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

Untersuchung von Zusammenhängen zwischen Stimmungen und Interaktionen von Meetings in Softwareprojekten

Bachelorarbeit

im Studiengang Informatik

von

David Zafer Jörg Schiller

Prüfer: Prof. Dr. Kurt Schneider
Zweitprüfer: Dr. Jil Klünder
Betreuer: Martin Obaidi

Hannover, 27.08.2021

Erklärung der Selbstständigkeit

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

Hannover, den 27.08.2021

David Zafer Jörg Schiller

Kurzfassung

Meetings sind ein wichtiger Bestandteil der Arbeitswelt und tragen zum Erfolg von Projekten teil. Dafür ist es wichtig, dass diese positiv ablaufen, da das andernfalls negative Auswirkungen auf die Leistungen der Teilnehmer und auf das Ergebnis des Projektes haben kann. Um zu überprüfen, ob diese Meetings gut ablaufen, gibt es act4teams und act4teams-SHORT. Mit diesen Methoden lassen sich die Meetings auf ihre Interaktionen untersuchen und somit die Interaktionsverteilung nachvollziehen. Auf Grundlage dieser Verteilung lassen sich eventuelle Probleme im Ablauf des Meetings finden und rechtzeitig beheben.
Auch das Nutzen von Sentiment-Analyse-Tools ist eine Möglichkeit, ein Meeting zu analysieren. Dabei handelt es sich um eine Untersuchung einzelner Texte auf ihre Sentimente, welche die Stimmung des Textes wiedergeben, meist in Form einer positiv, neutral oder negativ Kategorisierung. Das Meeting wird transkribiert und einem Stimmungsanalyse-Tool klassifiziert. Mit diesen Ergebnissen ist es ebenfalls möglich, den Verlauf eines Meetings zu erkennen und dort entsprechend rechtzeitig einzugreifen, falls etwas nicht wie gewünscht verläuft. Es muss allerdings darauf geachtet werden, dass die richtigen Werkzeuge für dieser Analyse benutzt werden, um eine bestmögliche Auswertung der Daten zu erhalten. Dazu werden die einzelnen Tools in der gewünschten Domäne der Softwareprojektmeetings geprüft und es wird überprüft, ob diese auf dem Datensatz einzeln oder in Kombination zusammen besser arbeiten.
In dieser Arbeit wird mit einem deutschen Datensatz aus Meetings von Softwareprojekten gearbeitet. Dieser ist mit Labeln aus den Kategorien von act4teams-SHORT versehen und wird mit Stimmungsanalyse-Tools untersucht. Anschließend werden die Resultate in einen Zusammenhang gesetzt. Dabei wird geprüft, worin diese Zusammenhänge gegebenenfalls bestehen.

Abstract

Meetings are an important part of the working world and contribute to the success of projects. Therefore it is important that they run positively, otherwise this can have a negative impact on the performance of the participants and on the outcome of the project. In order to check whether these meetings run well, act4teams and act4teams-SHORT are available. With these methods, the meetings can be examined for their interactions and thus the interaction distribution can be traced. Based on this distribution, any problems in the course of the meeting can be found and corrected in time.
Using sentiment analysis tools is another way to analyze a meeting. This involves examining individual texts for their sentiments, which reflects the mood of the text, usually in the form of a positive, neutral or negative categorization. The meeting is transcribed and classified to a sentiment analysis tool. With these results it is also possible to recognize the course of a meeting and to intervene there accordingly in time, if something does not go as desired. However, care must be taken to ensure that the right tools are used for this analysis in order to obtain the best possible evaluation of the data. For this purpose, the individual tools are examined in the desired domain of the software project meetings and it is checked whether they work better on the data set individually or in combination together.
In this work, we work with a German dataset of software project meetings. This is tagged with labels from the categories of act4teams-SHORT and is examined with sentiment analysis tools. Subsequently, the results are put into context. In doing so, it will be examined what these correlations are, if there are any.

Inhaltsverzeichnis

Kapitel 1

Einleitung

In der Softwareentwicklung spielen Meetings eine wesentliche Rolle [6]. Dabei ist die Effektivität des Meetings meist abhängig von den einzelnen Teilnehmern [6]. Negatives Verhalten der Teilnehmer kann die Effizienz des Meetings mindern [6]. Daher ist es wichtig, dass die Teambesprechungen möglichst effektiv abgehalten werden, damit das Projekt auf die bestmögliche Weise abgeschlossen werden kann [5]. Da die Softwareentwicklung stark von Interaktionen und menschlichen Anstrengungen abhängt, ist diese für die Emotionen der Teilnehmer abhängig [4]. Somit ist ein gutes Verständnis für die Emotionen und Interaktionen der Meetingteilnehmer für die effektive Zusammenarbeit von Vorteil, weil dies genutzt werden kann, um eine höhere Produktivität zu erreichen [4]. Um dieses Ziel zu erreichen, ist es hilfreich diese Meetings zu analysieren und die Ergebnisse auszuwerten. Dadurch können negative Einflüsse früh erkannt werden und es ist möglich, diese Besprechung durch geeignetes Entgegensteuern wieder auf Kurs zu bringen [6].
Eine Möglichkeit, Teambesprechungen in Softwareprojekten zu analysieren, bietet die Forschung von Kauffeld et al. [5], in welcher das Verfahren des *act4teams* vorgestellt wurde. Mit diesem Modell lassen sich die Interaktionen von Besprechungen in 44 Kategorien einteilen [5]. Da das allerdings ein aufwendiger Prozess ist, gibt es von Klünder et al. [6] das Modell von *act4teams-SHORT*, welches die Interaktionen in 11 Kategorien einteilt. Dieses wurde auf Grundlage von *act4teams* entwickelt und ermöglicht eine leichtere Anwendung. Dabei ist lediglich ein Beobachter während einer Besprechung notwendig, welcher das Meeting beobachtet und in die Kategorien einteilt [6]. Für act4teams wird geschultes Personal zur Durchführung benötigt. Diese transkribieren auf Grundlage einer Videoaufnahme das Meeting und Kategorisieren mit Hilfe diesem die einzelnen Aussagen [5]. Bei act4teams-SHORT dagegen, kann diese Kategorisierung mit weniger Erfahrung, direkt während des Meetings vorgenommen werden [7]. Mit Hilfe dieser Daten können die können die Interaktionen während eines Meetings analysiert werden [6] [7].
Eine weitere Möglichkeit die Besprechungen zu analysieren bietet die Methode der Stimmungsanalyse mit Hilfe von Sentiment-Analyse-Tools. Mit diesen Tools lassen sich Aussagen bezüglich ihrer Emotionen klassifizieren. Die Tools prüfen dabei die Dokumente auf Aussagen und klassifizieren diese. Die Aussagen werden somit in Sentimente klassifiziert, wie beispielsweise positiv, neutral und negativ [16].

Die Untersuchung von Teammeetings auf ihre Stimmung und Interaktion soll als Grundlage dienen, um einen Zusammenhang zwischen diesen zwei Aspekten zu finden und somit Rückschlüsse von den Interaktionen der Teilnehmer eines Meetings auf die Stimmung der Teilnehmer schließen zu können oder auch umgekehrt von der Stimmung einer Aussage auf die vorhandenen Interaktionen.

1.1 Problemstellung

Teammetings nehmen einen großen Platz in der Arbeitswelt ein. Die meisten Arbeitnehmer arbeiten in einer Teamumgebung und kommen somit auch mit Teammeetings in Kontakt [15]. Auch in der Softwareentwicklung spielen Teammeetings eine wichtige Rolle. Dabei ist die Kommunikation und die Koordination für die Effizienz bei einem Softwareprojekt entscheidend. Mitarbeiter erledigen dort ihre Arbeit und erfahren bei diesen auch Emotionen [15]. Ein negatives Beispiel für einen solchen Fall, bei dem das Meeting negativ beeinflusst werden kann, wären Mitarbeiter, die öfter dazwischen Reden und somit den Redefluss stören und unterbrechen. Dies kann die Stimmung von anderen Teilnehmern negativ beeinflussen [6].
Um negative Einflüsse in Softwaremeetings früh zu erkennen, gibt es verschiedene Methoden, diese zu untersuchen. Zum einen gibt es die Möglichkeit, die Interaktionen innerhalb eines Teammeetings zu untersuchen. Ein Verfahren um solche zu betrachten, ist die Verwendung von act4teams beziehungsweise in reduzierter Form act4teams-SHORT. Dadurch ist es möglich die einzelnen Interaktionen der Teilnehmer in Meetings zu Kategorisieren [5] [6]. Mit Hilfe dieser Kategorien ist es möglich, ein Meeting auszuwerten und Rückschlüsse über den Ablauf des Meetings zu ziehen. Somit ergibt sich ein Überblick über die Interaktionsverteilung, an welcher gesehen werden kann, ob ein Meeting proaktiv oder kontraproduktiv abläuft. Entsprechend dieser Verteilung kann ein Meeting gegebenenfalls wieder in die richtige Spur gebracht werden [15]. Eine weitere Möglichkeit ist es, die Teammeetings zu transkribieren und die Stimmung der Teilnehmer durch ihre Aussagen zu untersuchen. Erreicht werden kann das durch die Verwendung von Sentiment-Analyse-Tools. Diese nutzen die Aussagen der Teilnehmer und analysieren sie auf ihre Emotionen [15]. Wenn in einem Meeting über Probleme gesprochen wird, ohne dass Lösungsvorschläge genannt oder diskutiert werden, kann das schlechte Laune bei den Mitarbeitern auslösen [6]. Die Aussagen eines Meetings werden im Anschluss an dieses ausgewertet. Dafür wird ein Sentiment-Analyse-Tool genutzt, mit welchem Aussagen bezüglich ihrer Sentimente eingeteilt werden können. [16]. Somit kann von jeder Aussage das Sentiment bestimmt werden. Durch das Bestimmen der Stimmung der Aussagen von den Mitarbeitern, ist es möglich, auf die Stimmung während des Meetings Rückschlüsse zu ziehen [16]. Dies ist relevant, da negative Emotionen die Wahrnehmung beeinflussen können und so auch einen Einfluss auf die Leistung der Meetingteilnehmer haben [15].
Die Untersuchung der Interaktionen und der Stimmung in einem Meeting kann helfen Rückschlüsse über den Verlauf eines Meetings zu ziehen. Dabei stellt sich die Forschungsfrage:

Forschungsfrage: Gibt es Zusammenhänge zwischen den Interaktionen in einem Softwareprojektmeeting und den Stimmungen dieser Interaktionen und wo liegen die gegebenenfalls?

