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

Ermittlung von Erklärbarkeitsanforderungen zur Erhöhung der Nutzerakzeptanz eines Stimmungsanalysetools

Elicitation of Explainability Requirements to increase the User Acceptance of a Sentiment Analysis Tool

Masterarbeit

im Studiengang Informatik

von

Julian Voges

Prüfer: Prof. Dr. rer. nat. Kurt Schneider
Zweitprüfer: Dr. rer. nat. Jil Ann-Christin Klünder
Betreuer: Jakob Richard Christian Droste, M. Sc.

Hannover, 02.05.2024

Erklärung der Selbstständigkeit

Hiermit versichere ich, dass ich die vorliegende Masterarbeit 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 02.05.2024

Julian Voges

Zusammenfassung

Durch eine umfassende Etablierung von Stimmungsanalysetools könnten Schwierigkeiten und Probleme in der Kommunikation erkannt und bewältigt werden und somit zu einer reibungslosen und erfolgreichen Softwareentwicklung beitragen [46]. Obwohl in der Forschung die Vorteile von Stimmungsanalysetools deutlich werden, ist in der Praxis die Nutzung dieser nicht weit verbreitet [46]. Häufig basieren Stimmungsanalysetools auf komplexen und undurchsichtigen Algorithmen. Die fehlende Transparenz und Verständlichkeit kann zu fehlendem Vertrauen bei den Stakeholdern führen, die das Tool letztendlich verwenden sollen. Durch die Einführung von Erklärbarkeit kann jedoch die Nutzerakzeptanz und damit die Nutzung einer Software erhöht werden [12].

Nach kurzer Literatursichtung wurde deutlich, dass bereits viele Studien Erklärbarkeit untersuchen, wie z. B. die systematische Literaturanalyse von Chazette et al. [12]. Auch betrachten viele Studien Stimmungsanalysetools, wie z. B. die systematische Literaturanalyse von Obaidi und Klünder [46]. Jedoch wurde keine Studie gesichtet, welche die Einführung von Erklärbarkeit im Kontext von Stimmungsanalysetools untersucht.

Daher werden in dieser Arbeit Erklärbarkeitsanforderungen zur Erhöhung der Nutzerakzeptanz eines Stimmungsanalysetools ermittelt. Auf Grundlage der Literaturrecherche wird das Stimmungsanalysetool *RoBERTa* als das für diese Arbeit prädestinierteste Tool identifiziert. Um dieses Tool für die Anwendung im SE-Kontext nutzen zu können, wird *RoBERTa* mithilfe eines Datensatzes mit über 4000 Einträgen vortrainiert. Im Rahmen eines Workshops wird das Stimmungsanalysetool vorgestellt und Erklärbarkeitsanforderungen erhoben. Diese werden auf Realisierbarkeit überprüft und softwaretechnisch bzw. prototypisch im Stimmungsanalysetool *RoBERTa* umgesetzt. Danach wird mithilfe einer Online-Umfrage die Nutzerakzeptanz des Stimmungsanalysetools ohne und mit den jeweiligen Erklärungen erhoben. Bei der anschließenden Analyse der Umfrageergebnisse wird u. a. festgestellt, dass die Erklärungen *Beispiele* und *Schlüsselwörter* zu einer signifikanten mittelgroßen Erhöhung der Nutzerakzeptanz führen. Die Arbeit spricht Empfehlungen für die Industrie aus, wie z. B. die Umsetzung dieser Erklärungen unter Voraussetzung den vorliegenden Kontext auf Zielkonflikte zu untersuchen.

Abstract

Elicitation of Explainability Requirements to increase the User Acceptance of a Sentiment Analysis Tool

By comprehensively establishing sentiment analysis tools, difficulties and problems in communication could be recognized and overcome, thus contributing to smooth and successful software development [46]. Although the benefits of sentiment analysis tools are clear in research, their use is not widespread in practice [46]. Sentiment analysis tools are often based on complex and opaque algorithms. The lack of transparency and comprehensibility can lead to a lack of trust among the stakeholders who are ultimately supposed to use the tool. However, the implementation of explainability can increase user acceptance and thus the use of software [12].

After a brief review of the literature, it became clear that many studies are already investigating explainability, such as the systematic literature analysis by Chazette et al. [12]. Many studies also look at sentiment analysis tools, such as the systematic literature review by Obaidi and Klünder [46]. However, no study was reviewed that examined the implementation of explainability in the context of sentiment analysis tools.

Therefore, explainability requirements for increasing the user acceptance of a sentiment analysis tool are determined in this thesis. Based on the literature research, the sentiment analysis tool *RoBERTa* is identified as the most suitable tool for this work. In order to be able to use this tool in the SE context, *RoBERTa* is pre-trained using a data set with over 4000 entries. The sentiment analysis tool is presented in a workshop and explainability requirements are collected. These are checked for feasibility and implemented or prototypically implemented in the sentiment analysis tool *RoBERTa*. An online survey is then used to assess user acceptance of the sentiment analysis tool without and with the respective explanations. The subsequent analysis of the survey results shows, among other things, that the explanations *examples* and *keywords* lead to a significant medium increase in user acceptance. The work makes recommendations for the industry, such as the implementation of these explanations, provided that the present context is examined for conflicting objectives.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einleitung

1.1 Motivation

Durch die zunehmende Etablierung von Software im Alltag, wird auch die Entwicklung von Software immer komplexer. Um einen reibungslosen Ablauf in der Softwareentwicklung gewährleisten zu können, ist ein kollaboratives Arbeiten im Team besonders wichtig. Damit eine erfolgreiche und zufriedenstellende Softwareentwicklung durch eine reibungslose Kommunikation im Team gewährleistet wird, ist die Betrachtung der sozialen Aspekte ausschlaggebend [46].

Damit die Probleme und Schwierigkeiten in der Kommunikation bewältigt werden können, muss zuerst die Stimmung im Team identifizieren werden [46]. Hierfür eignen sich zum Beispiel Stimmungsanalysetools, die die Stimmung auf der Grundlage von textbasierter Kommunikation im Team erheben [46]. Mithilfe dieser Tools können zum Beispiel Teamleiter angemessene Maßnahmen durchführen, um die Stimmung im Team und somit auch die Softwareentwicklung zu verbessern.

1.2 Problemstellung

Obwohl in der Forschung die Vorteile bei der Nutzung von Stimmungsanalysetools deutlich werden [46], ist in der Praxis die Nutzung von Stimmungsanalysetools nicht weit verbreitet. Häufig basieren Stimmungsanalysetools auf komplexen und undurchsichtigen Algorithmen. Die fehlende Transparenz und Verständlichkeit kann zu fehlendem Vertrauen bei den Stakeholdern führen, die das Tool letztendlich verwenden sollen. Durch die Einführung von Erklärbarkeit kann jedoch die Nutzerakzeptanz und damit auch die Nutzung einer Software erhöht werden [12]. Daher wird in dieser Arbeit die Einführung von Erklärbarkeit zur Erhöhung der Nutzerakzeptanz von Stimmungsanalysetools betrachtet.

Nach kurzer Literatursichtung wurde deutlich, dass bereits viele Studien

sich mit Erklärbarkeit beschäftigen, wie z. B. die systematische Literaturanalyse von Chazette et al. [12]. Auch befassen sich viele Studien mit Stimmungsanalysetools, wie z. B. die systematische Literaturanalyse von Obaidi und Klünder [46]. Jedoch wurde keine Studie gesichtet, welche sich mit den Potenzialen bezüglich der Einführung von Erklärbarkeit bei Stimmungsanalysetools beschäftigt.

1.3 Lösungsansatz

Zuerst werden die Grundlagen der Stimmungsanalyse und Erklärbarkeit im Rahmen einer Literaturrecherche erkundet. Auf Grundlage der Literaturrecherche wird ein für diese Arbeit geeignetes Stimmungsanalysetool ausgewählt und mithilfe eines Workshops Erklärbarkeitsanforderungen im Kontext des Tools erhoben. Danach wird das Stimmungsanalysetool angepasst, sodass Erklärbarkeit unter Berücksichtigung der wichtigsten Anforderungen eingeführt wird. Abschließend wird eine Online-Umfrage durchgeführt, um eine Erhöhung der Nutzerakzeptanz durch die umgesetzten Erklärbarkeitsanforderungen feststellen zu können.

1.4 Ergebnisse der Arbeit

Auf Grundlage der Ergebnisse der Literaturrecherche wurde das Stimmungsanalysetool *RoBERTa* als das für diese Arbeit prädestinierteste Tool identifiziert. Um dieses Tool für die Anwendung im SE-Kontext nutzen zu können, wurde *RoBERTa* mithilfe eines Datensatzes vortrainiert. Im Rahmen eines Workshops wurde das Stimmungsanalysetool vorgestellt und Erklärbarkeitsanforderungen erhoben. Diese wurden auf Realisierbarkeit überprüft und softwaretechnisch bzw. prototypisch im Stimmungsanalysetool *RoBERTa* umgesetzt. Danach wurde mithilfe einer Online-Umfrage die Nutzerakzeptanz des Stimmungsanalysetools ohne und mit den jeweiligen Erklärungen erhoben. Bei der anschließenden Analyse der Umfrageergebnisse wurde u. a. festgestellt, dass die Erklärungen *Beispiele* und *Schlüsselwörter* zu einer signifikanten mittelgroßen Erhöhung der Nutzerakzeptanz führen. Die Arbeit spricht Empfehlungen für die Industrie aus, wie z. B. die Umsetzung dieser Erklärungen unter Voraussetzung den vorliegenden Kontext auf Zielkonflikte zu untersuchen.

1.5 Struktur der Arbeit

In diesem Kapitel 1 wird die Einleitung der Arbeit und hierzu u. a. die Motivation und Problemstellung beschrieben. Danach werden in Kapitel 2 die Grundlagen bezüglich Erklärbarkeit und Stimmungsanalysetools erläutert. In Kapitel 3 wird das Vorgehen der Literaturrecherche beschrieben und

begründet. Neben den Forschungsfragen werden in Kapitel 4 die Forschungsmethodik und Ergebnisse des Workshops dargestellt. In Kapitel 5 werden die Auswirkungen von Erklärbarkeit auf andere nichtfunktionale Anforderungen abgeschätzt. Die im Workshop erhobenen Erklärbarkeitsanforderungen werden softwaretechnisch bzw. prototypisch umgesetzt und dies in Kapitel 6 dokumentiert. Um einen möglichen Einfluss auf die Nutzerakzeptanz durch die umgesetzten Erklärbarkeitsanforderungen feststellen zu können, wird eine Umfrage durchgeführt und in Kapitel 7 die Forschungsmethodik und Ergebnisse erläutert. In Kapitel 8 werden die Ergebnisse der Arbeit diskutiert und zu diesem Zweck u. a. die Forschungsfragen beantwortet. Abschließend wird die Arbeit in Kapitel 9 zusammengefasst und ein Ausblick in die zukünftige Forschung skizziert.

Kapitel 2

Grundlagen und verwandte Arbeiten

In diesem Kapitel werden die Grundlagen der Arbeit beschrieben. Hierzu werden unter anderem die Begriffe *Stimmungsanalysetool* und *Erklärbarkeit* einheitlich definiert und der aktuelle Stand der Forschung erörtert.

2.1 Stimmungsanalysetools

In der Studie von Zucco et al. [64] wird die Stimmungsanalyse als der Prozess definiert, welcher Gefühle, Emotionen oder Meinungen aus Texten extrahiert. Folglich werden Stimmungsanalysetools definiert als Tools, welche die Stimmung auf der Grundlage von textbasierter Kommunikation analysieren.

2.1.1 Begriffliche Abgrenzung

In der Literatur wird ersichtlich, dass einige Studien wie z. B. Cagnoni et al. [8] die Begriffe Emotionsanalyse und Stimmungsanalyse gleichsetzen. Jedoch definieren Klünder und Karras [34] die Begriffe unterschiedlich. Die Studie definiert die Emotion als einen kurzfristigen und die Stimmung als langfristigen affektiven Zustand [34]. Im Gegensatz zur Einordnung der Stimmung in eine negative, neutrale oder positive Polarität wird die Emotion in deutlich umfangreichere Kategorien eingeordnet, wie z. B. Trauer, Wut, Freude und Liebe. Daher werden in dieser Arbeit die Begriffe Emotionsanalyse und Stimmungsanalyse voneinander abgegrenzt und die Definitionen von Klünder und Karras [34] in den Kontext dieser Arbeit übertragen.

2.1.2 Potenziale im Software Engineering

Um die Potenziale der Stimmungsanalyse im Projektmanagement zu untersuchen, werden in der Studie von Targiel [57] die Mailinglisten der

Open-Source Software Apache hinsichtlich der Stimmung analysiert. Die Ergebnisse der Studie zeigen, dass sich die Einstellungen der Entwicklerinnen und Entwickler durch externe Faktoren verändern, wie zum Beispiel die COVID-19 Pandemie [57]. Deshalb empfiehlt die Studie die Einführung von Stimmungsanalysetools im Projektmanagement, insbesondere hinsichtlich der COVID-19 Pandemie [57]. Auch wenn die Studie das Projektmanagement und nicht die Softwareentwicklung betrachtet, lassen sich die Ergebnisse auf den Kontext dieser Arbeit übertragen. Somit wird der Nutzen dieser Arbeit bestätigt.

Dhakad und Benedicenti [23] untersuchen die emotionale Ansteckung unter den Entwicklerinnen und Entwicklern von global verteilter Open-Source Software. Dabei wird untersucht, inwiefern sich die emotionale Ansteckung auf den Entwicklungsprozess auswirkt [23]. Die Ergebnisse zeigen, dass aus einer initial positiven bzw. negativen emotionalen Reaktion mehrere positive bzw. negative Reaktionen folgen [23]. Initial neutrale Reaktionen haben dagegen keine Auswirkungen auf die Folgereaktionen [23]. Auch wenn Klünder und Karras [34] Emotionen und Stimmungen unterschiedlich definieren, lassen sich Folgereaktionen auch bei Stimmungen vermuten. Daher bestätigen die Ergebnisse von Dhakad und Benedicenti [23] den Nutzen dieser Arbeit.

Die Studie von Klünder und Karras [34] untersucht die sozialen Aspekte von Meetings im Software-Engineering, da diese Auswirkungen auf die Stimmung im Team haben. Die Aussagen werden in positive, neutrale und negative Kategorien eingeordnet und mit der Stimmung im Team verglichen [34]. Karras und Klünder [34] kommen zum Ergebnis, dass eher die Gefühle und weniger die Stimmung durch die Aussagen im Meeting beeinflusst werden. Die Ergebnisse von Klünder und Karras [34] zeigen somit für die weitere Forschung, dass sich eher die Betrachtung von Emotionsanalysen und weniger von Stimmungsanalysen lohnt. In dieser Arbeit werden die Ergebnisse von Klünder und Karras [34] nicht weiter berücksichtigt, auch wenn dies ein guter Ansatz für die weitere Forschung ist.

In bestehender Literatur werden Ergebnisse der Stimmungsanalyse erst am Ende des Tages für die Entwicklerinnen und Entwickler ersichtlich [56]. Um dieses Problem zu beheben, wird in der Studie von Schroth et al. [56] eine Stimmungsanalyse in Echtzeit durch eine Nutzerstudie evaluiert. Durch die Ergebnisse der Studie wird die Nützlichkeit der Stimmungsanalyse in Echtzeit ersichtlich [56]. Daher wird in dieser Arbeit der Vorteil von Echtzeit bei der Auswahl eines geeigneten Stimmungsanalysetools berücksichtigt.

Die Studie von Kaur et al. [33] untersucht die Commit-Logs von Open-Source Projekten, um die Stimmung unter den Entwicklerinnen und Entwicklern solcher Projekte zu bestimmen. In den Commit-Logs wurden überwiegend neutrale Aussagen beobachtet [33]. Des Weiteren wird untersucht, inwieweit die Teamgröße, Commit-Logs und Umfang der Beiträge

die Stimmung im Team beeinflusst [33]. Die Beeinflussung von Teamgröße, Commit-Logs und Umfang der Beiträge auf die Stimmung konnte nachgewiesen werden [33]. Die Ergebnisse von Kaur et al. [33] werden in dieser Arbeit nicht berücksichtigt, da nicht die Einflussfaktoren der Stimmung, sondern nur dessen Analyse für die Arbeit relevant ist.

Da Stimmungsanalysetools meist über die Kommandozeile gesteuert werden, ist die Nutzung eher für technisch versierte Entwicklerinnen und Entwickler geeignet [50]. Um eine erleichterte Anwendung für alle Stakeholder zu erzielen, wird in der Bachelorarbeit von Olsen [50] eine grafische Oberfläche implementiert. Die zuvor vermutete verbesserte Bedienung des Stimmungsanalysetools wurde durch eine Nutzerstudie nachgewiesen [50]. Daher wird in dieser Arbeit ein Stimmungsanalysetool gewählt, welches mit einer grafischen Oberfläche arbeitet bzw. die Möglichkeit hat diese einfach zu implementieren.

Zwar bestehen Stimmungsanalysen auf Grundlage textbasierter Kommunikation, jedoch findet der Großteil der Kommunikation in Meetings statt [27]. Um auch diesen Teil der Kommunikation analysieren zu können, stellen Herrmann und Klünder [27] ein geeignetes Konzept vor. Dabei werden Audioaufnahmen von Meetings mithilfe von Spracherkennung verarbeitet und die transkribierten Aussagen im Stimmungsanalysetool weiterverarbeitet [27]. Die Ergebnisse von Herrmann und Klünder [27] sind ein guter Ansatz für die weitere Forschung, um die Anwendung von Stimmungsanalysetools auf Meetings zu erweitern. In dieser Arbeit wird das Konzept jedoch nicht berücksichtigt.

2.1.3 Erkennung von Ironie und Sarkasmus

Stimmungsanalysetools haben Probleme dabei Ironie unter den Entwicklerinnen und Entwicklern zu erkennen [52]. Dies wird durch die Ergebnisse von Obaidi und Klünder [46] bestätigt. Daher werden die Auswirkungen von Ironie auf die Stimmung untersucht [52]. Die Bachelorarbeit von Rakow [52] kommt zum Schluss, dass Ironie einen negativen, aber auch positiven Einfluss auf die Softwareentwicklung haben kann. Diese Auswirkungen werden jedoch unter den Entwicklerinnen und Entwicklern geringer bewertet als bei den Nicht-Entwicklerinnen und Nicht-Entwicklern [52]. Somit wird der Nutzen von Stimmungsanalysetools im Kontext der in dieser Arbeit betrachteten Domäne bestätigt.

Da die meisten Stimmungsanalysetools Ironie nicht oder nur schlecht erkennen, wird in der Masterarbeit von Schierholz [54] ein Verfahren zur Erkennung vorgestellt. Dazu wird ein Datensatz mit der Einordnung in *Ironie* und *nicht Ironie* erhoben [54]. Die Ergebnisse zeigen, dass die Erkennung von Ironie durch eine automatisierte Software möglich ist [54]. Dies ist ein guter Ansatz für die weitere Forschung, um die Nutzerakzeptanz von Stimmungsanalysetools zu erhöhen. Im Rahmen dieser Arbeit wird dies

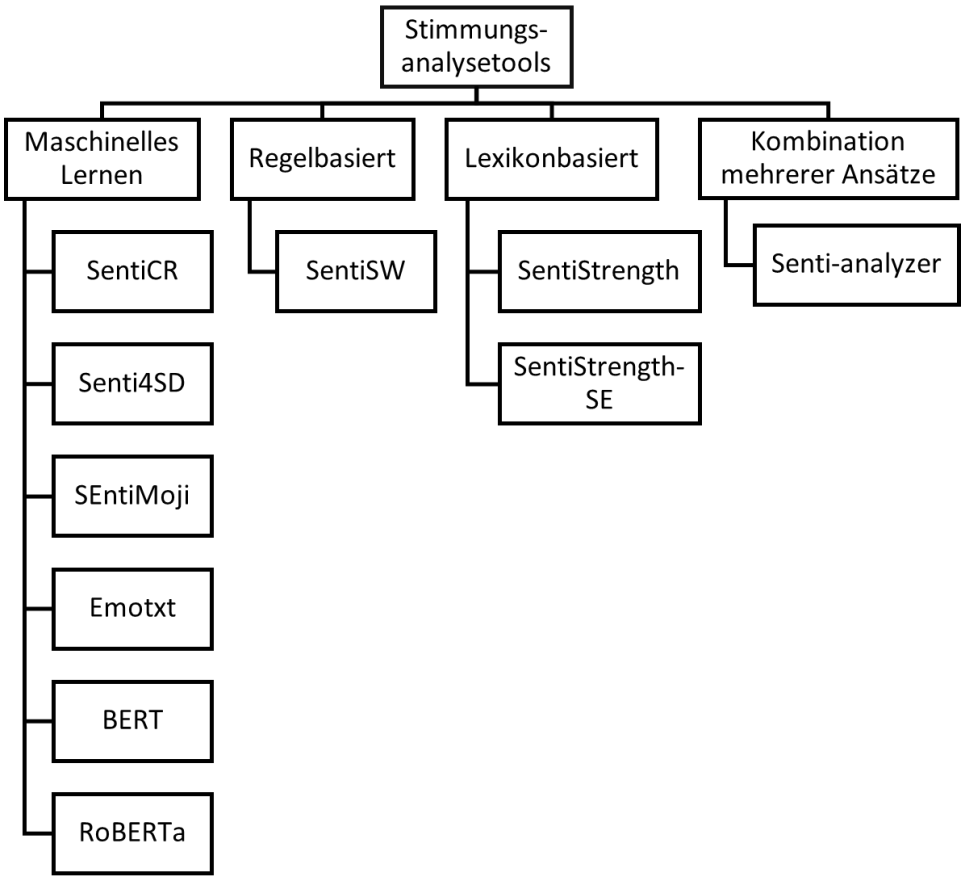

Abbildung 2.1: Zuordnung der Stimmungsanalysetools bezüglich der Ansätze

jedoch nicht umgesetzt.

Devika et al. [21] unterteilen die Stimmungsanalyse in die folgenden Ansätze: maschinelles Lernen, regelbasiert und lexikonbasiert. Die Stimmungsanalysetools werden den Ansätzen zugeordnet und dies in der Abbildung 2.1 dargestellt.

2.1.4 Maschinelles Lernen

Beim maschinellen Lernen wird der Algorithmus zuerst mit einem Trainingsdatensatz vortrainiert [21]. Anschließend kann dieser mit neuen unbekannten Eingaben arbeiten [21]. Die Support-Vector Maschine, N-Gramm Sentiment Analyse und Naïve-Bayes Methode sind unter anderem bekannte Verfahren, die auf maschinelles Lernen basieren und in der Stimmungsanalyse verwendet werden [21].

Anpassung an die Domäne Da die Ergebnisse nicht sehr präzise sind, ist die Anwendung von Stimmungsanalysetools bei textbasierten Software-Artefakten bisher problematisch [7]. Dies wird durch die Ergebnisse von Obaidi und Klünder [46] bestätigt. Um die Anwendung in diesem Kontext zu verbessern, werden in dieser Studie neuronale Netze verwendet und die Stimmungsanalyse an die Domäne des Software-Engineering angepasst [7]. Die Studie von Biswas et al. [7] zeigt eine verbesserte Präzision der Ergebnisse, da Oversampling und Undersampling der Daten verhindert wird. Um dies in der Arbeit zu berücksichtigen, wird das Stimmungsanalysetool ggf. an die Domäne angepasst.

Die zur Analyse von Texten auf Social-Media verwendeten Stimmungsanalysen können nicht ohne Weiteres auf den Bereich Software-Engineering übertragen werden [3]. Da jedoch der Bedarf für Stimmungsanalysen trotzdem besteht, werden in der Studie von Ahmed et al. [3] 2000 Review-Kommentare manuell kategorisiert und ein Datensatz daraus erstellt. Dieser wurde in sieben populären Stimmungsanalysetools angewandt [3].

SentiCR Da die Anwendung bestehender Stimmungsanalysetools zu sehr unpräzisen Ergebnissen führt, wird das Stimmungsanalysetool SentiCR entwickelt [3]. SentiCR arbeitet auf der Grundlage maschinellen Lernens und kann zu einer Wahrscheinlichkeit von 83% einen Kommentar mit negativer Stimmung identifizieren [3]. Somit ist dieses Stimmungsanalysetool für den Kontext dieser Arbeit geeignet.

