppen mit Hilfe der logistischen Regressionsanalyse möglich ist. Diese fünf Aussagen stechen ebenfalls hervor in der Auswertung der Ergebnisse der Umfrage. Eine sehr markante Aussage stellt die kurze Aussage „*lol :)*“, bei welcher es eine Separation der Wahrnehmungen gibt. Während eine Teilnehmergruppe diese Aussage als überwiegend positiv bewertet, entscheidet sich die andere Teilnehmergruppe eher für negativ [15]. Die jeweilige Wahrnehmung findet sich vor allem in den Freitext-Antworten zum Annotationskriterium wieder. Während ein alleinstehendes „*lol*“, das für das englischsprachige „*laughing out loud*“ steht, als negativ empfunden wird, lässt ein lächelndes Smiley diese Aussage eher positiv auf die Teilnehmer wirken [15].

Interessant ist zusammenfassend die Frage, ob Stimmungsanalysetools die unterschiedliche und subjektive Wahrnehmungen von Entwicklern miteinbeziehen müssen, um korrekte Ergebnisse bei der Bewertung der Stimmung von Aussagen zu erzielen. Um sich an die unterschiedlichen Wahrnehmungen anzupassen, müssen Stimmungsanalysetools kalibriert werden und im Zuge dessen die für die Separation der Wahrnehmung relevanten Merkmale bestimmt werden [15].

3.2 Zusammenhang von Stimmung und Bugs unter Softwareentwicklern

Die Entstehung von ungewolltem Verhalten in Code, sogenannter Bugs, ist ein nahezu unvermeidbarer Nebenfaktor in der Erstellung von Software. Nicht nur das erstmalige Entwickeln von Funktionen, sondern auch das Beheben bzw. Fixen von Fehlern in Software kann weitere Fehler hervorrufen, die das Nutzererlebnis beeinträchtigen, indem im schlimmsten Fall sogar Hauptfunktionen beeinflusst werden [6]. Nicht selten hat die Entstehung von Bugs auch einen Einfluss auf die Stimmung der für die Softwareerstellung primär zuständigen Softwareentwickler, wie Huq et al. [18] in ihrem Fachartikel „*Is Developer Sentiment Related to Software Bugs: An Exploratory Study on GitHub Commits*“ untersuchen.

In der Untersuchung wird zwischen vier Arten von Commits auf den bekannten Plattformen GitHub und Jira unterschieden und wie sie mit der Stimmung der Entwicklern zusammenhängen. Abbildung 3.1 stellt die Erkennung der verschiedenen Commits in Abhängigkeit der Zeit anhand einer Datei dar. Ebenfalls werden Zeilen in der Datei mit „++“ oder „– “ erkenntlich gemacht, in welchen Modifikationen ausschlaggebend für die Erkennung der Commitart sind [18]. Zu den vier Arten von Commits gehören Fix-inducing changes (FIC), Parents of FICs (pFIC), Fixing Changes (FC) und Fix-Inducing Fixes (FIF). Änderungen, die Bugs im Code enthalten

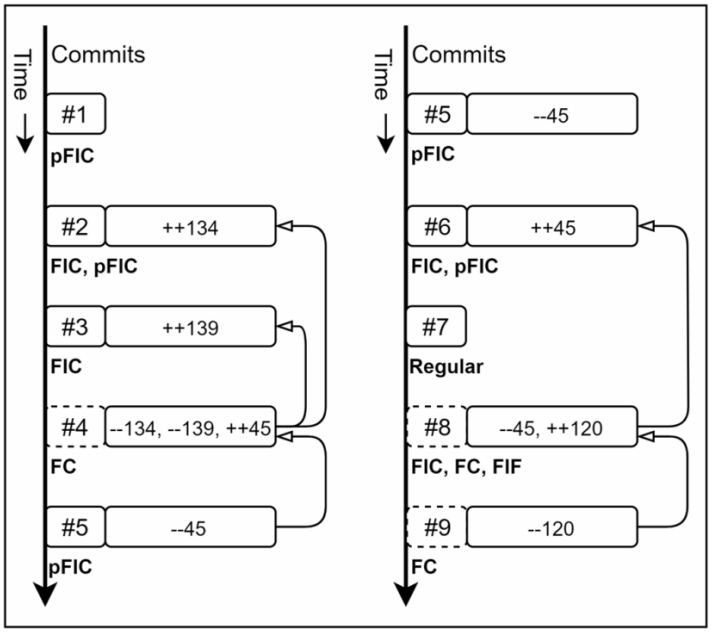

Abbildung 3.1: Kategorisierung der vier verschiedenen Arten von Commits

oder Bugs verursachen, werden FICs genannt. pFICs sind Initial-Commits, auf welchen FICs entstehen. In FCs hingegen werden Bugs behoben, welche in FICs aufkommen. Die letzte Kategorie bilden die FIFs, in welchen Bugs zwar behoben werden, jedoch weitere Bugs entstehen lassen [18].

Nach der Kategorisierung untersuchen Huq et al. [18] Commits aus diversen Repositories auf ihre Stimmung. Dabei kann die Stimmung als positiv, neutral und negativ aufgefasst werden. Begriffe wie „Bugs“ oder „Fehler“ werden von einigen Stimmungsanalysetools inhaltlich als negativ bewertet und werden demzufolge als Ausnahmen berücksichtigt [18]. Mit Hilfe des Stimmungsanalysetools Senti4SD wurde anschließend herausgefunden, dass die Entstehung von Bugs tatsächlich dazu beiträgt, die Stimmung von Entwicklern zu beeinflussen. In einem statistischen Vergleich der Commits mit normalen Commits tendieren alle vier Arten von Commits dazu, mehr negative Nachrichten zu beinhalten. Außerdem ist auffällig, dass die Commits mehr emotionale als neutrale Aussagen enthalten. Die einzige Ausnahme bilden FIFs, in welchen mehr neutrale als negative Nachrichten erkenntlich sind, wenn korrekte Fixes vorgenommen wurden [18].

Bugs als Einflussfaktoren in der Stimmung von Software-Entwicklern müssen berücksichtigt werden und daher der Entwicklungsprozess stets beobachtet werden, damit sich negative Stimmung nicht auf andere Entwickler überträgt oder sich anderweitig auf den Erfolg eines Softwareprojekts auswirkt [18].

3.3 Subjektivität in der Stimmungsanalyse

Ausgelöst durch die Subjektivität von Software-Entwicklern, können unterschiedliche Wahrnehmungen ein Hindernis für die Kommunikation in einem Softwareprojekt darstellen [31]. In einem weiteren Fachartikel von Herrmann et al. [16] „*On the subjectivity of emotions in software projects: How reliable are pre-labeled data sets for sentiment analysis?*“ wird untersucht, inwieweit sich die Wahrnehmungen von potenziellen Mitgliedern von Softwareunternehmen sich von einer bestimmten Wahrnehmung von ausgewählten Datensätzen unterscheiden. Dazu wurden zwei Datensätze mit jeweils 100 Aussagen betrachtet. Unterschieden wurde zwischen nach Richtlinien annotierten Datensatz aus dem Versionsmanagementsystem *Github* und einem ad hoc annotierten Datensatz aus dem Frageforum *Stack Overflow* [16].

Auch in diesem Fachartikel wurden Informatiker in einer Umfrage zu ihrer Wahrnehmung zu den gegebenen 100 Aussagen befragt, wobei die Sentiment-Polaritäten positiv, neutral oder negativ vergeben wurden. Im Gegensatz zu der Masterarbeit von Herrmann [15] wurden in diesem Fachartikel nach Durchführung der Umfrage die annotierten Sentiment-Polaritäten durch die Studienteilnehmer mit den annotierten Sentiment-Polaritäten aus den Datensätzen verglichen, um eine Übereinstimmung der Sentiment-Polaritäten zu untersuchen [16]. Es wurden die folgenden drei Fragestellungen aufgestellt:

- RQ1: Wie unterscheiden sich die durchschnittlichen Annotationen aller Studienteilnehmer von den annotierten Datensätzen?
- RQ2: Inwieweit unterscheiden sich die einzelnen Annotationen der jedes Studienteilnehmers von den annotierten Datensätzen?
- RQ3: Wie unterscheiden sich die Ergebnisse zwischen den nach Richtlinien und den ad hoc annotierten Datensätzen?

Um RQ1 zu beantworten, wurden zunächst die durchschnittlichen Anteile der Sentiment-Polaritäten sowie der Cohen's Koeffizient für die Zustimmung für jede Aussage berechnet und mit den vorbestimmten Sentiment-Polaritäten verglichen [16]. Für Fragestellung RQ2 wurden die durchschnittlichen Sentiment-Polaritäten und der Cohen's Koeffizient jedes Studienteilnehmers einzeln betrachtet. Für RQ3 wurde untersucht, ob es einen signifikanten Unterschied in den Cohen's Koeffizienten für die nach Richtlinien und ad hoc annotierten Datensätze gibt [16].

Nach einer Überprüfung der in diesem Fachartikel [16] relevanten Fragestellungen wurde zusammengefasst, dass signifikante Unterschiede in den Annotationen der Studienteilnehmer und den im Voraus bestimmten Sentiment-Polaritäten der Aussagen bestehen. Die Wahrnehmungen der Studienteilnehmer stimmen durchschnittlich nur zu 62.5% mit den Wahrnehmungen in den

Datensätzen überein. Bei der Betrachtung jedes einzelnen Informatikers gibt es sehr unterschiedliche Ergebnisse. Neben hohen Übereinstimmungen mit den vorbestimmten Sentiment-Polaritäten gibt es auch Studienteilnehmer, deren annotierte Sentiment-Polaritäten kaum mit den Sentiment-Polaritäten in den annotierten Datensätzen übereinstimmen. Wurden die durch die Studienteilnehmer annotierten Sentiment-Polaritäten mit den verschiedenen Datensätzen verglichen, so wurde eine höhere Übereinstimmung mit den nach Richtlinien annotierten Datensätzen aus *Github* festgestellt [16]. Zusammenfassend wurden die Erkenntnisse aufgestellt, dass trotz der teilweisen Übereinstimmung der Wahrnehmung mit den Datensätzen jeder einzelne Informatiker eine unterschiedliche Wahrnehmung aufweist [16]. Aus diesem Grund dürfen die unterschiedlichen Wahrnehmungen von Entwicklern für die Verwendung der Stimmungsanalyse nicht vernachlässigt werden.

3.4 Abgrenzung der Arbeit

Die Konzepte der genannten Arbeiten ähneln dieser Bachelorarbeit in vielerlei Hinsicht: Das Augenmerk ist dabei die Verwendung der Stimmungsanalyse, um sich einen Überblick über die Wahrnehmung von Entwicklern verschiedener Softwareunternehmen zu verschaffen [15]. Aus den verwandten Arbeiten geht besonders hervor, dass es Faktoren wie beispielsweise Bugs gibt, die zur Entstehung von emotionalen Wahrnehmungen beitragen [18]. Emotionale Nachrichten drücken nicht nur Zufriedenheit aus, sondern können ebenso Probleme ansprechen. Die Stimmungsanalyse ist daher von großer Relevanz, um die Möglichkeit zu bieten, frühzeitig negative Stimmungen zu erkennen und Einfluss auf den Erfolg des Projektes zu verhindern [13, 23]. Des Weiteren stellt sich die Vermutung auf, ob Stimmungsanalysetools die Wahrnehmungen jedes Entwicklers treffen können [15]. Je nach Umgebung müssen Stimmungsanalysetools Text verschieden behandeln oder die Berücksichtung von Ausnahmen wird gefordert, wie in Huqs Fachartikel erwähnt [18].

Diese Bachelorarbeit knüpft an die Erkenntnisse aus Herrmanns Arbeit [15] an und untersucht unter der Weiterverwendung dieser, ob eine Kalibrierung von Stimmungsanalysetools tatsächlich sinnvoll ist. Dazu wird mit Hilfe der logistischen Regressionsanalyse am selben Datensatz unter Eingabe der Wahrnehmungen verschiedener Entwickler eine Vorhersage getroffen und daraufhin untersucht. Eine solche Kalibrierung ist dann sinnvoll, wenn die Wahrnehmungen der Entwickler mit der Vorhersage übereinstimmen. Diese Arbeit bietet somit einen möglichen Ansatz für die Verbesserung von Stimmungsanalysetools hinsichtlich der Individualität von Wahrnehmungen und lässt sich daher von anderen Arbeiten abgrenzen.

Kapitel 4

Forschungsaufbau

In diesem Kapitel wird auf den Aufbau der in dieser Arbeit durchgeführten Studie eingeganen. Dabei wird erläutert, wie die in den Grundlagen genannten Methodiken angewandt werden, um die Datenbasis zu verarbeiten und eine Evaluation der Ergebnisse durchzuführen. Die nachstehende Abbildung 4.1 dient dazu, das Verständnis der relevanten Schritte im Forschungsaufbau zu erleichtern.