1.2 Lösungsansatz

Um eine Lösung für dieses Problem zu finden können die einzelnen Aspekte der Frage untersucht werden. Zum einen kann die Interaktion von Teammeetings in Softwareprojekten betrachtet werden. Zum Betrachten der Interaktionen wird auf einen Datensatz zurück gegriffen, welcher bereits mit act4teams Kategorien gelabelt ist. Im Rahmen der Dissertation von Klünder [7] wurde ein Verfahren entwickelt, mit welchem es möglich ist, act4teams zu act4teams-SHORT zu mappen. Dafür werden die einzelnen Kategorien aus act4teams zu den Kategorien aus act4teams-SHORT gemappt. Das genau Vorgehen wird in Kapitel 3.3 genauer beschriebenen. Den Datensatz selbst mit Labeln aus act4teams zu versehen ist nicht möglich, da es sich dabei um einen zeitlich aufwendigen Prozess handelt [6] und auch Erfahrungen und Einweisungen für das Labeln von act4teams notwendig sind [7]. Mit diesem Datensatz können nun Rückschlüsse auf den Interaktionsverteilung der Meetings gezogen werden, da diese mit Hilfe der act4teams-Kategorien bestimmt werden können.
Da der Datensatz einen großen Einfluss auf die Ergebnisse hat, muss darauf geachtet werden, dass dieser bestmöglich für die geplante Anwendung geeignet ist. Um das zu erreichen, wird der gleiche Datensatz für die Stimmungsanalyse angepasst. Die Anpassung wird durchgeführt, indem der Datensatz von der digitalen Formatierung angepasst wird und zum anderen durch eine Anpassung der inhaltlichen Form der Daten. Dabei ist darauf zu achten, dass nicht der Inhalt der Aussagen geändert wird, um das vom Tool klassifizierte Sentiment nicht zu beeinflussen. Die Daten müssen am Ende ein Format haben, welches nur die Aussagen der Meetingteilnehmer enthalten, da die Stimmung der Teilnehmer betrachtet werden soll. Die Beschreibung einer Aktion zum Beispiel, spiegelt nicht die Stimmung einer Aussage wieder. Auch muss darauf geachtet werden, dass der Datensatz möglichst frei von Rechtschreibfehlern ist, um eine bessere Leistung der Sentiment-Analyse-Tools zu gewährleisten.
Nach der Anpassung des Datensatzes wird eine Teilmenge bezüglich der Stimmungen der einzelnen Aussagen von diesem gelabelt. Diese Teilmenge wird anschließend auf verschiedene Sentiment-Analyse-Tools angewendet. Dabei werden bei jedem Tool die Ergebnisse auf Grundlage bekannter Metriken ausgewertet. Dies dient dazu, die beste Wahl eines Tools für die Domäne der Softwareprojekt-Meetings zu finden. Zum Schluss können die gewählten Tools genutzt werden, um den kompletten Datensatz zu klassifizieren. Diese Ergebnisse können dann in Zusammenhang mit den vorhandenen act4teams-SHORT Kategorien analysiert und ausgewertet werden. Dadurch soll am Ende festgestellt werden können, ob sich Zusammenhänge zwischen den Interaktionen und Stimmungen von Teammeetings bei Softwareprojekten feststellen lassen und wo diese gegebenenfalls liegen.

1.3 Struktur der Arbeit

Zunächst wird in Kapitel 2 auf die Grundlagen die zum Verständnis der Arbeit notwendig sind eingegangen. Weiterführend werden die Grundlagen von act4teams erklärt und wie dieses mit act4teams-SHORT zusammen hängt. Anschließend wird auf die Sentiment-Analyse und auf die Auswahl der hier verwendeten Tools eingegangen. Des weiteren werden die Messwerte und der Aufbau des Voting Classifier zur Weiterverarbeitung der Tools erklärt.
In Kapitel 3 wird das grundlegende Konzept der Arbeit erläutert. Dabei wird die Änderung von act4teams zu act4teams-SHORT behandelt. Darüber hinaus wird erklärt, wie der gegebene Datensatz angepasst wird und welche Tools ausgewählt werden. Zudem wird noch die Verwendung und Auswahl der ausgewählten Sentiment-Analyse-Tools erläutert und wie die Daten zusammen verarbeitet werden.
Über den Umgang mit den Ergebnissen wird in Kapitel 4 näher eingegangen. Insbesondere wird genauer erklärt, wie mit den Resultaten nach der Auswertung verfahren wird. Hier werden die gesammelten Ergebnisse entsprechend analysiert und in Zusammenhang gesetzt. Auf Grundlage dieser Resultate werden anschließend Ergebnisse aus diesen geschlussfolgert.
Kapitel 5 beschäftigt sich mit der Auseinandersetzung des gesamten Verfahrens, sowie die gewonnene Erkenntnis und nennt mögliche Verbesserungen, die vorgenommen werden können. Danach wird in Kapitel 6 auf verwandte Arbeiten eingegangen und wie sich diese von dieser Arbeit hier abgrenzen. Zur Abgrenzung wird dabei auf den Hauptteil der Arbeit eingegangen. Zum Schluss werden in Kapitel 7 alle Ergebnisse zusammengefasst und ein Ausblick auf zukünftige Anwendungsgebiete gegeben, die auf Grundlage dieser Arbeit möglich sind.

Kapitel 2

Grundlagen

In dieser Arbeit werden die Zusammenhänge zwischen Interaktionen und Stimmungen in Meetings von Softwareprojekten untersucht. Um zu verstehen, wo die Zusammenhänge dabei liegen, muss erst mal geklärt werden, worum es sich bei Interaktionen und Stimmung handelt. Im Nachfolgenden werden die Grundlagen erläutert, die zum Verständnis der Arbeit notwendig sind. Um mit den Interaktionen von Menschen arbeiten zu können, werden in Abschnitt 2.1 die Grundlagen von act4teams behandelt, mit welchen die Tätigkeiten von Personen in einem Meeting kategorisiert werden können [5]. Anschließend wird in Abschnitt 2.2 die Sentiment-Analyse erklärt, mit welcher es möglich ist, Aussagen in schriftlicher Form von Personen bezüglich ihrer Sentiments zu klassifizieren [16]. Dafür werden noch die Funktionsweisen der benutzten Tools vorgestellt, welche genutzt werden, um die Sentimente der Datensätze zu bestimmen. Um die daraus erhaltenen Informationen auswerten zu können, werden in Abschnitt 2.4 grundlegende Methoden zur Berechnung von Metriken erklärt. Im Anschluss werden noch in Abschnitt 2.5 die Python Voraussetzungen und der Voting-Classifier erklärt, welche notwendig sind, um die Sentiment-Analyse-Tools nutzen und auswerten zu können.

2.1 act4teams

Kaufeld et al. [5] stellen in ihrer Arbeit die Methode *act4teams* zur Analyse von Interaktionen vor [5]. Bei act4teams handelt es sich um eine Untersuchung von Meetings. Dabei wird die Zusammenarbeit von Teams untersucht, indem zwischenmenschliche Interaktionen und Prozesse betrachtet werden [7]. Entwickelt wurde act4teams an der Technischen Universität Braunschweig und steht mit der Abkürzung für *advanced interaction anaylsis for teams*® [5]. Es handelt sich um eine Methode, mit welcher auf Video aufgenommene Meetings analysiert werden können. Diese Methode misst dabei die Teamkompetenz, welche die Teamfähigkeiten beschreibt, die notwendig sind, um Aufgaben in einem Team erfolgreich zu bewältigen [**?**]. Somit ermöglicht act4teams eine Analyse der Teamkompetenz eines Teams bezüglich ihrer Stärken und Schwächen [7]. Um dies umzusetzen, werden die Aussagen aller Teilnehmer ausgewertet und in eine von vier Kompetenzbereichen aus act4Teams eingeordnet [7] [5].

2.1.1 Kategorisierung

Die vier Kompetenzbereichen von act4teams sind wie folgt gegliedert:

Problem fokussierte Kommunikation *(Professionelle Kompetenz)*:
Dabei handelt es sich um Aussagen, die direkt im Zusammenhang mit einem Problem stehen und dieses ausfzeigen. Auch Aussagen, die sich mit der Lösungsfindung von solchen Problemen befassen, gehören zu problemfokussierter Kommunikation.
Prozedurale Kommunikation *(Methodenkompetenz)*:
Bei dieser Art der Kommunikation geht es um die Strukturierung und Organisation des Meetings. Sowohl negative prozedurale Aussagen, als auch Aussagen, die sich in Details verlieren sind in diese Kategorie mit aufzunehmen.
Sozioemotionale Kommunikation *(Sozialkompetenz)*:
Dies sind Aussagen, die sich mit der zwischenmenschlichen Interaktionen befassen, die in den Teams stattfinden. Zu diesen gehören zum einen positive Aussagen, aber auch negative Aussagen wie Beleidigungen.
Aktionsorientierte Kommunikation *(Selbstkompetenz)*:
Aktionsorientierte Kommunikation beinhaltet Aussagen, die Bereitschaft eines Teams an Veränderung signalisieren. Dies betrifft Aussagen für Maßnahmenplanung, sowie Aussagen von Desinteresse über Veränderungen, Jammern und Schuldzuweisungen.

Diese vier Kompetenzbereiche lassen sich in funktionale und dysfunktionale Kommunikation aufteilen. Dabei handelt es sich bei funktionaler Kommunikation um positiv beitragende Aussagen und bei dysfunktionaler Kommunikation um negativ beitragende Aussagen. Problem fokussierte Kommunikation werden dabei als funktionale Kommunikation angesehen. Prozedurale, sozioemotionale und aktionsorientierte Kommunikation setzen sich zusammen aus funktionale und dysfunktionale Aussagen. Insgesamt lassen sich die vier Kompetenzbereiche in 12 Kompetenzaspekte gliedern. Diese 12 Kompetenzbereiche lassen sich somit in 44 Kategorien aufteilen [5]. Diese können in Form eines Codes den einzelnen Aussagen zugeordnet werden [7]. Die Aufteilung der einzelnen Kategorien kann im Anhang aus Tabelle A.1 entnommen werden.