Senti4SD Stimmungsanalysetools haben häufig das Problem stimmungsneutrale Aussagen fehlerhaft in die Kategorien positiv und negativ einzuordnen [9]. Um dies zu beheben, wird in der Studie von Calefato et al. [9] das auf maschinelles Lernen basierende Stimmungsanalysetool Senti4SD vorgestellt. Calefato et al. [9] erstellen auf Grundlage von über 4400 Beiträgen einen Goldstandard für die Stimmungsanalyse im Software-Engineering. Das Tool wird mithilfe des Goldstandard-Datensatzes speziell für die Anwendung im Software-Engineering trainiert [9]. Die Studie zeigt, dass Senti4SD tatsächlich die Validität der Ergebnisse im Bereich Software-Engineering verbessert [9]. Daher wird Senti4SD als ein für diese Arbeit potenziell geeignetes Stimmungsanalysetool betrachtet.

SEntiMoji Durch technische Fachbegriffe im Software-Engineering leidet die Präzision der Ergebnisse von Stimmungsanalysetools und anwendungsspezifische Anpassungen werden nötig [17]. Da Emojis allgegenwärtig und domänenübergreifend verwendet werden, sind diese einfacher zu interpretieren [17]. Daher wird in der Studie von Chen et al. [17] das auf maschinelles Lernen basierende Tool SEntiMoji vorgestellt, welches mithilfe von Emojis die Stimmung analysiert. Im Vergleich zu bestehenden Stimmungsanalyse-

tools arbeitet SEntiMoji präziser [17]. Somit wird SEntiMoji [17] ebenfalls als ein für diese Arbeit geeignetes Stimmungsanalysetool betrachtet, da es präziser arbeitet als z. B. Senti4SD [9].

Emotxt Eine Open-Source Software zur Analyse von Emotionen aus Texten, aber auch zum Training von benutzerdefinierten Klassifizierungsmodellen, wird bisher nicht angeboten [10]. Daher wird in der Studie von Calefato et al. [10] das Tool EmoTxt vorgestellt, welches dies anbietet und auf maschinelles Lernen basiert. Durch eine Evaluierung auf Grundlage mehrerer Datensätze wird eine stabile und gute Validität festgestellt [10]. Zwar verwenden Emotionsanalysen umfangreiche Kategorien, wie z. B. Trauer, Wut, Freude und Liebe. Jedoch können Polaritäten (negativ, neutral, positiv) der Stimmungsanalyse abgeleitet und im Rahmen dieser Arbeit verwendet werden. Daher wird Emotxt [10] als ein für diese Arbeit potenziell geeignetes Stimmungsanalysetool betrachtet.

BERT In der Studie von Devlin et al. [22] wird das auf maschinelles Lernen basierende Sprachrepräsentationsmodell BERT vorgestellt, das im Gegensatz zu aktuellen Modellen bidirektionales Labeln unterstützt. Da durch den Einsatz die Genauigkeit um minimal 1,5% bzw. maximal 7,7% erhöht wird, ist die Leistungsfähigkeit von BERT empirisch bewiesen [22]. BERT kann auch im Kontext dieser Arbeit als ein geeignetes Stimmungsanalysetool eingesetzt werden.

Um die Stimmung unter den Softwareentwicklerinnen und Softwareentwicklern zu verbessern, wird ein Assistent vorgestellt, der die Stimmung in Echtzeit analysiert und Verbesserungen in der Kommunikation vorschlägt [26]. Durch die Anwendung des Assistenten lassen sich einige Potenziale erschließen wie z. B., dass die Entwicklerinnen und Entwickler den Assistenten als angemessen beurteilen und die Stimmung im Team verbessert wird [26]. In der Masterarbeit von Herkenhof [26] wird das Stimmungsanalysetool BERT [22] verwendet, da dies natürliche Textsprache besser verarbeitet als z. B. SentiStrength-SE [31]. Somit bestätigen die Ergebnisse von Herkenhof [26] die Anwendbarkeit von BERT [22] als Stimmungsanalysetool.

RoBERTa In der Studie von Liu et al. [41] werden Designentscheidungen von BERT [22] kritisiert und die Weiterentwicklung RoBERTa vorgestellt. Liu et al. [41] berücksichtigen fehlende Hyperparameter und die Größe des Trainingsdatensatzes, wodurch BERT [22] signifikant untertrainiert wird. Die Ergebnisse der Studie [41] zeigen, dass RoBERTa effizienter arbeitet als BERT [22]. In dieser Arbeit kann RoBERTa [41] ebenfalls als ein Stimmungsanalysetool eingesetzt werden, da dies eine Weiterentwicklung von BERT [22] ist.

Liao et al. [39] führen ein textbasiertes Stimmungsanalysemodell ein, welches auf RoBERTa [41] basiert. Die Ergebnisse der Studie zeigen, dass bisherige Vergleichsmodelle übertroffen werden [39]. Somit wird die Anwendbarkeit von RoBERTa [41] als Stimmungsanalysetool bestätigt.

2.1.5 Lexikonbasiert

Der lexikonbasierte Ansatz verwendet ein Domänenwörterbuch mit den Polaritäten, die das Vorliegen einer positiven, neutralen oder negativen Stimmung repräsentieren [21]. Auf dessen Grundlage werden die einzelnen Wörter bzw. Sätze gewertet [21]. Diese werden unter anderem gewichtet und anschließend summiert [21]. Die Summe stellt die Polarität des vollständigen Eingabetextes dar [21].

SentiStrength Da die meisten Algorithmen für die Stimmungsanalyse kommerziell sind bzw. nicht das Verhalten des Nutzenden erkennen, stellen Thelwall et al. [58] SentiStrength vor. Der Algorithmus untersucht jedes einzelne Wort im Text und ordnet es mithilfe einer Nachschlagtabelle einer passenden Stimmungsstärke zu [58]. Dabei repräsentieren die Werte von $+5$ bis $+1$ positive und von -1 bis -5 negative Stimmungen [58]. Thelwall et al. [58] zeigen, dass positive Stimmungen mit einer Genauigkeit von 60,6% und negative Stimmungen mit 72,8% vorhergesagt werden. Somit wird SentiStrength [58] als ein für diese Arbeit potenziell geeignetes Stimmungsanalysetool betrachtet.

SentiStrength-SE Islam und Zibran [31] untersuchen die Schwierigkeiten bei der Nutzung von Stimmungsanalysetools im Software Engineering. Durch den Aufbau eines Domänenwörterbuchs und geeigneter Heuristiken wird das Stimmungsanalysetool SentiStrength-SE vorgestellt, welches eine Weiterentwicklung von SentiStrength [58] ist und speziell für die Nutzung im Software Engineering konzipiert ist [31]. Durch eine Evaluierung von Islam und Zibran [31] anhand eines Datensatzes mit 5.600 Jira-Einträgen wird ersichtlich, dass SentiStrength-SE deutlich bessere Ergebnisse im Software Engineering erzielt als SentiStrength [58]. Da in dieser Arbeit die Domäne des Software Engineering untersucht wird, ist SentiStrength-SE [31] ein für diese Arbeit potenziell geeignetes Stimmungsanalysetool.

act4teams® Meetings mit zu destruktiven Äußerungen können bei den Entwicklerinnen und Entwicklern zu Unzufriedenheit im Team führen [48]. Mithilfe des Tools act4teams® können diese ermittelt werden [48]. Die Masterarbeit von Obeidi [48] untersucht die automatisierte Stimmungsanalyse in Echtzeit mithilfe des lexikonbasierten Stimmungsanalysetools SentiStrength-SE [31] und act4teams®. Da in dieser Arbeit die gesamte Kommunikation im

Team betrachtet wird und act4teams® lediglich die Aussagen in Meetings kategorisiert, ist das Tool für den Kontext dieser Arbeit ungeeignet.

2.1.6 Regelbasiert

Beim regelbasierten Ansatz werden Regeln festgelegt, die im Eingabetext auf Vorhandensein überprüft werden und dann als positiv bzw. negativ gewertet werden [21]. Die Einzelwertungen werden zusammengefasst und eine Gesamtwertung für den vollständigen Eingabetext wird ausgegeben [21]. Danach wird abgefragt, ob die Gesamtwertung korrekt ist. Dies ist erforderlich, um durch überwachtes Lernen das System auf neue Eingaben zu trainieren [21].

SentiSW Bei der Ausweitung der Anwendung von Stimmungsanalysetools auf andere Entitäten entstehen meist Fehlklassifizierungen [24]. Daher stellen Ding et al. [24] das regelbasierte Stimmungsanalysetool SentiSW vor, welches neben der Stimmung auch die Entität bestimmt. Die Entität weist eine Genauigkeit von 75,15% bzw. die Stimmung von 77,19% auf und ist somit deutlich eiffizienter als bestehende Stimmungsanalysetools [24], wie z. B. SentiStrength [58] und SentiStrength-SE [31]. Somit wird auch SentiSW [24] als ein für diese Arbeit potenziell geeignetes Stimmungsanalysetool betrachtet.

2.1.7 Kombination mehrerer Ansätze

In der Literatur bestehende Stimmungsanalysetools verwenden u. a. eine Kombination aus mehreren zuvor genannten Ansätzen, wie z. B. die Studie von Herrmann et al. [29].

Senti-analyzer Die Studie von Herrmann et al. [29] untersucht das Stimmungsanalysetool Senti-analyzer, welches die verbale Kommunikation im Team transkribiert und die jeweiligen Wörter in negative, neutrale oder positive Aussagen kategorisiert. Das Tool Senti-analyzer [29] verwendet eine Kombination aus maschinellem Lernen und dem lexikonbasierten Ansatz, da dieses auf Grundlage von SentiStrength [58], SentiStrength-SE [31] und Senti4SD [24] arbeitet. Herrmann et al. [29] zeigen, dass 73,3% der Aussagen korrekt kategorisiert werden und vergleichen dies mit anderen Stimmungsanalysetools. Daher wird Senti-analyzer als ein potenziell geeignetes Stimmungsanalysetool im Kontext dieser Arbeit berücksichtigt.

2.1.8 Vergleich der Ansätze

Obaidi et al. [47] führen eine systematische Mapping-Studie auf der Grundlage von 106 Studien durch. Dabei werden bei der Auswahl von Stimmungsanalysetools vom Kontext abhängige Empfehlungen abgeleitet [47]. Die Präzisionswerte der Stimmungsanalysetools sind der Tabelle 2.1 zu

Stimmungsanalysetool	Durchschnittliche Präzision
BERT	0.94 [47]
RoBERTa	0.91 [47]
SEntiMoji	0.87 [17]
SentiSW	0.77 [24]
SentiCR	0.76 [47]
Emotxt	0.75 [47]
Senti4SD	0.74 [47]
SentiStrength-SE	0.73 [47]
SentiStrength	0.71 [47]

Tabelle 2.1: Durchschnittliche Präzision der Stimmungsanalysetools

entnehmen. Anhand der Präzisionswerte wird ersichtlich, dass der Einsatz von neuronalen Netzen der präziseste Ansatz und BERT das präziseste Werkzeug ist [47]. Daher sind in dieser Arbeit auf maschinelles Lernen basierende Stimmungsanalysetools, insbesondere BERT zu bevorzugen. Auch wird in der Studie abgeleitet, dass Stimmungsanalysetools Probleme bei der Erkennung von Ironie und Sarkasmus haben [47]. Dies ist eine Grenze von Stimmungsanalysetools, die auch in dieser Arbeit berücksichtig wird.

Die Studie von Herrmann et al. [28] verwendet ein Stimmungsanalysetool mit vorher gelabelten Datensätzen aus unterschiedlichen Quellen. Ziel der Studie ist die Überprüfung, inwieweit die Label von den Wahrnehmungen der Teilnehmenden des Softwareprojekts abweichen [28]. In der Studie mit 94 Teilnehmenden wird ersichtlich, dass teilweise große Differenzen zwischen dem Label und der Wahrnehmung bestehen [28]. Die Ergebnisse der Studie zeigen, dass die Übereinstimmung teilweise nur bei 62,5% liegt [28]. Auch zeigt die Studie, dass richtlinienbasierte Datensätze besser abschneiden als ad-hoc basierte Datensätze [28]. Da vorher gelabelte Datensätze zu einer höheren Fehlerquote führen, empfiehlt die Studie neuronale Netze zu bevorzugen [28]. Somit wird die Empfehlung von Obaidi et al. [47] bestätigt, dass auf maschinelles Lernen basierende Stimmungsanalysetools zu bevorzugen sind.

In der Metastudie von Abo et al. [1] wird nicht die Stimmung im Software-Engineering untersucht, sondern von Nutzenden des Internets. Hauptforschungsbereiche der Metastudie sind Validierungs-, Lösungs- und Evaluationsforschung [1]. Die Ergebnisse zeigen einen Anstieg der Publikationen im Bereich der Stimmungsanalyse, wobei der größte Bereich mit 76% auf Grundlage maschinellen Lernens arbeitet [1]. Dagegen arbeiten 14% auf Grundlage eines Lexikons und die restlichen 10% auf Grundlage beider Techniken [1]. Die Ergebnisse von Abo et al. [1] werden bei der Auswahl eines geeigneten Stimmungsanalysetools im Kontext dieser Arbeit berücksichtigt.

Um die durchschnittliche Stimmung im Internet zu bestimmen, werden

in der Studie von Devika et al. [21] Bewertungen auf Onlineplattformen mithilfe von Stimmungsanalysetools untersucht. Dabei werden verschiedene Methoden der Stimmungsanalyse miteinander verglichen und vorgestellt [21]. Bei dem Vergleich wird ersichtlich, dass die meisten Studien auf maschinelles Lernen basierende Stimmungsanalysetools verwenden und diese auch in den Bereichen Leistung, Effizienz und Genauigkeit die besten Ergebnisse erzielen [21].

Auch werden in der Studie von Devika et al. [21] die Vor- und Nachteile der Ansätze genannt. Der lexikonbasierte Ansatz hat den Vorteil, dass im Gegensatz zum maschinellen Lernen keine Daten gelabelt und trainiert werden müssen [21]. Dafür ist beim maschinellen Lernen ein Lexikon nicht erforderlich und die Ergebnisse sind genauer [21]. Jedoch funktioniert maschinelles Lernen meist nur in der Domäne, in der die Daten trainiert wurden [21].

Im Rahmen der Studien Calefato et al. [9] und Novielli et al. [45] werden Kommentare aus dem Kontext der Softwareentwicklung gesammelt und in Bezug auf die vorliegende Stimmung eingeordnet. Beide Studien stellen Datensätze mit mehreren tausenden Einträgen der weiteren Forschung zur Verfügung. Daher können diese in der Arbeit zum Trainieren eines Modells beim maschinellen Lernen verwendet werden. Folglich ist das Stimmungsanalysetool nach dem Trainieren nur geeignet für die Anwendung in der Softwareentwicklung. Dies ist jedoch akzeptabel, da in der Arbeit lediglich diese Domäne betrachtet wird.

In der Softwareentwicklung werden bereits erfolgreich Stimmungsanalysetools eingesetzt [44]. In der Bachelorarbeit von Moghadam [44] wird dessen Anwendung bei Softwarenutzenden untersucht und empfiehlt für die jeweilige Domäne das optimale Stimmungsanalysetool. Die Ergebnisse von Moghadam [44] zeigen, dass RoBERTa in allen untersuchten Domänen am effizientesten arbeitet. Daher ist RoBERTa das für diese Arbeit geeignetste Stimmungsanalysetool und wird im weiteren Verlauf der Arbeit verwendet.

2.2 Erklärbarkeit

Chazette et al. [12] definiert ein System bezüglich eines Aspektes im Kontext als erklärbar, wenn der Adressat durch die Bereitstellung von Informationen in die Lage versetzt wird diesen Aspekt in Bezug auf den Kontext zu verstehen. Somit kann für jeden Aspekt eines Systems eine Erklärung eingeführt werden. Die Definition von Chazette et al. [12] wird nachfolgend für den Begriff *Erklärbarkeit* in der Arbeit verwendet.

2.2.1 Ziele der Erklärbarkeit

Um die mangelnde Benutzerfreundlichkeit von Softwaresystemen zu erhöhen, empfiehlt die Studie von Mann et al. [43] Strategien zur Erhöhung der Trans-

parenz in Abhängigkeit zum Kontext und der Quelle der Intransparenz. In der Studie werden die Quellen der Intransparenz in die drei Hauptkategorien architektonisch, analytisch und sozio-technisch unterteilt und schlägt für jede Kategorie Verbesserungsvorschläge vor, wie zum Beispiel die Einführung von Erklärbarkeit [43]. Die Ergebnisse von Mann et al. [43] zeigen die Potenziale von Erklärbarkeit und bestätigen somit den Nutzen dieser Arbeit.

Da ethische, rechtliche, soziale und technische Hürden bei KI-Technologien zu bewältigen sind, ist die Einführung von vertrauenswürdiger KI sinnvoll [59]. Daher werden in der Studie von Thiebes et al. [59] Wohltätigkeit, Nicht-Bösartigkeit, Autonomie, Gerechtigkeit und Erklärbarkeit als die Grundprinzipien der vertrauenswürdigen KI konzeptionell vorgestellt und dessen Umsetzung demonstriert. Die Ergebnisse der Studie zeigen, dass das Konzept sinnvoll ist und einen Nutzen für die zukünftige Forschung bietet [59]. Da auch diese Studie die Potenziale von Erklärbarkeit zeigt, werden die Ergebnisse von Mann et al. [43] bestätigt.

2.2.2 Einfluss auf andere nichtfunktionale Anforderungen

Die Erklärbarkeit als nichtfunktionale Qualitätsanforderung kann die Softwarequalität verbessern, jedoch besteht die Annahme, dass Erklärbarkeit die Leistung und damit auch die Softwarequalität negativ beeinflusst [19]. Daher wird in der Studie von Crook et al. [19] untersucht, ob ein Zielkonflikt besteht und gegebenenfalls ein Kompromiss zwischen den Zielen gefunden werden kann. Damit ein Kompromiss gefunden wird, empfiehlt die Studie die Berücksichtigung von Ressourcenverfügbarkeit, Domänenmerkmalen und Risikobetrachtung [19]. Um in dieser Arbeit die Beeinflussung der Softwarequalität durch Zielkonflikte untersuchen zu können, werden die nichtfunktionalen Qualitätsanforderungen von Chazette et al. [12] im Rahmen eines Workshops auf Wichtigkeit untersucht. Danach wird mithilfe der Studie von Chazette et al. [12] überprüft, inwiefern sich die Erklärbarkeit auf andere nichtfunktionale Qualitätsanforderungen auswirkt. Um ein Kompromiss zwischen diesen zu finden, lohnt sich die Betrachtung der Ergebnisse von Crook et al. [19].

Obwohl nichtfunktionale Anforderungen häufig untereinander konkurrieren, sind sie wichtig für den Erfolg von Softwareprojekten [42]. Daher werden in der Metastudie die in der Literatur bestehenden Definitionen und Konflikte der nichtfunktionalen Anforderungen untersucht [42]. Zur Einordnung in absolute, relative und keine Konflikte stellt die Metastudie von Mairiza und Zowghi [42] einen umfassenden Katalog zur Verfügung. Um in dieser Arbeit zu überprüfen, inwiefern durch die Einführung von Erklärbarkeit andere nichtfunktionale Anforderungen beeinflusst werden, wird neben der Studie von Chazette et al. [12] auch der Konfliktkatalog von Mairiza und Zowghi [42] berücksichtigt.

Um den Bedarf nach Transparenz und dessen Erreichbarkeit durch

Erklärungen zu erheben, wird in der Studie von Chazette und Schneider [16] eine Umfrage durchgeführt. Die Studie untersucht, ebenso wie die Studie von Crook et al. [19], inwiefern sich Erklärbarkeit auf die Softwarequalität auswirkt [16]. Chazette und Schneider [16] kommen zum Ergebnis, dass leichtgewichtige nutzerzentrierte Erklärungen die Softwarequalität nicht negativ beeinflussen. Daher wird in dieser Arbeit der Design- und Implementierungsprozess so gestaltet, dass die Erklärungen leichtgewichtig und nutzerzentriert sind.

2.2.3 Mensch als Erklärer

In vielen Studien wird die Erklärbarkeit als eine wichtige nichtfunktionale Qualitätsanforderung betrachtet, jedoch bestehen in der Literatur nur wenige Studien wie die von Chazette et al. [12], welche die Definition von Erklärbarkeit untersuchen [5]. Daher wird in der Studie von Balasubramaniam et al. [5] eine einheitliche Definition erarbeitet. Dabei wird ersichtlich, dass Erklärbarkeit als ein wichtiger Bestandteil der Transparenz betrachtet und vor allem zur Förderung von Vertrauen in KI-Systeme eingesetzt wird [5]. Des Weiteren wird die Erklärbarkeit in die Komponenten Adressat, Inhalt, Kontext und Erklärer eingeteilt [5]. Da die Betrachtung der Komponenten bei der Einführung von Erklärbarkeit relevant ist, wird dies in der nachfolgenden Arbeit berücksichtigt. Um die Vertrauenswürdigkeit in die Erklärbarkeit zu erhöhen, empfiehlt die Studie den Menschen als Erklärer zu nutzen [5]. In dieser Arbeit wird nur die Software und nicht der Mensch als Erklärer betrachtet, da das Stimmungsanalysetool in der Lage sein soll sich selbst zu erklären.

Damit die Undurchsichtigkeiten bei der Nutzung von KI-Systemen bewältigt werden, empfehlen Hind et al. [30] die Einführung von Erklärbarkeit. In der Studie wird im Gegensatz zu bisherigen Studien nicht die Funktionsweise, sondern die Entscheidung des KI-Systems erklärt [30]. Die Studie von Hind et al. [30] zeigt auf Grundlage von zwei Beispielen, dass dieser Ansatz wirksam ist und im Allgemeinen angewendet werden kann. Hind et al. [30] empfiehlt, beim Design die Erklärungen einfach zu gestalten und Denkprozesse eines menschlichen Anwenders nachzubilden. Somit werden die Ergebnisse von Balasubramaniam et al. [5] bestätigt.

2.2.4 Erklärbare KI

Der Einsatz von KI-Technologien führt zu leistungsfähigen Vorhersagen, aber diese werden nicht durch das System erklärt [2]. Da die mangelnde Transparenz zu Schwierigkeiten bei der Nutzung von KI-Technologien führt, wird in der Forschung die erklärbare KI ausführlich betrachtet [2]. Um einen Überblick über den schnell wachsenden Forschungsbereich zu ermöglichen, werden in der Studie von Adadi und Berrada [2] die Ansätze,

Trends und Forschungsrichtungen auf Grundlage einer Literaturrecherche untersucht. In der Studie werden Erklärungsmethoden vorgeschlagen, wie z. B. Entscheidungsbäume, Regelwerke, Substitutionsmodelle und Teilabhängigkeitsdiagramme [2]. Die Erklärungsmethoden werden in dieser Arbeit berücksichtigt, da hierdurch Design und Implementierung der Erklärbarkeit vereinfacht wird. Auch kommt die Studie zum Ergebnis, dass Erklärungsmodelle menschenzentriert sein sollen [2]. Damit wird das Ergebnis der Studie von Balasubramaniam et al. [5] bestätigt.

Um die KI für die Anwendung nachvollziehbarer zu gestalten ist die Begründung besonders wichtig, also wie der KI-Algorithmus das Ergebnis berechnet [6]. Neben einer ausführlichen Begründung wird empfohlen eine verantwortliche Person für die Berechnung zu benennen, um die Vertrauenswürdigkeit der KI-Algorithmen zu erhöhen [6]. Dies ist insbesondere hilfreich bei Uneinigkeiten zwischen Menschen und KI [6]. Da das in dieser Arbeit verwendete Stimmungsanalysetool *RoBERTa* auf maschinelles Lernen basiert, lohnt sich die Betrachtung der Ergebnisse von Baum et al. [6].

Die Effizienz von Textklassifizierungen hat sich durch den Einsatz von neuronalen Netzen erhöht [61]. Das fehlende Vertrauen in neuronale Netze aufgrund von Undurchsichtigkeiten kann durch den Einsatz von Erklärbarkeit erhöht werden [61]. Daher wird in der Studie von Winkler und Vogelsang [61] ein Ansatz zur Veranschaulichung der Erklärbarkeit vorgestellt, der beschreibt, wie das neuronale Netz das Klassifizierungsergebnis berechnet. Die Ergebnisse der Studie zeigen, dass dieser Ansatz zu mehr Vertrauen in die Textklassifizierung führt [61]. Die Ergebnisse von Winkler und Vogelsang [61] bestätigen die Aussage von Chazette et al. [14] bezüglich der Einführung von Erklärungen über die Funktionsweise des Algorithmus. Da die Textklassifizierung des Stimmungsanalysetools *RoBERTa* auf Grundlage eines neuronalen Netzes arbeitet, sind die Ergebnisse von Winkler und Vogelsang [61] auf den Kontext dieser Arbeit übertragbar.