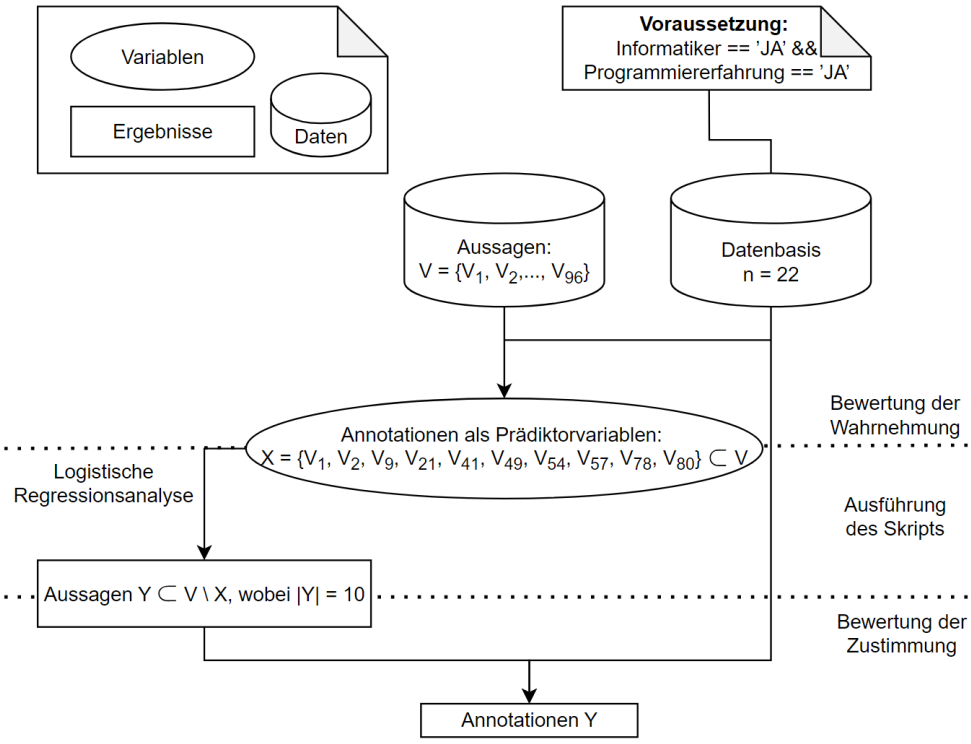

Abbildung 4.1: Überblick über die Durchführung der Studie und die Abhängigkeiten der verwendeten Methoden

Im Forschungsteil beinhaltet diese Arbeit die Durchführung einer Studie mit einer Datenbasis von insgesamt n = 22 Studienteilnehmern, wie Abbildung 4.1 darstellt. Ebenfalls gibt es insgesamt 96 Aussagen V_1 bis V_{96}, von welchen eine ausgewählte Teilmenge $X = \{V_1, V_2, V_9, V_{21}, V_{41}, V_{49}, V_{54}, V_{57}, V_{78}, V_{80}\}, X \subset V$ hinsichtlich ihrer jeweiligen Stimmung annotiert wurde. Verwendet wurde derselbe Datensatz wie in Herrmanns Arbeit [15] und dieser kann im Repositorium *Zenodo* eingesehen werden (vgl. Obaidi et al. [30]). Dazu wird das Erhebungsdesign der Studie in Kapitel 4.1 genauer erläutert. Die durch die Studienteilnehmer annotierten Aussagen werden als Prädiktorvariablen genutzt, um mit Hilfe der logistischen Regressionsanalyse die 22 Studienteilnehmer zwei Teilnehmergruppen nach Wahrnehmung zuzuordnen. Außerdem wird auf Basis der Wahrnehmung der jeweiligen Teilnehmergruppe eine Vorhersage für die Stimmung zehn anderer Aussagen $Y \subset V \setminus X$ aus demselben Datensatz V zu treffen (vgl. Kapitel 4.1.3). In einem weiteren Teil der Umfrage werden diese Aussagen den Studienteilnehmern präsentiert und nach einem anderem Kriterium bewertet. Die Entstehung der annotierten neuen Aussagen Y wird in Kapitel 4.1.4 thematisiert. In einem weiteren Kapitel (vgl. Kapitel 4.2) wird auf Funktionen in der Umfrage eingegangen, die relevant sind für die Durchführung der Umfrage. Wie in Abbildung 4.1 zu sehen ist, erhält man nach der vollständigen Durchführung der Umfrage die annotierten Aussagen Y. Die Art der Visualisierung der Ergebnisse wird in Kapitel 4.4 erklärt. Anschließend wird in Kapitel 4.3 beschrieben, wie das in den Grundlagen erläuterte Statistikverfahren auf die Ergebnisse in dieser Studie angewandt wird.

4.1 Erhebungsdesign

Für die Untersuchung, ob eine Kalibrierung von Stimmungsanalysetools notwendig ist, wird eine dreiteilige Umfrage durchgeführt. Die zugehörige englischsprachige Umfrage ist, so wie sie den Studienteilnehmern vorgestellt wurde, in den nächsten Kapiteln zu sehen. Die Abbildungen 4.2, 4.3 und 4.4 stellen dabei die drei Abschnitte in der Umfrage dar. In diesem Kapitel wird zunächst auf den Prozess der Datenerhebung eingegangen und in den darauffolgenden Kapiteln (vgl. 4.1.2, 4.1.3 und 4.1.4) das Erhebungsdesign der Umfrage beschrieben.

4.1.1 Datenerhebung

Für die Teilnahme an der Umfrage wurden ausschließlich Informatiker eingeladen und befragt. Unter den 22 Befragten befanden sich ebenfalls drei bereits Berufstätige in der Informatikbranche wie Software-Entwickler. Den Großteil der Studienteilnehmer machten jedoch 19 Informatik-Studenten aus.

Die zweite Voraussetzung, um an der Umfrage teilzunehmen, bildete die Programmiererfahrung. Während einige Studienteilnehmer erste Erfahrungen mit dem Programmieren im Studium machten, hatten andere Studienteilnehmer bereits mehrjährige Erfahrungen mit verschiedenen Programmiersprachen im Beruf oder auch in der privaten Nutzung. Die Vorbedingungen für die Teilnahme an der Umfrage wurden während der Durchführung erhoben und dadurch sichergestellt. Für die Befragung wurden primär dem Autor bekannte Personen wie Kommilitonen im selben Studiengang eingeladen, aber auch Beiträge auf Servern in Onlinediensten für Instant Messaging wie *Discord* erstellt, um Mitglieder für die Teilnahme anzuregen.

Die Durchführung der Umfrage selbst fand ebenfalls mittels *Discord* inmitten eines virtuellen Interviews statt. Dieser Ansatz hat den Grund, dass die Anzahl der Studienteilnehmer gering ist, und bietet den Vorteil, als Moderatorrolle bei Rückfragen aushelfen und die Durchführung der Umfrage beaufsichtigen zu können. So kann sich der Studienteilnehmer bei Fragen beim Verständnis an den Moderator wenden. Da die Studienteilnehmer größtenteils Deutsch als Erstsprache besitzen, wird durch die Moderatorrolle unter anderem eine Sprachbarriere verhindert. Die Moderatorrolle wurde von der Autorin übernommen. Eine anderen Ursache für die Durchführung mit Hilfe eines virtuellen Interviews auf der Plattform *Teamviewer* stellt die Umfrage selbst dar. Durchgeführt wird diese nämlich auf einem lokalen Servers auf dem Computer des Moderators. Der Studienteilnehmer hat dabei die Auswahl, auf den Computer des Moderators mit Hilfe von Remote-Desktop zuzugreifen oder dem Moderator mittels Screensharing zu beschreiben, welche Auswahl getroffen wurde. So kann der Moderator sicherstellen, dass keine Schwierigkeiten während der Umfrage auftreten. Außerdem wird generell jeder Studienteilnehmer dazu angeregt, seine Bewertungskriterien mit der Think-Aloud-Methode offen auszusprechen.

4.1.2 Bewertung der Aussagen

Wie bereits in Kapitel 4.1.1 beschrieben, bestand die in dieser Studie durchgeführte Umfrage aus drei Teilen: Die Bewertung von Aussagen nach ihrer Wahrnehmung, die Generierung neuer Aussagen mit ihren Wahrnehmungen und die Zustimmung zu den gegebenen Wahrnehmungen durch den Studienteilnehmer. Die Studienteilnehmer sahen die Umfrage auf einem lokalen Server, die mit Hilfe von HTML erstellt wurde. Die dazugehörige Webseite ist als Abbildung in 4.2 dargestellt. Zu Beginn wurden jedem Studienteilnehmer zehn Aussagen aus einem Datensatz von 96 insgesamt Aussagen aus Herrmanns Arbeit [15] präsentiert(vgl. Obaidi et al. [30]). Die zehn Aussagen, die jedem Studienteilnehmer im ersten Teil der Umfrage präsentiert wurden, sind zehn fest ausgewählte Aussagen aus dem Datensatz V, welche die Menge $X = \{V_1, V_2, V_9, V_{21}, V_{41}, V_{49}, V_{54}, V_{57}, V_{78}, V_{80}\}$ aus Abbildung 4.1 bilden. Diese Aussagen ergeben sich aus dem Ergebnis der

Dark Mode

You are given several statements. Please read them carefully and then decide if it's a negative, neutral or positive statement.

Statements	Negative	Neutral	Positive
Trust URI.	▢	▢	▢
(Hopefully with a good example.)	▢	▢	▢
why allocate a new String instance? def apply(name: String): Node = hash(name) def fromHash(hash: String): Node = hash	▢	▢	▢
Comparison time should be fast, so total run time should be only slightly more than sum of run time for each ordered query.	▢	▢	▢
Which we currently are saved from using CODE_FRAGMENT.	▢	▢	▢
I am trying to link my native library to FILE_NAME application but when I try to run it I get a CODE_FRAGMENT exception complaining about missing symbols (CODE_FRAGMENT).	▢	▢	▢
So, everything builds fine, but when we try to deploy the application to GFNUMBER we get the FILE_NAME file not found ''error.	▢	▢	▢
I can successfully call my service but when generating the response, it seems to crash.	▢	▢	▢
A bi thanks for this :) we all are really happy that now this is fully supported by Core. Now is missingvarehicles support and MaNGOS will rulez =D Congrats for all your work	▢	▢	▢
And it works like a charm now SMILE_FACE.	▢	▢	▢

Submit

Abbildung 4.2: Bewertung der Aussagen anhand ihrer jeweiligen Sentiment-Polarität im ersten Teil der Umfrage

in Herrmanns Arbeit [15] angewandten logistischen Regressionsanalyse und werden im zweiten Teil der Umfrage relevant. Die nachstehende Tabelle 4.1 listet die zehn fest ausgewählten Aussagen auf.

Im ersten Teil der Umfrage wiederum wurden die Studienteilnehmer dazu gebeten, sich die zehn Aussagen genau durchzulesen und jede Aussage danach zu bewerten, wie sie auf den Studienteilnehmer wirkt. Wichtig war dabei, dass die initiale Wahrnehmung zu jeder Aussage annotiert wird. Es bestand dabei wie oftmals bekannt in der Stimmungsanalyse [11] die Auswahl zwischen den Sentiment-Polaritäten positiv, neutral und negativ. Insgesamt wurden im ersten Teil der Umfrage also zehn Sentiment-Polaritäten von jedem Teilnehmer angegeben.

Die Studienteilnehmer wurden darüber informiert, die Freiheit zu haben, bei der Bewertung der Aussagen ein beliebiges Bewertungskriterium zu wählen. Außerdem wurde ausdrücklich darauf hingewiesen, dass die erste

Aussage	Text
V_1	Trust URI.
V_2	(Hopefully with a good example.)
V_9	why allocate a new String instance? def apply(name: String): Node = hash(name) def fromHash(hash: String): Node = hash
V_{21}	Comparison time should be fast, so total run time should be only slightly more than sum of run time for each ordered query.
V_{41}	Which we currently are saved from using CODE_FRAGMENT.
V_{49}	I am trying to link my native library to FILE_NAME application but when I try to run it I get a CODE_FRAGMENT exception complaining about missing symbols (CODE_FRAGMENT).
V_{54}	So, everything builds fine, but when we try to deploy the application to GFNUMBER we get the FILE_NAME file not found "error.
V_{57}	I can successfully call my service but when generating the response, it seems to crash.
V_{78}	A bi thanks for this :) we all are really happy that now this is fully supported by Core. Now is missingvarehicles support and MaNGOS will rulez =D Congrats for all your work
V_{80}	And it works like a charm now SMILE_FACE.

Tabelle 4.1: Zehn fest ausgewählte Prädiktoraussagen aus Datensatz von 96 Aussagen von Herrmanns Arbeit [15]

Wahrnehmung in der Bewertung wichtig ist und daher nicht die Notwendigkeit besteht, die Aussagen zu interpretieren. Durch die Anwesenheit des Moderators wurde die Möglichkeit gegeben, Hilfestellung in der Übersetzung der englischsprachigen Aussagen zu erhalten. Zu der Beeinflussung in der Bewertung der Wahrnehmung jeder Aussage gehörte ebenfalls die Verarbeitung der annotierten Aussagen, weshalb auch aufgeklärt wurde, dass es keine richtigen oder falschen Antworten gibt. Somit wurde jeder Studienteilnehmer darüber in Kenntnis gesetzt, dass er sich in einem Umfeld befindet, in dem er nicht beurteilt oder unter Druck gesetzt wird.

4.1.3 Vorhersage der Gruppenzugehörigkeit

Nachdem jede der zehn gegebenen Aussagen nach der ersten Wahrnehmung bewertet wurde, wurde nach einer Bestätigung auf den nächsten Teil der Umfrage weitergeleitet. Damit der Studienteilnehmer nicht den Eindruck erhält, dass die annotierten Aussagen berechnet werden, blieb die Logik in diesem Teil verborgen. Die folgende Abbildung 4.3 zeigt die zugehörige Seite.