2.1.2 Ablauf

Bei act4teams werden die Meetings auf Video aufgenommen und diese Videos werden anschließend transkribiert. Danach werden die einzelnen Aussagen in Sinneseinheiten unterteilt, welche ein einzelnes Wort, ein Teil eines Satzes oder auch ein ganzer Satz sein kann [7]. Diesen Einheiten werden anschließend einem Code zugewiesen aus einer der 44 act4teams Kategorien. Des weiteren werden den Meetingteilnehmern Buchstaben zu gewiesen, um die Sprecherfolge analysieren zu können [7]. Zusätlich zu den 44 Kategorien gibt es noch weitere Codes, welche den einzelnen Einheiten zugewiesen werden können. Dabei handelt es sich um Codes wie „Gemeinsames Lachen“, „Pause“ oder auch „unverständlich“ [7]. Die 44 Kategorien und ihre Codes sind im Anhang aus den Tabellen A.1,A.2, A.3 und A.4 zu entnehmen [7].

2.1.3 act4teams-SHORT

Neben act4teams haben Klünder et al. [6] in ihrer Arbeit die vereinfachte Methode *act4teams-SHORT* entwickelt. Das Ziel von act4teams-SHORT ist es, die Interaktion von Entwicklerteams in der Softwareentwicklung zu analysieren. Dabei soll dies vor allem in Echtzeit während des Meetings geschehen. Auch soll das Vorgehen dabei möglichst objektiv sein, indem es vom Beobachter unabhängig ist. Des weiteren sollen die Ergebnisse vergleichbar bleiben mit den Ergebnissen der act4teams Analyse und dabei soll keine Einarbeitungszeit für den Anwender dieses Verfahrens notwendig sein [7].
Aufgrund dieser Ziele kommt es zu einigen Einschränkungen bei der Anwendung von act4teams-SHORT. Zum einen werden bei der Anwendung nur Aussagen kodiert, die aus Sicht des Beobachters relevant sind. Somit gibt es eine selektive Kodiereung. Zum anderen werden nicht alle 44 Kategorien aus act4teams genutzt. Diese werden auf eine kleinere Anzahl reduziert. act4team-SHORT besteht aus 11 Kategorien, die 28 Kategorien aus act4teams abdecken. Es werden Kategorien teilweise zusammen gefasst oder eine act4teams Kategorie wird auf mehrere act4teams-SHORT Kategorien aufgeteilt. Somit ist act4teams nicht vollständig äquivalent zu act4teams-SHORT [7]. Um die übrigen Kategorien zuzuordnen, werden die übrigen Kategorien manuell zugeordnet. Wie dieses genau geschieht, wird in Kapitel 3.3 genauer beschrieben. Aus der nachfolgenden Tabelle, können die act4teams-SHORT Kategorien entnommen werden:

- **Problembenennung**: Nennen oder erläutern von Problemen
- **Problemvernetzung**: Analysieren von Ursachen und Folgen von Problemen
- **Lösungsbenennung**: Erläutern oder nennen von Lösungsvorschlägen
- **Lösungsvernetzung**: Analysieren von Lösungen oder Anforderungen von Folgen
- **Verknüpfung und Vernetzung**: Verknüpfung und Vernetzung von Problemen und Anforderungen
- **Destruktives Verhalten**: Tadeln, Lästern, Desintersse und störendes Verhalten
- **Methodisch-strukturierendesVerhalten**: Priorisieren von Aufgaben und Zusammenfassen oder wiederholen von Aussagen oder zum Thema zurückführen
- **Proaktives Verhalten**: Interesse an Veränderungen oder Planen von Maßnahmen und Aufgaben verteilen
- **Wissenstransfer**: Eigenes Wissen zu Informationen hinzugeben und dieses für Erklärungen nutzen
- **Informationsweitergabe**: Fakten und Informationen über Ereignisse mit dem Team teilen

- **Kollegiales Verhalten**: Andere einbinden und auf andere Beiträge eingehen, Wertschätzung oder Humor

Die Umwandlung der act4teams Kategorien zu act4teams-SHORT Kategorien kann im Anhang aus Tabelle A.1 entnommen werden [7].

2.2 Sentiment-Analyse

Sentiment-Analyse befasst sich in erster Linie mit dem Heraussuchen von Informationen aus gegebenen Texten, welche emotionsbezogen sind. [20]. Meist wird dabei auf die Erkennung von Subjektivität und die Polaritätserkennung abgezielt [18]. Bei der Erkennung von Subjektivität geht es um die Erkennung subjektiver Aussagen. Bei der Bestimmung der Polarität, wird meist versucht zu entscheiden, ob eine Aussage positiv, negativ oder neutral ist [18]. Einige Tools bestimmen auch die Stärke von Gefühlen innerhalb der Stimmung [18]. Für diese Arbeit wird die Erkennung der Polarität durch Klassifizierung in positiv, neutral und negativ genutzt. Aus einer Aussage werden Informationen zur Verarbeitung gesammelt und mit diesen wird eine Entscheidung getroffen was die Stimmung einer Aussage betrifft [18]. Mit dieser wird versucht, Rückschlüsse über die Stimmung die Person zu schließen, die diese Aussagen geäußert hat. [16]. Beispiele für die Klassifizierung von Aussagen in positiv, neutral oder negativ, können aus der Tabelle 2.1 entnommen werden.

Kommentare	Sentiment-Label
@DrabJay: excellent suggestion! Code changed. :-)	Positive
That really stinks! I was afraid of that...	Negative
A few but they all seem proprietary	Neutral

Tabelle 2.1: Beispiele für Klassifizierungen entnommen aus SentiStrength-SE [4]

Zwei der häufigsten Methoden in der Sentiment-Analyse basieren auf maschinellem Lernen oder auf einem lexikonbasiertem Verfahren [16]. Das lexikonbasierte Verfahren hat den Vorteil, dass es Wörterbücher nutzt, welche sich leicht in andere Domänen übertragen lassen [16]. Dabei beinhalten die Wörterbücher Wörter, welche ein Rating besitzen, auf dessen Grundlage die Polarität von Aussagen ermittelt werden kann [20]. Bei Domänen handelt es sich um die Themen oder Textgattungen, in welchen sich die vorhandenen Daten befinden [19]. Es muss darauf geachtet werden, was Wörter und Aussagen für unterschiedliche Domänen für Relevanz haben. Das Wörterbuch enthält je nach Domäne positive und negative vermerkte Wörter [16]. Beim maschinellen Lernen wird ein Klassifikator erstellt, der die Polarität von Aussagen auf Grundlage von gelernten Parametern bestimmt [16]. Dies hat den Vorteil, dass ein zum Beispiel positiv, neutral und negativ gelabelter Datensatz für eine Domäne genutzt werden kann, um leicht ein Tool mit diesem zu trainieren. Dabei ist dies auch ein Nachteil, denn wenn ein dTool für eine bestimmte Domäne trainiert wird, kann dieses nicht einfach übertragen werden [16]. Beide Methoden können gegebenen Aussagen

beispielsweise in positiv, negativ und gegebenenfalls auch in neutral einteilen [18]. Problematisch dabei ist, dass es für Wörter und Aussagen verschiedene Deutungen gibt, die abhängig vom Kontext sind [18]. Auch muss darauf geachtet werden, dass das Tool in anderen Domänen zu anderen Ergebnissen kommen kann, da sich dort auch der Kontext ändert [18].

2.3 Sentiment-Analyse-Tools

Bei Sentiment-Analyse-Tools handelt es sich um Tools, welche die Stimmungen von Aussagen ermitteln können. [18]. In den nachfolgenden Abschnitten werden die in der Arbeit hauptsächlich genutzten Sentiment-Analyse-Tools vorgestellt, sowie das deutsche Lexikon *SentiWS* [13]. Bei all diesen Tools handelt es sich um lexikonbasierte Tools. Diese nutzen Wörterbücher mit Wörtern, welche ein Rating zur Bestimmung der Polaritäten besitzen [20]. In Kapitel 3.4.1 wird genauer auf die Erläuterung der Auswahl der Sentiment-Analyse-Tools eingegangen.

2.3.1 SentiWS

Für die Sentiment-Analyse deutscher Texte stellen Remus et al. [13] in ihrer Arbeit das deutsche Wörterbuch *SentiWS* [13] vor. Dies beinhaltet positiv und negativ bewertete Wörter. Alle Wörter sind in ihrer Grundform enthalten und ebenso ihre zusätzliche Variationen der Formen im Plural für die Nomen und verschiedene Zeitformen für Verben [13]. Die Bewertungen der Grundformen können ohne Probleme auf ihre grammatikalischen Varianten übertragen werden. Daher kann es nicht dazu kommen, dass zum Beispiel Wörter unterschiedlicher Zeitformen nicht bewertet werden, da diese sonst unbekannt wären [20].
SentiWS nutzt mehrere Quellen sowie ihre semantische Ausrichtung. Eine der erste Quelle, die benutzt wurden, ist das „General Inquirer lexicon“, da dieses über eine große Lexikon-Reichweite, sowie breite Akzeptanz verfügt [13]. Es beinhaltet Wörter, die ins deutsche übersetzt und anschließend manuell bearbeitet wurden. In diesem Fall wurden Wörter verschoben, die unpassend waren oder ohne vorherige Polarität waren [13]. Die zweite Quelle ist die „Co-occurrence Analysis“. Hier erschließt sich eine besondere Art der Analyse von Produktbewertungen, bei dem die Verfasser der Bewertungen diese mit einem Hinweis versehen, ob diese stark positiv oder stark negativ sind. Das liefert wichtige Informationen für stimmungssentimentale Ausdrücke, die unter anderem für Produktbewertungen relevant sind. Die dritte Quelle ist das „German Collocation Dictionary“, die Wörter nach ihrer Ähnlichkeit gruppiert. [13] Dieses Wörterbuch ist ein weit verbreitetes deutsches Wörterbuch und wird in einigen Deutschen Sentiment-Analyse-Tools wie GerVADER verwendet [20].