Der Einsatz von künstlicher Intelligenz bietet viele Potenziale im Gesundheitswesen, jedoch herrscht bei den Nutzenden Unverständnis über das Verhalten [4]. Durch den Einsatz von Erklärbarkeit soll die Nutzerakzeptanz erhöht werden, jedoch ist aufgrund unterschiedlicher Definitionen die Umsetzung von Erklärbarkeit erschwert [4]. Daher werden in der Studie von Arbelaez et al. [4] Mindeststandards definiert, um die Anforderungen an die Erklärbarkeit zu ermitteln und somit die Bedürfnisse der Stakeholder zu befriedigen. Die Studie empfiehlt bei der Einführung von Erklärbarkeit kontextbezogene Mindeststandards [4]. In dieser Arbeit wird zwar nicht das Gesundheitsweisen, sondern die Softwareentwicklung als Domäne betrachtet. Jedoch lohnt sich auch in dieser Domäne die Betrachtung von kontextbezogenen Erklärungen. Somit werden die Ergebnisse von Chazette et al. [13] [14] und Hind et al. [30] bestätigt.

Um zu überprüfen, ob die in den internationalen Richtlinien geforderte Erklärbarkeit von KI-Systemen den Qualitätsanforderungen entspricht, wird

in der Studie ein multidisziplinärer Ansatz zur Überprüfung der Erklärbarkeit vorgestellt [38]. In der Studie von Langer et al. [38] führt die technische, psychologische, ethische und rechtliche Überprüfung zu einer verbesserten Qualität der Erklärbarkeit. Auch wenn die Ergebnisse von Langer et al. [38] einen guten Ansatz für die weitere Forschung darstellen, wird dies im Kontext der Arbeit nicht berücksichtigt.

2.2.5 Erklärbarkeit und Nutzerakzeptanz

Erklärbarkeit soll das Vertrauen der Stakeholder fördern, jedoch wird in neueren Studien gezeigt, dass dies nicht unbedingt zutrifft [32]. Daher müssen die Anforderungen an die Erklärbarkeit angepasst werden, sodass das Vertrauen der Stakeholder gefördert und ein vertrauenswürdiges System geschaffen wird [32]. Die Studie von Kastner et al. [32] zeigt, dass die Vertrauenswürdigkeit des zu erklärenden Systems wichtiger ist als das Vertrauen des Nutzenden selbst und empfiehlt daher, dass dies bei der Erklärbarkeit berücksichtigt wird. Daher wird dies in der Arbeit berücksichtigt, insbesondere bei der Erhebung der Erklärbarkeitsanforderungen.

Da nur wenige bestehende Studien die Erhebung von Erklärbarkeitsanforderungen und das Systemdesign untersuchen, wird in dieser Studie die Definition von Erklärbarkeit, ein konzeptionelles Modell, ein Wissenskatalog und ein Referenzmodell für erklärbare Systeme vorgestellt [13]. Die Ergebnisse der Studie sollen die Auswahl der passenden Erklärbarkeitsmethode und Evaluationsmetrik erleichtern [13]. In der neueren Studie von Chazette et al. [13] wird dieselbe Definition wie bei der älteren Studie von Chazette et al. [12] verwendet. Dadurch wird die Übernahme der Definition von Chazette et al. [12] in den Kontext dieser Arbeit bestätigt. Außerdem empfehlen Chazette et al. [13], dass folgende Aspekte erklärt werden: System im Allgemeinen, Argumentationsprozesse, innere Logik, Intention, Verhalten, Entscheidung, Leistung, Interna des Modells (z. B. Parameter) und Wissen über den Benutzer bzw. die Welt. Da durch die Betrachtung dieser Aspekte die Erhebung von Erklärbarkeitsanforderungen erleichtert wird, werden diese in der Arbeit berücksichtigt.

Da nur wenige Leitlinien bestehen, die die Definition und Operationalisierung von Erklärbarkeitsanforderungen vorgeben, wird in einer weiteren Studie von Chazette et al. [14] hierfür eine Leitlinie erstellt und anhand eines praktischen Beispiels auf Anwendbarkeit überprüft. Die Ergebnisse der Studie zeigen, dass die Leitlinie im industriellen Kontext verwendbar ist und die Erklärungen zu einer Verbesserung der Nutzungshäufigkeit, Systemakzeptanz und Nutzerzufriedenheit führen [14]. Auch kommt die Studie zum Ergebnis, dass die Einführung von Erklärungen immer zu bevorzugen ist als diese nicht einzuführen [14]. Daher lohnt es sich die Einführung von Erklärbarkeit in Stimmungsanalysetools im Rahmen dieser Arbeit zu untersuchen. Chazette et al. [14] zeigen, dass kontextbezogene Erklärungen zu einer

höheren Systemnutzung und Nutzerzufriedenheit führen und Erläuterungen über den Algorithmus die Systemakzeptanz erhöhen. Daher wird dies in der Arbeit bei der Erhebung von Erklärbarkeitsanforderungen berücksichtigt. Auch empfiehlt die Studie zu ermitteln, warum die Erklärbarkeit eingeführt, was dabei erklärt und wie sie dargestellt werden soll [14]. Die Fragen werden im Kontext dieser Arbeit beantwortet und fließen in die Erhebung der Erklärbarkeitsanforderung ein.

2.2.6 Praktische Umsetzung der Erklärbarkeit

Weil nur wenige Studien Praktiken zur Entwicklung von erklärbaren Systemen untersuchen, werden in der Studie von Chazette et al. [15] sechs Praktiken auf Grundlage einer Literaturrecherche und eines Interviews vorgestellt. Die Ergebnisse der Studie zeigen, dass die Praktiken realisierbar sind und in verschiedene Entwicklungsprozesse integriert werden können [15]. Chazette et al. [15] empfiehlt bei der Entwicklung von erklärbaren Systemen die Anwendung der sechs Praktiken Definition der Vision, Stakeholder-Analyse, Backend-Analyse, Trade-off-Analyse, Design der Erklärbarkeit und Bewertung. Die Praktiken werden in dieser Arbeit betrachtet, um den Designprozess von Erklärbarkeit zu erleichtern.

Um einen fehlenden systematischen Ansatz für das Design von Erklärbarkeit zu entwickeln, wird in der Studie von Köhl et al. [36] die Erhebung, Spezifikation und Verifikation von Erklärbarkeit betrachtet. Die Studie skizziert einen einheitlichen Zertifizierungsprozess für Erklärbarkeit mit passenden Entwicklungsmethoden [36]. Köhl et al. [36] empfiehlt die Messung des Verständnisses durch Tests. Daher werden in dieser Arbeit die umgesetzten Erklärungen im Rahmen der Nutzerstudie auf Verständnis überprüft. Auch kommt die Studie zum Ergebnis, dass die Auswahl der Entwicklungsmethode durch die Anforderungen beschränkt wird [36]. Nachdem in dieser Arbeit die erhobenen Erklärbarkeitsanforderungen erhoben wurden, werden auf dessen Grundlage und mithilfe der Studie von Köhl et al. [36] eine passende Entwicklungsmethode ausgesucht.

Die Studie von Rosenfeld und Richardson [53] schlägt eine Taxonomie für Erklärbarkeit im Kontext von Mensch-Agenten-Systemen vor, um die wichtigsten Fragen nach dem *Warum*, *Wer*, *Was*, *Wann* und *Wie* beantworten zu können. Daher wird in der Studie auf Grundlage einer Definition beschrieben, warum Erklärbarkeit notwendig ist, an wen sie gerichtet ist und welche Erklärungen entwickelt werden können [53]. Des Weiteren empfiehlt die Studie Maßnahmen zur Bewertung der Erklärbarkeit im ganzen System [53]. Im Gegensatz zur Studie von Chazette et al. [14] werden nicht nur die Fragen nach Inhalt *Warum*, Rahmenbedingungen *Was* und Bewertung der Erklärbarkeit *Wie*, sondern auch Adressaten *Wer* und Zeit *Wann* beantwortet [53]. Die ergänzenden Fragen werden im Kontext dieser Arbeit beantwortet und beeinflussen die Erhebung der

Erklärbarkeitsanforderungen.

Die Erklärbarkeit ist ein wichtiges Mittel geworden, um die Softwarequalität sicherzustellen [25]. Jedoch ist sich die Forschung unsicher, ob die Anforderungen an die Erklärbarkeit für jeden Nutzenden dieselben sind oder zwischen den Nutzergruppen variieren [25]. Da in bestehenden Studien nur bei bestimmten Systemen die Erklärbarkeitsanforderungen mithilfe von Personas erhoben werden, wird in der Studie von Droste et al. [25] das Potential von Personas im Allgemeinen untersucht. Die Ergebnisse der Studie zeigen, dass Personas geeignet sind im Allgemeinen den Erklärungsbedarf von Nutzenden abzuschätzen [25]. Da in dieser Arbeit der Erklärungsbedarf nicht abgeschätzt, sondern im Rahmen einer Bedarfsanalyse erhoben wird, sind Personas im Kontext der Arbeit ungeeignet. In der Studie von Droste et al. [25] wird die Erklärbarkeit in die Komponenten Kontext, Erklärer und Adressat unterteilt, wodurch die Studien von Rosenfeld und Richardson [53] und Balasubramaniam et al. [5] bestätigt werden.

Kapitel 3

Gestaltung der Literaturrecherche

Dieses Kapitel beschreibt die Gestaltung der Literaturrecherche. Auf Grundlage der Primärliteratur wird sowohl bezüglich Stimmungsanalysetools als auch Erklärbarkeit das Snowballing-Verfahren von Wohlin [62] durchgeführt.

3.1 Auswahlkriterien

Damit eine Studie im Rahmen der Literaturrecherche untersucht wird, muss diese alle Inklusionskriterien und darf keines der Exklusionskriterien erfüllen.

Inklusionskriterien

1. Die Studie untersucht ein für diese Arbeit relevantes Thema, wie z. B. Stimmungsanalysen oder Erklärbarkeit.
2. Die Studie ist frei zugänglich.

Damit diese Arbeit auf der Grundlage der Literatur durchgeführt werden kann, müssen die Studien ein für diese Arbeit relevantes Thema untersuchen. Um die Ergebnisse dieser Arbeit reproduzieren zu können, müssen die Studien frei zugänglich sein.

Exklusionskriterien

1. Die Studie wurde vor dem Jahr 2016 publiziert.
2. Die Studie wurde in einer anderen Sprache als Englisch oder Deutsch verfasst.

Damit der aktuelle Forschungsstand untersucht wird, müssen die Studien aktuell sein. Daher werden Studien nicht berücksichtigt, wenn diese vor dem Jahr 2016 publiziert wurden. Die Studien müssen in den Sprachen Deutsch oder Englisch verfasst wurden, um den Inhalt dieser zu verstehen.

3.2 Snowballing-Verfahren

In der Studie von Wohlin [62] werden einheitliche Richtlinien für die Durchführung von systematischen Literaturrecherchen vorgeschlagen. Die Richtlinien basieren auf gesammelten Erfahrungen bei der Durchführung von verschiedenen systematischen Literaturrecherchen [62]. Bei der Snowballing-Suche werden Vorwärts- und Rückwärtsiterationen durchgeführt, bis keine neuen Studien mehr gefunden werden [62]. Während der Iteration in die Vorwärtsrichtung werden alle Studien betrachtet, die die Primärstudie zitieren [62]. Dagegen werden bei der Iteration in die Rückwärtsrichtung alle Studien betrachtet, die von der Primärstudie zitiert werden [62]. Dies ist eine gute Alternative zur Datenbank-Suche, um über ein bereits gut untersuchtes Forschungsgebiet zu recherchieren [62]. Da sowohl Erklärbarkeit als auch Stimmungsanalysetools in der Forschung intensiv untersucht wurden, eignet sich in dieser Arbeit die Anwendung der Snowballing-Suche. Daher werden bei der Durchführung der Literraturrecherche die Richtlinien von Wohlin [62] angewendet. Jedoch wird in dieser Arbeit nur einmal in die Vorwärts- und Rückwärtsrichtung iteriert, da schon nach einer Iteration keine neuen Konzepte mehr gefunden wurden und dies nach Wolfswinkel et al. [63] ausreichend ist. Der Prozess der Literaturrecherche und der in dieser Arbeit noch folgenden Forschungsmethodik ist in der Abbildung 3.1 dargestellt.

Des Weiteren werden in der Studie von Lin [40] verschiedene Literaturrecherche-Techniken in unterschiedlichen Kontexten angewendet. Bei der Auswertung kommt die Studie zum Ergebnis, dass die Snowballing-Suche effizienter als die Datenbank-Suche ist [40]. Dies unterstützt die Entscheidung für die Snowballing-Suche.

3.3 Primärliteratur

Da Obaidi und Klünder [46] eine Metaanalyse mit 80 Studien durchführen, eignet sich die Studie für einen Einblick in die umfangreiche Forschung im Bereich der Stimmungsanalysetools. Daher wird für die Literaturrecherche bezüglich Stimmungsanalysetools die Snowballing-Suche von Wohlin [62] angewendet und für dessen Start die Studie von Obaidi und Klünder [46] als Primärstudie betrachtet.

Die Studie von Chazette et al. [12] entwickelt eine einheitliche Definition und ein Modell von Erklärbarkeit auf der Grundlage bestehender Studien in der Forschung. Auch werden Auswirkungen auf die Erklärbarkeit durch Qualitätsaspekte erhoben und in einem Katalog zusammengefasst [12]. Da in der Studie der aktuelle Forschungsstand durch eine Metaanalyse zusammengefasst wird [12], eignet sich die Studie, um einen Einblick in die Forschung im Bereich der Erklärbarkeit zu ermöglichen. Daher wird für die Literaturrecherche bezüglich Erklärbarkeit die Studie von Chazette et al. [12]

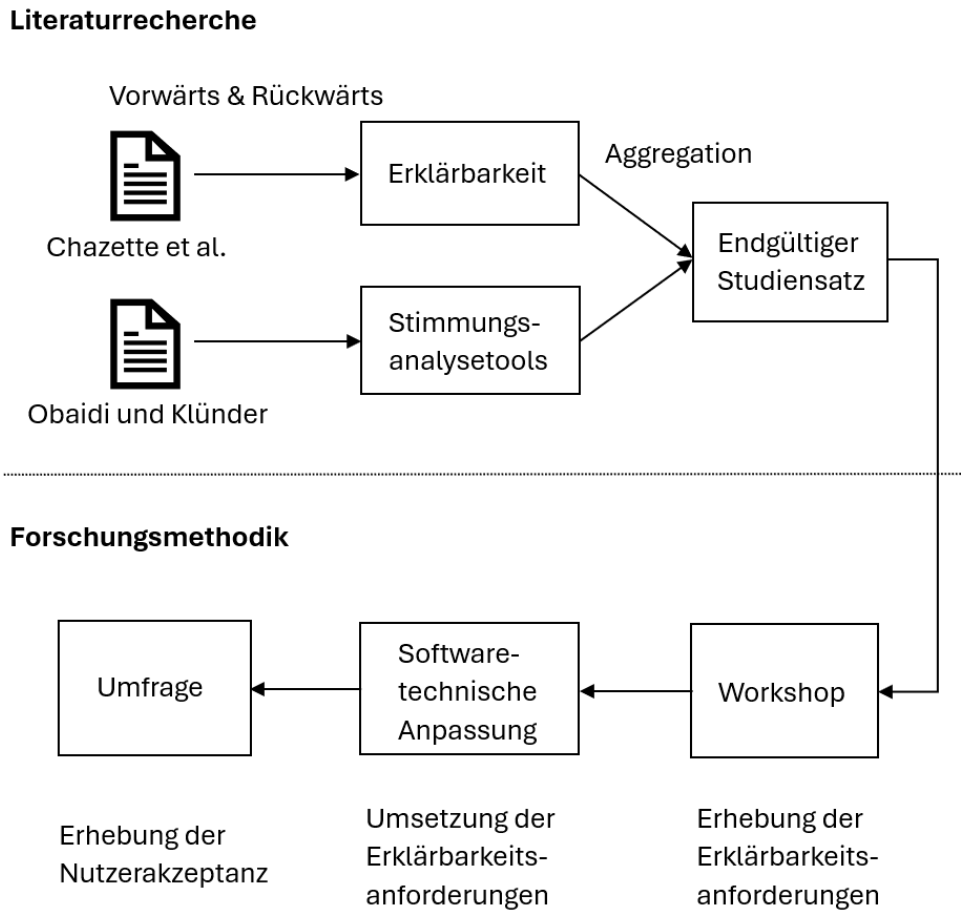

Abbildung 3.1: Prozess der Literaturrecherche und Forschungsmethodik

als Primärstudie für die Snowballing-Suche genutzt.

3.4 Ergebnisse des Snowballing-Verfahrens

Die Ergebnisse des Snowballing-Verfahrens bezüglich Stimmungsanalysetools sind der Tabelle 3.1 und Erklärbarkeit der Tabelle 3.2 zu entnehmen. Eine detaillierte Auflistung der Ergebnisse ist in dem Anhang A dargestellt. In den Tabellen 3.1 und 3.2 wird für die jeweilige Iterationsrichtung die Anzahl der Studien angegeben, die insgesamt untersucht wurden bzw. für diese Arbeit relevant sind. Auch ist die Anzahl der untersuchten Studien zu entnehmen, die Stimmungsanalysen und maschinelles Lernen bzw. Erklärbarkeit und Erklärbarkeitsmethoden thematisieren.

Richtung	**Insgesamt**	**Relevant**	**Stimmungs-analyse**	**Maschinelles Lernen**
Vorwärts	28	23	20	11
Rückwärts	18	14	11	12

Tabelle 3.1: Ergebnisse bezüglich Stimmungsanalysetools

Während der Vorwärtssuche wurden insgesamt 28 Studien untersucht, von denen 23 für diese Arbeit relevant sind. Dagegen wurden bei der Rückwärtssuche 18 Studien gesichtet, wobei von denen 14 relevant sind. Da sowohl während der Vorwärts- als auch Rückwärtssuche die meisten der untersuchten Studien für diese Arbeit relevant sind, wird die Annahme bestätigt die Studie von Obaidi und Klünder [46] als prädestinierte Startstudie im Bereich der Stimmungsanalysetools zu betrachten.

Richtung	**Insgesamt**	**Relevant**	**Erklärbar-keit**	**Erklärbarkeits-methoden**
Vorwärts	20	17	18	9
Rückwärts	15	12	10	6

Tabelle 3.2: Ergebnisse bezüglich Erklärbarkeit

Innerhalb der Literaturrecherche im Bereich der Erklärbarkeit wurden insgesamt 20 Studien bei der Vorwärts- und 15 Studien bei der Rückwärtssuche gesichtet. Von denen sind 17 bei der Vorwärts- und 12 bei der Rückwärtssuche relevant. Da ebenfalls im Bereich der Erklärbarkeit die meisten der untersuchten Studien für diese Arbeit relevant sind, bestätigt sich die Annahme die Studie von Chazette et al. [12] als prädestinierte Startstudie im Bereich der Erklärbarkeit zu betrachten.

Kapitel 4

Workshop

In diesem Kapitel werden die Forschungsfragen aus der Forschungslücke abgeleitet. Um die Forschungsfragen in dieser Arbeit beantworten zu können, wird eine geeignete Forschungsmethodik beschrieben. Hierzu wird u. a. der Workshop als eine geeignete Forschungsmethode und *RoBERTa* als ein geeignetes Stimmungsanalysetool betrachtet. Danach werden Durchführung und Ergebnisse des Workshops dargestellt.

4.1 Forschungsfragen

FF1: Erklärbarkeitsanforderungen In der Literaturrecherche wurden nur Studien gefunden, die Erklärbarkeit [12] und Stimmungsanalysetools [46] getrennt voneinander betrachten. Um trotzdem die Einführung von Erklärbarkeit bei Stimmungsanalysetools realisieren zu können, werden die Anforderungen an die Erklärbarkeit erhoben. Da der Workshop eine geeignete Forschungsmethode zur Erhebung der Anforderungen darstellt, wird dieser in der Arbeit eingesetzt. Die Anforderungen an die Erklärbarkeit werden am Beispiel des Stimmungsanalysetools *RoBERTa* erhoben, da dies das weitverbreitetste und effizienteste Tool ist [44] [47] [21]. Da die Erhebung der Anforderungen durch die Rahmenbedingungen beeinflusst wird, ist die erste Forschungsfrage wie folgt definiert:

FF1: Welche Anforderungen an die Erklärbarkeit lassen sich bei der Nutzung des Stimmungsanalysetools *RoBERTa* im Rahmen eines Workshops erheben?

FF2: Einfluss der Erklärbarkeit auf die Nutzerakzeptanz Durch die Einführung von Erklärbarkeit kann die Transparenz und damit auch die Benutzerfreundlichkeit erhöht werden [43]. Daher wird in dieser Arbeit untersucht, ob die Nutzerakzeptanz des Stimmungsanalysetools durch die Einführung erhöht wird. Somit könnten ggf. Nutzeffekte durch eine erhöhte

Nutzung des Stimmungsanalysetools entstehen. Da die Umsetzung der Erklärbarkeit bei dem Stimmungsanalysetool *RoBERTa* durch die ermittelten Anforderungen zu realisieren ist, wird der Einfluss auf die Nutzerakzeptanz am Beispiel des Stimmungsanalysetools *RoBERTa* untersucht. Aufgrund der Rahmenbedingungen ist die zweite Forschungsfrage wie folgt definiert:

FF2: Welchen Einfluss hat die Umsetzung der Erklärbarkeitsanforderungen auf die Nutzerakzeptanz des Stimmungsanalysetools *RoBERTa*?

FF3: Wichtigkeit der nichtfunktionalen Anforderungen Die Einführung von Erklärbarkeit kann die Softwarequalität sowohl positiv als auch negativ beeinflussen [19]. Um ggf. eine negative Beeinflussung der Softwarequalität bestimmen zu können, werden in dieser Arbeit die Zielkonflikte untersucht. Mithilfe der Konfliktkataloge von Chazette et al. [12] und Mairiza und Zowghi [42] wird ersichtlich, inwiefern Erklärbarkeit andere nichtfunktionale Anforderungen beeinflusst. Um die Konsequenzen der Beeinflussung abschätzen zu können, werden die nichtfunktionalen Anforderungen von Chazette et al. [12] im Rahmen eines Workshops auf Wichtigkeit überprüft. Da die Erklärbarkeit im Stimmungsanalysetool *RoBERTa* umgesetzt wird, werden die Zielkonflikte in dessen Kontext untersucht. Daher wird die dritte Forschungsfrage wie folgt definiert:

FF3: Wie wichtig werden die nichtfunktionalen Anforderungen des Stimmungsanalysetools *RoBERTa* im Rahmen eines Workshops bewertet?

4.2 Forschungsmethodik

Für die Erhebung von Erklärbarkeitsanforderungen eignet sich die Durchführung eines Workshops, da Erklärbarkeitsanforderungen sehr umfangreich sein können und im Workshop viele unterschiedliche Ideen gesammelt werden. Auch eignet sich der Workshop eher als zum Beispiel eine Umfrage, da durch Gruppenarbeit effektiver gearbeitet wird, als wenn nur eine Person beteiligt ist. Jedoch ist die Durchführung eines Workshops für die Teilnehmenden deutlich zeitintensiver als zum Beispiel bei einer Umfrage. Insgesamt ist der Workshop für die Erhebung von Erklärbarkeitsanforderungen prädestiniert, da die Vorteile überwiegen.

Die Zielsetzung des Workshops ist die Erhebung von Anforderungen bei einer möglichen Einführung von Erklärbarkeit im Stimmungsanalysetool *RoBERTa*. Des Weiteren sollen die in der Literaturrecherche ermittelten nichtfunktionalen Anforderungen von den Teilnehmenden im Kontext des Stimmungsanalysetools *RoBERTa* auf Gültigkeit überprüft und priorisiert werden.