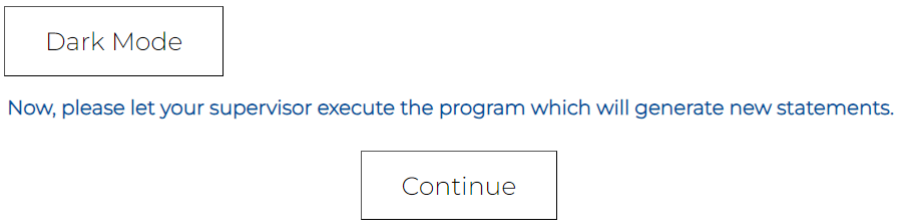

Abbildung 4.3: Zwischenteil der Umfrage für die Generierung neuer Aussagen

Darin wurde zwar erwähnt, dass ein Programm neue Aussagen generiert, jedoch wurde nicht weiter ausgeführt, ob und inwieweit der erste Teil der Umfrage mit der Generierung zusammenhängt.

Die Logik hinter dem ausgeführten Skript wird nachfolgend beschrieben. Zunächst werden die Antworten des Studienteilnehmers im ersten Teil der Umfrage abgespeichert und danach verarbeitet. Dazu wird das logistische Regressionsmodell in Verbindung mit den Ergebnissen aus Herrmanns Arbeit [15] verwendet. In seiner Masterarbeit hat Herrmann [15] zwei Teilnehmergruppen mit unterschiedlichen Wahrnehmungen identifiziert, welche in dieser Bachelorarbeit wiederverwendet werden. Mit Hilfe der logistischen Regressionsanalyse wird der Studienteilnehmer je nach annotierten Aussagen entweder der ersten oder der zweiten Teilnehmergruppe zugewiesen. Dabei stellen die zehn Aussagen aus dem ersten Umfrageteil die Prädiktorvariablen dar. Ausgewählt wurden die fünf Aussagen $V_{21}, V_{49}, V_{57}, V_{78}$ und V_{80}, mit welchen sich eine vollständige Separation der Teilnehmergruppen und damit eine Vorhersagegenauigkeit von 100% treffen lässt [15]. Für die Sicherstellung einer korrekten Vorhersage und für die Länge der Umfrage wurden weitere fünf Aussagen hinzugefügt, welche neben den ersten fünf Aussagen die größte Relevanz für die Vorhersagegenauigkeit besitzen [15].

Anschließend werden zehn Aussagen aus dem Datensatz mit 96 Aussagen exklusive der Aussagen X ausgewählt. Abhängig von der Gruppenzugehörigkeit werden den Studienteilnehmern die Aussagen mit unterschiedlichen Sentiment-Polaritäten negativ, neutral und positiv gezeigt, wobei -1 eine negative, 0 eine neutrale und 1 eine positive Wahrnehmung darstellt. Die Sentiment-Polaritäten der 96 Aussagen repräsentieren dabei die Wahrnehmung der jeweiligen Teilnehmergruppe und haben ihren Ursprung in der Datenanalyse aus Herrmanns Arbeit [15], wobei jeweils der Median der jeweiligen Gruppe für die einzelnen Aussagen bestimmt wurde. Das Ziel bei der Durchführung des zweiten Teils der Umfrage ist es herauszufinden, ob die Wahrnehmungen mit den durch das Skript bestimmten Wahrnehmungen übereinstimmt. Somit wird die Notwendingkeit einer Kalibrierung der Stimmungsanalyse bewertet.

Bei der Auswahl der neuen Aussagen muss beachtet werden, dass die dem

Studienteilnehmer bekannten Aussagen nicht erneut in Erscheinung treten dürfen, weil diese bereits annotiert wurden. Damit jeder Studienteilnehmer unterschiedliche Aussagen erhält, wurde das Zufallsprinzip gewählt. Die Auswahl dieses Ansatzes birgt jedoch die Notwendigkeit einer Variabilität zwischen den Wahrnehmungen der Aussagen. Um dies zu bewerkstelligen, wurde die folgende Bedingung eingebaut: Die zehn neuen Aussagen müssen jeweils drei positive, drei negative, zwei neutrale und zwei Aussagen mit einer beliebigen Wahrnehmung enthalten.

Das Ziel dieser Arbeit ist es zu zeigen, ob eine Kalibrierung von Stimmungsanalysetools sinnvoll ist. Aufgrund dessen ist es nicht nur wichtig, zu überprüfen, ob die Wahrnehmungen der Studienteilnehmer mit den vorhergesagten Wahrnehmungen übereinstimmen. Neben der Überprüfung der Aufmerksamkeit bei der Bewertung der Wahrnehmungen der Aussagen muss ebenfalls die Zustimmung zu Wahrnehmungen, die nicht repräsentativ für die jeweilige Teilnehmergruppe sind, überprüft werden. Dazu werden die Sentiment-Polaritäten zweier Aussagen invertiert. Eine negativ vorhergesagte Aussage wird also bewusst positiv präsentiert und eine positiv vorhergesagte Aussage wiederum negativ. Durch diese Vorangehensweise wird beispielsweise leicht erkannt, dass eine vollständige Zustimmung zu allen Aussagen sehr unwahrscheinlich ist und die Möglichkeit in Betracht gezogen werden kann, dass die Aussagen nicht aufmerksam oder gar nicht gelesen wurden.

4.1.4 Angabe der Zustimmung

Nach der Ausführung des Skripts durch den Moderator und einer Bestätigung gelangte der Studienteilnehmer zu dem zweiten Teil der Umfrage. Der zweite Teil der durchgeführten Umfrage umfasste die im Kapitel 4.1.3 erwähnten zehn neuen Aussagen Y, die mit Hilfe des ausgeführten Skripts bestimmt wurden. Die obige Abbildung 4.4 dient zur Veranschaulichung des auf der neuen Webseite enthaltenen nächsten Umfrageteils.

Wie man sehen kann, wurden dem Studienteilnehmer in diesem Umfrageteil zehn weitere Aussagen präsentiert, wobei jeder Aussager nun eine der drei Sentiment-Polaritäten positiv, neutral oder negativ zugewiesen wurde. Diese Sentiment-Polaritäten stellen die vorhergesagten Wahrnehmungen aus dem Zwischenteil der Umfrage dar. Da die Zustimmung jedes Studienteilnehmers zu den vorhergesagten Wahrnehmungen überprüft werden soll, fiel die Entscheidung für die Bewertung auf eine zweistufige Likert-Skala im zweiten Umfrageteil. Diese enthielt die Auswahl der englischsprachigen Möglichkeit, der gegebenen Sentiment-Polarität mit „*Agree*“ zuzustimmen oder mit „*Disagree*“ abzulehnen (vgl. 4.4). Im Falle einer Ablehnung wurde der Studienteilnehmer dazu gebeten, seine Wahrnehmung zu der gegebenen Aussage zu annotieren. Wurde beispielsweise eine positiv gewertete Aussage

Dark Mode

In the following, some other statements are presented to you. As you can see, the polarities negative, neutral or positive are already given. Please note if you agree with them and if not, rate what polarity would match better.

Statements	Polarity	Agree	Disagree	Negative	Neutral	Positive
I think they should be the other way around. It's the `consistentHash` that should be in the atomic since you don't want to overwrite it with a stale `consistentHash`. Overwriting the `consistentHashRoutees` with stale values will just trigger a new updat	Neutral	☐	☐	☐	☐	☐
This impl is weird. Uses system identityHashCode for hashCode AND equals?	Positive	☐	☐	☐	☐	☐
I have looked and found that there are some packages that would automatically enter my keyring on login but that isn't really an option.	Neutral	☐	☐	☐	☐	☐
Now we're getting to the good part.	Positive	☐	☐	☐	☐	☐
lol :)	Positive	☐	☐	☐	☐	☐
Still, we need a potentially unlimited number of buffers. What should we do with them if we don't need them any more if we don't let them be garbage collected?	Neutral	☐	☐	☐	☐	☐
Ops, that seems to have [failed the build] (https://travis-ci.org/rails/rails/builds/6066789) = (	Negative	☐	☐	☐	☐	☐
If I run the code in the GUI, it just hangs.	Negative	☐	☐	☐	☐	☐
OMG stupid me	Neutral	☐	☐	☐	☐	☐
There is a good example showing how to put a file onto WebDAV server.	Negative	☐	☐	☐	☐	☐

Submit

Abbildung 4.4: Bewertung der Zustimmung zu den gegebenen Aussagen samt der vorgeschlagenen Sentiment-Polarität im zweiten Teil der Umfrage

abgelehnt, hatte der Studienteilnehmer die Auswahl, die Aussage als neutral oder negativ zu werten.

Wie in Kapitel 4.1.3 beschrieben, enthält der zweite Teil der Umfrage zwei Aussagen mit invertierten emotionalen Sentiment-Polaritäten. Die invertierte Sentiment-Polarität zu positiv stellt negativ und zu negativ wiederum positiv dar. Damit der Studienteilnehmer die invertierten Sentiment-Polaritäten nicht leicht identifizieren kann, wurde eine zufällige Ordnung der Aussagen sichergestellt. Notwendig ist diese Maßnahme besonders, da die Aussagen aus der Datenbasis eine feste Ordnung besitzen und bei der Invertierung der Sentiment-Polaritäten stets die erste negativ und die erste positiv bewertete Aussage betrachtet wurden. Somit wurden durch die zufällige Ordnung der präsentierten Aussagen zwei aufeinanderfolgende invertierte Sentiment-Polaritäten verhindert. Wurde die Zustimmung zu der gegebenen Sentiment-Polarität jeder Aussage abgegeben, so wurden die angegebenen Daten nach einer Bestätigung abgespeichert und damit die Umfrage beendet.

4.2 Datenverarbeitung

Bei der Verarbeitung aller eingegebenen Daten werden beide Teile der Umfrage betrachtet. Im ersten Teil der Umfrage werden für jeden Studienteilnehmer die Sentiment-Polaritäten zu jeder Aussage mit einer eindeutigen Identifikationsnummer gespeichert. Für den zweiten Teil der Umfrage werden die durch das Skript bestimmten Sentiment-Polaritäten und die Stärke der Zustimmung mit diesen zu jeder Aussage gespeichert. Die Stärke der Zustimmung liegt in $[0, 2]$ und wird wie folgt bestimmt. Die Zustimmung zu der gegebenen Sentiment-Polarität bedeutet die stärkste Zustimmung und erhält somit den Wert 2. Ein Wert von 0 oder 1 ergibt sich für das Ankreuzen der Antwortmöglichkeit „*Disagree*“. Wurde eine vorgegebene emotionale Wahrnehmung abgelehnt und die invertierte Sentiment-Polarität als eigene Wahrnehmung angegeben, so entsprach diese Angabe einem Wert von 0. Das Speichern des Wertes 1 wurde hingegegen durch zwei Fälle verursacht: Im ersten Fall enthielt die jeweilige Aussage eine neutrale Wahrnehmung als Vorgabe. Somit konnte nur eine emotionale Sentiment-Polarität als eigene Wahrnehmung gewählt werden. Im zweiten Fall wurde entgegen des ersten Falls eine emotionale Wahrnehmung vor- und eine neutrale Wahrnehmung als eigene angegeben. Damit nicht dieselben Wahrnehmung bei einer Ablehnung gewählt werden und keine falschen Werte für die Stärke der Zustimmung bestimmt werden, erscheinen die vorgegebenen Wahrnehmung als ausgegraut. Ebenfalls können die subjektiven Wahrnehmungen nicht selektiert werden, solange keine der beiden Optionen „*Agree*“ und „*Disagree*“ oder eine Zustimmung angegeben wurde. Des Weiteren wurde aufgrund der zufälligen Ordnung der Aussagen gespeichert, ob die zugehörige vorgegebene Wahrnehmung invertiert ist oder nicht, damit eine einfache Analyse der invertierten Aussagen möglich ist. Die finale Datei enthält schließlich alle Aussagen für jeden Studienteilnehmer, wobei die genannten Daten für die jedem einzelnen Studienteilnehmer zehn präsentierten Aussagen gespeichert werden.

Um die Notwendigkeit einer Vervollständigung fehlender Daten zu vermeiden, wurde außerdem die Funktion eines ausgegrauten Buttons hinzugefügt. Während der Button im ersten Umfrageteil bei zehn Kreuzen wählbar ist, ist die Anzahl der Kreuze im zweiten Umfrageteil variabel. Bei jedem Setzen oder Abwählen eines Kreuzes wird daher die Wählbarkeit des Buttons so aktualisiert, dass die benötigten Daten vollständig sind. Dazu wird die Bedingung festgelegt, dass für jede Aussage ein Kreuz für der Zustimmung oder zwei Kreuze jeweils für die Angabe der Ablehnung und der eigene Wahrnehmung gesetzt werden müssen. Insgesamt müssen also zwischen zehn und zwanzig Kreuze gesetzt werden.

4.3 Evaluation der Ergebnisse

Dieses Kapitel thematisiert die Vorangehensweise bei der Analyse der Wahrnehmung der einzelnen Teilnehmergruppen aus der durchgeführten Studie. Dabei wird erläutert, wie das in Kapitel 2 beschriebene statistische Testverfahren auf die annotierten Aussagen während und nach der Durchführung der Umfrage in Abbildung 4.1 angewandt wird.