2.3.2 GerVADER

In ihrer Arbeit stellen Tymann et al.[20] das deutsche Sentiment-Analyse-Tool *GerVADER* [20] vor. Dabei handelt es sich um ein Sentiment-Analyse-Tool, welches extra für die deutsche Sprache entwickelt wurde [20]. Zur Nutzung wurde das Wörterbuch SentiWS genutzt, welches eins für die deutsche Sprache

entwickeltes und angepasstes Wörterbuch ist [20]. Im Zuge der Entwicklung wird bei GerVADER eine Rechtschreibkorrektur hinzugefügt, um so mögliche falsch geschriebene Wörter zu erkennen und Fehler bei der Klassifizierung zu vermeiden. [20]. Auch wurde das Wörterbuch um weitere deutsche Wörter erweitert, durch die Aufnahme des Wörterbuches von dem Sentiment-Analyse-Tool VADER [3] [20].
Um die Funktionsfähigkeit von GerVADER zu verstehen, ist ein genauer Blick auf seinen englischen Vorgänger VADER notwendig [20]. Für die deutsche Sprache wurde GerVADER aus VADER entwickelt [20]. Das Sentiment-Analyse-Tool VADER wurde speziell für Social Media Plattformen entwickelt und verwendet dafür einen lexikonbasierten Ansatz. Das Sentiment-Analyse-Tool wurde für kleine Wortbereiche entwickelt, was zur Folge hat, dass längere und komplexere Texte falsch bewertet werden könnten [20].

2.3.3 SentiStrength

Das Analysetool *SentiStrength* wird von Thelwall et al.[19] in ihrer Arbeit vorgestellt. Dies ist ein Lexikon basierter Klassifizierer, der zusätzliche Informationen und Regeln verwendet, um in englischen Texte die Emotionen zu bewerten. Hierfür werden Noten auf einer Skala von -5 bis +5 vergeben, wofür -5 für äußerst negativ und +5 für äußerst positiv steht. Für die Bewertung werden zwei Skalen verwendet, da kurze Texte sowohl positive als auch negative Gefühle zeitgleich enthalten können[19]. SentiStrength bewertet dabei sowohl die einzelnen Wörter als auch die allgemeine Stimmung des gesamten Satzes [19].

2.3.4 TextBlobDe

Bei *textblob-de* [1] handelt es sich ein Sentiment-Analyse-Tool zum Klassifizieren von deutschen Texten. Dieses Tool wurde 2019 auf Github zur Verfügung gestellt in Form einer Python Erweiterung [1][1]. Grundlage von *textblob-de* [1] ist das Analyse Tool *TextBlob* von 2013 auf Github[2] [2]. Das Tool benutzt ein Lexikon basiertes Verfahren zur Klassifikation von Aussagen, wobei das vorhandene Wörterbuch nicht Domänen spezifisch ist. Dabei nutzt das Tool Lemmatisierung und Tokenisierung, um bei einer eingegebene Aussagen die Polarität zu bestimmen, nimmt aber keine Korrektur der Rechtschreibung vor [1].

2.4 Metriken zur Evaluation

Für die Evaluation der Daten der Klassifkatoren von Sentiment-Analysis-Tools gibt es verschiedene Metriken. Mit diesen lässt sich Performance und Genauigkeit eines Tools im Bezug auf die vorhandenen Datensätze bestimmen und einordnen.

[1]https://github.com/markuskiller/textblob-de
[2]https://github.com/sloria/textblob

Dazu werden im Verlauf der Arbeit Recall, Precision, F-Score, Accuracy und Fleiss' Kappa für alle genutzten Tools bestimmt.

2.4.1 Precision, Recall, F-Wert, Accuracy

Im Folgenden werden Recall, Precision, F-Score und Accuracy erläutert, welche für die Auswertung der Ergebnisse der Sentiment-Analyse-Tools notwendig sind. Für die Bestimmung der Werte dieser Metriken ist die Kenntnis von verschiedenen Verfahren notwendig. Diese Metriken werden für jedes Ergebnis der Sentiment-Analyse-Tools berechnet und ausgewertet.

True Positive (TP): Vorhersagen, die korrekt in positiv eingeordnet sind.
False Positive (FP): Vorhersagen, die in positiv eingeordnete sind, allerdings nicht in positiv hinein gehören.
True Negative (TN): Vorhersagen, die korrekt in negativ eingeordnet sind.
False Negative (FN): Vorhersagen, die in negativ eingeordnete sind, allerdings nicht negativ sind [17].

		Predicted	
		P	N
Actual	P	TP	FN
	N	FP	TN

Tabelle 2.2: Confusion Matrix

Recall: Das Verhältnis von korrekt positiv erkannten Vorhersagen im Verhältnis zu allen wirklichen positiven Elementen, wodurch die Vollständigkeit der Positiven Daten gemessen werden kann.

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

Precision: Das Verhältnis von korrekt positiv erkannten Vorhersagen im Verhältnis zu der gesamten Anzahl an Positiv klassifizierten Vorhersagen, womit die Genauigkeit gemessen werden kann.

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

F1-Score: Kombiniert Precision und Recall und fasst diese gewichtet zusammen für die einzelnen Klassen.

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

Accuracy: Das Verhältnis zwischen den korrekt klassifizierten Vorhersagen und der Anzahl aller Daten über alle Klassen hinweg. Dabei handelt es sich um die Genauigkeit bei der Klassifizierung.

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

2.4.2 Fleiss' Kappa

Bei der Beurteilung von Daten zwischen mehreren Beurteilern kann es zu verschiedenen Ergebnissen kommen. Um zu überprüfen in wie weit die Beurteiler miteinander und ob die Beurteiler übereinstimmen, gibt es das statistische Verfahren *Fleiss' Kappa*. Mit diesem kann für zwei oder mehr als zwei Beurteilern die Zuverlässigkeit zwischen diesen bestimmt werden. Dabei steht der *Kappa-Wert* für das Ausmaß der Übereinstimmung zwischen den Beurteiler [14].
Fleiss' Kappa kann mit den folgenden 3 Gleichungen bestimmt werden:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - P_e}. \tag{2.1}$$

$$\bar{P} = \frac{1}{Nn(n-1)} \sum_{i=1}^{N} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1) \tag{2.2}$$

$$\bar{P}_e = \sum_{j=1}^{k} (\frac{1}{Nn} \sum_{i=1}^{N} n_{ij})^2 \tag{2.3}$$

Dabei ist N die Gesamtzahl der Daten und n die Anzahl der Bewertungen pro Bewerter [17].
Nach Landis und Koch [9] liegt eine schlechte Übereinstimmung für einen Kappa-Wert unter 0 vor. Der Wert sollte im Bereich von 0 bis 1 liegen, wobei es hier Abstufungen gibt, wie gut die Übereinstimmung ist [9].

2.5 Programmierung

Für die Nutzung der Sentiment-Analyse-Tools wurde in dieser Arbeit die Programmiersprache Python Version 3.8.10 benutzt. Auf Grundlage dieser Version wird in dieser Arbeit alle benötigten Softwaretools selbst geschrieben, sowie auch der Voting-Classifier.

2.5.1 Voting-Classifier

Es gibt die Möglichkeit die Ergebnisse mehrerer Sentiment-Analyse-Tools zusammen zu tun und diese gemeinsam eine Entscheidung treffen zu lassen. Diese können dann zusammen gemeinsam über die Polarität eines Textes abstimmen. Dies ist möglich mit Hilfe eines Voting-Classifiers. Hierbei wird meist per Mehrheitsabstimmung über das Ergebnis entschieden [11].
In dieser Arbeit wird auch ein Voting-Classifier genutzt, der auf Grundlage der Mehrheitsentscheidung ein Entscheidungen fällt. Dafür werden die Ergebnisse der einzelnen Sentiment-Analyse-Tools genommen und diese werden an den Voting-Classifier gegeben. Dabei werden für jeden Eintrag die Labels der jeweiligen Tools eingelesen und verglichen. Da es sich im Fall dieser Arbeit um drei Sentiment-Analyse-Tools handelt, wird somit das Label übernommen, für welches zwei oder mehr Tools sich entschieden haben. Die Entscheidung für die Wahl der Labels wird somit auf Grundlage einer Mehrheitsentscheidung getroffen.

Bei der Mehrheitsentscheidung ist es auch möglich, dass es zu einer gleichen Verteilung der Klassifizierungen kommt und es somit keine absolute Mehrheit auf einem gibt. Dieser Fall kann das nur auftreten, wenn alle Tools eine unterschiedliche Entscheidung getroffen haben. Beim Auftreten dieses Falls wird von dem Voting-Classifier eine zufällige Entscheidung getroffen, welche der Entscheidungen gewählt werden soll[11].
Der Voting-Classifier in dieser Arbeit nimmt einen Datensatz der von drei verschiedenen Sentiment-Analyse-Tools drei bereits Klassifizierte wurde und erstellt auf Grundlage von diesen eine neue Datei mit den Ergebnissen. Da der Voting-Classifier bei Gleichstand zufällige Entscheidungen trifft, wird in einer zusätzlichen Datei festgehalten, wie viele zufällige Entscheidungen getroffen werden. In der nachfolgenden Abbildung 2.1 kann der Aufbau des Voting-Classifiers schematisch entnommen werden.

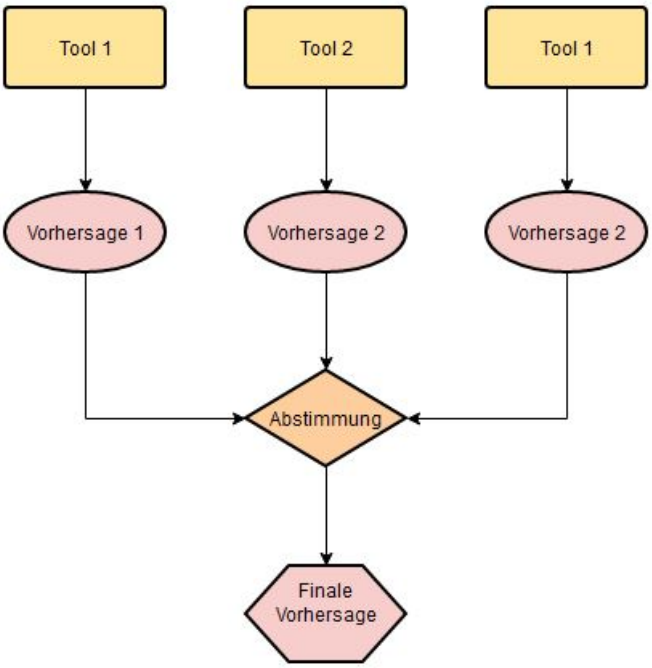

Abbildung 2.1: Modell des Voting-Classifiers

Kapitel 3

Konzept und Implementierung

Das Konzept dieser Arbeit untersucht den Zusammenhang zwischen Interaktionen und Stimmungen in Meetings von Softwareprojekten. Um dieses umzusetzen, werden die Interaktionen und Stimmungen zuerst einzeln betrachtet und mit Hilfe ausgewählter Verfahren untersucht. Die Interaktionen werden durch die Einordnung in die Kategorien von act4teams-SHORT nach Klünder [7] analysiert. Die Untersuchung der Stimmungen in den Meetings geschieht über die Klassifizierung durch Sentiment-Analyse-Tools. In diesem Kapitel wird vorgestellt, wie die einzelnen Schritte vorbereitet und wie diese im einzelnen ablaufen werden.