Da im Workshop Erklärbarkeitsanforderungen im Kontext der Softwareentwicklung erhoben werden, müssen bei den Teilnehmenden Kenntnisse und Erfahrungen in der Softwareentwicklung vorliegen. Weil die Stimmung im ganzen Team für die Leistung relevant ist, wird der Workshop mit Teilnehmenden unabhängig von ihrer Berufserfahrung durchgeführt.

Die zuvor durchgeführte Literaturrecherche kann zu einer voreingenommenen Moderation des Workshops führen. Daher ist während der Moderation auf ein neutralen Still zu achten.

Ablauf

Um das Interesse der Teilnehmenden zu fördern, müssen am Anfang des Workshops Ziel und Ablauf verständlich kommuniziert werden. Auch sollte die Motivation des Themas und dessen wissenschaftliche Einordnung erläutert werden.

Damit die Teilnehmenden einen ersten Eindruck in das Stimmungsanalysetool erhalten, wird anschließend die Anwendung des Tools durch den Moderator anhand eines Beispiels vorgestellt. Um ein tiefergehendes Verständnis bei den Teilnehmenden zu erzielen, werden die Teilnehmenden in Kleingruppen das Tool durch eigene Beispielsätze testen. Hierbei soll auch die Stimmung von ironisch formulierten Sätzen analysiert werden, da Stimmungsanalysetools meistens Schwierigkeiten haben, Ironie zu detektieren.

Um Missverständnisse bei den Teilnehmenden zu vermeiden, wird danach durch den Moderator eine einheitliche Definition des Begriffs *Erklärbarkeit* vorgegeben und anhand eines Beispiels verdeutlicht.

Im nächsten Schritt diskutieren die Teilnehmenden in Kleingruppen, inwiefern Erklärbarkeit im Stimmungsanalysetool zu gestalten ist, um einen Mehrwert bei der Nutzung des Tools zu erzielen. Anschließend werden die Ergebnisse der Kleingruppen im Plenum zusammengefasst, diskutiert und dokumentiert.

Damit die nichtfunktionalen Anforderungen und Erklärbarkeitsanforderungen voneinander thematisch abgegrenzt werden, wird danach der Workshop für eine kurze Zeit pausiert. Außerdem soll nach den zeitintensiven Aufgaben somit die Konzentration der Teilnehmenden aufrechterhalten werden.

Um die nichtfunktionalen Anforderungen von Chazette et al. [12] im Kontext des Stimmungsanalysetools auf Wichtigkeit zu überprüfen, werden anschließend diese in den Kleingruppen priorisiert. Die Ergebnisse der Kleingruppen werden danach wieder im Plenum zusammengefasst, diskutiert und dokumentiert.

Da in der Studie von Chazette et al. [12] insgesamt 57 nichtfunktionale Anforderungen erhoben werden und dessen Priorisierung durch den zeitlichen Rahmen des Workshops nicht möglich ist, werden bei der Priorisierung nur die 18 auf den nutzerbezogenen nichtfunktionalen Anforderungen berück-

sichtigt. Diese sind für den Workshop prädestiniert, da diese direkt von den Nutzenden abhängig sind und im Workshop vor allem die Nutzersicht des Stimmungsanalysetools betrachtet wird.

Daher werden nur folgende Anforderungen berücksichtigt: Benutzererfahrung, Genauigkeit des mentalen Modells, wahrgenommene Nützlichkeit, wahrgenommener Wert, Bewusstsein für Datenschutz, Benutzerbewusstsein, Benutzerzufriedenheit, Benutzerfreundlichkeit, Mensch-Maschine Kooperation, Benutzerkontrolle, Benutzereffektivität, Benutzereffizienz, Benutzerleistung, Leitfaden, Entdeckung von Wissen, Lernfreundlichkeit, Überprüfbarkeit und Unterstützung der Entscheidungsfindung.

Zwecks Beantwortung offener inhaltlicher Fragen und Einholung von Verbesserungsvorschlägen bezüglich der Moderation, wird am Ende das Feedback der Teilnehmenden eingeholt. Der zeitliche Aufwand des Ablaufs wird geschätzt und ist der Tabelle 4.1 zu entnehmen.

Aufgabe	**Zeit (in Minuten)**
Einführung	5
Vorstellung des Stimmungsanalysetools	5
Definition des Begriffs *Erklärbarkeit*	5
Erhebung von Erklärbarkeitsanforderungen	40
Pause	15
Priorisierung der nichtfunktionalen Anforderungen	20
Feedback	10
Puffer	5
Gesamtdauer: 1 Stunde und 45 Minuten	

Tabelle 4.1: Zeitplan des Workshops

4.3 Durchführung

Die elf Teilnehmenden des Workshops schreiben oder haben bereits eine Abschlussarbeit im Studiengang *Informatik* geschrieben und haben somit das nötige Vorwissen im Bereich der Softwareentwicklung. Die Rohdaten des Workshops sind dem Anhang B zu entnehmen.

4.4 Ergebnisse

Erklärbarkeitsanforderungen
Von den Teilnehmenden wurden Erklärbarkeitsanforderungen im Kontext des Stimmungsanalysetools *RoBERTa* erhoben und priorisiert. Die Ergebnisse sind der Tabelle 4.2 zu entnehmen.

Priorität	**Erklärbarkeitsanforderungen**
Wichtig	Genauigkeit der Ergebnisse, Grenzen der Software (z. B. wenn etwas nicht berechenbar ist), Verteilung der Wahrscheinlichkeiten bei der Kategorisierung
Eher wichtig	Keywords (welche Wörter führen zum Ergebnis, Einfluss von Satzzeichen und Smileys, positive/negative Wörter grün/rot markieren), Kontext (mit Beispielsätzen das Gefühl bekommen, wie die Software das Ergebnis berechnet)
Neutral	Größe des Trainingskorpus, Schritt für Schritt die Berechnung erklären
Eher unwichtig	Bewertung der Software (wenn Ergebnis vom eigenen Empfinden abweicht)
Unwichtig	-

Tabelle 4.2: Priorisierte Erklärbarkeitsanforderungen

Eine von den Teilnehmenden als wichtig bewertete Erklärbarkeitsanforderung ist die *Genauigkeit der Ergebnisse*. Hierdurch soll den Nutzenden des Stimmungsanalysetools erklärt werden, mit welcher Genauigkeit das Modell vortrainiert wurde.

Die Erklärbarkeitsanforderung *Grenzen der Software* wird von den Teilnehmenden des Workshops ebenso als wichtig betrachtet. Durch eine Sensibilisierung über die Grenzen der Software soll u. a. ersichtlich werden, welche Eingaben nicht berechenbar sind.

Im Rahmen des Workshops wurde die Erklärbarkeitsanforderung *Wahrscheinlichkeiten bei der Kategorisierung* ebenfalls als wichtig bewertet. Das Stimmungsanalysetool soll nicht nur die vorliegende Stimmung angeben, sondern auch mit welcher Wahrscheinlichkeit eine positive, neutrale oder negative Stimmung vorliegt.

Die Teilnehmenden des Workshops betrachten die Erklärbarkeitsanforderung *Schlüsselwörter* als eher wichtig. Durch die farbliche Markierung von positiven und negativen Wörtern soll ersichtlich werden, welche Wörter zum Ergebnis geführt haben.

Ebenso wird die Erklärbarkeitsanforderung *Kontext* von den Teilnehmenden als eher wichtig bewertet. Mithilfe von Beispielen soll den Nutzenden veranschaulicht werden, in welchem Kontext das Tool nutzbar ist.

Die Erklärbarkeitsanforderung *Größe des Trainingskorpus* wird von den Teilnehmenden als neutral betrachtet. Damit die Nutzenden die Validität des Modells überprüfen können, soll die Größe des Trainingsdatensatzes angegeben werden.

Im Rahmen des Workshops wird die Erklärbarkeitsanforderung *Schritt*

für Schritt die Berechnung erklären ebenfalls als neutral bewertet. Durch die schrittweise Berechnung soll die Funktionsweise des Tools erklärt werden.

Eine von den Teilnehmenden als eher unwichtig bewertete Erklärbarkeitsanforderung ist die *Bewertung der Software*. Wenn das eigene Empfinden vom Ergebnis abweicht, sollen die Nutzenden die Möglichkeit haben die Software zu bewerten.

Nichtfunktionale Anforderungen
Auch wurden im Workshop die wichtigsten nichtfunktionalen Anforderungen von Chazette et al. [12] im Kontext des Stimmungsanalysetools *RoBERTa* priorisiert. Diese sind der Tabelle 4.3 zu entnehmen.

Priorität	**Nichtfunktionale Anforderungen**
Wichtig	Überprüfbarkeit, Benutzereffektivität
Eher wichtig	Lernfreundlichkeit, Benutzereffizienz, Benutzerbewusstsein, Benutzerzufriedenheit, Bewusstsein für Datenschutz, wahrgenommener Wert, wahrgenommene Nützlichkeit, Unterstützung der Entscheidungsfindung
Neutral	Leitfaden, Mensch-Maschine Kooperation, Benutzerkontrolle, Benutzerfreundlichkeit
Eher unwichtig	Benutzererlebnis, Genauigkeit des mentalen Modells
Unwichtig	Entdeckung von Wissen, Benutzerleistung

Tabelle 4.3: Priorisierte nichtfunktionale Anforderungen

Die nichtfunktionale Anforderung *Überprüfbarkeit* wird von Chazette et al. [12] beschrieben als die Fähigkeit das Überprüfen des Systems zu ermöglichen. Neben der *Überprüfbarkeit* wird auch die *Benutzereffektivität* von den Teilnehmenden als wichtig bewertet.

Chazette et al. [12] definieren die *Lernfreundlichkeit* als die Fähigkeit den Nutzenden die Funktionsweise des Systems zu erläutern. Diese Anforderung wird im Rahmen des Stimmungsanalysetools *RoBERTa* als eher wichtig bewertet.

In der Studie von Chazette et al. [12] werden die nichtfunktionalen Anforderungen *wahrgenommener Wert* und *wahrgenommene Nützlichkeit* eines Systems untersucht. Beide Anforderungen werden im Kontext des Stimmungsanalysetools als eher wichtig betrachtet.

Ebenso beschreiben Chazette et al. [12] die *Unterstützung der Entscheidungsfindung* als eine nichtfunktionale Anforderung. Diese wird von den Teilnehmenden als eher wichtig bewertet.

Die nichtfunktionale Anforderung *Bewusstsein für Datenschutz* wird von Chazette et al. [12] definiert als das Sensibilisieren der Nutzenden bezüglich des Datenschutzes. Neben dieser Anforderung werden auch die Anforderungen *Benutzereffizienz*, *-bewusstsein* und *-zufriedenheit* im Rahmen des Workshops als eher wichtig beurteilt.

Chazette et al. [12] beschreiben die nichtfunktionale Anforderung *Leitfaden* als das Bereitstellen von Informationen über mögliche Lösungsansätze. Diese Anforderung wird im Kontext des Stimmungsanalysetools *RoBERTa* als neutral bewertet.

Auch definieren Chazette et al. [12] die nichtfunktionale Anforderung *Mensch-Maschine Kooperation* als die Interaktion von Mensch und Maschine. Sowohl diese Anforderung als auch die Anforderungen *Benutzerkontrolle* und *-freundlichkeit* werden von den Teilnehmenden als neutral beurteilt.

In der Studie von Chazette et al. [12] wird die nichtfunktionale Anforderung *Genauigkeit des mentalen Modells* beschrieben als die Übereinstimmung der von den Nutzenden subjektiv interpretierten und der tatsächlichen Funktionsweise der Software. Neben dieser wird auch die Anforderung *Benutzererlebnis* im Kontext des Stimmungsanalysetools als eher unwichtig betrachtet.

Chazette et al. [12] untersuchen die nichtfunktionalen Anforderungen *Entdecken von Wissen* und *Benutzerleistung*. Beide Anforderungen werden im Rahmen des Workshops als unwichtig bewertet.

Kapitel 5

Auswirkungen auf verwandte NFAs

In diesem Kapitel werden die Auswirkungen von Erklärbarkeit auf andere nichtfunktionale Anforderungen (NFAs) im Kontext des Stimmungsanalysetools abgeschätzt. Um diese Beeinflussungen abschätzen zu können, werden u. a. die im Workshop priorisierten nichtfunktionalen Anforderungen und der Konfliktkatalog von Chazette et al. [12] berücksichtigt. Die Ergebnisse der Abschätzung sind der Tabelle 5.1 zu entnehmen.

5.1 Abschätzung der Beeinflussungen

Sehr hohe Priorisierung
Sowohl die *Überprüfbarkeit* als auch *Benutzereffektivität* wurde bei der Nutzung des Stimmungsanalysetools als sehr wichtig bewertet. Chazette et al. [12] zeigen, dass durch die Erklärbarkeit die *Überprüfbarkeit* positiv und die *Benutzereffektivität* sowohl positiv als auch negativ beeinflusst wird.

Hohe Priorisierung
Die Studie von Chazette et al. [12] zeigt, dass durch die Erklärbarkeit die im Workshop als wichtig bewerteten Anforderungen *Lernfreundlichkeit*, *Benutzerbewusstsein*, *Benutzerzufriedenheit*, *Bewusstsein für Datenschutz*, *wahrgenommener Wert*, *wahrgenommene Nützlichkeit* und *Unterstützung der Entscheidungsfindung* positiv beeinflusst werden. Auch wird durch die Studie von Chazette et al. [12] ersichtlich, dass durch die Erklärbarkeit die ebenfalls wichtig bewertete Anforderung *Benutzereffizienz* sowohl positiv als auch negativ beeinflusst wird.

Neutrale Priorisierung
Im Rahmen des Workshops wurden die Anforderungen *Leitfaden*, *Mensch-Maschine Kooperation*, *Benutzerkontrolle* und *Benutzerfreundlichkeit* in

Wichtigkeit	**Nichtfunktionale Anforderung**	**Beeinflussung durch Erklärbarkeit [12]**
Wichtig	Überprüfbarkeit	positiv
	Benutzereffektivität	positiv und negativ
Eher wichtig	Lernfreundlichkeit	positiv
	Benutzereffizienz	positiv und negativ
	Benutzerbewusstsein	positiv
	Benutzerzufriedenheit	positiv
	Bewusstsein für Datenschutz	positiv
	wahrgenommener Wert	positiv
	wahrgenommene Nützlichkeit	positiv
	Unterstützung der Entscheidungsfindung	positiv
Neutral	Leitfaden	positiv
	Mensch-Maschine Kooperation	positiv
	Benutzerkontrolle	positiv
	Benutzerfreundlichkeit	positiv und negativ
Eher unwichtig	Benutzererlebnis	positiv und negativ
	Genauigkeit des mentalen Modells	positiv
Unwichtig	Entdeckung von Wissen	positiv
	Benutzerleistung	positiv

Tabelle 5.1: Abschätzung der Auswirkungen von Erklärbarkeit auf andere nichtfunktionale Anforderungen

Bezug auf die Wichtigkeit als neutral bewertet. Mithilfe des Konfliktkatalogs von Chazette et al. [12] wird erkennbar, dass durch Erklärbarkeit ein positiver Effekt auf die Anforderungen *Leitfaden*, *Mensch-Maschine Kooperation* und *Benutzerkontrolle* bzw. ein sowohl positiver als auch negativer Effekt auf die *Benutzerfreundlichkeit* vorliegt.

Niedrige Priorisierung
Sowohl das *Benutzererlebnis* als auch die *Genauigkeit des mentalen Modells* wurde im Kontext des Stimmungsanalysetools als unwichtig bewertet. Chazette et al. [12] zeigen, dass die Erklärbarkeit die *Genauigkeit des mentalen Modells* positiv und das *Benutzererlebnis* sowohl positiv als auch negativ beeinflusst.

Sehr niedrige Priorisierung
Die nichtfunktionalen Anforderungen *Entdeckung von Wissen* und *Benutzerleistung* wurden innerhalb des Workshops als sehr unwichtig bewertet. Mithilfe des Konfliktkatalogs von Chazette et al. [12] wird ersichtlich, dass die Erklärbarkeit sowohl die *Entdeckung von Wissen* als auch das *Benutzererlebnis* positiv beeinflusst.

5.2 Übertragung in den Kontext dieser Arbeit

Bei der Erhebung der Wichtigkeit werden nur 18 nutzerbezogene und nicht alle 57 nichtfunktionale Anforderungen von Chazette et al. [12] berücksichtigt. Zwar lassen sich die Beeinflussungen von Erklärbarkeit nur auf die in dieser Arbeit berücksichtigten Anforderungen untersuchen, jedoch zeigen sich bei 15 von diesen positive und bei drei sowohl positive als auch negative Effekte. Da die Potenziale von Erklärbarkeit ersichtlich werden, wird der Nutzen dieser Arbeit bestätigt.

Kapitel 6

Softwaretechnische Umsetzung

Um die Nutzerakzeptanz der Erklärungen im Rahmen der Umfrage erheben zu können, müssen diese den Teilnehmenden visuell dargestellt werden. Daher beschreibt dieses Kapitel die softwaretechnische bzw. prototypische Umsetzung der im Workshop erhobenen Erklärbarkeitsanforderungen. Neben der Modifikation wird auch die Implementierung des Stimmungsanalysetools *RoBERTa* erläutert.

6.1 Stimmungsanalysetool RoBERTa

RoBERTa wird als ein prädestiniertes Tool betrachtet, da dies das effizienteste und weitverbreitetste Stimmungsanalysetool ist [44] [47] [21]. Somit könnten die Ergebnisse dieser Arbeit größtenteils auf die Industrie übertragen werden. Daher wurde bei der Erhebung der Erklärbarkeitsanforderungen und wird bei der in dieser Arbeit noch folgenden softwaretechnischen Umsetzung und Erhebung der Nutzerakzeptanz das Stimmungsanalysetool *RoBERTa* betrachtet.

Mithilfe der Entwicklungsumgebung *JupyterLab* [35] und der *GitHub*-Dokumentation von *RoBERTa* [41] wird das Stimmungsanalysetool implementiert und mit einem Datensatz aus der Forschung vortrainiert. In diesem Datensatz von Calefato et al. [9] wurden über 4000 Beiträge aus *Stack Overflow* hinsichtlich der vorliegenden Polarität kategorisiert. Somit ist dieser für die Stimmungsanalyse im Software-Engineering prädestiniert. Um einen Einblick in die Entwicklungsumgebung *JupyterLab* zu ermöglichen, wird diese in der Abbildung 6.1 dargestellt.

Damit die Teilnehmenden des Workshops die Möglichkeit haben eigene Beispielsätze auf die vorliegende Stimmung zu analysieren, wird im *JupytherLab*-Dokument neben der Implementierung von *RoBERTa* eine grafische Oberfläche implementiert. Auch werden in diesem Dokument die Erklärungen softwaretechnisch umgesetzt, um im Rahmen einer Online-Umfrage die Nutzerakzeptanz mithilfe von Screenshots erheben zu können.

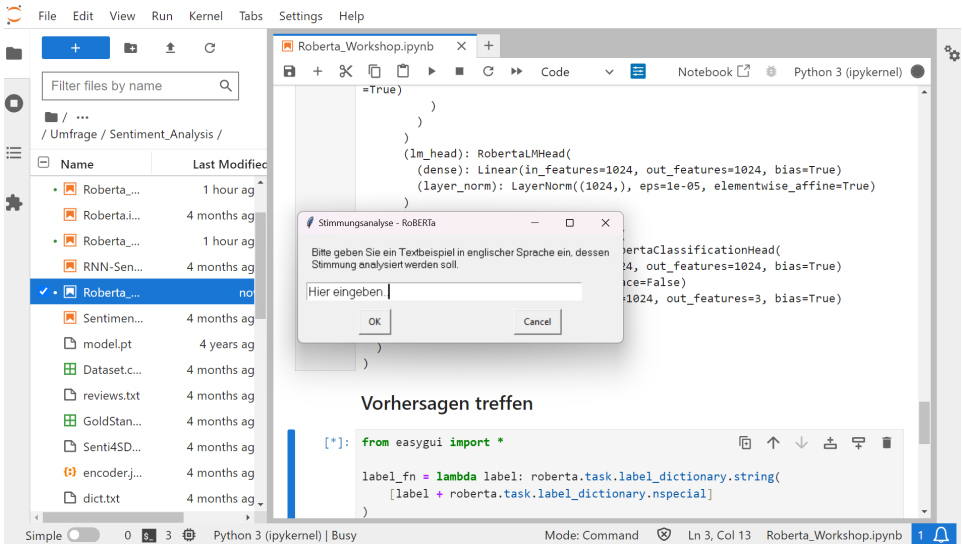

Abbildung 6.1: Entwicklungsumgebung *JupyterLab*

6.2 Umgesetzte Anforderungen

Bei der Umsetzung der sehr hoch priorisierten Erklärbarkeitsanforderung *Grenzen der Software* wird das Erkennen von nicht berechenbaren Eingaben aufgrund der hohen Komplexität der Implementierung nicht weiter betrachtet. Um diese Erklärbarkeitsanforderung trotzdem erfüllen zu können, wird bei der Eingabe im Tool auf dessen ausschließliche Anwendbarkeit in der Softwareentwicklung und das Nichterkennen von Ironie und Sarkasmus hingewiesen. Viele Studien zeigen, dass Stimmungsanalysetools Schwierigkeiten bei der Erkennung von Ironie und Sarkasmus haben. Die umgesetzte Erklärbarkeitsanforderung ist in der Abbildung 6.2 dargestellt.

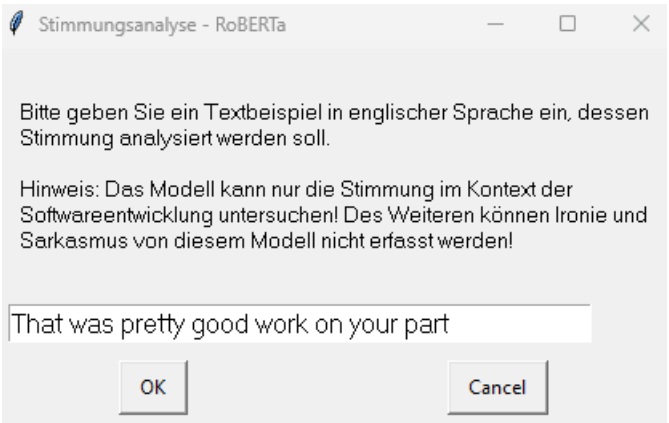

Abbildung 6.2: Hinweisen auf die Grenzen des Tools

Die ebenfalls sehr hoch priorisierte Erklärbarkeitsanforderung *Genauigkeit der Ergebnisse* lässt sich durch die triviale Ausgabe der Genauigkeit beim Tainieren des Modells umsetzen.

Um die hoch priorisierte Erklärbarkeitsanforderung *Keywords* erfüllen zu können, müssen für die Stimmung positive Wörter grün und negative Wörter rot dargestellt werden. Aufgrund des hohen Implementationsaufwands durch die notwendige Einführung einer *class activation map*-Methode, wird der Ansatz nicht weiter betrachtet. Um trotzdem in der Nutzerstudie die Bedeutung der Erklärbarkeitsanforderung untersuchen zu können, wird diese prototypisch umgesetzt und ist der Abbildung 6.3 zu entnehmen.

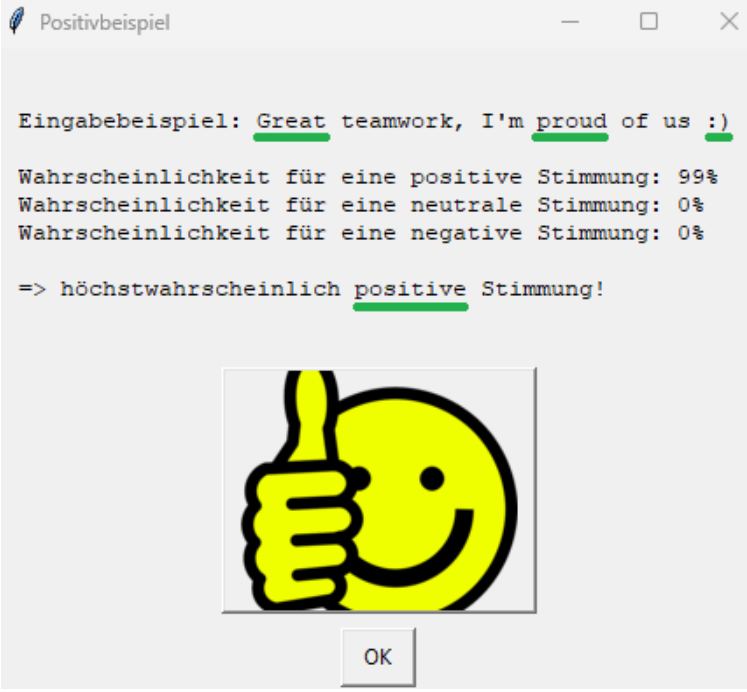

Abbildung 6.3: Identifikation und Markierung von Schlüsselwörtern

Damit die hoch priorisierte Erklärbarkeitsanforderung *Verteilung der Wahrscheinlichkeiten bei der Kategorisierung* erfüllt ist, wird bei der Modifikation des Tools nicht nur die Stimmung in negativ, neutral und positiv kategorisiert, sondern auch deren Wahrscheinlichkeiten berechnet und ausgegeben.