4.3.1 Hypothesentests für die Wahrnehmung der Prädiktoraussagen

Ziel dieser Arbeit ist es zu zeigen, ob es signifikante Unterschiede in der Zustimmung zu gegebenen Aussagen durch die verschiedenen Teilnehmergruppen gibt. Da im ersten Teil dazu zehn Aussagen nach ihrer Sentiment-Polarität annotiert werden, ist es interessant zu sehen, ob sich die Wahrnehmungen der verschiedenen Teilnehmergruppen in dieser Arbeit mit denen der beiden Teilnehmergruppen von Herrmanns Arbeit gleichen [15]. Infolgedessen besteht eine Notwendigkeit für die Verwendung eines statistischen Testverfahrens auf die Annotationen der zehn fest gewählten Aussagen $X = \{V_1, V_2, V_9, V_{21}, V_{41}, V_{49}, V_{54}, V_{57}, V_{78}, V_{80}\}$ aus der Datenbasis von 96 Aussagen V_1 bis V_2.

	Definition der Nullhypothese
$H1_0$	Es gibt keine Unterschiede in der Wahrnehmung der Aussagen X zwischen den beiden Teilnehmergruppen

Tabelle 4.2: Definition der Nullhypothese $H1_0$

Für die Überprüfung der zehn genannten Aussagen X wurde, wie in Tabelle 4.2 dargestellt, die zu überprüfende Nullhypothese $H1_0$ aufgestellt. Da lediglich eine einzelne Nullhypothese überprüft wird, muss das nach Fisher bekannte Signifikanzniveau von $\alpha = 0.05$ [10] nicht angepasst werden. Liegt der Werte der Nullhypothesen $H1_0$ nach Durchführung des Mann-Whitney-U-Tests nun unter dem Signifikanzniveau α, so muss diese abgelehnt werden.

4.3.2 Hypothesentests für die Wahrnehmung der einzelnen Sentiment-Polaritäten der Prädiktoraussagen

Da in dieser Arbeit die Anzahl der annotierten Aussagen von 96 auf insgesamt zehn Aussagen $X = \{V_1, V_2, V_9, V_{21}, V_{41}, V_{49}, V_{54}, V_{57}, V_{78}, V_{80}\}$ verringert wurde, kann der Mann-Whitney-U-Test für die Gesamtwahrnehmung der ersten zehn Aussagen durchaus auch nicht signifikant ausfallen. Daher wurde die einzelne Wahrnehmung der drei verschiedenen Sentiment-

Polaritäten $P \subset \{negativ, neutral, positiv\}$ durch die beiden Teilnehmergruppen auf signifikante Unterschiede überprüft. Dadurch kann überprüft werden, ob sich die Wahrnehmung zwischen den verschiedenen Sentiment-Polaritäten anders verhält. Die Tabelle 4.3 visualisiert die aufgestellte übergeordnete Nullhypothese $H2_0$ und die drei untergeordneten Nullhypothesen $H2(negativ)_0$, $H2(neutral)_0$, $H2(positiv)_0$.

	Definition der Nullhypothesen
$H2_0$	Es gibt keine Unterschiede in den Anteilen der Sentiment-Polaritäten zwischen den beiden Teilnehmergruppen
$H2(P)_0$	Es gibt keine Unterschiede in den Anteilen der Sentiment-Polarität P zwischen den beiden Teilnehmergruppen, wobei $P \subset \{negativ, neutral, positiv\}$

Tabelle 4.3: Definition der übergeordneten Nullhypothese $H2_0$ und der drei untergeordneten Nullhypothesen $H2(P)_0$

Das Signifikanzniveau $\alpha = 0.05$ nach Fisher [10] wurde hier mit Hilfe der Bonferroni-Korrektur [2] auf $\alpha = 0.05/3 = 0.016\overline{6}$ herabgesetzt, da die zehn zu annotierenden Aussagen X drei Sentiment-Polaritäten besitzen können. Sobald das Signifikanzniveau von einer Nullhypothese $H2(P)_0$ den genannten Wert α unterschreitet, so muss auch die übergeordnete Nullhypothese $H2_0$ verworfen werden.

4.3.3 Hypothesentests für die Zustimmung zu einer invertierten Wahrnehmung

Wie man in der obenstehenden Tabelle 4.4 sehen kann, wurden zwei Nullhypothesen aufgestellt. Diese sollen überprüft werden, um zu zeigen, ob es statistisch signifikante Unterschiede in der Zustimmung zu den gegebenen Wahrnehmungen im zweiten Teil der Umfrage gibt.

Jeder Studienteilnehmer erhält zehn zufällige Aussagen aus der Menge von 96 Aussagen V_1 bis V_{96} ohne die Prädiktorvariablen $X \subset Y$. Aufgrund der Überprüfung von zwei untergeordneten Aussagen für die jeweilige Gruppe eines jeden Studienteilnehmers musste das durch Fisher gewählte Signifikanzniveau von $\alpha = 0.05$ [10] mit der Bonferroni-Korrektur [2] heruntergesetzt werden auf $\alpha = 0.05/2 = 0.025$. Liegen die Ergebnisse des Mann-Whitney-U-Tests von mindestens einer der zwei untergeordneten Nullhypothesen $H3(G_1)_0$ oder $H3(G_2)_0$ unter dem Signifikanzniveau, so muss demzufolge die untergeordnete Hypothese $H3(G_1)_0$ oder $H3(G_2)_0$ sowie die übergeordnete Nullhypothese $H3_0$ abgelehnt werden.

	Definition der Nullhypothesen
$H3_0$	Es gibt keine Unterschiede in der Zustimmung zwischen den invertierten und nicht invertierten Sentiment-Polaritäten für beide Teilnehmergruppen
$H3(G_1)_0$	Es gibt keine Unterschiede in der Zustimmung von $Y_i, i \in \{V_1, V_2, ..., V_{96}\} = V, Y \subset V \setminus X, \lvert Y \rvert = 10$ zwischen den invertierten und nicht invertierten Sentiment-Polaritäten für die erste Teilnehmergruppe
$H3(G_2)_0$	Es gibt keine Unterschiede in der Zustimmung von $Y_i, i \in \{V_1, V_2, ..., V_{96}\} = V, Y \subset V \setminus X, \lvert Y \rvert = 10$ zwischen den invertierten und nicht invertierten Sentiment-Polaritäten für die zweite Teilnehmergruppe

Tabelle 4.4: Definition der übergeordneten Nullhypothese $H3_0$ und der zwei untergeordneten Nullhypothesen $H3(G_1)_0$ und $H3(G_2)_0$ für die erste und zweite Teilnehmergruppe(G_1 und G_2)

4.3.4 Hypothesentests für die Zustimmung zur gegebenen Wahrnehmung

Nun soll überprüft werden, ob ein signifikanter Unterschied in der Zustimmung zur gegebenen Wahrnehmung zwischen den zwei Teilnehmergruppen besteht. Dazu wird bei der Durchführung des Mann-Whitney-U-Tests nun außer Acht gelassen, ob die emotionalen Sentiment-Polaritäten invertiert sind oder nicht. Tabelle 4.5 stellt die aufgestellte Nullhypothese dar.

	Definition der Nullhypothese
$H4_0$	Es gibt keine Unterschiede in der Zustimmung zwischen den verschiedenen Teilnehmergruppen

Tabelle 4.5: Definition der Nullhypothese $H4_0$

Das Signifikanzniveau musste aufgrund der Überprüfung der Zustimmung für jeden Studienteilnehmer in einer Nullhypothese $H4_0$ nicht angepasst werden und bleibt daher bei dem durch Fisher gewählten Wert von $\alpha = 0.05$. Es wird auch hier die Zustimmung der zehn Aussagen aus dem zweiten Teil der Umfrage geprüft, die eine Teilmenge aus der Datenbasis von 96 Aussagen V_1 bis V_{96} ohne die Prädiktorvariablen $X \subset Y$ darstellen. Dabei wurden jedem Studienteilnehmer dieselben Sentiment-Polaritäten gezeigt: Jeweils drei positive und drei negative, zwei neutrale sowie zwei zufällige Sentiment-Polaritäten. Ist nun das Signifikanzniveau der Nullhypothese $H4_0$ unter dem Signifikanzniveau α, so muss die Nullhypothese abgelehnt werden. Ein signifikanter Unterschied lässt darauf schließen, dass die Wahrnehmung einer

der beiden Teilnehmergruppen häufiger mit der gegebenen Wahrnehmung übereinstimmt als die der jeweils anderen Teilnehmergruppe.

4.4 Visualisierung der Ergebnisse

Die Visualisierung der Ergebnisse der in dieser Arbeit durchgeführten Studie erfolgt mittels Histogramme. Durch Histogramme wird die allgemeine Häufigkeitsverteilung etwas deutlicher dargestellt [4].

Zu überprüfen gilt vor allem die Verteilung der durch die Studienteilnehmer angegebenen Sentiment-Polaritäten im ersten Teil der Umfrage. Interessant ist dabei, ob sich die Verteilung der Sentiment-Polaritäten ähnlich verhält wie in Herrmanns Masterarbeit [15] für die zwei ermittelten Teilnehmergruppen. Zudem ist es wichtig, die Stärke der Zustimmung als Verteilung im zweiten Umfrageteil in einem Histogramm zu sehen. Besonders die Verteilung der Zustimmung zu den invertierten sowie nicht invertierten Sentiment-Polaritäten ist hierbei in einem Vergleich wichtig. Es sollen alle Studienmitglieder, danach die erste und schließlich die zweite Gruppe in Betracht gezogen werden, die in den drei Mann-Whitney-U-Tests untersucht wurden, um einen eventuellen Unterschied in der Wahrnehmung festzustellen. Für die Evaluierung bietet die Betrachtung der Verteilung der Sentiment-Polaritäten bei einer Zustimmung zu der gegebenen Polarität zudem einen guten Einblick in die Wahrnehmung der verschiedenen Gruppen. Daher soll überprüft werden, welche Sentiment-Polarität am häufigsten vertreten ist, sofern der Studienteilnehmer der gegebenen Polarität zugestimmt hat. Um eine Erkenntnis darüber zu erhalten, ob die Kalibrierung von Stimmungsanalysetools sinnvoll ist, soll die Wichtigkeit des Vergleiches zwischen Antworten für die invertierten und nicht invertierten Sentiment-Polaritäten hervorgehoben werden.

Kapitel 5

Evaluation

Dieses Kapitel trägt die Ergebnisse aus der im Kapitel 4 durchgeführten Studie zusammen. Neben der Visualisierung werden die Ergebnisse hierbei auch im Hinblick auf die Wahrnehmung und Zustimmung statistisch ausgewertet.

5.1 Evaluation der Wahrnehmung der Prädiktoraussagen

In diesem Kapitel werden die Ergebnisse des ersten Teils der durchgeführten Umfrage vorgestellt. Dabei wird auf die Bewertung der Wahrnehmungen der teilgenommenen Entwickler eingegangen, die mit Hilfe der verwendeten statistischen Testverfahren auf Signifikanzen überprüft wurden.

5.1.1 Ergebnisse des Shapiro-Wilk-Tests für die Wahrnehmung

Um ein geeignetes Statistikverfahren für die untersuchten Daten zu bestimmen, wurden die in Kapitel 2.3 genannten Vorbedingungen der Normalverteilung und Homoskedastizität überprüft. Für die Normalverteilung wurde der Shapiro-Wilk-Test auf die Annotationen angewandt [36]. Fällt dieser Test für mindestens eine der beiden Teilnehmergruppen signifikant aus, so sind die Daten nicht normalverteilt und der Mann-Whitney-U-Test muss angewandt werden [36]. Ebenfalls müssen die Daten ordinalskaliert sein [21]. Aufgrund der Überprüfung von drei Sentiment-Polaritäten (positiv, neutral, negativ) mit unklarem Abstand, aber einer Rangordnung, ist das ordinale Skalenniveau gegeben.

Liefert der Shapiro-Wilk-Test kein signifikantes Ergebnis, dann sind die Daten nicht normalverteilt und es wird mit Hilfe des Bartlett-Tests [1] die Vorbedingung der Homoskedastizität für die Verwendung des Zwei-Stichproben-t-Tests [37] oder des Welch-Tests [42] überprüft.

Merkmal	**G**	**W**	**p**	**Statistischer Test**
Anteil negativer Annotationen	1	0.8763	0.0515	Mann-Whitney-U-Test
	2	0.7109	**0.0030**	
Anteil neutraler Annotationen	1	0.9217	0.2326	Bartlett-Test
	2	0.8733	0.1623	
Anteil positiver Annotationen	1	0.7855	**0.0033**	Mann-Whitney-U-Test
	2	0.9009	0.2945	

Tabelle 5.1: Ergebnisse des Shapiro-Wilk-Tests für die Verteilung der Sentiment-Polaritäten für beide Teilnehmergruppen ($G = 1$ und $G = 2$)

Merkmal	χ^2	**p**	**Statistischer Test**
Anteil neutraler Annotationen	0.3044	0.5811	Zwei-Stichproben-t-Test

Tabelle 5.2: Ergebnisse des Bartlett-Tests für die neutralen Annotationen

In der nachstehenden Tabelle 5.1 werden die Ergebnisse des Shapiro-Wilk-Tests aufgeführt. Überprüft wurde die Normalverteilung der Anteile der jeweiligen Sentiment-Polaritäten negativ, neutral und positiv für die erste und zweite Teilnehmergruppe. Das Signifikanzniveau liegt bei $\alpha = 0.05$ [36] und wird, wie in der Tabelle zu sehen ist, für mindestens eine der beiden Teilnehmergruppen für die annotierten Sentiment-Polaritäten positiv und negativ unterschritten. Während für die negativen Annotationen die zweite Gruppe mit einem p-Wert von 0.0033 eine Signifikanz aufweist und für keine Normalverteilung spricht, ist dies bei den positiven Annotationen für die erste Gruppe mit einem p-Wert von 0.003 der Fall. Die Entscheidung für die statistische Analyse fällt aufgrund der nicht normalverteilten Daten folglich auf den Mann-Whitney-U-Test [36].