3.1 Idee

Die grundlegende Idee bei dieser Arbeit ist es, einen vorhanden Datensatz erst einmal auf die Interaktionen und anschließend auf die Stimmungen zu untersuchen. Der Datensatz wird im genauer im folgenden Kapitel erklärt werden. Dazu wird der Datensatz nach der Arbeit von Künder et al. [6] auf act4teams-SHORT Kategorien untersucht und nach diesen gruppiert, um somit Rückschlüsse auf die Interaktionsverteilung zu bekommen. Im nächsten Schritt werden basierend auf der Verteilung der act4-teams-SHORT Kategorien Sentiment-Analyse-Tools auf dem Datensatz ausgeführt. Dabei werden mehrere Tools getestet werden, um zu ermitteln, welche am besten auf dem vorhandenen Datensatz arbeiten. Anschließend werden diese genutzt, um den kompletten Datensatz zu analysieren und somit Ergebnisse bezüglich der Stimmung der Aussagen zu erhalten. Zum Abschluss werden diese Daten auf Zusammenhänge untersucht, worauf in Kapitel 4 weiter eingegangen wird.

3.2 Datensatz

Der Datensatz, der für diese Arbeit verwendet wird, stammt aus den Softwareprojekten der Leibniz Universität Hannover aus den Jahren 2012/13 und 2013/14, welches vom Fachgebiet Software Engineering abgehalten wird. Die die Daten in dem Datensatz sind auf Deutsch festgehalten. Der Datensatz besteht aus 38 einzelnen Meetingtranskripten, welche mit zusätzlichen Informationen, wie Label für die act4teams-Kategorien, dem Teilnehmer in anonymisierter Form

und der Dauer der Aussage, in Form von .csv Datei abgespeichert sind. Diese Softwaremeetings wurden auf Video festgehalten und im Anschluss auf Grundlage der Videos transkribiert. An einigen Stellen des Datensates wurden sind auch Interaktionen der Meetingteilnehmer aufgenommen wurden und entsprechend gelabelt wurden. Hiermit sind Aussagen gemeint, welche die Bewegungen oder Aktivität eines Meetingteilnehmers beschreiben. Somit handelt es sich um einen deutschen Datensatz, welcher in der Domäne der Teammeetings in Softwareprojekten einzuordnen ist. Allerdings ist der Datensatz in einer Form, welche es notwendig macht, diesen zu bearbeiten. Diese Anpassungen werden benötigt, um den Datensatz für die Sentiment-Analyse-Tools zu optimieren.
Der Datensatz wurde durch geschultes und erfahrenes Personal nach act4teams-Kategorien gelabelt. Diese haben die Transkripte auf Grundlage von Videoaufnahmen erstellt und auf Grundlage dieser, die act4teams-Kategorien für jede Sinneseinheit bestimmt. Der Datensatz ist somit nach dem Verfahren von Kauffeld et al. [5] für die Kategorisierung von act4teams erstellt worden.

3.2.1 Entfernen von Einträge

Wie bereits in Abschnitt 3.2 erwähnt, ist es notwendig Interaktionen und Stimmungen aus Meetings in Softwareprojekten auf ihren Zusammenhang zu untersuchen und dafür den Datensatz anzupassen. Im Datensatz sind in den Transkripten Markierungen der Form „*(?)*“, „*(??)*“ und „*??*“ eingesetzt. Diese werden genutzt, um Stellen im Transkript zu markieren, welche akustisch nicht verstanden wurden. Für die spätere weitere Bearbeitung des Datensatzes ist es wichtig, dass diese Transkripte nicht mehr enthalten sind. Das liegt daran, dass bei der Sentiment-Analyse die Gespräche der Meeting-Teilnehmer auf ihre Stimmung analysiert werden sollen [16]. Die genutzten Sentiment-Analyse-Tools sind für die Analyse von Texten ausgelegt. Daher werden diese Art von Kommentaren als Aussagen eines Teilnehmer des Meetings aufgefasst und entsprechend klassifiziert. Die Tools können diese Markierungen als eine Verstärkung der Polarität der Aussage interpretieren und dadurch eventuell bei ihrer Klassifizierung durch diese Markierungen beeinflusst werden.[19].
Bei den Transkripten des Datensatzes handelt es sich um Daten, die die Gespräche der Meeting-Teilnehmer wiedergeben sollen. Somit haben alle Formen von Transkripten, die nicht ausschließlich die verbale Kommunikation eines Meeting-Teilnehmers wiedergeben, Auswirkung auf das Ergebnis der Klassifizierung, weil diese durch die Tools mit ausgewertet werden. Das kann am Ende eventuell dazu führen, dass es zu falschen Labels kommt und die Klassifizierung negativ beeinflusst oder verfälscht. Deshalb können Aussagen, die Tätigkeiten eines Teilnehmers beschreiben, auch das Ergebnis beeinflussen, da diese eine Beschreibung vom Schreiber der Transkripte sind. Von dieser Anpassung sind hauptsächlich die Kategorien Visualisierung, Seitengespräche und Unterbrechung betroffen, welche am meisten solcher Einträge enthalten. Der daraus resultierende Datensatz wird anschließend benutzt, um ihn für die weiterführende Arbeit zu nutzen. Beispiele solcher Einträge können aus der nachfolgenden Tabelle 3.1 entnommen werden.

Eintrag im Datensatz	Erklärung
ja (zeigt auf C)	Aussage + Beschreibung einer Aktion
(lachen)	Eine Person lacht
UNV	Aussage ist unverständlich
das sind diese (?) Felder	Aussage mit einem fehlendem Wort

Tabelle 3.1: Einträge aus dem originalen Datensatz

Zum Schluss werden aus dem Datensatz noch Einträge entfernt, welche keinen Text enthalten. Ohne ein Transkript ist es nicht möglich die Daten auf ihre Stimmung zu untersuchen und somit kann kein Zusammenhang zwischen der Stimmung und den Interaktionen untersucht werden.

3.2.2 Anpassen der Rechtschreibung und Grammatik

Um richtige Ergebnisse bei der Klassifizierung zu gewährleisten, darf die Korrektur von Grammatik und Rechtschreibung nicht vernachlässigt werden. Die in den Softwaremeetings benutzte Umgangssprache spielt dabei eine große Rolle. Daher sollte bereits vor der Arbeit mit den Sentiment-Analyse-Tools der Datensatz hinsichtlich dieser beiden Punkte untersucht werden. Im Datensatz wird dazu „*ne*“ durch „*nein*“ oder „*oder*“ ersetzt, wenn die Bedeutung aus dem Kontext eindeutig ersichtlich ist. Tools wie TextBlobDe haben Probleme mit der Einordnung dieses Wortes als „*eine*“ oder als „*nein*“ und dies führt dazu, dass dieses Wort nicht betrachtet und ausgewertet wird. Damit wird es nicht für die Klassifizierung berücksichtigt und beeinflusst eventuell die Klassifizierung. [1].
Des weiteren werden bei den Transkripten die Abkürzungen „*villt*“ und „*eig*“ durch „*vielleicht*“ und „*eigentlich*“ ausgetauscht. Diese Abkürzungen treten häufig in dem Datensatz auf und sind nicht in allen Wörterbüchern der Tools hinterlegt. Zudem haben Huong et. al. [10] in ihrer Arbeit gezeigt, dass die Performance der Sentiment-Analyse-Tools in Texten schlechter ist, in denen mehr Abkürzungen enthalten sind. Dabei lässt sich die Performance durch das Ausschreiben der Abkürzungen wieder verbessern [10].
Zum Schluss werden in dem Datensatz noch die Rechtschreibfehler korrigiert. Da die Transkripte auf Video- und Audioaufnahmen basieren, sollten die transkribierten Wörter frei von Rechtschreibfehlern sein. Die Rechtschreibfehler sind von Seiten der Schreiber mit eingebracht worden und können somit im Datensatz bereinigt werden. Diese Anpassungen sind auch notwendig für die Analyse-Tools, weil diese nicht immer über eine Korrektur der Rechtschreibung verfügen, wie beispielsweise TextBlobDe [2]. Um eine bessere Performance der Tools zu gewährleisten, sind Anpassungen für eine bessere Kategorisierung der Wörter bei der Klassifizierung, notwendig. Da es sich hierbei um Änderungen am ursprünglichen Datensatz handelt, bringt das Folgen mit sich, die in Kapitel 5 genauer behandelt werden. In der nachfolgenden Tabelle 3.2 sind Beispiele für solche Einträge in dem Datensatz zu sehen.