Bei der Umsetzung der ebenfalls hoch priorisierten Erklärbarkeitsanforderung *Kontext*, wird vor Eingabe des auf Stimmung zu analysierenden Textes jeweils ein Positiv-, Neutral- und Negativbeispiel vorgestellt. Dies soll den Nutzenden der Software helfen den Kontext der Software besser zu verstehen. Die Implementierung stellt sich als trivial aber umfangreich dar.

Um die mittelhoch priorisierte Erklärbarkeitsanforderung *Größe des*

Trainingskorpus umzusetzen, wird die Größe des Trainingsdatensatzes berechnet und ausgegeben. Die Größe des Trainingsdatensatzes wird neben der Genauigkeit angegeben, um den Nutzenden die Überprüfbarkeit der Validität des Modells zu ermöglichen. Die Implementierung ist wie bei der Umsetzung der Erklärbarkeitsanforderung *Genauigkeit der Ergebnisse* trivial. Die beiden umgesetzten Erklärbarkeitsanforderungen sind in der Abbildung 6.4 dargestellt.

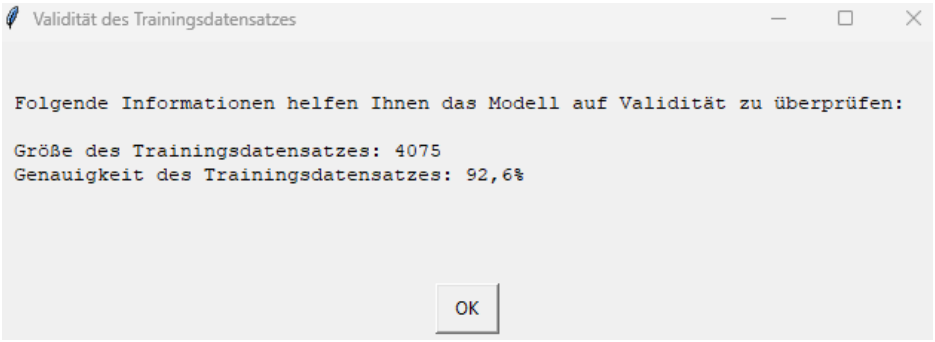

Abbildung 6.4: Überprüfbarkeit der Validität des Modells

Bei der Umsetzung der ebenfalls mittelhoch priorisierten Erklärbarkeitsanforderung *Schritt für Schritt die Berechnung erklären* wird nach Ausgabe der Ergebnisse dem Nutzenden die Möglichkeit gegeben, Informationen über die Funktionsweise des Tools zu erhalten. Hier wird darauf hingewiesen, dass die Berechnung des Ergebnisses aufgrund des Vorliegens einer nicht deterministischen künstlichen Intelligenz nicht ersichtlich ist. Da das Modell mit einem Datensatz aus der Softwareentwicklung vortrainiert wurde, wird auch auf die Limitierung des Modells auf den Kontext der Softwareentwicklung hingewiesen. Die umgesetzte Erklärung ist der Abbildung 6.5 zu entnehmen.

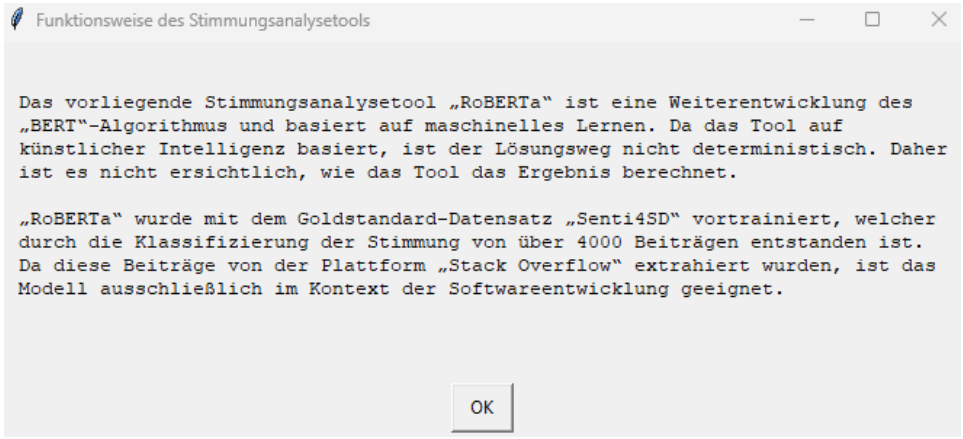

Abbildung 6.5: Beschreibung der Funktionsweise des Tools

Die Abbildungen der in diesem Kapitel nicht dargestellten Erklärungen sind dem Anhang C zu entnehmen.

6.3 Ausgesparte Anforderung

Um die Erklärbarkeitsanforderung *Bewertung der Software* umsetzen zu können, muss der Nutzende die Möglichkeit haben, die Software zu bewerten, wenn das eigene Empfinden von dem Ergebnis des Tools abweicht. Aufgrund der niedrigen Priorität und dem hohen Implementationsaufwand wird die Erklärbarkeitsanforderung ausgeschlossen.

Kapitel 7

Umfrage

Die Umfrage wird als geeignete Forschungsmethode betrachtet, um einen möglichen Einfluss auf die Nutzerakzeptanz durch die umgesetzten Erklärbarkeitsanforderungen feststellen zu können. Daher werden in diesem Kapitel Forschungsmethodik und Ergebnisse der Umfrage beschrieben. Die Ergebnisse werden u. a. mithilfe von statistischen Tests ausgewertet und nachfolgend dokumentiert.

7.1 Forschungsmethodik

Nachdem die Erklärbarkeit im Stimmungsanalysetool softwaretechnisch umgesetzt wurde, können die Auswirkungen der Erklärbarkeit auf die Nutzerakzeptanz untersucht werden. Um dies im Rahmen einer Nutzerstudie untersuchen zu können, wird eine Online-Umfrage durchgeführt. Um zuverlässig auf die Grundgesamtheit schließen zu können, ist eine hohe Stichprobengröße nötig. Da die Online-Umfrage für die Teilnehmenden einfacher durchzuführen ist als z. B. bei einem Workshop, ist die Online-Umfrage für das Erreichen einer hohen Stichprobengröße prädestiniert.

Auch ist die Auswertung der Ergebnisse bei einer Online-Umfrage durch einen standardisierten Fragebogen einfacher zu realisieren als z. B. bei einem Workshop, da bei einem Workshop die Ergebnisse gegebenenfalls noch zusammengefasst werden müssen. Des Weiteren wird durch eine Online-Umfrage die Anonymität der Teilnehmenden gewährleistet, wodurch ehrliche Antworten wahrscheinlicher werden. Jedoch könnten somit die Teilnehmenden diese mehrmals durchführen. Da die Vorteile überwiegen, ist die Online-Umfrage für die Erhebung der Nutzerakzeptanz prädestiniert. Somit wird eine Online-Umfrage durchgeführt und der vollständige Fragebogen der Umfrage ist dem Anhang D zu entnehmen.

Wilcoxon-Test
Der Test überprüft die Gleichheit der zentralen Tendenzen von zwei gepaarten Stichproben [60]. Unter Berücksichtigung der Rangsummen wird die Richtung und Höhe der Differenzen berechnet [60]. Da im Gegensatz zum t-Test keine intervallskalierten Daten vorausgesetzt werden, ist dieser für die Arbeit prädestiniert und wird nachfolgend verwendet. Für die Durchführung des Wilcoxon-Tests müssen folgende Voraussetzungen erfüllt sein [60]:

1. Abhängigkeit der Messungen [60]
2. Die unabhängige Variable ist nominalskaliert und hat zwei Ausprägungen [60]
3. Die abhängige Variable ist mindestens ordinalskaliert [60]
4. Die Verteilungsform der Differenzen ist symmetrisch [60]

Erfüllung der Voraussetzungen:

1. Die Messungen sind abhängig, da jede/-r Teilnehmer/-in das Stimmungsanalysetool sowohl ohne als auch mit den Erklärungen bewertet.
2. Die unabhängige Variable ist ordinalskaliert und damit auch nominalskaliert, da zwischen den Ausprägungen der Zustimmung eine Rangordnung existiert.
3. Die abhängige Variable ist ordinalskaliert, da zwischen den Ausprägungen der Zustimmung eine Rangordnung existiert.
4. Aufgrund der Stichprobengröße von 30 wird eine Normalverteilung angenommen. Daher ist die Verteilungsform der Differenzen symmetrisch.

Da alle Voraussetzungen des Wilcoxon-Tests erfüllt sind, darf dieser verwendet werden.

Hypothesen des Wilcoxon-Tests:

H_0: Die zentralen Tendenzen der zwei Stichproben sind identisch.

H_1: Die zentralen Tendenzen der zwei Stichproben unterscheiden sich.

Spearman'sche Rangkorrelationskoeffizient
Der Korrelationskoeffizient überprüft den Zusammenhang zwischen zwei Stichproben und berechnet dessen Richtung und Stärke [55]. Im Gegensatz zum Pearson-Korrelationskoeffizienten wird kein linearerer Zusammenhang angenommen, wodurch dieser bei ordinalskalierten Daten angewendet werden kann [55]. Daher wird der Spearman'sche Rangkorrelationskoeffizient im Rahmen dieser Arbeit verwendet. Um diesen durchführen zu dürfen, müssen folgende Voraussetzungen erfüllt sein [55]:

1. Die Variablen müssen mindestens ordinal skaliert sein [55]
2. Paarweise Beobachtungen [55]

Erfüllung der Voraussetzungen:

1. Die Selbsteinschätzung und Zustimmung sind ordinalskaliert, da zwischen dessen Ausprägungen eine Rangordnung besteht.
2. Die Beobachtungen sind paarweise, da jede Zeile einen Teilnehmenden und jede Spalte eine Frage darstellt.

Da alle Voraussetzungen des Spearman-Rangkorrelationskoeffizienten erfüllt sind, darf dieser berechnet und verwendet werden.

Hypothesen der Spearman-Korrelation:

H_0: Zwischen den beiden Stichproben besteht kein Zusammenhang.

H_1: Zwischen den beiden Stichproben besteht ein Zusammenhang.

Signifikanzniveau
Das Signifikanzniveau beschreibt die Wahrscheinlichkeit einer fehlerhaften Ablehnung der Nullhypothese [20]. Somit wird bei einem Signifikanzniveau von 0,05 mit einer Wahrscheinlichkeit von 5% ein zufällig entstandenes Ergebnis als statistisch signifikant gewertet [20]. Sowohl für den Wilcoxon-Test als auch den Spearman-Rangkorrelationskoeffizienten wird ein Signifikanzniveau von 0,05 festgelegt (7.1), da dieser Fehler im Rahmen der Arbeit akzeptiert wird.

$$\text{Signifikanzniveau } \alpha = 0,05 \tag{7.1}$$

7.2 Ergebnisse und statistische Auswertung

Eigenschaften der Stichprobe
Insgesamt haben 30 Personen vollständige Angaben gemacht, wobei 29 Personen männlich sind bzw. eine weiblich ist. Damit eine Normalverteilung erreicht wird, müssen mindestens 30 Stichproben vorliegen [37]. Da die Umfrage von 30 Personen durchgeführt wurde, sind die Ergebnisse normalverteilt. Das durchschnittliche Alter der Teilnehmenden beträgt 26 Jahre und die Standardabweichung ungefähr 3,5 Jahre.

Selbsteinschätzung
Neben Alter und Geschlecht werden weitergehende demografische Daten erhoben. So wird von den Teilnehmenden eingeschätzt, wie ausgeprägt dessen Erfahrungen in den Bereichen Softwareentwicklung und künstlicher Intelligenz sind. Auch wird erhoben, wie ausgeprägt sich die Teilnehmenden als technisch versiert einschätzen. Die Ergebnisse sind der Abbildung 7.1 zu entnehmen.

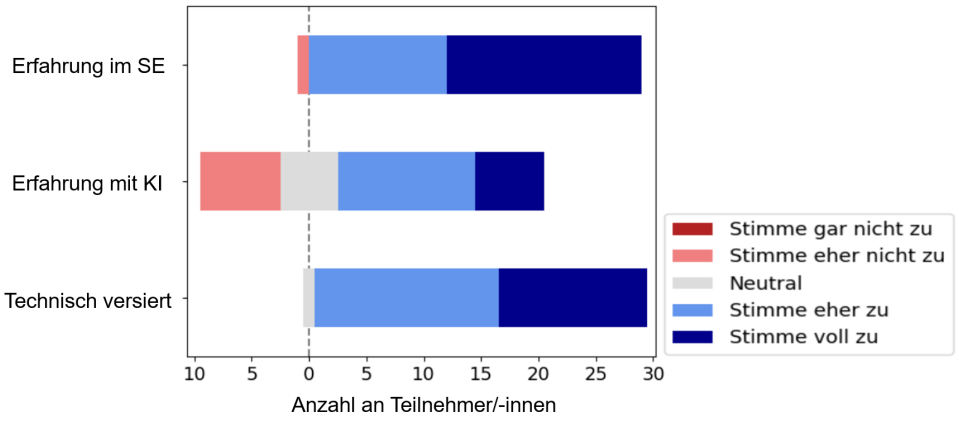

Abbildung 7.1: Selbsteinschätzung

Unter Ausschluss einer Person stimmen alle Teilnehmenden zu, dass sie Erfahrungen in der Softwareentwicklung haben. 60% der Teilnehmenden stimmen zu bzw. 23% stimmen eher nicht zu, dass sie Erfahrungen mit künstlicher Intelligenz haben. Bis auf eine Person stimmen alle Teilnehmenden zu, dass sie technisch versiert sind.

7.2.1 Verteilung der Antworten auf Fragen zur Nutzerakzeptanz

Ohne Erklärbarkeit
Um die Nutzerakzeptanz des Tools ohne Erklärbarkeit erheben zu können, wird ein Verständnis über das Tool vorausgesetzt. Daher wird dieses mithilfe

von Screenshots vorgestellt und anschließend die Einstellung gegenüber dem Tool erhoben. Hierzu wird von den Teilnehmenden bewertet, wie ausgeprägt das Tool verstanden wird bzw. hilfreich ist. Auch wird erhoben, wie ausgeprägt die Nutzungsabsicht des Tools ist. Die Ergebnisse sind in der Abbildung 7.2 dargestellt.

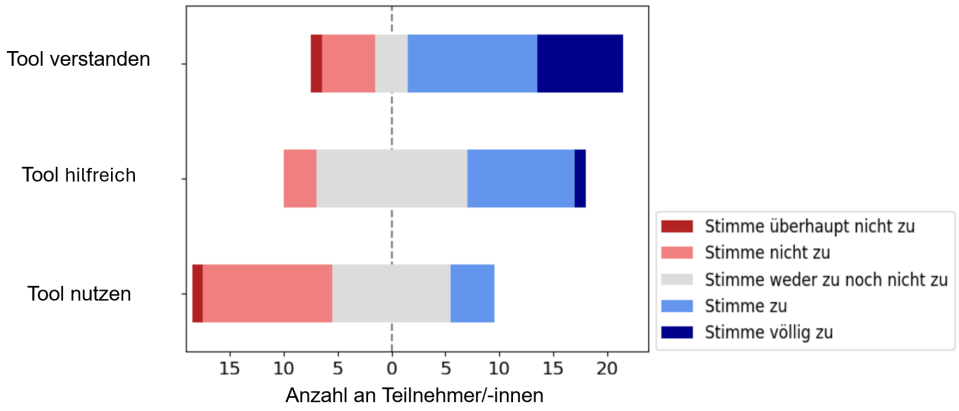

Abbildung 7.2: Nutzerakzeptanz des Tools ohne Erklärung

69% der Teilnehmenden geben an das Tool ohne Erklärung verstanden zu haben bzw. 21% das Tool nicht verstanden zu haben. Drei Personen stimmen weder zu noch nicht zu das Tool verstanden zu haben.

39% der Teilnehmenden bewerten das Tool ohne Erklärung als hilfreich bzw. 11% als nicht hilfreich. Die restlichen 50% stimmen weder zu noch nicht zu das Tool als hilfreich zu empfinden.

14% der Teilnehmenden geben an das Tool ohne Erklärung zu nutzen bzw. 47% das Tool nicht zu nutzen. 39% geben an das Tool weder zu nutzen noch nicht zu nutzen.

Beispiele

Um die Nutzerakzeptanz des Tools mit der Erklärung *Beispiele* erheben zu können, wird zuerst die Erklärung vorgestellt. Danach wird von den Teilnehmenden bewertet, wie ausgeprägt die Erklärung bzw. das Tool mit der Erklärung verstanden wird. Auch wird erhoben, wie ausgeprägt die Nutzungsabsicht des Tools mit der Erklärung ist. Die Ergebnisse sind der Abbildung 7.3 zu entnehmen.

29 Personen geben an die Erklärung zu verstehen und eine Person gibt an diese nicht zu verstehen.

87% der Teilnehmenden geben an das Tool mithilfe der Erklärung zu verstehen bzw. eine Person das Tool mithilfe der Erklärung nicht zu verstehen. Die restlichen 10% geben an weder das Tool mithilfe der Erklärung zu verstehen noch nicht zu verstehen.

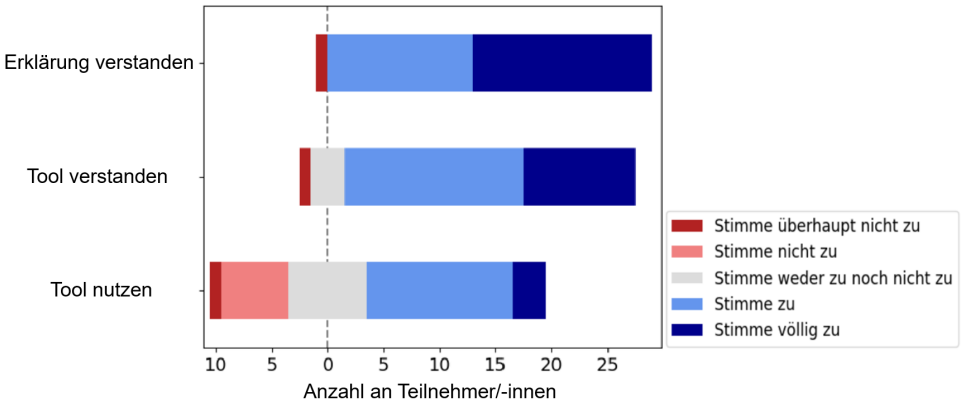

Abbildung 7.3: Nutzerakzeptanz des Tools mit Beispielen

53% der Teilnehmenden geben an das Tool mit der Erklärung zu nutzen bzw. 23% das Tool mit der Erklärung nicht zu nutzen. 24% geben an weder das Tool mit der Erklärung zu nutzen noch nicht zu nutzen.

Schlüsselwörter
Wie bei der vorherigen Erklärung wird mithilfe von Screenshots ein Verständnis über die Erklärung bewirkt. Danach wird wieder erhoben, wie ausgeprägt die Erklärung bzw. das Tool mit der Erklärung verstanden wird. Auch wird erneut erhoben, wie ausgeprägt die Teilnehmenden das Tool mit der Erklärung nutzen würden. Die Ergebnisse sind der Abbildung 7.4 zu entnehmen.

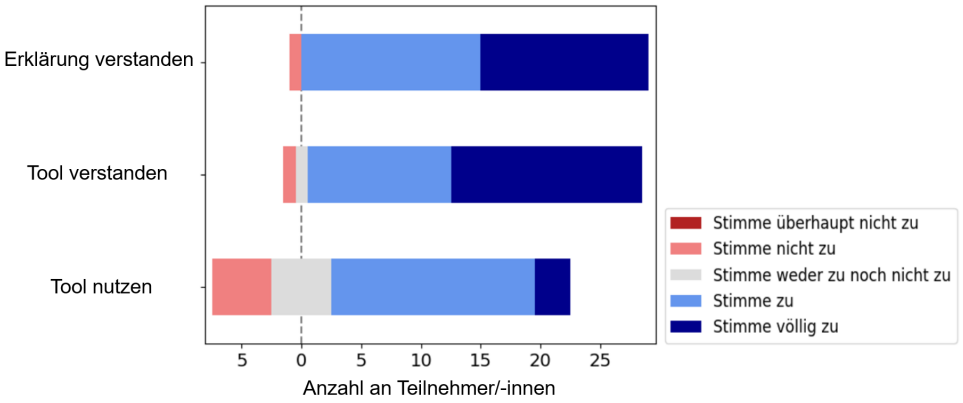

Abbildung 7.4: Nutzerakzeptanz des Tools mit Schlüsselwörtern

29 Personen geben an die Erklärung zu verstehen und eine Person gibt an diese nicht zu verstehen.

93% der Teilnehmenden geben an das Tool mithilfe der Erklärung zu verstehen bzw. eine Person das Tool mithilfe der Erklärung nicht zu

verstehen. Eine weitere Person gibt an das Tool mithilfe der Erklärung weder zu verstehen noch nicht zu verstehen.

67% der Teilnehmenden geben an das Tool mit der Erklärung zu nutzen bzw. 16% das Tool mit der Erklärung nicht zu nutzen. Die restlichen 17% geben an das Tool mit der Erklärung weder zu nutzen noch nicht zu nutzen.

Größe des Trainingsdatensatzes

Übereinstimmend zu den anderen Erklärungen wird zuerst das Verständnis der Erklärung durch Screenshots ermöglicht. Ein weiteres Mal wird erhoben, wie ausgeprägt die Erklärung bzw. das Tool mit der Erklärung verstanden wird. Auch wird erneut erhoben, wie ausgeprägt die Nutzungsabsicht des Tools mit der Erklärung ist. Die Ergebnisse sind in der Abbildung 7.5 dargestellt.

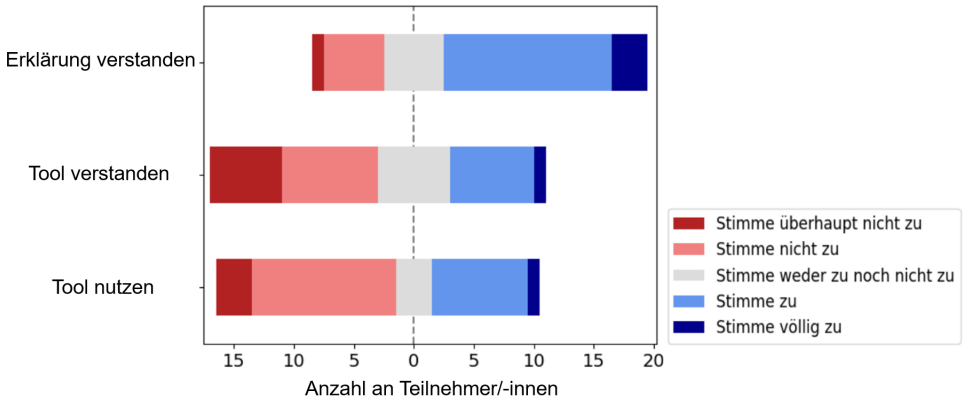

Abbildung 7.5: Nutzerakzeptanz des Tools mit Größe des Trainingsdatensatzes

61% der Teilnehmenden geben an die Erklärung zu verstehen bzw. 21% die Erklärung nicht zu verstehen. Die restlichen 18% geben an weder die Erklärung zu verstehen noch nicht zu verstehen.

29% der Teilnehmenden geben an das Tool mithilfe der Erklärung zu verstehen bzw. 50% das Tool mithilfe der Erklärung nicht zu verstehen. 21% geben an das Tool mithilfe der Erklärung weder zu verstehen noch nicht zu verstehen.

33% der Teilnehmenden geben an das Tool mit der Erklärung zu nutzen bzw. 56% das Tool mit der Erklärung nicht zu nutzen. Die restlichen 11% geben an das Tool mit der Erklärung weder zu nutzen noch nicht zu nutzen.