Das Ergebnis des Shapiro-Wilk-Tests fällt für die Aussagen, die neutral bewertet wurden, jedoch nicht signifikant aus, da der p-Wert für beide Gruppen das Signifikanzniveau $\alpha = 0.05$ [36] überschreitet. Die Daten für die neutralen Annotationen sind daher normalverteilt und es muss die weitere Bedingung der Homoskedastizität überprüft werden, um sich für die Verwendung des Zwei-Stichproben-t-Tests oder dem Welch-Test zu entscheiden. Die Ergebnisse des durchgeführten Bartlett-Tests wurden in Tabelle 5.2 aufgeführt [1]. Wie man sehen kann, liegt auch der p-Wert des Bartlett-Tests nicht unter dem Signifikanzniveau von $\alpha = 0.05$ [1].

Infolgedessen weisen die Standardvarianzen für die neutralen Annotationen Ähnlichkeiten auf und der Zwei-Stichproben-t-Test muss für die statistische Untersuchung verwendet werden [37].

5.1.2 Hypothesenprüfung von $H1_0$

In Kapitel 4.3 wurde die Nullhypothese $H1_0$ aufgestellt. Diese Nullhypothese besagt, dass es keine signifikanten Unterschiede in den Wahrnehmungen beider Teilnehmergruppen gibt. Mit der Durchführung des Mann-Whitney-U-Tests soll also die allgemeine Wahrnehmung der zehn Aussagen X im ersten Teil der Umfrage zwischen beiden Teilnehmergruppen auf statistische Signifikanz überprüft werden. Eine Signifikanz spräche dafür, dass eine Teilnehmergruppe positiver oder negativer denkt als die jeweils andere. Tabelle 5.3 stellt die Ergebnisse aus dem Mann-Whitney-U-Test dar.

Merkmal	**G**	**ø**	**σ**	**U**	**p**	**Interpretation**
Wahrnehmung	1	0.15	0.7552	5490	0.7968	Nicht signifikant
	2	0.1625	0.8433			

Tabelle 5.3: Ergebnisse des Mann-Whitney-U-Tests für die Wahrnehmung der zehn Aussagen im ersten Teil der Umfrage für beide Teilnehmergruppen ($G = 1$ und $G = 2$)

Anhand des Durchschnitts der gesamt annotierten Sentiment-Polaritäten kann bereits festgestellt werden, dass es keine großen Unterschiede in den Wahrnehmungen zwischen der ersten und zweiten Gruppe gibt. Während der Durchschnitt für Gruppe 1 $= 0.15$ beträgt, liegt der Wert für Gruppe 2 bei $= 0.1625$. Für beide Gruppen lässt sich also eine ähnliche Wahrnehmung vermuten. Die Ergebnisse des Mann-Whitney-U-Tests geben jedoch einen genaueren Einblick in die Unterschiede in den Wahrnehmungen. Das Signifikanzniveau für die Nullhypothese $H1_0$ liegt für diesen Test bei $\alpha = 0.05$ und wird von dem p-Wert des Ergebnisses unterschritten. Da die Nullhypothese $H1_0$ damit abgelehnt werden kann, gibt es also keinen signifikanten Unterschied zwischen den Wahrnehmungen beider Teilnehmergruppen. Dies kann einerseits daran liegen, dass statt 96 Aussagen zehn verwendet werden. Andernfalls ist die Datenbasis von n=22 Aussagen im Vergleich zu n=94 aus der Studie von Obaidi et al. [30] gering. Eine weitere Möglichkeit stellen die Wahrnehmungen beider Teilnehmergruppen für jeweils andere Aussagen aus X dar. Beispielsweise kann die Wahrnehmung der ersten Teilnehmergruppe für die ersten fünf Aussagen genauso positiv ausfallen wie die Wahrnehmung der zweiten Teilnehmergruppe für die letzten fünf Aussagen des ersten Umfrageteils, obwohl der Durchschnitt bei Betrachtung aller Aussagen für dieselbe Wahrnehmung spricht.

5.1.3 Hypothesenprüfung von $\mathbf{H2}_0$ und $\mathbf{H2}(P)_0$

Um eine zusätzliche Überprüfung der einzelnen Sentiment-Polaritäten zwischen beiden Teilnehmergruppen im ersten Teil der Umfrage im Hinblick auf eine statistische Signifikanz durchzuführen, wurden Nullhypothesen aufgestellt. Während die übergeordnete Nullhypothese $\mathrm{H2}_0$ keine Unterschiede zwischen den Anteilen der Sentiment-Polaritäten ausdrückt, so unterteilen die drei untergeordneten Nullhypothesen $\mathrm{H2}(P)_0$ die Überprüfung in die jeweiligen Sentiment-Polaritäten negativ, neutral und positiv. Die nachstehende Tabelle 5.4 visualisiert die Ergebnisse des auf die negativ und positiv bewerteten Aussagen X angewandten Mann-Whitney-U-Tests aus dem ersten Umfrageteil. Ebenfalls werden die Ergebnisse aus dem Zwei-Stichproben-t-Test dargestellt, welcher auf die neutral bewerteten Aussagen angewandt wurde.

Anteile in %	**G**	**ø**	**σ**	**U**	**p**	**Interpretation**
Negativ	1	22.14	10.13	35	0.1435	Nicht signifikant
	2	28.75	12.69			
Positiv	1	37.14	11.61	48	0.5954	Nicht signifikant
	2	45.00	20.62			
Anteile in %	**G**	**ø**	**σ**	**T**	**p**	**Interpretation**
Neutral	1	40.71	12.80	2.2849	0.0334	**Signifikant**
	2	26.25	14.95			

Tabelle 5.4: Ergebnisse des Mann-Whitney-U-Tests und des Zwei-Stichproben-t-Tests für die Anteile der Sentiment-Polaritäten der zehn Aussagen im ersten Teil der Umfrage für beide Teilnehmergruppen ($G = 1$ und $G = 2$)

Aufgrund der Überprüfung von drei Sentiment-Polaritäten wurde das Signifikanzniveau mittels der Bonferroni-Korrektur auf den Wert $\alpha = 0.05/3 = 0.016\overline{6}$ [10] für die untergeordneten Nullhypothesen $\mathrm{H2}(P)_0$ heruntergesetzt. Die Tabelle 5.4 zeigt genauso wie bei der Überprüfung der Gesamtwahrnehmung der zehn Aussagen, dass es keine signifikanten Unterschiede in der Wahrnehmung der Sentiment-Polaritäten gibt, da der p-Wert für jede Sentiment-Polarität unter dem Signifikanzniveau α liegt. Auffallend ist jedoch für die Häufigkeitsverteilung für die neutral bewerteten Aussagen, dass sich der Durchschnitt zwischen beiden Teilnehmergruppen stärker unterscheidet als für die anderen Sentiment-Polaritäten. Bei der Anwendung des Mann-Whitney-U-Tests für die neutral bewerteten Aussagen überschreitet der p-Wert das Signifikanzniveau α für die untergeordnete Nullhypothese $\mathrm{H2}(neutral)_0$, liegt jedoch trotzdem unter dem Wert $\alpha = 0.05$. Damit muss die untergeordnete Nullhypothese $\mathrm{H2}(neutral)_0$ zwar

abgelehnt werden, die übergeordnete Nullhypothese $H2_0$ aber nicht. Es gibt folglich eine unterschiedliche Wahrnehmung für die neutralen Sentiment-Polaritäten, jedoch keine Unterschiede für alle Sentiment-Polaritäten zwischen den beiden Teilnehmergruppen. Da im ersten Umfrageteil dieser Arbeit genauso wie in Herrmanns Arbeit [15] die Wahrnehmungen der Studienteilnehmer überprüft wurden, ist es interessant, die Ergebnisse miteinander zu vergleichen. Während die Wahrnehmungen der neutralen und positiven Sentiment-Polaritäten in Herrmanns Arbeit [15] Signifikanzen aufweisen, unterscheiden sich die Wahrnehmungen in dieser Arbeit nur für die neutralen Sentiment-Polaritäten. Untersucht man jedoch die Häufigkeitsverteilung der verschiedenen Annotationen, sieht man eine gewisse Ähnlichkeit zwischen den Daten dieser Arbeit und denen von Herrmanns Arbeit. Für die negativ bewerteten Aussagen unterscheiden sich die Anteile kaum und liegen bei ca. 20% in beiden Arbeiten [15]. Auch bei den neutral und positiv bewerteten Aussagen findet sich eine ähnliche Verteilung zwischen der ersten und zweiten Gruppe wieder. In Herrmanns Arbeit gab die erste Teilnehmergruppe ca. 8% häufiger eine neutrale Sentiment-Polarität und ca. 10% häufiger eine positive Sentiment-Polarität für die vorliegenden Aussagen an als die zweite Teilnehmergruppe [15]. Die Differenz in der Häufigkeit zwischen der ersten und zweiten Teilnehmergruppe ähnelt für die neutrale Wahrnehmung der positiven in Herrmanns Arbeit mit einem Wert von 10%. Die Differenz für die positive Wahrnehmung entspricht einer Differenz von 8%.

Abschließend lassen die Ergebnisse für den ersten Umfrageteil trotz der Ähnlichkeiten zu den Häufigkeitsverteilungen in Herrmanns Arbeit [15] darauf schließen, dass es keine Unterschiede in den Wahrnehmungen zwischen der ersten und zweiten Teilnehmergruppe gibt.

5.1.4 Visualisierung der Wahrnehmung der Aussagen

Für die Visualisierung der Ergebnisse aus dem ersten Teil der durchgeführten Umfrage wurde ein normiertes Histogramm erstellt.

Die Häufigkeitsverteilung der für die zehn Aussagen X annotierten Sentiment-Polaritäten wird in der dazugehörigen Abbildung 5.1 dargestellt. Das Histogramm zeigt den Vergleich der Wahrnehmungen der ersten (blaue Farbe) sowie zweiten (orange Farbe) Teilnehmergruppe nach ihrer Sentiment-Polarität negativ, neutral und positiv auf. Die Balken im Histogramm zeigen, wie häufig die jeweilige Sentiment-Polarität in Prozent innerhalb der eigenen Teilnehmergruppe vorkommt. Aus dem Histogramm lässt sich ableiten, dass in der ersten Gruppe verhältnismäßig weniger negative als positive oder neutrale Wahrnehmungen annotiert wurden, während die Wahrnehmungen in der zweiten Gruppe eher ausgeglichen vertreten sind. Des Weiteren kann unter Berücksichtigung der Ergebnisse aus der Tabelle 5.4 in Kapitel 5.1.3 festgestellt werden, dass sich die neutralen Wahrnehmungen beider Teilnehmergruppen etwas stärker unterscheiden als die emotionalen

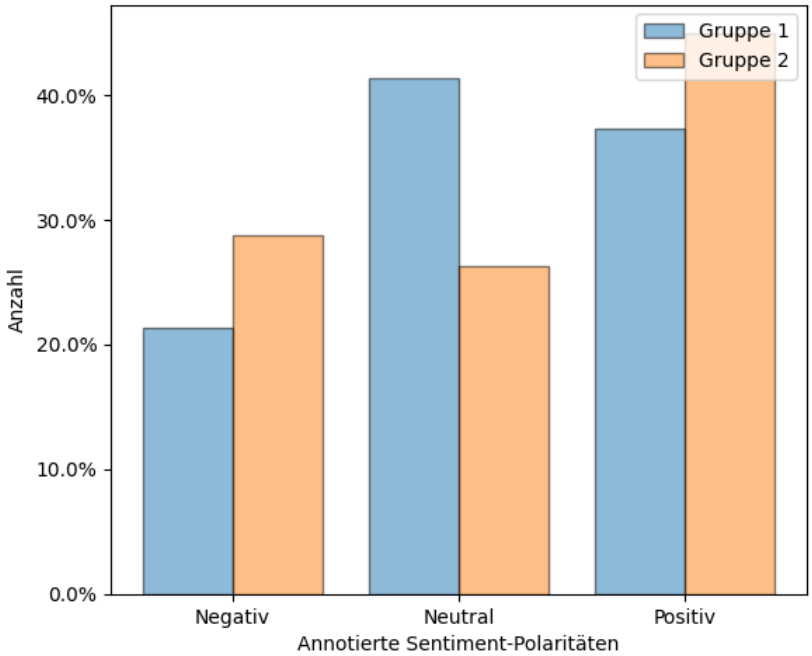

Abbildung 5.1: Anzahl der annotierten Sentiment-Polaritäten zwischen Gruppe 1 und Gruppe 2 im ersten Teil der Umfrage

Wahrnehmungen. Insgesamt beträgt die Differenz der Anteile der neutralen Annotationen wie bereits beschrieben 10%.

5.2 Evaluation der Zustimmung zu den gegebenen Wahrnehmungen

Dieses Kapitel stellt den Hauptteil dieser Arbeit dar und präsentiert die Ergebnisse des zweiten Teils der durchgeführten Studie. Die vorgestellten Ergebnisse werden im Hinblick auf die Zustimmung der Studienteilnehmer zu den gegebenen Wahrnehmungen evaluiert und daraufhin in Verbindung mit der Kalibrierung von Stimmungsanalysetools gebracht.