Eintrag im Datensatz	Grund
wer weiß, also vllt gibts	Abkürzung
weil das finde ich eig auch nicht wichtig	Abkürzung
das wird ne nummer sien	Umgangssprache + Rechtschreibfehler

Tabelle 3.2: Einträge aus dem original Datensatz

3.3 Mapping von act4teams zu act4teams-SHORT

Der Datensatz, der für diese Arbeit zur Verfügung steht, ist bereits mit Codewörtern aus act4teams nach der Arbeit von Kauffeld et al. [5] gelabelt. Allerdings wurden für das Labeln der act4teams Kategorien nur Audiodateien der Meetingaufnahmen benutzt. Somit weicht es von dem richtigen Verfahren ab, bei welchem das Zuordnen auf Grundlage der Videoaufnahmen geschehen sollte. Da allerdings auf diesem Faktor kein Einfluss genommen werden kann, wird in dieser Arbeit mit diesem Konzept weiter gearbeitet.
Klünder [7] hat in ihrer Dissertation ein Mapping von act4teams zu act4teams-SHORT vorgestellt. Um dieses nutzen zu können, müssen in dem Datensatz die einzelnen Kategorien ausfindig gemacht werden und nach den act4teams-Kategorien sortiert werden. Nachdem der Datensatz nach den Labels umsortiert worden ist, ist es möglich, diese zu act4teams-SHORT zusammen zu fassen. Beim Mapping sollten folgende Dinge beachtet werden. Ein Aspekt ist, dass es nicht für alle 44 Kategorien aus act4teams ein Äquivalent in act4teams-SHORT existiert. Somit können einige Kategorien nicht direkt zugeordnet werden [7]. Diese Kategorien sollten durchgegangen werden und den einzelnen act4teams-SHORT-Kategorien zugeordnet werden. Ein weiterer Aspekt ist, dass einige Kategorien aus act4teams sich auf zwei oder mehr Kategorien in act4teams-SHORT aufteilen [7]. Somit werden diese Daten auf mehrere Kategorien verteilt.
Für den Fall, dass eine Kategorie aus act4teams auf in mehreren Kategorien in act4teams-SHORT vorhanden ist, wird diese in alle Kategorien in act4teams-SHORT übertragen. Dies ist zum Beispiel für die Kategorie „*Organisationales Wissen*“ der Fall. Diese Kategorie ist sowohl der Kategorie „*Wissentransfer*“ als auch „*Informationsweitergabe*“ zugeordnet. Für den Fall, dass die Einträge einer Kategorie einzeln aufgeteilt werden müssen, werden die act4teams Kategorien beibehalten und später parallel untersucht.
Dabei wird das Mapping von Klünder [7] erweitert und ist aus der nachfolgenden Tabelle 3.3 zu entnehmen.

Kategorie bei act4teams-SHORT	**Kategorie bei act4teams**
Problembennenung	Problem, Problemerläuterung, Organisationales Wissen
Problemvernetzung	Verknüpfung Problemanalyse
Lösungsbenennung	Lösungsvorschlag, Lösungserläuterung
Lösungsvernetzung	Verknüpfung mit Lösung, Sollentwurf
Verknüpfung und Vernetzung	Problem zur Lösung, Kosten-Nutzen-Abwägung
Destruktives Verhalten	Tadel/Abwertung, Schuldigensuche, Jammern, Kein Interesse an Veränderungen, Seitengespräche, Betonung autoritärer Elemente, Abbruch, Verlieren in Details und Beispielen, Unterbrechung, Ablehnung
Methodisch-strukturierendes Verhalten	Zielorientierung, Priorisieren, Verfahrensvorschlag, Verfahrensfrage, Aufgabenverteilung, Zusammenfassung, Klärung/Konkretisierung, Visualisierung, Zeitmanagement
Proaktives Verhalten	Interesse an Veränderungen, Eigenverantwortung, Maßnahmenplanung
Wissensaustausch	OrganisationalesWissen,Wissen Wer
Informationsweitergabe	OrganisationalesWissen,Wissen Wer
Kollegiales Verhalten	Atmosphärische Auflockerung, Ermunternde Ansprache, Lob, Unterstützung, Rückmeldung, Aktives Zuhören, Lachen

Tabelle 3.3: Mapping der Kategorien act4teams zu act4teams-SHORT

3.4 Auswahl der Sentiment-Analysis-Tools

Für die richtige Analyse der Stimmung ist die Wahl eines Sentiment-Analyse-Tools von Bedeutung, welches eine gute Performance für den vorhandenen Datensatz hat. Islam et al. [4] haben in ihrer Arbeit gezeigt, das die Erstellung eines Tools für die Domäne des Software Engineerings, die eine bessere Performance liefert. Damit hat die Domäne hat einen Einfluss auf die Wahl des Tools, da die Tools weniger Genauigkeit aufweisen, auf fremden Domänen[12]. Da Daten aus Meetings von Softwareprojekten ausgewertet werden sollen, ist die Wahl eines Tools für den Bereich der Softwareentwicklung von Vorteil. Hinzu kommt, dass der Datensatz Transkripte aus deutschen Meetings enthält und daher ein deutsches Tool gewählt werden sollte. Um zu ermitteln, welches Tool sich am besten eignet, werden mehrere Sentiment-Analyse-Tools getestet.

3.4.1 Konzept zur Auswahl

Um zu prüfen, welches Tool sich am besten für den Datensatz eignet, werden Tools der Sentiment-Analyse ausgewählt und für diese die Accuracy bestimmt. Unter diesen wurden für diese Arbeit drei Tools ausgewählt und geprüft, welche am besten auf dem vorhanden Datensatz arbeiten. Untersucht werden hierbei die Tools GerVader [20], SentiStrength [19] und TextBlobDe[1].
Bei SentiStrength ist es notwendig vor der Verwendung eine Anpassung des Wörterbuches vorzunehmen, da SentiStrength mit der englischen Sprache arbeitet [19]. Da der Datensatz allerdings auf deutsch ist, muss dieses ausgetauscht werden. Dafür tausche ich das Wörterbuch von SentiStrength durch eine von der Austrian Research Institute for AI[2] übersetzte Version aus. Diese Änderung ermöglicht die Nutzung von SentiStrength, für deutsche Datensätze und kann somit für den Datensatz dieser Arbeit genutzt werden.
Zum bestimmen der Performance eines Tools wird ein Datensatz benötigt, welcher bereits gelabelt ist. Da der Datensatz aus deutschen Transkripten besteht, wird zum Bestimmen der Accuracy ein Teil des eigentlichen Datensatzes genutzt. Es sind keine öffentliche Daten in der Domäne in der deutschen Sprache erhältlich. Daher erstelle ich selbst einen Testdatensatz zum analysieren der Genauigkeit der Tools. Dafür habe ich 300 zufällige Einträge gewählt und manuell mit positiv, neutral und negativ gelabelt. Ich habe dabei darauf geachtet, dass am Ende in dem Testdatensatz 100 positive, 100 neutrale und 100 negative Einträge enthalten sind. Damit kann sichergestellt werden, dass die Testdaten in der gleichen Domäne wie der Datensatz liegen und auch die Sprache berücksichtigt wird. Die damit berechneten Metriken sind für die Accuracy der Tools sind in folgenden Tabellen 3.4 dargestellt. In der darauf folgenden Tabelle 3.5 werden zur genaueren Analyse Precision, Recall und F1-Score abgebildet. Die höchsten Werte aller Tools in einer Kategorie werden dabei fett markiert.

Tool	**Accuracy**
GerVADER	57,33%
SentiStrength	56,00%
TextBlobDe	50,67%

Tabelle 3.4: Accuracy deutscher Sentiment-Analyse-Tools

[1]https://github.com/markuskiller/textblob-de
[2]http://www.ofai.at/research/interact/

Tool	Sentiment	Precision	Recall	F1
GerVADER	Positiv	61,48%	**83,00%**	**70,64%**
	Neutral	**47,11%**	57,00%	51,58%
	Negativ	72,73%	**32,00%**	**44,44%**
SentiStrength	Positiv	**71,58%**	68,00%	69,74%
	Neutral	44,57%	**78,00%**	**56,73%**
	Negativ	**73,33%**	22,00%	33,85%
TextBlobDe	Positiv	67,00%	67,00%	67,00%
	Neutral	40,94%	70,00%	51,66%
	Negativ	51,72%	15,00%	23,26%

Tabelle 3.5: Precision, Recall und F1 deutscher Sentiment-Analyse-Tools

Zu jedem Tool werden wichtige Metriken erhoben, welche zum Vergleich der Tools genutzt werden. Bei der Bewertung der Tools wird als wichtigster Faktor die Accuracy genutzt, da diese die Genauigkeit bei der Klassifizierung von Daten repräsentiert [17]. Für ähnliche Werte bei der Accuracy wird zur weiteren Betrachtung der F1-Score dazu gezogen, da dieser Precision und Recall mit in die Berechnung einbezieht [17]. Die Auswertung erfolgt in Kapitel 3.4.2.
In der Arbeit von Tymann et al. [21] wurde untersucht, ob ein deutsches Tool genauso gut wie ein englisches Tool auf dem gleichen Datensatz ist. Dafür wurde der Datensatz ins Englische übersetzt. Das englische Tool schnitt dabei mit dem übersetzen Datensatz besser ab, als das deutsche Tool mit deutschen Datensatz [21]. Dies kann zu einer erweiterten Betrachtung des vorhandenen Datensatzes bei der späteren Untersuchung genutzt werden. Analog dazu wird in dieser Arbeit der Datensatz mit Hilfe von *DeepL* übersetzt und gleiche Bedingungen erfüllt[3], wie das in der Arbeit von Tyman et al. [21] genutzte Tool. Diese Übersetzungen werden per Hand durch stichprobenhafte Kontrolle und schnelles durchgehen der übersetzten Einträge auf Korrektheit überprüft. Für diesen Ansatz werden die gleiche Werte erhoben, wie für die deutschen Tools. Anhand der Werte kann auch hier gesehen werden, dass es bei der Übersetzung eine Verbesserung in der Performance erkennbar ist. Für den Vergleich habe ich zum einen das Tool SentiStrength. Bei diesem Tool habe ein englischen Wörterbuch verwendet. Als zweites Tool habe ich das SentiStrength-SE benutzt. Dieses Tool basiert auf SentiStrength und ist auf die Domäne des Software Engineerings angepasst. Als letztes wird SpacyTextBlob benutzt, welches ein englisches lexikon basiertes Sentiment-Analyse-Tool ist. Die Auswertung der Daten dieser Tools folgt in Kapitel 4. Aus der nachfolgenden Tabelle 3.6 kann die Accuracy für die englischen Tools entnommen werden. In der darauf folgenden Tabelle 3.7 werden zur genaueren Analyse Precision, Recall und F1-Score dargestellt. Die höchsten Werte aller Tools in einer Kategorie werden dabei fett markiert.