Genauigkeit des Modells

Erneut wird mithilfe von Screenshots die Erklärung vorgestellt, um bei den Teilnehmenden ein Verständnis zu ermöglichen. Um die Erklärungen in der Arbeit miteinander vergleichen zu können, werden bei all diesen einheitliche

Fragen gestellt. Somit wird erneut erhoben, wie ausgeprägt die Erklärung bzw. das Tool mit der Erklärung verstanden wird. Daneben wird noch einmal erhoben, wie ausgeprägt die Teilnehmenden das Tool mit der Erklärung nutzen würden. Die Ergebnisse sind in der Abbildung 7.6 zu entnehmen.

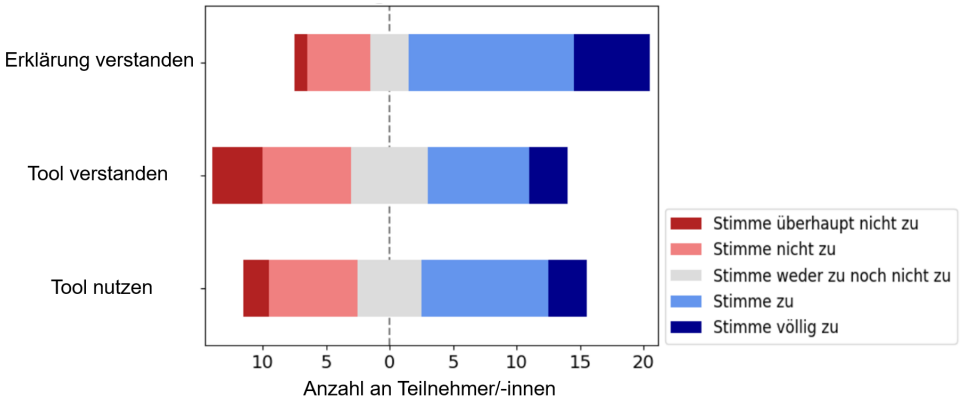

Abbildung 7.6: Nutzerakzeptanz des Tools mit Genauigkeit des Modells

68% der Teilnehmenden geben an die Erklärung zu verstehen bzw. 21% die Erklärung nicht zu verstehen. Die restlichen 11% geben an weder die Erklärung zu verstehen noch nicht zu verstehen.

39% der Teilnehmenden geben an das Tool mithilfe der Erklärung zu verstehen und ebenfalls 39% das Tool nicht zu verstehen. Die restlichen 22% geben an weder das Tool mithilfe der Erklärung zu verstehen noch nicht zu verstehen.

48% geben an das Tool mit der Erklärung zu nutzen und 33% das Tool mit der Erklärung nicht zu nutzen. Die restlichen 19% geben an weder das Tool mit der Erklärung zu nutzen noch nicht zu nutzen.

Grenzen der Software

Wie bei den zuvor beschriebenen Erklärungen wird zuerst ein Verständnis mithilfe von Screenshots bewirkt. Erneut wird erhoben, wie ausgeprägt die Erklärung bzw. das Tool mit der Erklärung verstanden wird. Danach wird erneut erhoben, wie ausgeprägt die Teilnehmenden das Tool mit der Erklärung nutzen würden. Die Ergebnisse sind in der Abbildung 7.7 dargestellt.

Alle Personen geben an die Erklärung zu verstehen und keine Person gibt an diese nicht zu verstehen.

89% geben an das Tool mithilfe der Erklärung zu verstehen und eine Person gibt an das Tool nicht zu verstehen. Zwei Personen geben an weder das Tool mithilfe der Erklärung zu verstehen noch nicht zu verstehen.

57% geben an das Tool mit der Erklärung zu nutzen bzw. 14% das Tool nicht zu nutzen. 29% geben an weder das Tool mit der Erklärung zu nutzen

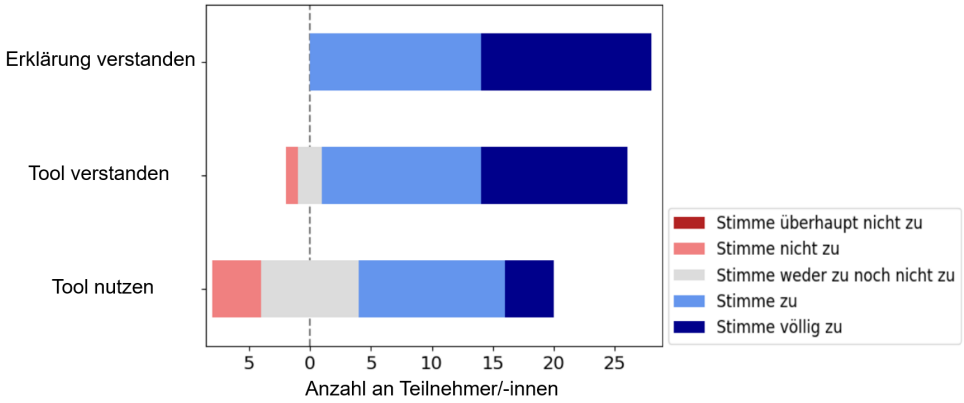

Abbildung 7.7: Nutzerakzeptanz des Tools mit Grenzen der Software

noch nicht zu nutzen.

Wahrscheinlichkeiten für die Stimmungen
Um ein Verständnis über die Erklärung zu erzielen, wird auch diese mithilfe von Screenshots vorgestellt. Zum wiederholten Male wird erhoben, wie ausgeprägt die Erklärung bzw. das Tool mit der Erklärung verstanden wird. Auch wird abermals erhoben, wie ausgeprägt die Teilnehmenden das Tool mit der Erklärung nutzen würden. Die Ergebnisse sind der Abbildung 7.8 zu entnehmen.

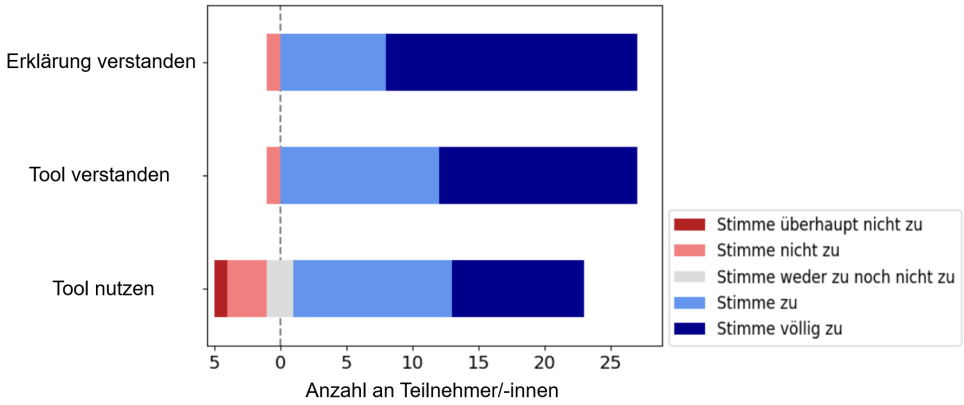

Abbildung 7.8: Nutzerakzeptanz des Tools mit Wahrscheinlichkeiten für die Stimmungen

27 Personen geben an die Erklärung zu verstehen und eine Person gibt an diese nicht zu verstehen.

27 Personen geben an das Tool mithilfe der Erklärung zu verstehen und eine Person gibt an das Tool nicht zu verstehen.

79% geben an das Tool mit der Erklärung zu nutzen bzw. 14% das Tool nicht zu nutzen. 7% geben an weder das Tool mit der Erklärung zu nutzen noch nicht zu nutzen.

Funktionsweise des Tools
Analog zu den vorherigen Erklärungen wird diese mithilfe von Screenshots vorgestellt. Von den Teilnehmenden wird letztmalig bewertet, wie ausgeprägt die Erklärung bzw. das Tool mit der Erklärung verstanden wird. Auch wird zum letzten Mal erhoben, wie ausgeprägt die Teilnehmenden das Tool mit der Erklärung nutzen würden. Die Ergebnisse sind in der Abbildung 7.9 dargestellt.

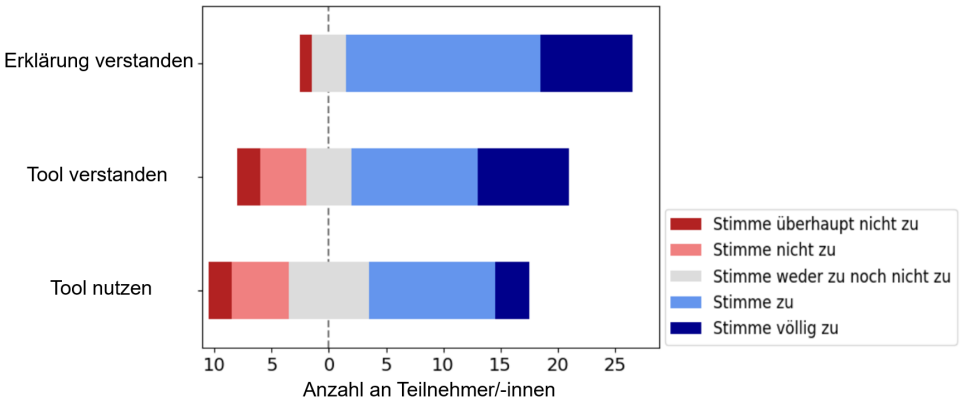

Abbildung 7.9: Nutzerakzeptanz des Tools mit Funktionsweise des Tools

86% geben an die Erklärung zu verstehen und eine Person gibt an diese nicht zu verstehen. Drei Personen geben an weder die Erklärung zu verstehen noch nicht zu verstehen.

65% geben an das Tool mithilfe der Erklärung zu verstehen bzw. 21% das Tool nicht zu verstehen. 14% geben an weder das Tool mithilfe der Erklärung zu verstehen noch nicht zu verstehen.

50% geben an das Tool mit der Erklärung zu nutzen bzw. 25% das Tool nicht zu nutzen. 25% geben an weder das Tool mit der Erklärung zu nutzen noch nicht zu nutzen.

7.2.2 Vergleich der zentralen Tendenzen

Um die Nutzerakzeptanz zwischen dem Tool ohne und mit den jeweiligen Erklärungen vergleichen zu können, werden neben der Darstellung der Verteilungen die zentralen Tendenzen berechnet. Hierbei werden Mediane und nicht Mittelwerte berechnet, da ordinal- und keine intervallskalierte Werte vorliegen. Sowohl die Medianwerte der Nutzerakzeptanz ohne als auch mit den jeweiligen Erklärungen sind der Tabelle 7.1 zu entnehmen.

Auch wenn die Fragen voneinander abweichen, sind diese in dem Kontext vergleichbar.

	Tool verstanden	**Tool hilfreich**	**Tool nutzen**
	Stimme weder zu noch nicht zu	Stimme zu	Stimme weder zu noch nicht zu
Erklärung	**Erklärung verstanden**	**Tool mithilfe der Erklärung verstanden**	**Tool mit der Erklärung nutzen**
Beispiele	Stimme völlig zu	Stimme zu	Stimme völlig zu
Schlüssel-wörter	Stimme völlig zu	Stimme völlig zu	Stimme völlig zu
Größe des Trainings-datensatzes	Stimme zu	Stimme weder zu noch nicht zu	Stimme zu
Genauigkeit des Modells	Stimme weder zu noch nicht zu	Stimme nicht zu	Stimme nicht zu
Grenzen der Software	Stimme völlig zu	Stimme völlig zu	Stimme völlig zu
Wahrschein-lichkeiten für die Stimmungen	Stimme zu	Stimme zu	Stimme zu
Funktions-weise des Tools	Stimme weder zu noch nicht zu	Stimme völlig zu	Stimme zu

Tabelle 7.1: Medianwerte der Nutzerakzeptanz ohne und mit den entsprechenden Erklärungen

Im Vergleich zu ohne Erklärung sind die Medianwerte bei den Erklärungen *Beispiele*, *Schlüsselwörter*, *Grenzen der Software*, *Wahrscheinlichkeiten für die Stimmungen* und *Funktionsweise des Tools* höher bzw. konstant. Diese sind bei der Erklärung *Genauigkeit des Modells* geringer bzw. konstant und bei der Erklärung *Größe des Trainingsdatensatzes* sowohl höher als auch geringer.

Tool verstanden/Erklärung verstanden
Aus der Tabelle 7.1 ist zu entnehmen, dass sich die Medianwerte bezüglich Tool und Erklärung verstanden unterscheiden. Um diese Unterschiede auf statistische Signifikanz zu überprüfen, wird der Wilcoxon-Test [60] durchgeführt. Somit werden die Hypothesen wie folgt definiert:

H_0: Die Mediane bezüglich Tool und Erklärung verstanden sind identisch.

H_1: Die Mediane bezüglich Tool und Erklärung verstanden unterscheiden sich.

Erklärung	**Effekt-stärke r**	**p-Wert**	**Statistische Signifikanz**
Beispiele	0.29367	0.00171	Ja
Schlüsselwörter	0.295	0.01067	Ja
Größe des Trainings-datensatzes	0.22736	0.39644	Nein
Genauigkeit des Modells	0.09543	0.73375	Nein
Grenzen der Software	0.155	0.00150	Ja
Wahrscheinlichkeiten für die Stimmungen	0.08533	0.00152	Ja
Funktionsweise des Tools	0.31878	0.06771	Nein

Tabelle 7.2: Vergleich zwischen keiner Erklärung und den jeweiligen Erklärungen bezüglich Tool verstanden (Ergebnisse des Wilxocon-Tests)

Die Ergebnisse des Wilcoxon-Tests [60] sind der Tabelle 7.2 zu entnehmen. Mithilfe der Ergebnisse und des Signifikanzniveaus wird ersichtlich, dass nur bei den Erklärungen *Beispiele*, *Schlüsselwörter*, *Grenzen der Software* und *Wahrscheinlichkeiten für die Stimmungen* statistisch signifikante Unterschiede existieren. Daher dürfen nur bei diesen die Nullhypothese abgelehnt werden.

Effektstärke r	**Interpretation**
$\|r\| < 0.1$	kein oder sehr kleiner Effekt [18]
$\|r\| = 0.1$	kleiner Effekt [18]
$\|r\| = 0.3$	mittlerer Effekt [18]
$\|r\| = 0.5$	großer Effekt [18]

Tabelle 7.3: Interpretation der r-Werte

Unter Zuhilfenahme der Tabellen 7.2 und 7.3 wird deutlich, dass bei den Erklärungen *Beispiele* und *Schlüsselwörter* signifikante mittelgroße Unterschiede existieren. Dagegen ist der Effekt des signifikanten Unterschieds bei der Erklärung *Grenzen der Software* gering und bei der Erklärung *Wahrscheinlichkeiten für die Stimmungen* sogar sehr gering.

Tool ist hilfreich/Tool mithilfe der Erklärung verstanden
Auch ist der Tabelle 7.1 zu entnehmen, dass sich die Mediane bezüglich Tool ist hilfreich und Tool mithilfe der Erklärung verstanden unterscheiden. Um einen zufällig entstandenen Unterschied der Medianwerte feststellen zu können, wird der Wilcoxon-Test [60] durchgeführt. Aus dem Kontext lassen sich die Hypothesen wie folgt ableiten:

> H_0: Die Mediane bezüglich Tool ist hilfreich und Tool mithilfe der Erklärung verstanden sind identisch.
>
> H_1: Die Mediane bezüglich Tool ist hilfreich und Tool mithilfe der Erklärung verstanden unterscheiden sich.

Erklärung	**Effekt-stärke r**	**p-Wert**	**Statistische Signifikanz**
Beispiele	0.257	0.00064	Ja
Schlüsselwörter	0.3905	0.00019	Ja
Größe des Trainings-datensatzes	0.32892	0.00281	Ja
Genauigkeit des Modells	0.35082	0.09710	Nein
Grenzen der Software	0.2305	0.00047	Ja
Wahrscheinlichkeiten für die Stimmungen	0.12707	0.00004	Ja
Funktionsweise des Tools	0.34192	0.24961	Nein

Tabelle 7.4: Vergleich zwischen keiner Erklärung und den jeweiligen Erklärungen bezüglich Tool ist hilfreich (Ergebnisse des Wilxocon-Tests)

Die Ergebnisse des Wilcoxon-Tests [60] sind in der Tabelle 7.4 dargestellt. Mithilfe dieser wird ersichtlich, dass unter Ausschluss von *Genauigkeit des Modells* und *Funktionsweise des Tools* bei allen Erklärungen statisch signifikante Unterschiede bestehen. Daher dürfen bei den Erklärungen *Beispiele*, *Schlüsselwörter*, *Größe des Trainingsdatensatzes*, *Grenzen der Software* und *Wahrscheinlichkeiten für die Stimmungen* die Nullhypothese abgelehnt werden.

Unter Berücksichtigung der Tabellen 7.3 und 7.4 wird sichtbar, dass bei den Erklärungen *Beispiele*, *Schlüsselwörter*, *Größe des Trainingsdatensatzes* und *Grenzen der Software* signifikante mittelgroße Unterschiede bestehen. Im Vergleich dazu ist der Effekt des signifikanten Unterschieds bei der Erklärung *Wahrscheinlichkeiten für die Stimmungen* gering.

Tool nutzen/Tool mit der Erklärung nutzen
Der Tabelle 7.1 ist zu entnehmen, dass sich die Mediane bezüglich Tool (mit der Erklärung) nutzen unterscheiden. Erneut wird der Wilcoxon-Test [60] durchgeführt, um die Unterschiede der Medianwerte auf statistische Signifikanz zu überprüfen. Daher werden die Hypothesen wie folgt definiert:

> H_0: Die Mediane bezüglich Tool (mit der Erklärung) nutzen sind identisch.
>
> H_1: Die Mediane bezüglich Tool (mit der Erklärung) nutzen unterscheiden sich.

Erklärung	**Effekt-stärke r**	**p-Wert**	**Statistische Signifikanz**
Beispiele	0,2262	0.00384	Ja
Schlüsselwörter	0,2475	0.00062	Ja
Größe des Trainings-datensatzes	0,3414	0.82925	Nein
Genauigkeit des Modells	0,0478	0.03081	Ja
Grenzen der Software	0,2231	0.00024	Ja
Wahrscheinlichkeiten für die Stimmungen	0,22081	0.00007	Ja
Funktionsweise des Tools	0,19893	0.01134	Ja

Tabelle 7.5: Vergleich zwischen keiner Erklärung und den jeweiligen Erklärungen bezüglich Tool nutzen (Ergebnisse des Wilxocon-Tests)

Die Ergebnisse des Wilcoxon-Tests [60] sind der Tabelle 7.5 zu entnehmen. Unter dessen Zuhilfenahme wird ersichtlich, dass bei allen Erklärungen bis auf *Größe des Trainingsdatensatzes* statistisch signifikante Unterschiede existieren. Daher dürfen bei diesen die Nullhypothese abgelehnt werden.

Anhand der Tabellen 7.3 und 7.5 wird erkennbar, dass die Erklärungen *Beispiele*, *Schlüsselwörter*, *Grenzen der Software* und *Wahrscheinlichkeiten für die Stimmungen* signifikante mittelgroße Unterschiede aufweisen. Dagegen ist der Effekt des signifikanten Unterschieds bei der Erklärung *Funktionsweise des Tools* gering und bei der Erklärung *Genauigkeit des Modells* sehr gering.

7.2.3 Zusammenhang zwischen demografischen Daten und Tool nutzen

Um einen möglichen Zusammenhang zwischen den demografischen Daten und Tool nutzen detektieren zu können, werden potenzielle Korrelationen mithilfe des Spearman Rangkorrelationskoeffizienten [55] berechnet.

Erfahrung mit künstlicher Intelligenz
Nachfolgend wird überprüft, ob ein Zusammenhang zwischen der Erfahrung mit KI und Tool nutzen ohne und mit den jeweiligen Erklärungen existiert. Durch die Anwendung des Rangkorrelationskoeffizienten [55] in diesem Kontext ergeben sich folgende Hypothesen:

H_0: Zwischen Erfahrung mit KI und Tool nutzen besteht kein Zusammenhang.

H_1: Zwischen Erfahrung mit KI und Tool nutzen besteht ein Zusammenhang.

	Korrelations-koeffizient	**p-Wert**	**Statistische Signifikanz**
Ohne Erklärung	0.32739	0.08298	Nein
Beispiele	0.01426	0.94038	Nein
Schlüsselwörter	-0.16660	0.37891	Nein
Größe des Trainings-datensatzes	0.61005	0.00044	Ja
Genauigkeit des Modells	0.10322	0.59416	Nein
Grenzen der Software	0.01479	0.93818	Nein
Wahrscheinlichkeiten für die Stimmungen	-0.07282	0.70217	Nein
Funktionsweise des Tools	0.00522	0.97857	Nein

Tabelle 7.6: Zusammenhang zwischen Erfahrung mit KI und Tool nutzen (Ergebnisse des Spearman-Tests)

Unter Berücksichtigung des in dieser Arbeit festgelegten Signifikanzniveaus existiert nur ein statistisch signifikanter Zusammenhang zwischen Erfahrung mit KI und Tool mit der Erklärung *Größe des Trainingsdatensatzes* nutzen. Dies ist die einzige Erklärung, bei der die Nullhypothese abgelehnt werden kann und damit ein Zusammenhang besteht. Mithilfe der Tabelle 7.3 wird ersichtlich, dass dieser statisch signifikante Zusammenhang stark positiv ist. Die Ergebnisse sind in der Tabelle 7.6 dargestellt.

Technisch versiert
Danach wird untersucht, ob ein Zusammenhang zwischen technisch versiert und Tool nutzen besteht. Daher werden die Hypothesen wie folgt definiert:

> H_0: Zwischen technisch versiert und Tool nutzen besteht kein Zusammenhang.
>
> H_1: Zwischen technisch versiert und Tool nutzen besteht ein Zusammenhang.

Die Nullhypothese kann sowohl ohne als auch mit den entsprechenden Erklärungen nicht abgelehnt werden. Da die Ergebnisse statistisch nicht signifikant sind, werden diese in der Arbeit nicht weiter betrachtet und sind der Tabelle E.1 im Anhang zu entnehmen.

Erfahrung in der Softwareentwicklung
Auch wird überprüft, ob ein Zusammenhang zwischen Erfahrung im SE und Tool nutzen existiert. Somit werden die Hypothesen wie folgt definiert:

> H_0: Zwischen Erfahrung im SE und Tool nutzen besteht kein Zusammenhang.
>
> H_1: Zwischen Erfahrung im SE und Tool nutzen besteht ein Zusammenhang.

Unter Berücksichtigung des Signifikanzniveaus wurde kein statistisch signifikanter Zusammenhang festgestellt. Daher kann die Nullhypothese sowohl ohne als auch mit den entsprechenden Erklärungen nicht abgelehnt werden. Die Ergebnisse sind im Anhang in der Tabelle E.2 dargestellt.

Kapitel 8

Diskussion der Ergebnisse

In diesem Kapitel werden zuerst die Forschungsfragen mithilfe der Ergebnisse aus dem Workshop und der Umfrage beantwortet. Darüber hinaus werden die Ergebnisse im Kontext dieser Arbeit interpretiert. Auch werden die Grenzen dieser Arbeit und die Validität in Bezug auf Einschränkungen untersucht.

8.1 Beantwortung der Forschungsfragen

Erklärbarkeitsanforderungen

FF1: Welche Anforderungen an die Erklärbarkeit lassen sich bei der Nutzung des Stimmungsanalysetools *RoBERTa* im Rahmen eines Workshops erheben?

Von den Teilnehmenden des Workshops hat sich die *Genauigkeit der Ergebnisse* als eine wichtige Erklärbarkeitsanforderung herausgestellt. Durch dessen Einführung soll ersichtlich werden mit welcher Genauigkeit das Modell vortrainiert wurde.

Auch haben sich die *Grenzen der Software* als eine wichtige Erklärbarkeitsanforderung erwiesen. Mithilfe dieser soll u. a. erkennbar werden, welche Eingaben nicht berechenbar sind.

Eine ebenfalls von den Teilnehmenden als wichtig betrachtete Erklärbarkeitsanforderung ist die Angabe der *Wahrscheinlichkeiten für die Stimmungen*. So soll nicht nur das Vorliegen einer negativen, neutralen oder positiven Stimmung, sondern auch dessen Wahrscheinlichkeiten ausgegeben werden.