5.2.1 Ergebnisse des Shapiro-Wilk-Tests für die Zustimmung

Dazu soll der Mann-Whitney-U-Test verwendet werden. Um diesen anwenden zu können, müssen die zu überprüfenden Daten ordinalskaliert sein [21]. Da die Annotationsmöglichkeiten die Stärke der Zustimmung darstellen und keine klaren Abstände aufweisen, sind diese Daten ordinalskaliert. Auch hier muss der Shapiro-Wilk-Test angewandt werden, um die weitere Bedingung zu erfüllen, dass die Daten nicht normalverteilt sind. Fällt der Shapiro-Wilk-Test signifikant aus, muss also der Mann-Whitney-U-Test angewandt

werden. Andernfalls muss erneut überprüft werden, welcher von beiden statistischen Testverfahren (Zwei-Stichproben-t-Tests oder Welch-Test) für die Normalverteilung in Frage kommt [36].

Auch in dem zweiten Teil der zu dieser Arbeit gehörigen Umfrage wurde der Shapiro-Wilk-Test auf die Ergebnisse angewandt [36]. Mit diesem Test wird zwischen der Verwendung der in Kapitel 2 beschriebenen Statistikverfahren im Hinblick auf die Vorbedingung der Normalverteilung der Daten unterschieden. Die zu überprüfenden Merkmalswerte sind in diesem Umfrageteil die jeweilige Stärke der Zustimmung zu den gegebenen Sentiment-Polaritäten negativ, neutral und positiv der zehn Aussagen Y aus $V = \{V_1, V_2, ..., V_{96}\}$ ohne die Prädiktoraussagen $X = \{V_1, V_2, V_9, V_{21}, V_{41}, V_{49}, V_{54}, V_{57}, V_{78}, V_{80}\}$. Relevant ist hierbei der Vergleich der Zustimmungen zu den invertierten sowie nicht invertierten Sentiment-Polaritäten, da ein nicht signifikantes Ergebnis dafür spricht, dass den gegebenen Sentiment-Polaritäten unabhängig von der Invertierung genauso häufig zugestimmt wird. Tabelle 5.5 stellt die Ergebnisse aus dem Shapiro-Wilk-Test für die zehn Aussagen $Y \subset V \setminus X$ dar, wobei je ein Test für beide Teilnehmergruppen ($G = 1 + 2$) und ein weiterer für jede einzelne Teilnehmergruppe ($G = 1$ und $G = 2$) durchgeführt wurde.

Betrachtet man die p-Werte für die gesamten Zustimmungen, unabhängig von der jeweiligen Teilnehmergruppe, so erkennt man, dass durch eine Unterschreitung des Signifikanzniveaus $\alpha = 0.05$ für die invertierten sowie nicht invertierten Wahrnehmungen keine Normalverteilung der Daten vorliegt. Aufgrunddessen muss der Mann-Whitney-U-Test angewandt werden.

Unterteilt man die Ergebnisse beider Teilnehmergruppen in die erste und zweite Teilnehmergruppe, so erhält man ein ähnliches Ergebnis. Für die erste Teilnehmergruppe spricht der p-Wert mit < 0.0001 unabhängig von der Invertierung deutlich für eine Signifikanz. In der zweiten Teilnehmergruppe ist der p-Wert für die Zustimmungen zu den invertierten Wahrnehmungen nicht deutlich unter dem gewählten Signifikanzniveau $\alpha = 0.05$, jedoch sprechen beide p-Werte ebenfalls für eine Signifikanz und damit keine Normalverteilung der Daten. Zusammenfassend muss also für alle Zustimmungen der Mann-Whitney-U-Test verwendet werden [36].

5.2.2 Hypothesenprüfung von $\mathbf{H3}_0$ und $\mathbf{H3}(G)_0$

Für die Untersuchung der Signifikanz der Unterschiede zwischen den Zustimmungen zu den invertierten und nicht invertierten Sentiment-Polaritäten in dieser Arbeit wurden die übergeordnete Nullhypothese $\mathrm{H3}_0$ und die zwei untergeordneten Nullhypothesen $\mathrm{H3}(G)_0$ aufgestellt. Während die übergeordnete Nullhypothese $\mathrm{H3}_0$ keinen signifikanten Unterschied für die Zustimmungen zwischen den invertierten und nicht invertierten Sentiment-Polaritäten für alle Studienteilnehmer angibt, so untersucht man für die untergeordneten Nullhypothesen $\mathrm{H3}(1)_0$ und $\mathrm{H3}(2)_0$ die Aussage, dass es

Merkmal	**G**	**I**	**W**	**p**	**Statistischer Test**
Anteile der Zustimmungen	1+2	Inv	0.7456	< **0.0001**	Mann-Whitney-U-Test
		NInv	0.6645	< **0.0001**	
Anteile der Zustimmungen	1	Inv	0.6657	< **0.0001**	Mann-Whitney-U-Test
		NInv	0.6239	< **0.0001**	
Anteile der Zustimmungen	2	Inv	0.8190	**0.0049**	Mann-Whitney-U-Test
		NInv	0.7170	< **0.0001**	

Tabelle 5.5: Ergebnisse des Shapiro-Wilk-Tests für die Verteilung der Zustimmungen für beide Teilnehmergruppen, Gruppe 1 und Gruppe 2 ($G = 1 + 2$, $G = 1$ und $G = 2$) für jeweils invertierte und nicht invertierte Sentiment-Polaritäten (I=Inv und I=NInv) im zweiten Umfrageteil

keinen signifikanten Unterschied der Zustimmungen für die jeweils erste oder zweite Teilnehmergruppe gibt. Wie in Kapitel 5.2.1 bereits erläutert wurde, wird für die Überprüfung der Mann-Whitney-U-Test auf die Annotationen der Zustimmungen zu den zehn Aussagen $Y \subset V \setminus X$ nach Invertierung aus dem zweiten Umfrageteil angewandt. In der nachstehenden Tabelle 5.6 werden die Ergebnisse des Mann-Whitney-U-Tests aufgeführt.

G	**I**	**ø**	**σ**	**U**	**p**	**Interpretation**
1+2	Inv	0.5909	0.6848	1239	< **0.0001**	**Signifikant**
	NInv	1.5966	0.5557			
1	Inv	0.3929	0.5567	271.5	< **0.0001**	**Signifikant**
	NInv	1.6607	0.527			
2	Inv	0.9375	0.7474	311	**0.0076**	**Signifikant**
	NInv	1.4844	0.5861			

Tabelle 5.6: Ergebnisse des Mann-Whitney-U-Tests für die Zustimmung zu den invertierten und nicht invertierten Sentiment-Polaritäten (I=Inv und I=NInv) im zweiten Teil der Umfrage jeweils für beide Teilnehmergruppen, Gruppe 1 und Gruppe 2 ($G = 1 + 2$, $G = 1$ und $G = 2$)

Das Signifikanzniveau wurde in der Untersuchung der Nullhypothesen $H3_0$ und $H3(G)_0$ auf $\alpha = 0.05/2 = 0.025$ [2] gesetzt. Für beide Teilnehmer-

gruppen zeigt die Tabelle zwischen den invertierten und nicht invertierten Wahrnehmungen einen Unterschied für die Durchschnitte der Stärke der Zustimmung. Eine 2 entspricht einer Zustimmung zu der gegebenen Sentiment-Polarität, eine 1 einer teilweisen Zustimmung und eine 0 wiederum einer Ablehnung, sofern eine emotionale Wahrnehmung vorgegeben ist (negativ oder positiv). Die Zustimmung zu invertierten Wahrnehmungen liegt für beide Teilnehmergruppen der Durchschnitt zwischen einer teilweisen Zustimmung und einer Ablehnung. Für nicht invertierte Wahrnehmungen liegt der Durchschnitt jedoch zwischen einer vollständigen und einer teilweisen Zustimmung. Die Differenz zwischen beiden Invertierungen (I=Inv und I=NInv) liegt bei einem Wert von ungefähr 1. Auch der Mann-Whitney-U-Test bestätigt die Unterschiede in den durchschnittlichen Stärken der Zustimmungen. Der dazugehörige p-Wert bei < 0.0001 und unterschreitet damit das Signifikanzniveau α.

Überprüft man die Verteilungen der ersten und zweiten Teilnehmergruppe, sieht man in der Tabelle 5.6, dass sich die Zustimmungsstärken auch dort zwischen den invertierten und nicht invertierten Sentiment-Polaritäten unterscheiden. Für die erste Teilnehmergruppe verhält sich der Durchschnitt für die Invertierung und ohne Invertierung ähnlich wie für beide Teilnehmergruppe zusammen. Die Differenz der beiden Gruppen ist etwas größer als die Differenz für beide Teilnehmergruppen, beträgt jedoch ebenfalls ungefähr 1. Auch hier kann mit Hilfe des Mann-Whitney-U-Tests durch die Unterschreitung des Signifikanzniveaus α ein signifikanter Unterschied in der Zustimmung zu den gegebenen Polaritäten festgestellt werden. Bei der Betrachtung der Durchschnitte der invertierten sowie nicht invertierten Sentiment-Polaritäten für die zweite Teilnehmergruppe fällt auf, dass häufiger eine teilweise Zustimmung als eine Ablehnung angegeben wurde für die invertierten Sentiment-Polaritäten, was sich von den anderen Ergebnissen unterscheidet. Auch der p-Wert fällt hier höher aus: Mit einem Wert von 0.0076 nähert sich der Wert dem Signifikanzniveau α, unterschreitet diesen jedoch und gibt damit einen signifikanten Unterschied in den Zustimmungen an.

Damit kann also abschließend die übergeordnete Nullhypothese $H3_0$ sowie die untergeordneten Nullhypothesen $H3(G)_0$ abgelehnt werden. Daraus folgt, dass die Wahrnehmungen der Aussagen durch die Studienteilnehmer mit den auf Basis des ersten Umfrageteils bestimmten Sentiment-Polaritäten übereinstimmen.

5.2.3 Hypothesenprüfung von $H4_0$

Um zu überprüfen, wie sich die Häufigkeit der Zustimmung für beide Teilnehmergruppen im zweiten Teil der Umfrage verhält, wurde die Nullhypothese $H4_0$ aufgestellt. Mit Hilfe dieser soll überprüft werden, ob eine Teilnehmergruppe häufiger den gegebenen Sentiment-Polaritäten zustimmt

als die anderen Teilnehmergruppe, unabhängig davon, ob die Sentiment-Polaritäten invertiert wurden oder nicht. Die Ergebnisse des auf die zehn Aussagen $Y \subset V \setminus X$ angewandten Mann-Whitney-U-Tests werden in Tabelle 5.7 dargestellt.

Anteil in %	**G**	**ø**	**σ**	**U**	**p**	**Interpretation**
Zustimmungen	1	55.71	14.98	70.5	0.3296	Nicht signifikant
	2	47.50	17.85			

Tabelle 5.7: Ergebnisse des Mann-Whitney-U-Tests für die Zustimmung zu den Sentiment-Polaritäten im zweiten Teil der Umfrage für beide Teilnehmergruppen ($G = 1$ und $G = 2$), unabhängig von der Invertierung

Bereits die Anteile der durch die Teilnehmergruppen angegebenen vollständigen Zustimmungen zeigen keine starken Unterschiede. Die Zustimmung der ersten Teilnehmergruppe liegt bei einem Wert von ungefähr 55%. Die zweite Teilnehmergruppe stimmte etwas seltener vollständig der gegebenen Sentiment-Polarität zu. Der Wert hierbei ist etwas niedriger mit ungefähr 47%. Der Mann-Whitney-U-Test verstärkt die Vermutung der ähnlichen Zustimmungshäufigkeit, da der p-Wert unter dem nicht veränderten Signigikanzniveau von $\alpha = 0.05$ liegt. Daher kann die Nullhypothese $H4_0$ abgelehnt werden und geschlussfolgert werden, dass sich die Anteile für die Zustimmung für beide Teilnehmergruppen ähneln.

5.2.4 Visualisierung der Zustimmung zu den gegebenen Wahrnehmungen

Um die Signifikanzen in den Unterschieden der Wahrnehmungen der Studienteilnehmer im zweiten Teil der durchgeführten Umfrage zu visualisieren, wurden die zwei Histogramme 5.2 und 5.3 erstellt. Während Abbildung 5.2 zeigt, wie sehr die Wahrnehmungen der Studienteilnehmer mit den für die zehn Aussagen Y bestimmten Sentiment-Polaritäten übereinstimmen, unterteilt Abbildung 5.3 die Häufigkeitsverteilungen der Studienteilnehmer in die zweite Teilnehmergruppe. Da sich die Häufigkeitsverteilung für beide Teilnehmergruppen mit denen der ersten Gruppe ähnelt, wurde keine Abbildung zu dem enstsprechenden Histogramm beigefügt.