[3]https://www.deepl.com/

Tool	**Accuracy**
SentiStrength	67,00%
SentiStrength-SE	63,67%
SpacyTextBlob	62,67%
VADER	54,00%

Tabelle 3.6: Accuracy englischer Sentiment-Analyse-Tools

Tool	**Sentiment**	**Precision**	**Recall**	**F1**
SentiStrength	Positiv	79,17%	76,00%	**77,55%**
	Neutral	**53,47%**	77,00%	63,11%
	Negativ	80,00%	**48,00%**	**60,00%**
SentiStrength-SE	Positiv	**91,43%**	64,00%	75,29%
	Neutral	48,48%	**96,00%**	**64,43%**
	Negativ	**96,88%**	31,00%	46,97%
SpacyTextBlob	Positiv	67,97%	**87,00%**	76,32%
	Neutral	52,78%	57,00%	54,81%
	Negativ	68,75%	44,00%	53,66%

Tabelle 3.7: Precision, Recall und F1 englischer Sentiment-Analyse-Tools

3.4.2 Nutzung des Voting-Classifiers

Nach dem Bestimmen aller Metriken werden die Tools nochmal mit Hilfe eines Voting-Classifiers analysiert. Dessen Aufbau wurde in Kapitel 2.5.1 beschrieben. Mit diesem wird untersucht, ob die Performance der Tools auf dem Datensatz durch die Kombination der einzelnen Tools noch verbessert werden kann. Dafür werden die einzelnen Ergebnisse der Tools vom Voting-Classifier untersucht und diese verarbeitet. Dieser Vorgang wird sowohl mit den deutschen, als auch mit den englischen Tools durchgeführt. Die Ergebnisse der Durchläufe sind in der folgenden Tabelle 3.8 aufgeführt.

Tool	**Sentiment**	**Precision**	**Recall**	**F1**
Deutsche Tools	Positiv	66,94%	83,00%	74,11%
	Neutral	50,00%	61,00%	54,95%
	Negativ	62,96%	34,00%	44,16%
Englische Tools	Positiv	86,52%	77,00%	81,48%
	Neutral	36,02%	85,00%	50,60%
	Negativ	80,00%	40,00%	53,33%

Tabelle 3.8: Precision, Recall und F1 der Voting-Classifier

Durch die Nutzung des Voting-Classifiers, zeigt sich für beide Sprachen bei den Ergebnissen eine Steigerung der Accuracy. Für die deutschen Tools liegt die Accuracy mit dem Voting-Classifier bei 59,33% und ist somit um 10% gestiegen im Verhältnis zum besten deutschen einzelnen Tool. Für die englischen Tools liegt die Accuracy mit dem Voting-Classifier bei 67,33% und ist somit um 0,33% im

Verhältnis zum besten einzelnen englischen Tool gestiegen. Dabei ist auch eine Steigerung von Precision, Recall und F1 vorhanden. Somit steigt bei beiden die Accuracy gegenüber der besten Tools der jeweiligen Sprache. Des weiteren werden die Tools noch bezüglich ihres Fleiss' Kappa-Wertes untersucht, um zu prüfen, wie einig sich die Tools des jeweiligen Voting-Classifiers bei der Klassifizierung sind. Nach der Arbeit von Landis et al. [9], sollte der Wert größer als 0 sein, damit bei den Tools eine Einigkeit bei der Klassifizierung vorhanden ist. Der Wert für die drei deutschen Tools liegt bei 0,27, wobei es sich um einen angemessenen Wert handelt nach Landis et al. [9]. Bei den englischen Tools liegt dieser Wert bei 0,21 und ist somit auch ein angemessener Wert nach Landis et al. [9].
Durch die Nutzung des Voting-Classifiers werden somit die Ergebnisse verbessert. Zudem liegt die Einigkeit zwischen den einzelnen Tools in einem angemessenen Bereich. Daher wird für die Untersuchung der Forschungsfrage die jeweiligen Voting-Classifier genutzt. Der englischsprachige Voting-Classifier wird zu der Analyse mit hinzu gezogen, da sich gezeigt hat, dass dieser eine 8% höhere Accuracy aufweist. Auch der F1-Wert für positive Aussagen liegt um 7,37% höher und der F1-Wert für negative Aussagen liegt um 9,17% höher als bei dem deutschsprachigen Voting-Classifer. Somit weißt der Voting-Classifier auf Basis der englischsprachigen Tools eine bessere Genauigkeit auf.

Kapitel 4

Auswertung

In diesem Kapitel werden die gewonnen Daten dargestellt und der Zusammenhang zwischen den Stimmungen und den Interaktionen betrachtet. Dafür werden die nach Kapitel 3 vorhandenen Daten untersucht und analysiert. Dafür wurde die Interaktionen mit den Stimmungen dieser gegenübergestellt. Dies geschieht durch die Untersuchung der einzelnen act4teams-SHORT-Kategorien und der Untersuchung der Sentimente in den einzelnen Kategorien. Diese werden anschließend übersichtlich dargestellt und ausgewertet. Im Abschnitt 4.1 werden dabei die einzelnen Kategorien von act4teams-SHORT betrachtet. Anschließend wird in Abschnitt 4.2 genauer untersucht, wie die Verteilung dieser Daten ist und wie diese zustande kommen.

4.1 Zusammenführung der Interaktionen mit den Stimmungen

Zur Betrachtung potenzieller Zusammenhänge zwischen den Stimmungen und den Interaktionen innerhalb von Softwareprojektmeetings werden zunächst die act4Teams-SHORT Kategorien untersucht. Für diesen Ansatz werden in erster Linie die Ergebnisse des Voting-Classifiers mit den deutschen Tools genutzt. Grund dafür ist, dass eine Übersetzung der Aussagen zu einer Veränderung der Daten führen kann, welche die Klassifikation dieser Daten beeinflusst, aufgrund von Variationen der Grammatik und den Satzbau der Sprachen [22]. Allerdings zeigt die Arbeit von Tymann et al.[21], dass die Übersetzung sinnvoll sein kann, da die Performance der englischsprachigen Tools besser sein kann als die der deutschen Tools[21]. In Kapitel 3.4.1 wird darauf genauer eingegangen.
Die Einträge in diesen Kategorien werden auf Grundlage der Ergebnisse des deutschen Voting-Classifiers in einen Zusammenhang gesetzt, da bei diesen keine Übersetzung notwendig ist, welches ein Beeinflussen der Klassifizierung bezüglich der Sprache ausschließt. Die einzelnen Kategorien der Interaktionen und die dazugehörigen Sentimente werden hierbei gegenübergestellt. Dafür werden die Einträge aus dem Datensatz nach den einzelnen act4teams-Kategorien sortiert. Die Kategorien „*Reputation*“, „*Phrase*“, „*Ich-Botschaft*“ und „*Frage*“ werden einzeln aufgelistet, da hier ein Mapping zu act4teams-SHORTS, wie in dem Kapitel 3.3 beschrieben, nicht möglich ist. Die Stimmungen der einzelnen Kategorien

werden in den Polaritäten positiv, neutral und negativ eingeteilt. Der genaue Aufbau des deutschen und englischen Voting-Classifiers, wurden in Kapitel 3.4.2 vorgestellt. Dazu werden die Ergebnisse der einzelnen Kategorien ausgewertet auf Grundlage des deutschen Voting-Classifiers und mit den Ergebnissen des englischen Voting-Classifiers verglichen. Dabei wird geprüft, in wie weit die Voting-Classifier im Ergebnis übereinstimmen. Bei einem Unterschied der Verteilung zwischen positiv, neutral und negativ des deutschen Voting-Classifiers und dem englischen Voting-Classifier, werden die Ergebnisse des englischen Voting-Classifiers zur Betrachtung hinzugezogen. Sollten beim Voting-Classifier, welcher die deutschen Tools nutzt, mehr positive als negative Aussagen auftreten und beim Voting-Classifier auf Basis der englischsprachigen Tools das Gegenteil der Fall sein, wird dies genauer betrachtet. Analog dazu werden auch solche Fälle mit neutralen Aussagen betrachtet. Die genaue Verteilung für den Datensatz, ausgewertet mit dem Voting-Classifier, basierend auf den deutschen Voting-Classifier, ist aus der nachfolgenden Tabelle 4.1 zu entnehmen.

Kategorien der Interaktionen	**Alle**	**Positiv**	**Neutral**	**Negativ**
Gefühle	1	0(0,00%)	0(0,00%)	1(100,00%)
Reputation	6	4(66,67%)	2(33,33%)	0(0,00%)
Phrase	63	5(7,94%)	54(85,71%)	4(6,35%)
Problemvernetzung	117	11(9,40%)	99(84,62%)	7(5,98%)
Verknüpfungen und Vernetzungen	196	36(18,37%)	125(63,78%)	35(17,86%)
Ich-Botschaft	345	56(16,23%)	277(80,29%)	12(3,48%)
Destruktives Verhalten	645	68(10,54%)	531(82,33%)	46(7,13%)
Proaktives Verhalten	868	213(24,54%)	607(69,93%)	48(5,53%)
Lösungsvernetzung	1440	345(23,96%)	1042(72,36%)	53(3,68%)
Frage	1693	128(7,56%)	1528(90,25%)	37(2,19%)
Methodisch strukturierendes Verhalten	1754	276(15,74%)	1431(81,58%)	47(2,68%)
Lösungsbenennung	2367	421(17,79%)	1883(79,55%)	63(2,66%)
Wissentransfer	3379	431(12,76%)	2818(83,40%)	130(3,85%)
Informationsweitergabe	3379	431(12,76%)	2818(83,40%)	130(3,85%)
Problembennenung	3857	499(12,94%)	3148(81,62%)	210(5,44%)
Kollegiales Verhalten	4895	509(10,40%)	4337(88,60%)	49(1,00%)

Tabelle 4.1: Verteilung der Sentimente in den einzelnen Interaktionskategorien beim deutschen Voting-Classifier

Auf Grundlage dieser Daten wird im folgenden Kapitel 4.2 untersucht, wie die Interaktionen und die Sentimente zusammenhängen. Dazu wird jede einzelne Kategorie betrachtet und dabei genauer untersucht, wie die Verteilung der Sentimente auf dieser ist und wie diese Werte zustande kommen. Die Daten des Voting-Classifiers, welcher die englischsprachigen Tools nutzt, können aus folgender Tabelle 4.2 entnommen werden.