Innerhalb des Workshops hat sich die als eher wichtig bewertete Erklärbarkeitsanforderung *Schlüsselwörter* manifestiert. Durch die grüne Markierung von positiven Wörtern bzw. rote Markierung von negativen Wörtern soll ersichtlich werden, welche Wörter zum Ergebnis geführt haben. Aufgrund einer möglichen Rot-Grün-Sehschwäche der Nutzenden, können auch andere Farben zur Markierung der Wörter verwendet werden.

Eine ebenso von den Teilnehmenden als eher wichtig betrachtete Erklärbarkeitsanforderung ist die Einführung von *Beispielen*. Mithilfe dieser soll u. a. die Anwendung des Tools und die Interpretation der Ergebnisse verdeutlicht werden.

Im Rahmen des Workshops hat sich die neutral bewertete Erklärbarkeitsanforderung *Größe des Trainingsdatensatzes* manifestiert. Damit die Validität des Modells überprüft werden kann, soll die Größe des beim Training verwendeten Datensatzes angegeben werden.

Die Erläuterung über die *Funktionsweise des Tools* hat sich ebenfalls als eine neutral bewertete Erklärbarkeitsanforderungen herausgestellt. Durch diese soll die schrittweise Berechnung des Ergebnisses erklärt werden.

Von den Teilnehmenden hat sich die als eher unwichtig betrachtete *Bewertung der Software* als Erklärbarkeitsanforderung erwiesen. Dies soll ermöglicht werden, wenn z. B. das eigene Empfinden von dem Ergebnis abweicht.

Einfluss der Erklärbarkeit auf die Nutzerakzeptanz

FF2: Welchen Einfluss hat die Umsetzung der Erklärbarkeitsanforderungen auf die Nutzerakzeptanz des Stimmungsanalysetools *RoBERTa*?

Unter Ausschluss der Erklärung *Genauigkeit des Modells* sind die Medianwerte bezüglich Tool mit der Erklärung nutzen bei allen Erklärungen höher als beim Tool ohne Erklärung. Um diese Unterschiede auf statistische Signifikanz zu überprüfen, wird der Wilcoxon-Test durchgeführt. Die Ergebnisse zeigen, dass die mittelgroßen positiven Unterschiede der Medianwerte in Bezug auf die Erklärungen *Beispiele*, *Schlüsselwörter*, *Grenzen der Software* und *Wahrscheinlichkeiten für die Stimmungen* statistisch signifikant sind. Im Gegensatz dazu ist der Effekt des signifikanten positiven Unterschieds bei der Erklärung *Funktionsweise des Tools* gering und bei der Erklärung *Genauigkeit des Modells* sehr gering.

Bis auf die Erklärungen *Größe des Trainingsdatensatzes* und *Genauigkeit des Modells* sind die Medianwerte bezüglich Tool mithilfe der Erklärung verstanden bei allen Erklärungen höher bzw. konstant im Vergleich zu Tool ist ohne Erklärung hilfreich. Die Ergebnisse des Wilcoxon-Tests zeigen, dass die mittelgroßen positiven Unterschiede der Medianwerte in Bezug auf die Erklärungen *Beispiele*, *Schlüsselwörter*, *Größe des Trainingsdatensatzes* und *Grenzen der Software* statistisch signifikant sind. Auch ist der geringe positive Unterschied der Medianwerte bezüglich der Erklärung *Wahrscheinlichkeiten für die Stimmungen* statistisch signifikant.

Die Medianwerte bezüglich Erklärung verstanden sind bei allen Erklärungen höher bzw. konstant im Vergleich zu Tool ohne Erklärung verstanden. Durch die Durchführung des Wilcoxon-Tests wird ersichtlich,

dass die mittelgroßen positiven Unterschiede der Medianwerte in Bezug auf die Erklärungen *Beispiele* und *Schlüsselwörter* statistisch signifikant sind. Demgegenüber ist der Effekt des signifikanten positiven Unterschieds bei der Erklärung *Grenzen der Software* gering und bei der Erklärung *Wahrscheinlichkeiten für die Stimmungen* sehr gering.

Wichtigkeit der nichtfunktionalen Anforderungen

FF3: Wie wichtig werden die nichtfunktionalen Anforderungen des Stimmungsanalysetools *RoBERTa* im Rahmen eines Workshops bewertet?

Die nichtfunktionalen Anforderungen *Überprüfbarkeit* und *Benutzereffektivität* wurden im Kontext des Stimmungsanalysetools als wichtig bewertet. Dagegen wurden die Anforderungen *Lernfreundlichkeit*, *Benutzereffizienz*, *Benutzerbewusstsein*, *Benutzerzufriedenheit*, *Bewusstsein für Datenschutz*, *wahrgenommener Wert*, *wahrgenommene Nützlichkeit* und *Unterstützung der Entscheidungsfindung* von den Teilnehmenden als eher wichtig betrachtet. Darüber hinaus wurden im Rahmen des Workshops die nichtfunktionalen Anforderungen *Leitfaden*, *Mensch-Maschine Kooperation*, *Benutzerkontrolle* und *-freundlichkeit* als neutral beurteilt. Des Weiteren wurden das *Benutzererlebnis* und die *Genauigkeit des mentalen Modells* als eher unwichtig bzw. das *Entdecken von Wissen* und die *Benutzerleistung* als unwichtig betrachtet.

8.2 Interpretation der Ergebnisse

Empfehlungen für die Industrie

Die Potenziale durch die Einführung von Erklärbarkeit werden ersichtlich, da bei einigen der in dieser Arbeit umgesetzten Erklärungen eine statistisch signifikante Erhöhung der Nutzerakzeptanz nachweisbar ist. Somit könnte durch dessen Umsetzung in der Industrie die Nutzung von Stimmungsanalysetools erhöht werden. Durch eine umfassendere Etablierung von Stimmungsanalysetools in der Industrie könnten Schwierigkeiten und Probleme in der Kommunikation erkannt und bewältigt werden und somit zu einer reibungslosen und erfolgreichen Softwareentwicklung beitragen.

Durch die prototypische Umsetzung der Erklärung *Schlüsselwörter* wurde eine signifikante mittelgroße Erhöhung der Nutzerakzeptanz des Tools festgestellt. Eine softwaretechnische Implementierung in der Industrie wird empfohlen, da dies wahrscheinlich eine Erhöhung der Nutzerakzeptanz bewirken würde.

In der Arbeit wurde gezeigt, dass neben der Erklärung *Schlüsselwörter* auch die Einführung der Erklärung *Beispiele* zu einer signifikanten mittelgroßen Erhöhung der Nutzerakzeptanz führt. Daher wird für die Industrie ebenso eine Empfehlung für dessen Einführung ausgesprochen.

Bei den restlichen der in dieser Arbeit umgesetzten Erklärungen wurde zwar keine Abnahme der Nutzerakzeptanz festgestellt, jedoch sind die Effekte gering oder nicht signifikant. Somit könnte die Einführung der Erklärungen durchaus die Erhöhung der Nutzerakzeptanz bewirken, aber eine explizite Empfehlung für die Industrie wird aus den zuvor genannten Gründen nicht ausgesprochen.

Mögliche Zielkonflikte durch Erklärbarkeit
Da sich nichtfunktionale Anforderungen untereinander beeinflussen können, empfehlen Chazette et al. [12] mögliche Zielkonflikte durch die Einführung von Erklärbarkeit zu berücksichtigen. Daher wurden diese mithilfe des Konfliktkatalogs von Chazette et al. [12] und dem Workshop abgeschätzt. In dieser Arbeit wurde festgestellt, dass die nutzerzentrierten nichtfunktionalen Anforderungen durch die Erklärbarkeit überwiegend positiv beeinflusst werden und somit die Voraussetzungen für deren Einführung erfüllt sind. Aufgrund möglicher Zielkonflikte empfiehlt diese Arbeit bei der Umsetzung von Erklärungen den vorliegenden Kontext auf Zielkonflikte zu untersuchen. Dies ist sowohl unter Berücksichtigung des Konfliktkatalogs von Chazette et al. [12] als auch von Mairiza und Zowghi [42] möglich.

Zusammenhang zwischen Erfahrung mit KI und Tool nutzen
In der Arbeit wurde ein signifikanter stark positiver Zusammenhang zwischen Erfahrung mit KI und Tool mit der Erklärung *Größe des Trainingsdatensatzes* nutzen festgestellt. Damit die Verwendung eines Systems akzeptiert wird, empfiehlt die Studie von Pieters [51] bei der Einführung von Erklärbarkeit die richtige Art an Informationen bereitzustellen. Durch die Erfahrung mit künstlicher Intelligenz könnte das nötige Vorwissen bestehen, um mithilfe der Erklärung *Größe des Trainingsdatensatzes* die Validität des Modells überprüfen zu können. Dies könnte zu einem erhöhten Vertrauen in das Tool führen und somit die *Größe des Trainingsdatensatzes* die richtige Art an Information nach Pieters [51] repräsentieren. Somit würde die in dieser Arbeit beobachtete Korrelation durch die Studie von Pieters [51] erklärt werden.

Vergleich der Ergebnisse des Workshops und der Umfrage
Auch wenn die Erklärungen *Schlüsselwörter* und *Beispiele* im Rahmen des Workshops als weniger wichtig bewertet wurden als z. B. die Erklärung *Genauigkeit des Modells*, bewirkt die Umsetzung dieser eine ausgeprägtere Nutzerakzeptanz als bei den höher priorisierten Erklärungen. Ebenso ist auffällig, dass die im Workshop als wichtig betrachtete Erklärung *Genauigkeit des Modells* eine sehr geringe bzw. keine Erhöhung der Nutzerakzeptanz herbeiführt. Daraus wird ersichtlich, dass die subjektiv beurteilte Wichtigkeit der Erklärung nicht unbedingt mit einer Erhöhung der Nutzerakzeptanz

durch dessen Umsetzung korreliert.

8.3 Limitierungen und Einschränkungen

Das innerhalb der Literaturrecherche verwendete Snowballing-Verfahren von Wohlin [62] wird bereits nach einer Iteration beendet, da bereits innerhalb der ersten Iteration nach einiger Recherche keine neuen Konzepte gefunden werden konnten. Somit fehlt möglicherweise für diese Arbeit relevante Literatur. Für die Einarbeitung in die Grundlagen der Stimmungsanalyse und Erklärbarkeit ist die Literaturrecherche eine geeignete Forschungsmethode. Da Studien dazu neigen eher positive als negative Ergebnisse zu publizieren, entsteht ein Publikationsbias. Die Literaturrecherche und die in der Literaturrecherche berücksichtigten Metastudien untersuchen lediglich publizierte Studien, wodurch dessen Ergebnisse eventuell überschätzt werden. Auch wird in dieser Arbeit dazu tendiert eher positive als negative Ergebnisse hervorzuheben, wodurch die Nutzerakzeptanz des Stimmungsanalysetools möglicherweise überschätzt wird.

Da die Teilnehmenden des Workshops sich möglicherweise gegenseitig beeinflussen, werden die Meinungen schneller angepasst. Eigene Meinungen werden eventuell nicht berücksichtigt und folglich werden die Ergebnisse des Workshops verzerrt. Auch könnten die Teilnehmenden die Aufgaben nachlässig bearbeiten aufgrund der zeitlichen Begrenzung des Workshops, wodurch wiederum die Ergebnisse verzerrt werden.

Im Rahmen des Workshops wird die Wichtigkeit der nichtfunktionalen Anforderungen von Chazette et al. [12] im Kontext des Stimmungsanalysetools bewertet. Mithilfe des Konfliktkatalogs von Chazette et al. [12] und der im Workshop priorisierten nichtfunktionalen Anforderungen können zwar die Auswirkungen der Erklärbarkeit auf andere nichtfunktionale Anforderungen abgeschätzt werden. Um jedoch die Beeinflussungen präziser beurteilen zu können, müssen diese empirisch erhoben werden.

Aufgrund der niedrigen Priorität und dem hohen Implementationsaufwands wird die Erklärung *Bewertung der Software* bei der softwaretechnischen Umsetzung ausgeschlossen. Daher wird diese bei der Erhebung der Nutzerakzeptanz im Rahmen der Umfrage nicht berücksichtigt und eine Auswertung der Erklärung ist folglich nicht möglich.

Damit die Teilnehmenden die Fragen der Umfrage ehrlich beantworten, wird bei der Durchführung dessen Anonymität gewährleistet. Somit könnte die Umfrage von einer Person mehrfach durchgeführt werden und die Ergebnisse der Umfrage werden möglicherweise verzerrt. Zudem wird die Umfrage u. a. bei Kommilitoninnen und Kommilitonen im persönlichen Umfeld beworben, um eine hohe Stichprobengröße zu erwirken. Durch eine voreingenommene Beantwortung werden die Ergebnisse möglicherweise verzerrt. Darüber hinaus kann aufgrund von z. B. Zeitmangel der Teilneh-

menden die Umfrage unpräzise und unehrlich durchgeführt werden, wodurch die Ergebnisse verzerrt werden. Somit kann die in der Umfrage angegebene Nutzerakzeptanz stark von der tatsächlichen Nutzerakzeptanz abweichen.

Die Umfrage wurde von 30 Teilnehmenden vollständig durchgeführt und eine Normalverteilung kann somit angenommen werden [37]. Daher ist die Stichprobengröße ausreichend groß, um auf die Grundgesamtheit schließen zu können. Eine höhere Stichprobengröße würde jedoch zu valideren Ergebnissen führen. Das Geschlechterverhältnis in der Umfrage weicht vom Geschlechterverhältnis von der Gesamtmenge an Mitarbeitenden in der Softwareentwicklung [11] ab. Die Ergebnisse der Umfrage sind für Frauen weniger aussagekräftig, da diese in der Stichprobe unterrepräsentiert sind.

Kapitel 9

Zusammenfassung und Ausblick

In diesem Kapitel werden die Ergebnisse der Arbeit abschließend zusammengefasst und die nächsten Schritte der zukünftigen Forschung skizziert.

9.1 Zusammenfassung

Ziel der Arbeit war die Ermittlung von Erklärbarkeitsanforderungen zur Erhöhung der Nutzerakzeptanz eines Stimmungsanalysetools. Um die Grundlagen bezüglich Stimmungsanalysetools und Erklärbarkeit zu erforschen, wurde eine ausführliche Literaturrecherche durchgeführt. Dabei wurde u. a. ersichtlich, dass Erklärbarkeit sich auf andere nichtfunktionale Anforderungen auswirkt und *RoBERTa* das präziseste und weitverbreitetste Stimmungsanalysetool ist.

Die Erklärbarkeitsanforderungen wurden im Kontext des Stimmungsanalysetools *RoBERTa* erhoben, da dies das für diese Arbeit prädestinierteste Tool ist. Um im Rahmen des Workshops die Anwendung des Tools im SE-Kontext zu ermöglichen, wurde *RoBERTa* mithilfe eines Datensatzes mit über 4000 Einträgen vortrainiert. Das Stimmungsanalysetool wurde im Rahmen eines Workshops vorgestellt und dessen Erklärbarkeitsanforderungen erhoben. Diese werden nachfolgend absteigend nach dessen beurteilter Wichtigkeit dargestellt: *Genauigkeit der Ergebnisse*, *Grenzen der Software*, *Wahrscheinlichkeiten für die Stimmungen*, *Schlüsselwörter*, *Beispiele*, *Größe des Trainingsdatensatzes*, *Funktionsweise des Tools* und *Bewertung der Software*.

Da nichtfunktionale Anforderungen sich untereinander beeinflussen, wurden die Auswirkungen von Erklärbarkeit auf verwandte nichtfunktionale Anforderungen abgeschätzt. Dazu wurden 18 nutzerzentrierte nichtfunktionale Anforderungen von Chazette et al. [12] innerhalb des Workshops bezüglich dessen Wichtigkeit bewertet. Nachfolgend werden die auf das

Stimmungsanalysetool *RoBERTa* bezogenen nichtfunktionalen Anforderungen absteigend nach dessen Wichtigkeit dargestellt: *Überprüfbarkeit*, *Benutzereffektivität*, *Lernfreundlichkeit*, *Benutzereffizienz*, *Benutzerbewusstsein*, *Benutzerzufriedenheit*, *Bewusstsein für Datenschutz*, *wahrgenommener Wert*, *wahrgenommene Nützlichkeit*, *Unterstützung der Entscheidungsfindung*, *Leitfaden*, *Mensch-Maschine Kooperation*, *Benutzerkontrolle*, *Benutzerfreundlichkeit*, *Benutzererlebnis*, *Genauigkeit des mentalen Modells*, *Entdecken von Wissen* und *Benutzerleistung*. Mithilfe dieser priorisierten nichtfunktionalen Anforderungen und des Konfliktkatalogs von Chazette et al. [12] wurden die potenziellen Auswirkungen durch die Einführung von Erklärbarkeit geschätzt. In der Arbeit wurde festgestellt, dass die nichtfunktionalen Anforderungen durch Erklärbarkeit überwiegend positiv beeinflusst werden. Da die Voraussetzungen der Erklärbarkeit erfüllt sind, wurden die Erklärbarkeitsanforderungen im nächsten Schritt umgesetzt.

Die Umsetzung der Erklärbarkeitsanforderungen wurde auf Realisierbarkeit im Kontext des Stimmungsanalysetools überprüft und ggf. diese softwaretechnisch bzw. prototypisch umgesetzt. *RoBERTa* wurde softwaretechnisch angepasst, sodass unter Ausschluss der Erklärung *Bewertung der Software* alle Erklärungen umgesetzt wurden. Um die Nutzerakzeptanz des Stimmungsanalysetools ohne und mit den jeweiligen Erklärungen erheben und vergleichen zu können, wurde eine Online-Umfrage durchgeführt.

Bei der anschließenden Analyse der Umfrageergebnisse wurde u. a. ermittelt, dass die Absicht das Tool mit den jeweiligen Erklärungen *Beispiele*, *Schlüsselwörter*, *Grenzen der Software* und *Wahrscheinlichkeiten für die Stimmungen* zu nutzen signifikant mittelgroß höher ist als die Absicht das Tool ohne Erklärung zu nutzen. Auch wurde u. a. festgestellt, dass das Verständnis des Tools mithilfe der jeweiligen Erklärungen *Beispiele*, *Schlüsselwörter*, *Größe des Trainingsdatensatzes* und *Grenzen der Software* signifikant mittelgroß höher ist als das Tool ohne Erklärung als hilfreich zu empfinden. Zudem wurde u. a. ermittelt, dass das Verständnis der jeweiligen Erklärungen *Beispiele* und *Schlüsselwörter* signifikant mittelgroß höher ist als das Verständnis des Tools ohne Erklärung.

Mithilfe von statistischen Tests wurde ersichtlich, dass durch die Umsetzung mancher Erklärbarkeitsanforderungen tatsächlich eine signifikante Erhöhung der Nutzerakzeptanz nachweisbar ist. Daher wurde das Ziel dieser Arbeit im Kontext des Stimmungsanalysetools *RoBERTa* bestätigt. Auch wurden in dieser Arbeit Empfehlungen ausgesprochen, wie z. B. die Umsetzung der Erklärungen *Beispiele* und *Schlüsselwörter* in der Industrie unter Voraussetzung den vorliegenden Kontext auf Zielkonflikte zu untersuchen. Insgesamt stellt die Einführung von Erklärbarkeit einen vielversprechenden Ansatz dar, um die Nutzerakzeptanz von Stimmungsanalysetools zu erhöhen und damit auch dessen Nutzung in der Industrie voranzutreiben.

9.2 Ausblick

Aufgrund der Empfehlungen von Balasubramaniam et al. [5] könnten in zukünftigen Arbeiten Erklärbarkeitsanforderungen unter Berücksichtigung den Menschen als Erklärer zu nutzen erhoben werden. Weil in der Praxis auch andere Stimmungsanalysetools außer *RoBERTa* verwendet werden, könnten in zukünftigen Arbeiten die Erklärbarkeitsanforderungen bezüglich dieser erhoben werden.

Da die Ergebnisse der Umfrage für Männer valider sind als für Frauen, könnten für validere Ergebnisse der Arbeit die Umfrage mit dem Fokus auf Frauen fortgeführt werden. Angesichts der möglicherweisen starken Abweichung der in dieser Arbeit ermittelten Nutzerakzeptanz des Stimmungsanalysetools von der tatsächlichen Nutzung, könnten in zukünftigen Arbeiten eine Erhöhung der Nutzung durch die Implementierung der Erklärungen untersucht werden.

Weil in dieser Arbeit die Auswirkungen von Erklärbarkeit auf andere nichtfunktionale Anforderungen lediglich abgeschätzt wurden, wäre die konkrete Erhebung dieser eine Aufgabe für zukünftige Forschungen. Da in dieser Arbeit eine signifikante mittelgroße Erhöhung der Nutzerakzeptanz durch die prototypisch umgesetzte Erklärung *Schlüsselwörter* festgestellt wurde, könnten zukünftige Arbeiten diese softwaretechnisch durch die Anpassung des Stimmungsanalysetools *RoBERTa* umsetzten.

Aufgrund der Empfehlung von Chazette und Schneider [16] erfolgte in dieser Arbeit zwar eine leichtgewichtige und nutzerzentrierte Umsetzung der Erklärbarkeitsanforderungen. Um jedoch die Softwarequalität von Stimmungsanalysetools nicht negativ zu beeinflussen, könnten in zukünftigen Forschungen der Fokus auf diese Anforderungen intensiviert werden.

Die Ergebnisse lassen sich eventuell auf andere Stimmungsanalysetools übertragen, vorausgesetzt diese funktionieren ähnlich wie *RoBERTa*. Doch inwiefern dies zutrifft, könnte in der weiterführenden Forschung untersucht werden.

Anhang A

Literaturrecherche

In der folgenden Tabelle werden die in dieser Arbeit verwendeten Techniken der Literaturrecherche und die mithilfe der Techniken ermittelte Literatur dargestellt.

Recherchetechnik	**Ermittelte Literatur**	
Startstudien	Obaidi und Klünder [46]	Chazette et al. [12]
Snowballing	[40] [47] [28] [27] [29] [34] [56] [33] [48] [23] [57] [52] [26] [49] [50] [54] [1] [3] [7] [8] [9] [10] [17] [21] [22] [24] [31] [39] [44] [45] [58] [64]	[4] [32] [5] [19] [38] [43] [13] [14] [15] [25] [2] [16] [61] [59] [30] [42] [6] [36] [53]
Datenbanksuche	[11] [18] [20] [35] [37] [41] [51] [55] [60] [62] [63]	

Tabelle A.1: Ergebnisse der Recherchetechniken

Anhang B

Workshop

Folgend werden die Ergebnisse aus dem Workshop dargestellt. Die erste Abbildung veranschaulicht die erhobenen und priorisierten Erklärbarkeitsanforderungen. In der zweiten Abbildung werden die priorisierten nichtfunktionalen Anforderungen dargestellt.