In beiden Abbildungen werden die Zustimmungen zu den invertierten und die Zustimmungen zu den nicht invertierten Sentiment-Polaritäten miteiander verglichen, wobei die blauen Balken die Zustimmungsstärken zu den invertierten und die orangen Balken die Zustimmungsstärken zu den nicht invertierten Sentiment-Polaritäten darstellen. Durch die Visualisierung werden also die Signifikanzen deutlich, welche bei der Überprüfung der Nullhypothese $H3_0$ in Kapitel 5.2.2 über die Zustimmungen der Studienteilnehmer entstanden sind. Im Hinblick auf die Häufigkeitsverteilung wird

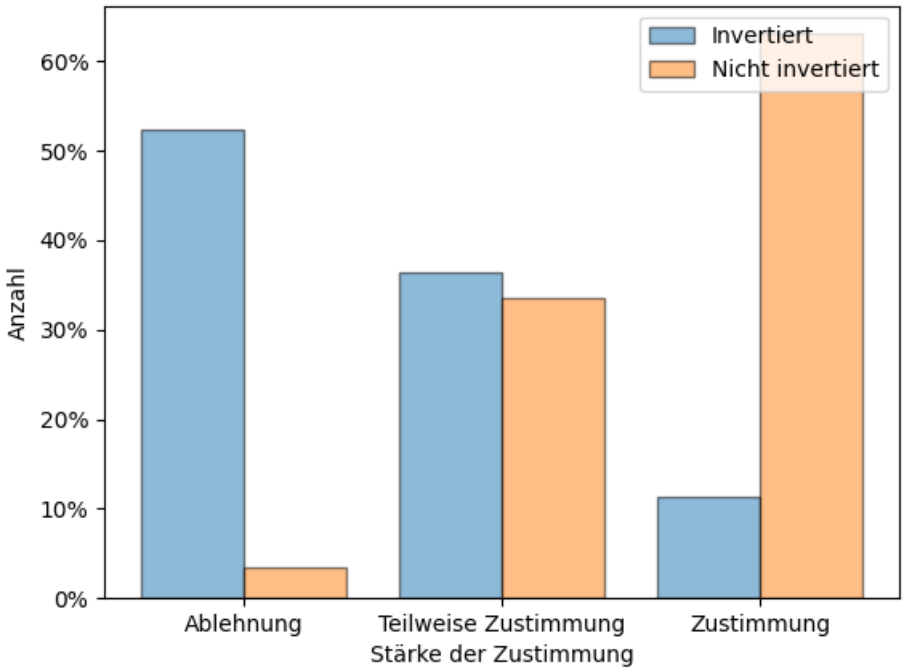

Abbildung 5.2: Verteilung der Zustimmungsstärken für beide Teilnehmergruppen zwischen invertierten und nicht invertierten Polaritäten

in Abbildung 5.2 erkennbar, dass den invertierten Sentiment-Polaritäten häufiger abgelehnt wurde mit einem Anteil von ungefähr 52%, unabhängig von der Gruppenzugehörigkeit. Eine teilweise Zustimmung ist ebenfalls mit einer geringeren Häufigkeit vorhanden. Eine Ablehnung zu der gegebenen Sentiment-Polarität bildet die Ausnahme mit einem Anteil von 11.36%. Vergleicht man die Zustimmungen für die invertierten Sentiment-Polaritäten mit denen der nicht invertierten, so fällt auf, dass sich die Häufigkeitverteilung sich im Verlauf spiegelt. Hier wurden vermehrt vollständige Zustimmungen zu den gegebenen Wahrnehmungen angegeben und der Anteil beträgt sogar ca. 63%. Die teilweise Zustimmung und Ablehnung liegen bei niedrigeren Werten von ca. 33.5 und 3.4%. Man bedenke, dass durch die Invertierung von genau zwei von zehn Sentiment-Polaritäten für jeden Studienteilnehmer die Gesamtanzahl von angegebenen Zustimmungen geringer ausfällt. Infolgedessen ist es interessant, dass der Anteil für der vollständigen Zustimmung zu den nicht invertierten Sentiment-Polaritäten höher ist als der für die invertierten Sentiment-Polaritäten, da eine größere Menge an Daten aussagekräftiger ist. Für die erste Teilnehmergruppe wurden die Häufigkeitsverteilungen ebenfalls berechnet und stellen eine ähnliche Verteilung dar.

Betrachtet man die Häufigkeitsverteilung der jeweiligen Zustimmungsstärken für die zweite Teilnehmergruppe im Histogramm 5.3, fällt zu Beginn

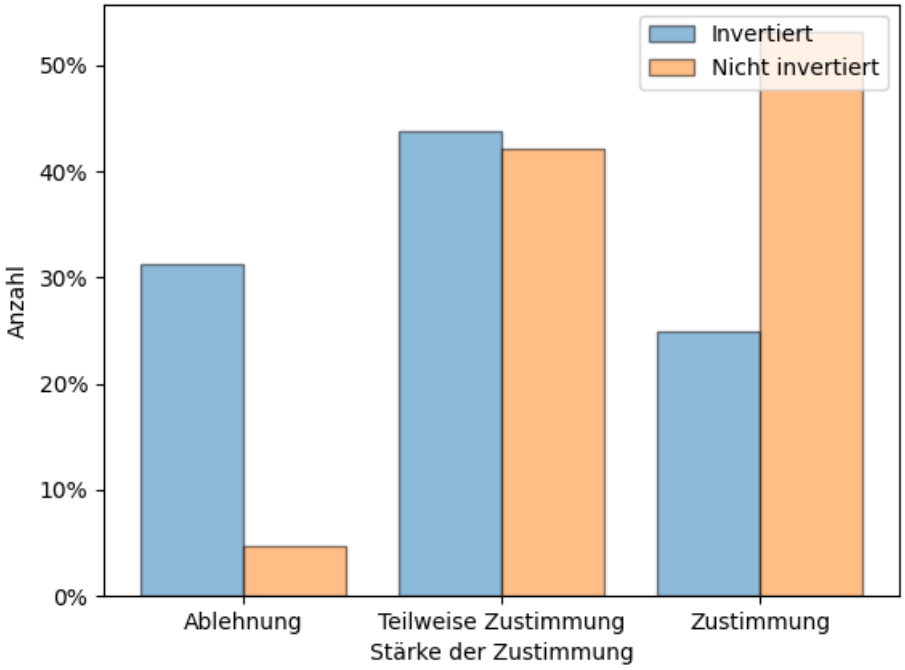

Abbildung 5.3: Verteilung der Zustimmungsstärken für Gruppe 2 zwischen invertierten und nicht invertierten Polaritäten

kein Unterschied zu den Anteilen der Zustimmungen für alle Studienteilnehmer in 5.2 auf. Die Zustimmung zu den nicht invertierten Sentiment-Polaritäten verhält sich ähnlich mit einem Anteil von ca. 53 und 42% für die vollständige und teilweise Zustimmung, wobei diese Anteile die Mehrheit darstellen und den Anteil der Ablehnung mit ca. 4.69% übertreffen. Für die invertierten Sentiment-Polaritäten kann hingegen festgestellt werden, dass die meisten Teilnehmer aus der zweiten Gruppe weder vollständig abgelehnt noch zugestimmt haben. Der Anteil der teilweisen Zustimmung beträgt hierbei 43.75% und den zweithöchste Anteil bildet die Ablehnung zu den gegebenen Sentiment-Polaritäten mit 31.25%. Jedoch bestehen auch hier signifikante Unterschiede in den Zustimmungen zwischen den invertierten sowie nicht invertierten Sentiment-Polaritäten, wie im dazugehörigen Mann-Whitney-U-Test in Kapitel 5.2.2 gezeigt wurde.

5.2.5 Visualisierung der Wahrnehmung bei Zustimmung

Um einen weiteren Einblick in die Zustimmung zu den Sentiment-Polaritäten aus dem zweiten Umfrageteil zu erhalten, wurde die nachstehende Abbildung 5.4 erstellt.

Die Histogramme visualisieren die Häufigkeitsverteilungen der Sentiment-Polaritäten bei Zustimmung zu den durch das in 4.1.3 erläuterte

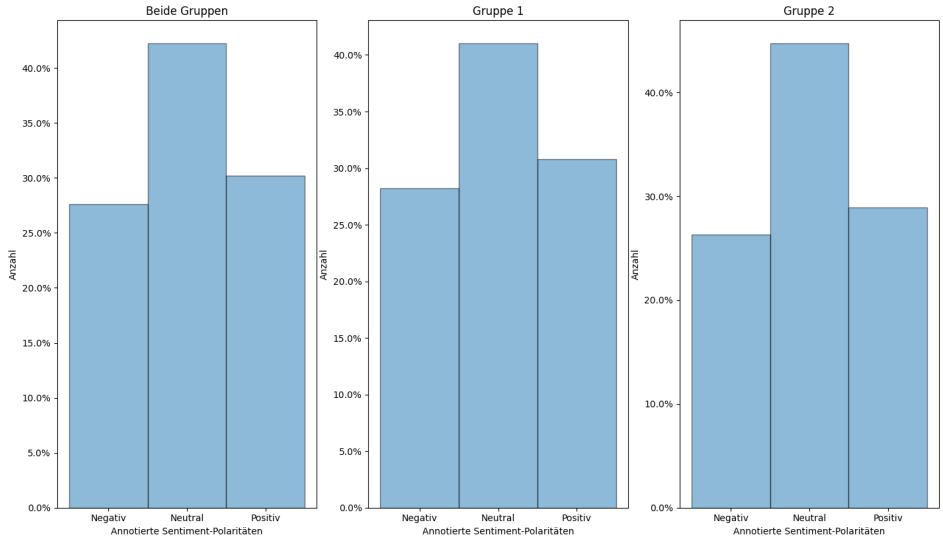

Abbildung 5.4: Anteile der jeweiligen Sentiment-Polaritäten bei Zustimmung für beide Teilnehmergruppen, Gruppe 1 und Gruppe 2

Skript bestimmten Sentiment-Polaritäten. Wie man in 5.4 sehen kann, geben die Studienteilnehmer beider Teilnehmergruppen eher eine Zustimmung an, wenn die gegebene Aussage mit einer neutralen Wahrnehmung bewertet wurde. Beispielweise liegt der Anteil neutraler Wahrnehmungen bei ungefähr 42% für alle Studienteilnehmer. Die Anteile für die emotionalen Wahrnehmungen haben für alle Teilnehmer eine ähnliche Verteilung, wobei die positiven Wahrnehmungen mit 2% überwiegen.

Kapitel 6

Diskussion

In diesem Kapitel werden die in Kapitel 5 vorgestellten Ergebnisse bewertet und in Verbindung mit dem Einsatz der Stimmungsanalyse in individuellen Softwareunternehmen gesetzt. Ebenfalls wird auf die Schwierigkeiten zur Studiendurchführung und Glaubwürdigkeit der Ergebnisse eingegangen.

6.1 Interpretation der Ergebnisse

Bisher wird die Stimmungsanalyse nicht in in individuellen Softwareunternehmen genutzt [8], da die bestehenden Stimmungsanalysetools nicht auf den Gebrauch abgestimmt sind. Neben der Missinterpretation von fachspezifischen Begriffen stellt sich die Subjektivität der Softwareentwickler nach Softwareunternehmen als Problematik heraus [8]. Erkenntnisse aus der Arbeit von Herrmann [15] besagen, dass verschiedene Softwareentwickler unterschiedliche Wahrnehmungen im Hinblick auf die Interpretation von Aussagen besitzen können. Die Zielsetzung dieser Arbeit bestand darin zu überprüfen, ob eine Kalibrierung von Stimmungsanalysetools durch die Softwareentwickler sinnvoll ist. Im Hinblick auf die Kalibrierung für den Einsatz von Stimmungsanalysetools in Projekten von Software-Unternehmen haben die Ergebnisse dieser Arbeit eine große Bedeutung. Diese zeigen einen signifikanten Unterschied in der Zustimmung zu den Wahrnehmungen von Aussagen, die auf Basis der im Voraus annotierten Aussagen bestimmt wurden. Unabhängig von der Gruppenzugehörigkeit wurde den individuell vorgestellten Aussagen zugestimmt. Dem Anteil der Aussagen, der mit einer invertierten Sentiment-Polarität präsentiert wurde, kamen die Studienteilnehmer hauptsächlich mit einer Ablehnung entgegen. Im Hinblick auf die Ergebnisse im ersten Teil der Umfrage wurde festgestellt, dass kein signifikanter Unterschied zwischen der allgemeinen Verteilung der Wahrnehmungen der beiden Teilnehmergruppen besteht, jedoch eine Signifikanz bei neutralen Aussagen. Mögliche Erklärungen für die stellen die geringe Anzahl an Teilnehmern oder dieselbe Wahrnehmung im Durchschnitt

der Wahrnehmungsverteilungen dar (vgl. 5.1.2). Aufgrunddessen kann durch die Übereinstimmung der Wahrnehmungen der Studienteilnehmer mit den gegebenen Wahrnehmungen durchaus gesagt werden, dass eine Kalibrierung von Stimmungsanalysetools sinnvoll ist. Durch die Identifikation von fünf relevanten Aussagen aus dem gesamten Datensatz durch Herrmann [15] für die korrekte Vorhersage der Wahrnehmungen auf Basis der jeweiligen Gruppenzugehörigkeit muss ein kleiner Aufwand betrieben werden. Die manuelle Annotation der spezifischen Aussagen durch die Entwickler muss im Voraus erfolgen, damit der Einsatz in Projekten möglich ist.

6.2 Art der Studiendurchführung

Die durchgeführte Studie brachte einige Schwierigkeiten mit sich. Zum Einen stellt die lokale Durchführung einen großen Aufwand dar. In einer Studie mit einer größeren Anzahl von Teilnehmern kann diese Art der Durchführung nicht verwendet werden, da sie durch die Anwesenheit eines Moderators einen erheblichen Zeitaufwand darstellt. Außerdem besteht die Notwendigkeit für den Studienteilnehmer, bei der Durchführung der Umfrage mittels Remote-Desktop teilnehmen zu müssen. Um diese Einschränkungen zu beheben, muss in einer größeren Studie auf die Verwendung einer Webseite zurückgegriffen werden, die das durchgeführte Skript im Zwischenteil der Umfrage automatisch durchführt. Zum Anderen bringt die Moderatorrolle neben der Möglichkeit von Rückfragen und Hilfestellungen gewisse Nachteile mit sich. Für die Anwesenheit des Moderators ist es essentiell, dass die Einführung in die Umfrage für jeden Studienteilnehmer dieselbe ist. Ebenfalls darf der Moderator bei Rückfragen keine subjektiven Erklärungen abgeben. Dies hat den Grund, dass der Studienteilnehmer keineswegs in der Bewertung der Aussagen beeinflusst werden darf, da sonst nicht die eigene Wahrnehmung untersucht wird.

6.3 Aussagekraft der Ergebnisse

Wie bereits in Kapitel 6.1 erwähnt, stellt die geringe Anzahl von insgesamt 22 Studienteilnehmern eine Einschränkung im Hinblick auf die Aussagekraft der Ergebnisse dar. Die Ergebnisse lassen sich nicht auf alle Software-Entwickler individueller Unternehmen verallgemeinern. Dieses Phänomen wird auch als *Convenient Sampling* bezeichnet. Eine weitere Schwierigkeit fällt ebenfalls unter diesen Begriff und bildet die Variabilität der Studienteilnehmer. Es wurden hauptsächlich Informatikstudenten befragt und kleiner Teil machten Software-Entwickler aus. Um eine bessere Aussagekraft der Ergebnisse zu erreichen, müssen Software-Entwickler aus individuellen Unternehmen betrachtet werden. Zuletzt führt der fehlende Kontext zu den Aussagen zu unterschiedlichen Wahrnehmungen durch die Studienteilnehmer.

Kapitel 7

Zusammenfassung und Ausblick

Dieses Kapitel fasst die Vorangehensweise in dieser Arbeit zusammen. Neben der kurzen Beschreibung der zugehörigen Studie wird besonders auf die dadurch entstandenen Ergebnisse eingegangen. Für den Abschluss dieser Arbeit werden die Erkenntnisse dieser Arbeit auf die Anwendung der Stimmungsanalyse Softwareprojekten von individuellen Unternehmen bezogen.

7.1 Zusammenfassung

In Anbetracht der Ergebnisse aus einer ähnlichen Arbeit von Herrmann [15] im Bereich der Stimmungsanalyse wurde die Notwendigkeit einer Kalibrierung von Stimmungsanalysetools untersucht. Im Zuge dessen wurde eine Umfrage durchgeführt, in welcher die 22 Studienteilnehmer Aussagen nach ihrer wahrgenommenen Sentiment-Polaritäten negativ, neutral oder positiv annotiert haben. Auf Basis der jeweiligen Wahrnehmung wurden die Studienteilnehmer daraufhin durch die logistische Regressionsanalyse einer von zwei Teilnehmergruppen zugeordnet. Die zu annotierenden Aussagen enthielten den Anteil des Datensatzes von 96 Aussagen, der relevant für die korrekte Zuordnung zu der jeweiligen Teilnehmergruppe war, wie in Herrmanns Arbeit [15] festgestellt wurde. Diese stellten Aussagen aus den fachspezifischen Quellen wie *Github* und *Stack Overflow* dar. Im weiteren Verlauf der Umfrage wurden jedem Studienteilnehmer neue Aussagen präsentiert, welche je nach Gruppenzugehörigkeit bereits mit anderen Sentiment-Polaritäten annotiert wurden. Um die Kalibrierung im ersten Teil der Umfrage zu überprüfen, gaben die Studienteilnehmer nun an, inwieweit sie der gegebenen Wahrnehmung zustimmen.

Die Ergebnisse zeigen, dass die Wahrnehmungen der Studienteilnehmer häufiger mit den gegebenen Wahrnehmungen übereinstimmen, da

überwiegend vollständige Zustimmungen als eine Ablehnungen angegeben wurden. Unabhängig von der Gruppenzugehörigkeit stellte die Häufigkeit der Zustimmungen einen Anteil von ca. 63% dar und überwog damit die Ablehnung um ganze Als besonders stellte sich ebenfalls die Überprüfung der Unterschiede der Zustimmungen zu invertierten emotionalen Wahrnehmungen heraus und der Vergleich mit den Zustimmungen zu nicht invertierten Wahrnehmungen. Wurde beispielsweise eine negative Sentiment-Polarität angegeben, so wurde eine positive als invertierte Sentiment-Polarität angezeigt. Eine statistische Untersuchung ergab einen signifikanten Unterschied zwischen den Zustimmungen zu den invertierten und nicht invertierten Wahrnehmungen. Aus diesem Grund erscheint eine Kalibrierung von Stimmungsanalysetools als sinnvoll, um korrekte Vorhersagen von Wahrnehmungen in dem Bereich der Informatikbranche zu treffen.

7.2 Ausblick

Im Sinne der Etablierung der Stimmungsanalyse in Softwareprojekten individueller Softwareunternehmen muss bedacht werden, dass ein Training von den zu verwendenden Stimmungsanalysetools erfolgen muss. Dies liegt einerseits daran, dass bestehende Tools nicht auf die Analyse der technischen Begriffe spezialisiert sind [8], und andererseits, dass die Subjektivität von verschiedenen Software-Entwicklern nicht außer Acht gelassen werden darf, wie verwandte Arbeiten zeigen [15]. Um die Korrektheit der Vorhersage von Wahrnehmungen durch Stimmungsanalysetools sicherzustellen, bietet die Kalibrierung mit Hilfe der manuellen Annotierung von Aussagen durch die Software-Entwickler des jeweiligen Unternehmens eine mögliche Ausführungsform. Durch die Notwendigkeit einer geringen Anzahl von Aussagen für die Annotierung vor der Verwendung des Stimmungsanalysetools [15] stellt die Sicherstellung dieser Bedingung keinen allzu hohen Aufwand dar. Soll dieser Betriebs- sowie Zeitaufwand jedoch minimiert werden, so müsste in zukünftigen Arbeiten eine Untersuchung der für die verschiedenen Wahrnehmungen spezifischen Separationsmerkmale stattfinden. Solche Merkmale könnten beispielsweise durch die nähere Betrachtung der einzelnen Tätigkeitsbereiche in Softwareunternehmen eines jeden Studienteilnehmers festgestellt werden. Da voraussichtlich mehrere Merkmale im Zusammenspiel für die verschiedenen Wahrnehmungen von Entwicklern sorgen, stellt sich die Identifikation der Merkmale als eine besondere Herausforderung in der Zukunft der Verwendung der Stimmungsanalyse in Softwareunternehmen heraus. In Anbetracht dessen sollen die Ergebnisse dieser Arbeit als Basis für die Weiterverwendung in der Entwicklung vonr Stimmungsanalysetools in der Softwareentwicklung dienen, da gezeigt wurde, dass eine Kalibrierung in der von Stimmungsanalyse durchaus sinnvoll für die korrekte Sicherstellung der Korrektheit von Vorhersagen ist.

Literaturverzeichnis

[1] Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series a - Mathematical and Physical Sciences*, 160:268–282, 1937.

[2] *Bonferroni Correction*, pages 227–227. Springer Netherlands, Dordrecht, 2008.

[3] *t-Test*, pages 2043–2043. Springer Netherlands, Dordrecht, 2008.

[4] *Histogram*, pages 681–681. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.

[5] *Ordinal Scale*, pages 978–979. Springer International Publishing, Cham, 2021.

[6] Q. U. Ain, T. Rana, and Aamana. A study on identifying, categorizing and reporting usability bugs and challenges. In *2023 International Conference on Communication Technologies (ComTech)*, pages 53–68, 2023.

[7] J. Behnke. *Das Logit-Modell*, pages 23–35. Springer Fachmedien Wiesbaden, Wiesbaden, 2015.

[8] F. Calefato, F. Lanubile, F. Maiorano, and N. Novielli. Sentiment polarity detection for software development. In *Proceedings of the 40th International Conference on Software Engineering*, ICSE '18, page 128, New York, NY, USA, 2018. Association for Computing Machinery.

[9] G. Destefanis, M. Ortu, D. Bowes, M. Marchesi, and R. Tonelli. On measuring affects of github issues' commenters. In *Proceedings of the 3rd International Workshop on Emotion Awareness in Software Engineering*, SEmotion '18, page 14–19, New York, NY, USA, 2018. Association for Computing Machinery.

[10] R. A. Fisher. *Statistical Methods for Research Workers*, pages 66–70. Springer New York, New York, NY, 1992.

[11] S. R. Goniwada. *Sentiment Analysis*, pages 165–184. Apress, Berkeley, CA, 2023.

[12] D. Graziotin, X. Wang, and P. Abrahamsson. Happy software developers solve problems better: psychological measurements in empirical software engineering. 2014. San Diego, CA, USA.

[13] D. Graziotin, X. Wang, and P. Abrahamsson. How do you feel, developer? an explanatory theory of the impact of affects on programming performance. 2015. San Diego, CA, USA.

[14] E. Guzman, D. Azócar, and Y. Li. Sentiment analysis of commit comments in github: An empirical study. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, MSR 2014, page 352–355, New York, NY, USA, 2014. Association for Computing Machinery.

[15] M. Herrmann. Analyzing the Perception of Sentiments in Software Projects Using Exploratory Data Analysis. Master's thesis, 2023.

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

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

[18] S. F. Huq, A. Z. Sadiq, and K. Sakib. Is developer sentiment related to software bugs: An exploratory study on github commits. In *2020 IEEE 27th International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 527–531, 2020.

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

[20] M.-A. Issa and K. L. Nadal. *Homoscedasticity*, pages 752–752. Springer US, Boston, MA, 2011.

[21] W. Kirch, editor. *Mann Whitney (U) Test*, pages 884–884. Springer Netherlands, Dordrecht, 2008.

[22] R. E. Kraut and L. A. Streeter. Coordination in software development. *Commun. ACM*, 38(3):69–81, mar 1995. New York, NY, USA.

[23] F. Lanubile. *Collaboration in Distributed Software Development*, pages 174–193. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[24] B. Liu. *Opinion Mining*, pages 1986–1990. Springer US, Boston, MA, 2009.

[25] T. W. MacFarland and J. M. Yates. *Mann–Whitney U Test*, pages 103–132. Springer International Publishing, Cham, 2016.

[26] H. B. Mann and D. R. Whitney. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50 – 60, 1947.

[27] I. R. McChesney and S. Gallagher. Communication and co-ordination practices in software engineering projects. *Information and Software Technology*, 46(7):473–489, 2004.

[28] G. Nahler. *ordinal scale*, pages 127–127. Springer Vienna, Vienna, 2009.

[29] N. Novielli, F. Calefato, and F. Lanubile. Towards discovering the role of emotions in stack overflow. In *Proceedings of the 6th International Workshop on Social Software Engineering*, SSE 2014, page 33–36, New York, NY, USA, 2014. Association for Computing Machinery.

[30] M. Obaidi, M. Herrmann, L. Chazette, and J. Klünder. Dataset: Sentisurvey for sentiment analysis in software projects. Zenodo, 2022.

[31] M. Ortu, G. Destefanis, B. Adams, A. Murgia, M. Marchesi, and R. Tonelli. The jira repository dataset: Understanding social aspects of software development. In *Proceedings of the 11th International Conference on Predictive Models and Data Analytics in Software Engineering*, PROMISE '15, New York, NY, USA, 2015. Association for Computing Machinery.

[32] G. Rebala, A. Ravi, and S. Churiwala. *Natural Language Processing*, pages 117–125. Springer International Publishing, Cham, 2019.

[33] C. Sammut and G. I. Webb, editors. *Logistic Regression*, pages 631–631. Springer US, Boston, MA, 2010.

[34] L. Schroth, M. Obaidi, A. Specht, and J. Klünder. On the potentials of realtime sentiment analysis on text-based communication in software projects. In R. Bernhaupt, C. Ardito, and S. Sauer, editors, *Human-Centered Software Engineering*, pages 90–109, Cham, 2022. Springer International Publishing.

[35] M. Schwab, editor. *Mann–Whitney U-test*, pages 1764–1764. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[36] S. S. SHAPIRO and M. B. WILK. An analysis of variance test for normality (complete samples)†. *Biometrika*, 52(3-4):591–611, 12 1965.

[37] Student. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908.

[38] J. F. Sánchez-Rada and C. A. Iglesias. Social context in sentiment analysis: Formal definition, overview of current trends and framework for comparison. *Information Fusion*, 52:344–356, 2019.

[39] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010.

[40] T. R. Tulili, A. Capiluppi, and A. Rastogi. Burnout in software engineering: A systematic mapping study. *Information and Software Technology*, 155:107116, 2023.

[41] J. R. Turner. *Regression Analysis*, pages 1633–1634. Springer New York, New York, NY, 2013.

[42] B. L. Welch. The generalization of 'Student's' problem when several different population variances are involved. *Biometrika*, 34(1-2):28–35, 01 1947.

[43] L. Zhang and B. Liu. *Sentiment Analysis and Opinion Mining*, pages 1152–1161. Springer US, Boston, MA, 2017.