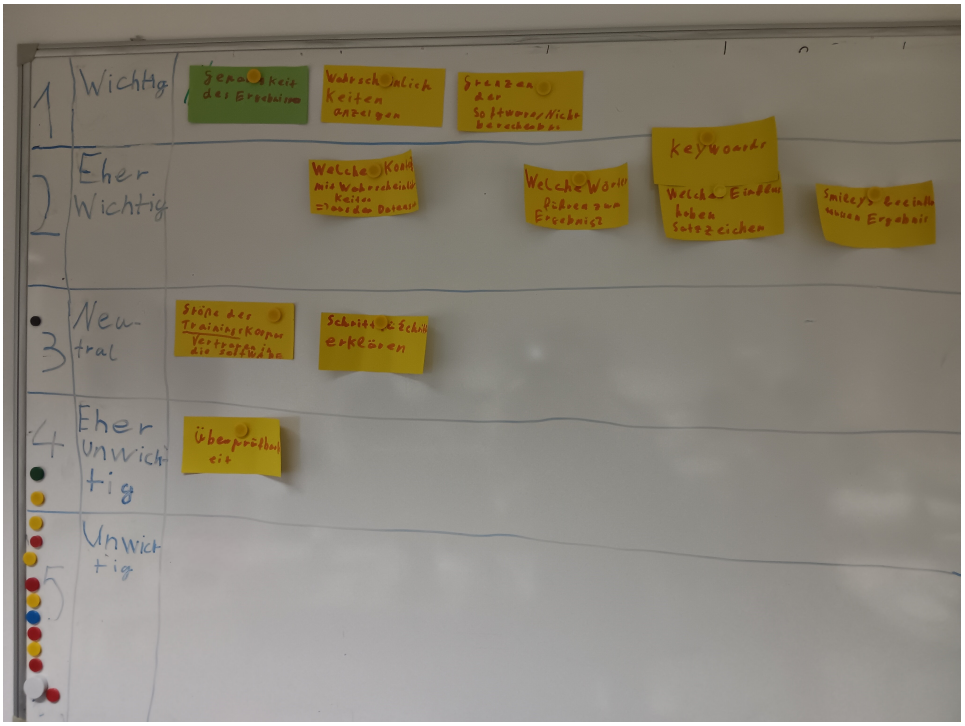

Abbildung B.1: Erklärbarkeitsanforderungen

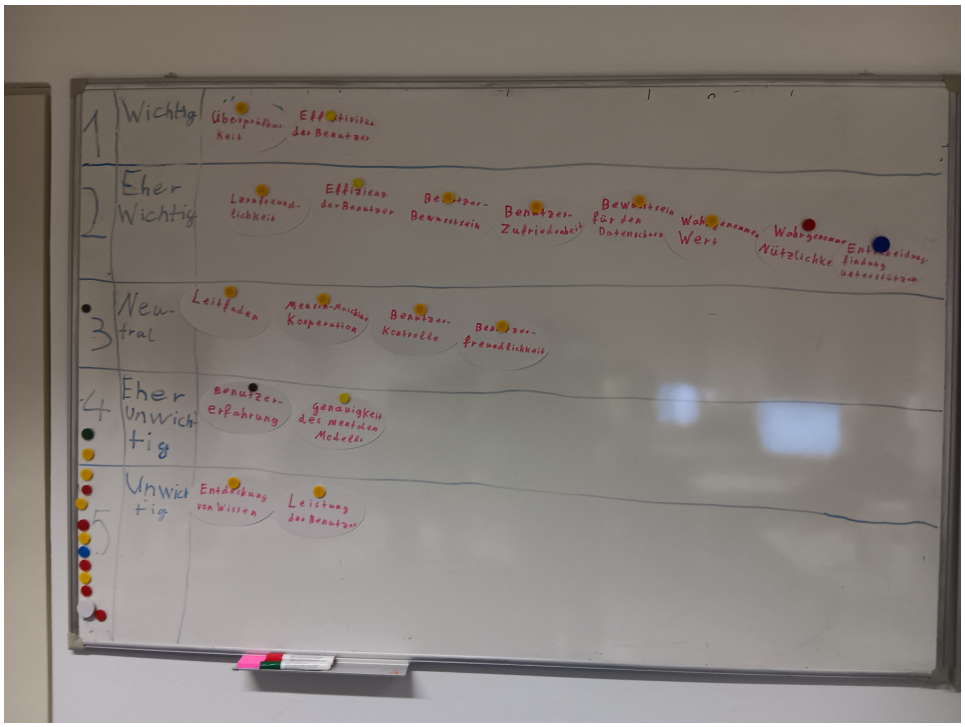

Abbildung B.2: Nichtfunktionale Anforderungen

Anhang C

Softwaretechnische Umsetzung

In Kapitel 6 ausgeschlossene Abbildungen werden in diesem Anhang ergänzend dargestellt. Die ersten beiden Abbildungen veranschaulichen die Ein- und Ausgabe des Stimmungsanalysetools ohne Erklärung. Daneben wird in den Abbildungen C.3 und C.4 die softwaretechnische Umsetzung eines Neutral- und Negativbeispiels dargestellt.

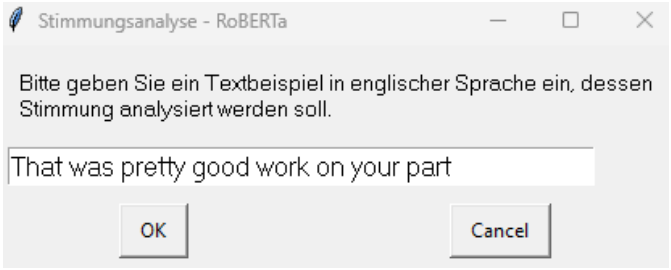

Abbildung C.1: Eingabe des Tools ohne Erklärung

Vorhersagen treffen

[5]:
```
from easygui import *

label_fn = lambda label: roberta.task.label_dictionary.string(
    [label + roberta.task.label_dictionary.nspecial]
)

information = "Bitte geben Sie ein Textbeispiel in englischer Sprache ein, dessen Stimmung analysiert werden soll."
title = "Stimmungsanalyse - RoBERTa"
d_text = "Hier eingeben.."

text = enterbox(information, title, d_text)

tokens = roberta.encode(text)
pred = label_fn(roberta.predict('se_sentiment_analysis_head', tokens).argmax().item())
print(pred)
```

```
2
```

Abbildung C.2: Ausgabe des Tools ohne Erklärung

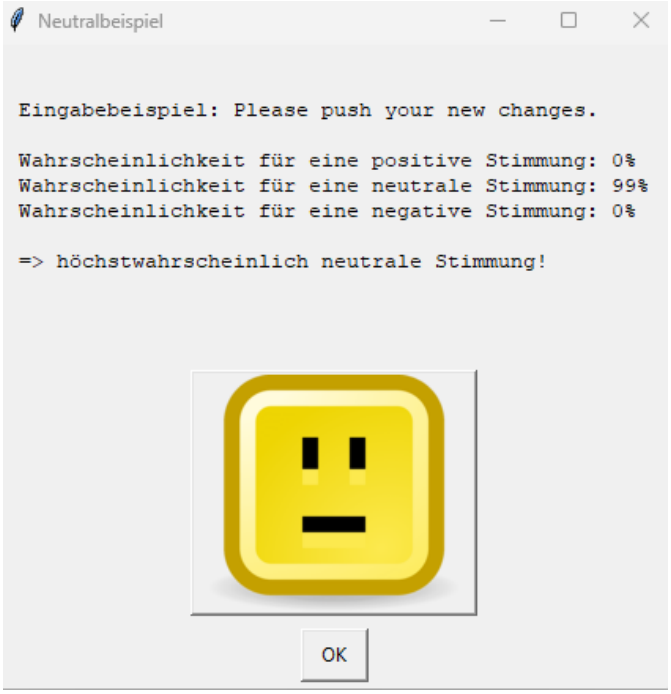

Abbildung C.3: Neutralbeispiel

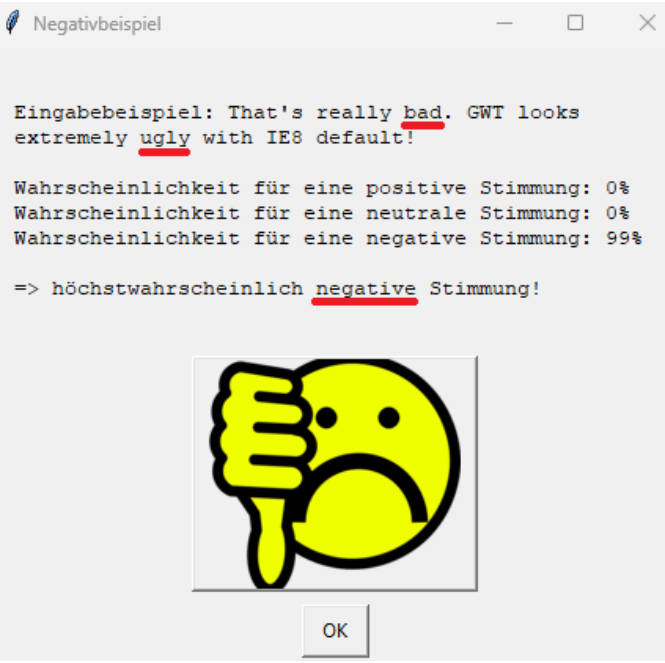

Abbildung C.4: Negativbeispiel

Anhang D

Fragebogen der Umfrage

Fragen	**Antwortmöglichkeiten**
Nutzerakzeptanz ohne Erklärung	
Vorstellung: Stimmungsanalysetool ohne Erklärung. Haben Sie dies verstanden?	Ja/Nein
Ich habe verstanden, wie das Tool funktioniert	Stimme völlig zu Stimme zu Stimme weder zu noch nicht zu Stimme nicht zu Stimme überhaupt nicht zu
Ich finde das Tool hilfreich	Stimme völlig zu Stimme zu Stimme weder zu noch nicht zu Stimme nicht zu Stimme überhaupt nicht zu
Ich könnte mir vorstellen das Tool zu verwenden	Stimme völlig zu Stimme zu Stimme weder zu noch nicht zu Stimme nicht zu Stimme überhaupt nicht zu
Nutzerakzeptanz mit Erklärung	
Für **jede Erklärung** *Beispiele*, *Schlüsselwörter*, *Größe des Trainingsdatensatzes*, *Genauigkeit des Modells*, *Grenzen der Software*, *Wahrscheinlichkeiten für die Stimmungen* und *Funktionsweise des Tools* **eine Durchführung**	
Vorstellung: Stimmungsanalysetool mit Erklärung. Haben Sie dies verstanden?	Ja/Nein

Ich finde die Erklärung verständlich	Stimme völlig zu Stimme zu Stimme weder zu noch nicht zu Stimme nicht zu Stimme überhaupt nicht zu
Ich finde die Erklärung hilfreich, um das Tool zu verstehen	Stimme völlig zu Stimme zu Stimme weder zu noch nicht zu Stimme nicht zu Stimme überhaupt nicht zu
Ich könnte mir vorstellen das Tool mit der Erklärung zu nutzen	Stimme völlig zu Stimme zu Stimme weder zu noch nicht zu Stimme nicht zu Stimme überhaupt nicht zu
Priorisierung der Erklärungen	
Erneute Vorstellung aller Erklärungen. Haben Sie dies verstanden?	Ja/Nein
Priorisierung nach der Wichtigkeit der Erklärungen	Reihenfolge
Demografische Daten	
Ich habe Erfahrungen in der Softwareentwicklung	Stimme voll zu Stimme eher zu Neutral Stimme eher nicht zu Stimme gar nicht zu
Ich habe Erfahrungen mit künstlicher Intelligenz	Stimme voll zu Stimme eher zu Neutral Stimme eher nicht zu Stimme gar nicht zu
Ich halte mich für technisch versiert	Stimme voll zu Stimme eher zu Neutral Stimme eher nicht zu Stimme gar nicht zu
Haben Sie schon einmal an einem Softwareprojekt teilgenommen?	Ja/Nein
Bitte geben Sie ihr Geschlecht an	männlich/weiblich/divers
Bitte geben Sie ihr Alter an	18 – 99
Sonstiges	
Feedback	Offene Antwortmöglichkeit

Tabelle D.1: Fragebogen der Umfrage

Anhang E

Auswertung der Umfrageergebnisse

In diesem Anhang werden die in Kapitel 7 ausgeschlossenen Tabellen ergänzend dargestellt. Die Tabellen E.1 und E.2 dokumentieren die Ergebnisse des Spearman-Tests bezüglich eines möglichen Zusammenhangs zwischen Tool nutzen und Erfahrung im SE bzw. technisch versiert sein.

	Korrelations-koeffizient	**p-Wert**	**Statistische Signifikanz**
Ohne Erklärung	0.24751	0.19550	Nein
Beispiele	-0.25817	0.16837	Nein
Schlüsselwörter	-0.01271	0.94687	Nein
Größe des Trainings-datensatzes	0.34105	0.07021	Nein
Genauigkeit des Modells	-0.18944	0.32499	Nein
Grenzen der Software	0.14033	0.45951	Nein
Wahrscheinlichkeiten für die Stimmungen	0.14615	0.44093	Nein
Funktionsweise des Tools	-0.09864	0.61070	Nein

Tabelle E.1: Zusammenhang zwischen technisch versiert und Tool nutzen (Ergebnisse des Spearman-Tests)

	Korrelations-koeffizient	**p-Wert**	**Statistische Signifikanz**
Ohne Erklärung	0.22919	0.23172	Nein
Beispiele	0.03162	0.86827	Nein
Schlüsselwörter	-0.15746	0.40596	Nein
Größe des Trainings-datensatzes	0.11966	0.53641	Nein
Genauigkeit des Modells	-0.07145	0.71263	Nein
Grenzen der Software	0.06040	0.75119	Nein
Wahrscheinlichkeiten für die Stimmungen	0.01497	0.93742	Nein
Funktionsweise des Tools	0.18129	0.34662	Nein

Tabelle E.2: Zusammenhang zwischen Erfahrung im SE und Tool nutzen (Ergebnisse des Spearman-Tests)

Anhang F

Inhalte auf dem USB-Stick

Der dieser Arbeit beigefügte USB-Stick beinhaltet die folgenden Ordner und Dateien:

- Eine *PDF*-Datei von diesem Dokument
- Das zur Erstellung von diesem Dokument verwendete *LaTeX*-Archiv
- Der Ordner „Literatur“ beinhaltet die Ergebnisse der Literaturrecherche und die dabei berücksichtigten Studien
- Der Ordner „RoBERTa“ beinhaltet u. a. die folgenden Dateien:
  - Die Implementierung des Stimmungsanalysetools *RoBERTa* (*JupyterLab*-Dokument und *PDF*-Version)
  - Das trainierte Modell von *RoBERTa*
  - Die zur Durchführung des Workshops modifizierte Version von *RoBERTa* (*JupyterLab*-Dokument und *PDF*-Version)
  - Die modifizierte Version von *RoBERTa* mit den hinzugefügten Erklärungen (*JupyterLab*-Dokument und *PDF*-Version)
- Der Fragebogen der Umfrage im *PDF*-Format
- Der Datensatz der Umfrage im *CSV*-Format
- Der Ordner „Umfrageanalyse“ beinhaltet die folgenden Dateien:
  - Der Wilcoxon-Test im *R*-Format
  - Die restliche Umfrageanalyse als *JupyterLab*-Dokument
  - Ergebnisse der Umfrageanalyse im *PDF*-Format

Literaturverzeichnis

[1] M. E. M. Abo, R. G. Raj, A. Qazi, and A. Zakari. Sentiment analysis for arabic in social media network: A systematic mapping study. *arXiv preprint arXiv:1911.05483*, 2019.

[2] A. Adadi and M. Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

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

[4] L. Arbelaez Ossa, G. Starke, G. Lorenzini, J. E. Vogt, D. M. Shaw, and B. S. Elger. Re-focusing explainability in medicine. *Digital health*, 8:20552076221074488, 2022.

[5] N. Balasubramaniam, M. Kauppinen, K. Hiekkanen, and S. Kujala. Transparency and explainability of ai systems: ethical guidelines in practice. In *International Working Conference on Requirements Engineering: Foundation for Software Quality*, pages 3–18. Springer, 2022.

[6] K. Baum, S. Mantel, E. Schmidt, and T. Speith. From responsibility to reason-giving explainable artificial intelligence. *Philosophy & Technology*, 35(1):12, 2022.

[7] E. Biswas, K. Vijay-Shanker, and L. Pollock. Exploring word embedding techniques to improve sentiment analysis of software engineering texts. In *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*, pages 68–78. IEEE, 2019.

[8] S. Cagnoni, L. Cozzini, G. Lombardo, M. Mordonini, A. Poggi, and M. Tomaiuolo. Emotion-based analysis of programming languages on stack overflow. *ICT Express*, 6(3):238–242, 2020.

[9] 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*, pages 128–128, 2018.

[10] F. Calefato, F. Lanubile, and N. Novielli. Emotxt: a toolkit for emotion recognition from text. In *2017 seventh international conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 79–80. IEEE, 2017.

[11] G. Catolino, F. Palomba, D. A. Tamburri, A. Serebrenik, and F. Ferrucci. Gender diversity and women in software teams: How do they affect community smells? In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*, pages 11–20. IEEE, 2019.

[12] L. Chazette, W. Brunotte, and T. Speith. Exploring explainability: a definition, a model, and a knowledge catalogue. In *2021 IEEE 29th international requirements engineering conference (RE)*, pages 197–208. IEEE, 2021.

[13] L. Chazette, W. Brunotte, and T. Speith. Explainable software systems: from requirements analysis to system evaluation. *Requirements Engineering*, 27(4):457–487, 2022.

[14] L. Chazette, V. Klös, F. Herzog, and K. Schneider. Requirements on explanations: a quality framework for explainability. In *2022 IEEE 30th International Requirements Engineering Conference (RE)*, pages 140–152. IEEE, 2022.

[15] L. Chazette, J. Klünder, M. Balci, and K. Schneider. How can we develop explainable systems? insights from a literature review and an interview study. In *Proceedings of the International Conference on Software and System Processes and International Conference on Global Software Engineering*, pages 1–12, 2022.

[16] L. Chazette and K. Schneider. Explainability as a non-functional requirement: challenges and recommendations. *Requirements Engineering*, 25(4):493–514, 2020.

[17] Z. Chen, Y. Cao, X. Lu, Q. Mei, and X. Liu. Sentimoji: an emoji-powered learning approach for sentiment analysis in software engineering. In *Proceedings of the 2019 27th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering*, pages 841–852, 2019.

[18] J. Cohen. Statistical power analysis. *Current directions in psychological science*, 1(3):98–101, 1992.

[19] B. Crook, M. Schlüter, and T. Speith. Revisiting the performance-explainability trade-off in explainable artificial intelligence (xai). In *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, pages 316–324. IEEE, 2023.

[20] A. P. Dempster and M. Schatzoff. Expected significance level as a sensitivity index for test statistics. *Journal of the American Statistical Association*, 60(310):420–436, 1965.

[21] M. D. Devika, C. Sunitha, and A. Ganesh. Sentiment analysis: a comparative study on different approaches. *Procedia Computer Science*, 87:44–49, 2016.

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

[23] R. Dhakad and L. Benedicenti. Analyzing emotional contagion in commit messages of open-source software repositories. In *CS & IT Conference Proceedings*, volume 13. CS & IT Conference Proceedings, 2023.

[24] J. Ding, H. Sun, X. Wang, and X. Liu. Entity-level sentiment analysis of issue comments. In *Proceedings of the 3rd International Workshop on Emotion Awareness in Software Engineering*, pages 7–13, 2018.

[25] J. Droste, H. Deters, J. Puglisi, and J. Klünder. Designing end-user personas for explainability requirements using mixed methods research. In *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, pages 129–135. IEEE, 2023.

[26] N. J. Herkenhoff. Entwicklung eines assistenten zur verbesserung der textbasierten kommunikation in entwicklungsteams. *Masterarbeit, Leibniz Universität Hannover*, 2023.

[27] M. Herrmann and J. Klünder. From textual to verbal communication: towards applying sentiment analysis to a software project meeting. In *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, pages 371–376. IEEE, 2021.

[28] 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.

[29] M. Herrmann, M. Obaidi, and J. Klünder. Senti-analyzer: joint sentiment analysis for text-based and verbal communication in software projects. *arXiv preprint arXiv:2206.10993*, 2022.

[30] M. Hind, D. Wei, M. Campbell, N. C. Codella, A. Dhurandhar, A. Mojsilović, K. Natesan Ramamurthy, and K. R. Varshney. Ted: Teaching ai to explain its decisions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 123–129, 2019.

[31] 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.

[32] L. Kästner, M. Langer, V. Lazar, A. Schomäcker, T. Speith, and S. Sterz. On the relation of trust and explainability: Why to engineer for trustworthiness. In *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, pages 169–175. IEEE, 2021.

[33] R. Kaur, K. K. Chahal, and M. Saini. Analysis of factors influencing developers' sentiments in commit logs: Insights from applying sentiment analysis. *e-Informatica Software Engineering Journal*, 16(1):220102, 2022.

[34] J. Klünder and O. Karras. Meetings and mood–related or not? insights from student software projects. In *Proceedings of the 16th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, pages 148–158, 2022.

[35] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. E. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. B. Hamrick, J. Grout, S. Corlay, et al. Jupyter notebooks-a publishing format for reproducible computational workflows. *Elpub*, 2016:87–90, 2016.

[36] M. A. Köhl, K. Baum, M. Langer, D. Oster, T. Speith, and D. Bohlender. Explainability as a non-functional requirement. In *2019 IEEE 27th International Requirements Engineering Conference (RE)*, pages 363–368. IEEE, 2019.

[37] J. Krithikadatta. Normal distribution. *Journal of Conservative Dentistry and Endodontics*, 17(1):96–97, 2014.

[38] M. Langer, K. Baum, K. Hartmann, S. Hessel, T. Speith, and J. Wahl. Explainability auditing for intelligent systems: a rationale for multi-disciplinary perspectives. In *2021 IEEE 29th international requirements engineering conference workshops (REW)*, pages 164–168. IEEE, 2021.

[39] W. Liao, B. Zeng, X. Yin, and P. Wei. An improved aspect-category sentiment analysis model for text sentiment analysis based on roberta. *Applied Intelligence*, 51:3522–3533, 2021.

[40] B. Lin, N. Cassee, A. Serebrenik, G. Bavota, N. Novielli, and M. Lanza. Opinion mining for software development: a systematic literature review. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 31(3):1–41, 2022.

[41] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[42] D. Mairiza and D. Zowghi. Constructing a catalogue of conflicts among non-functional requirements. In *Evaluation of Novel Approaches to Software Engineering: 5th International Conference, ENASE 2010, Athens, Greece, July 22-24, 2010, Revised Selected Papers 5*, pages 31–44. Springer, 2011.

[43] S. Mann, B. Crook, L. Kästner, A. Schomäcker, and T. Speith. Sources of opacity in computer systems: Towards a comprehensive taxonomy. In *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, pages 337–342. IEEE, 2023.

[44] M. G. Moghadam. Automatisierter vorschlag für ein stimmungsanalysetool basierend auf nutzerangaben zu einem entwicklerteam. *Bachelorarbeit, Leibniz Universität Hannover*, 2023.

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

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

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

[48] Z. Obeidi. Automatisierte erkennung von destruktiven äußerungen in meetings von softwareprojekten. *Masterarbeit, Leibniz Universität Hannover*, 2021.

[49] R. Ochsner. Erstellung eines deutschen datensatzes zur stimmungsanalyse von entwickleraussagen. *Bachelorarbeit, Leibniz Universität Hannover*, 2022.

[50] T. Olsen. Entwicklung einer grafischen benutzeroberfläche zur analyse von stimmungen in entwicklungsteams. *Bachelorarbeit, Leibniz Universität Hannover*, 2022.

[51] W. Pieters. Explanation and trust: what to tell the user in security and ai? *Ethics and information technology*, 13:53–64, 2011.

[52] A. Rakow. Analyse der verwendung von ironie in natürlicher sprache in softwareprojekten. *Bachelorarbeit, Leibniz Universität Hannover*, 2021.

[53] A. Rosenfeld and A. Richardson. Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems*, 33:673–705, 2019.

[54] M. C. Schierholz. Automatische erkennung von ironie in der kommunikation von softwareentwicklungsteams. *Masterarbeit, Leibniz Universität Hannover*, 2022.

[55] P. Schober, C. Boer, and L. A. Schwarte. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768, 2018.

[56] L. Schroth, M. Obaidi, A. Specht, and J. Klünder. On the potentials of realtime sentiment analysis on text-based communication in software projects. In *International Conference on Human-Centred Software Engineering*, pages 90–109. Springer, 2022.

[57] K. S. TARGIEL. The impact of the covid-19 pandemic on the level of sentiment in it projects implemented in the open source formula. *Scientific Papers of Silesian University of Technology. Organization & Management/Zeszyty Naukowe Politechniki Slaskiej. Seria Organizacji i Zarzadzanie*, (163), 2022.

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

[59] S. Thiebes, S. Lins, and A. Sunyaev. Trustworthy artificial intelligence. *Electronic Markets*, 31:447–464, 2021.

[60] F. Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution*, pages 196–202. Springer, 1992.

[61] J. P. Winkler and A. Vogelsang. "what does my classifier learn?" a visual approach to understanding natural language text classifiers. In *Natural Language Processing and Information Systems: 22nd International Conference on Applications of Natural Language to Information Systems, NLDB 2017, Liège, Belgium, June 21-23, 2017, Proceedings 22*, pages 468–479. Springer, 2017.

[62] C. Wohlin. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th international conference on evaluation and assessment in software engineering*, pages 1–10, 2014.

[63] J. F. Wolfswinkel, E. Furtmueller, and C. P. Wilderom. Using grounded theory as a method for rigorously reviewing literature. *European journal of information systems*, 22(1):45–55, 2013.

[64] C. Zucco, B. Calabrese, G. Agapito, P. H. Guzzi, and M. Cannataro. Sentiment analysis for mining texts and social networks data: Methods and tools. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(1):e1333, 2020.