Kategorien der Interaktionen	**Alle**	**Positiv**	**Neutral**	**Negativ**
Gefühle	1	0(0,00%)	0(0,00%)	1(100,00%)
Reputation	6	1(16,67%)	4(66,67%)	1(16,67%)
Phrase	63	7(11,11%)	51(80,95%)	5(7,94%)
Problemvernetzung	117	7(5,98%)	97(82,91%)	13(11,11%)
Verknüpfungen und Vernetzungen	196	12(6,12%)	151(77,04%)	33(16,84%)
Ich-Botschaft	345	42(12,17%)	287(83,19%)	16(4,64%)
Destruktives Verhalten	645	26(4,03%)	545(84,50%)	74(11,47%)
Proaktives Verhalten	868	87(10,02%)	753(86,75%)	28(3,23%)
Lösungsvernetzung	1440	150(10,42%)	1217(84,51%)	73(5,07%)
Frage	1693	56(3,31%)	1602(94,62%)	35(2,07%)
Methodisch strukturierendes Verhalten	1754	106(6,04%)	1606(91,56%)	42(2,39%)
Lösungsbenennung	2367	125(5,28%)	2165(91,47%)	77(3,25%)
Wissentransfer	3379	169(5,00%)	3056(90,44%)	154(4,56%)
Informationsweitergabe	3379	169(5,00%)	3056(90,44%)	154(4,56%)
Problembennenung	3857	200(5,19%)	3439(89,16%)	218(5,65%)
Kollegiales Verhalten	4895	345(7,05%)	4480(91,25%)	70(1,43%)

Tabelle 4.2: Verteilung der Sentimente in den einzelnen Interaktionskategorien beim englischen Voting-Classifier

4.2 Auswertung der Daten

In diesem Kapitel wird auf die einzelnen act4teams-Kategorien-SHORT beziehungsweise auf die act4teams-Kategorien, welche nicht zu act4teams-SHORT überführt werden konnten, wie in Abschnitt 3.3 beschrieben, eingegangen. Dies beinhaltet somit elf Kategorien aus act4teams-SHORT und fünf Kategorien aus act4teams. Der Datensatz enthält zusätzlich noch Kategorien, welche als erweiterte Kategorien in act4teams hinzu kommen. Diese wurden ausführlicher in den Grundlagen in Abschnitt 2.1.2 beschrieben. Die Kategorie „*Abgebrochener Satz*“ besteht aus abgebrochenen Sätzen, welche sich inhaltlich keiner anderen act4teams-Kategorie zuordnen lassen. Somit ist eine Auswertung bezüglich act4teams-SHORT nicht möglich. Die restlichen dieser Kategorien enthalten weniger als 35 Aussagen und werden wegen einer zu geringen Menge an Daten nicht weiter betrachtet. Die beiden Kategorien „*Gefühle*“ und „*Reputation*“ enthalten bei dem betrachteten Datensatz eine nur sehr kleine Menge von weniger als 10 Aussagen. Das führt dazu, dass diese für die weitere Auswertung nicht betrachtet werden und somit auch keine Aussage über diese beiden Kategorien getroffen wird, da die vorhandene Datenmenge zu klein ist, um diese Kategorie genau beurteilen zu können.
Bei den Polaritäten handelt es sich um eine Verteilung auf einer Nominalskala. Generell lässt sich damit der Modus ermitteln, welcher für alle Kategorien sich bei neutral liegt. Somit lässt sich feststellen, dass für alle Interaktionenkategorien

gilt, dass der größte Teil dieser Interaktionen mit einer neutralen Stimmung der Meetingteilnehmer ausgeführt wurde. Des weiteren wird für die Stimmungen noch der Median betrachtet. Für die Betrachtung des Medians, ist es möglich, die einzelnen Polaritäten als Ordinalskala aufzufassen. Dafür werden positiv, neutral und negativ jeweils Zahlenwerte zugeordnet. Positiv entspricht einem Wert von 1, neutral einem Wert von 0 und negativ einem Wert von -1. Mit dieser Anpassung kann für jede Kategorie der Median bezüglich der Polaritätswerte der Aussgagen bestimmt werden. Auch hier wird festgestellt, dass für alle Kategorien der Interaktionen, der Median bei einem neutralen Polaritätswert liegt. Somit ist der zentrale Wert im Bereich neutraler Stimmung. Im folgenden Abschnitt 4.2.1 werden zur genaueren Untersuchung die einzelnen Kategorien genauer betrachtet.

4.2.1 Auswertung der einzelnen Kategorien

Da alle Kategorien den größten Anteil an Aussagen im neutralen Bereich haben, werden hauptsächlich die Verteilungen und der Anteil der positiven und negativen Aussagen innerhalb einer Kategorie betrachtet.

Phrase:
Zunächst wird die Kategorie „*Phrase*“ betrachtet. Bei dieser Kategorie sind es 7,94% positive und 6,35% negativen klassifizierte Aussagen. Daher sind zu ungefähr gleichen Anteilen positive und negative Aussagen vorhanden. Aussagen aus diesem Datensatz, welche der act4teams-Kategorie „*Phrasen*“ zugeordnet sind, sind zu 85,71% neutral. Somit lässt sich erkennen, dass der Anteil an positiven und negativen Aussagen in der Kategorie Phrasen in etwa gleich groß ist. Diese Annahme wird durch die Klassifizierung des englischen Voting-Classifiers unterstützt.

Problemvernetzung:
Bei der Kategorie der „*Problemvernetzung*“ ist der Anteil an positiven klassifizierten Aussagen mit 9,40% höher als die der negativen mit 5,98%. Der Voting-Classifier auf Basis der englischsprachigen Tools dagegen, klassifiziert 5,98% der Aussagen negativ und 11,11% positiv. Dieser Unterschied tritt vor allem bei den Klassifizierungen der neutralen Aussagen auf. Einige Aussagen, welche vom Voting-Classifer mit den englischsprachigen Tools neutral gelabelt wurden, wurden vom deutschen als positiv gelabelt. Von dem deutschen Voting-Classifier ausgegangen, haben die Teilnehmer beim Analysieren von Ursachen und Folgen von Problemen eine mehr positive als negative Stimmung und umgekehrt beim englischen Voting-Classifer.

Verknüpfungen und Vernetzungen:
Bei dieser Kategorie ist der Anteil an positiv und negativ klassifizierten Aussagen ungefähr gleich verteilt mit 18,37% positiven und 17,86% negativen Aussagen. Der Anteil an neutralen Aussagen liegt hierbei bei 63,78%. Der Voting-Classifer mit den englischsprachigen Tools hat dagegen weniger Aussagen als positiv klassifiziert. Dieser klassifiziert 6,12% der Aussagen positiv und 16,84% negativ. Nach dem englischen Voting-Classifier enthält diese Kategorie somit mit mehr

negative als positive Aussagen mit einem Anteil von 77,04% an neutralen Aussagen.

Ich-Botschaft:
In dieser Kategorie sind nach dem deutschen Voting-Classifier 16,23% positive und 3,48% negative Aussagen vorhanden. Das zeigt, dass es sich bei den getätigten Ich-Botschaften im Datensatz um Aussagen handelt, welche eine positive Stimmung vertreten, neben dem Hauptteil an neutralen Aussagen. Diese Annahme wird durch die Kategorisierung des Voting-Classifiers mit den englischen Tools unterstützt.

Destruktives Verhalten:
Bei der Klassifizierung von „*Destruktives Verhalten*“ klassifizierte der deutsche Voting-Classifier 10,54% der Aussagen als positiv und 7,13% der Aussagen als negativ. Beim englischen Voting-Classifier dagegen ist der Anteil an negativ klassifizierten Aussagen mit 11,47% höher als die der positiv klassifizierten Aussagen mit 4,03%. Aufgrund dieses Unterschiedes sollten beide Voting-Classifier hier berücksichtigt werden.

Proaktives Verhalten:
Aussagen, welche der act4teams-SHORT Kategorie „*Proaktives Verhalten*“ zugeordnet sind, sind nach dem deutschen Voting-Classifier 24,54% positiv. Da es sich um proaktives Verhalten handelt, ist eine höhere Anzahl an Aussagen mit positiver Stimmung zu erwarten. Der Anteil an Aussagen mit negativer Stimmung liegt bei 5,53%. Dieses Ergebnis spiegeln auch die englischsprachigen Tools im Voting-Classifier wieder. Für Proaktives Verhalten zeigt sich daher, dass die Aussagen dieser Kategorie mehr positive als negative Aussagen enthalten. Aber auch hier bleibt der Hauptteil der Aussagen mit 69,93% neutral.

Lösungsvernetzung:
Nach dem deutsche Voting-Classifier handelt es sich bei dieser Kategorie bei 23,96% der Aussagen, um positiv klassifizierte Aussagen. Dagegen sind 3,68% der Aussagen negativ klassifiziert. Somit sind Aussagen in diesem Datensatz, welche mit der Analyse von Lösungen zu tun haben, öfter von positiver als negativer Stimmung unter der Berücksichtigung, dass die meisten Aussagen in dieser Kategorie neutral klassifiziert sind. Diese Tendenz spiegelt auch der Voting-Classifier mit den englischsprachigen Tools wieder.

Frage:
Bei Aussagen, welche sich mit Fragen innerhalb des Meetings beschäftigen, handelt es sich bei 7,56% um Aussagen mit einer positiven Stimmung und bei 2,19% um Aussagen mit negativer Stimmung. Daher haben die Teilnehmer beim stellen von Fragen eine eher gute Stimmung als eine negative. Dabei ist zu beachten, dass der Großteil der Aussagen mit 90,25% neutral sind. Eine ähnliche Verteilung der Polaritäten lässt sich auch beim englischen Voting-Classifier erkennen.

Methodisch-strukturierendes Verhalten und Lösungsbenennung:
Aussagen in den Meeting, welche sich mit methodisch-strukturierendem Verhalten befassen, sind mehr positiv als negativ. Die Aussagen sind zu 15,74% positiv klassifiziert und zu 2,68% negativ. Diese Ausrichtung der Verteilung wird auch durch den Voting-Classifier auf Basis der englischsprachigen Tools wieder gegeben. Insgesamt lässt sich sagen, dass die Stimmung der Teilnehmer hier deutlich mehr positiv ist als negativ. Allerdings ist hier auch nicht außer Acht zu lassen, dass 81,58% der Aussagen als neutral einzuordnen sind.

Wissenstransfer und Informationsweitergabe:
Da sich „*Wissenstransfer*“ und „*Informationsweitergabe*“ aus den gleichen Kategorien zusammensetzen, haben auch beide die gleiche Verteilung der Daten. Das genauere Vorgehen für diese Zusammenfassung wurde in Kapitel 3.3 beschrieben. Die Analyse der Aussagen mit den Sentiment-Analyse-Tools zeigen, dass es sich beim „*Wissenstransfer*“ und der „*Informationsweitergabe*“ um Interaktionen in Meetings handelt, bei welchen 12,76% der Aussagen positiv und 3,85% negativ einzuordnen sind. Der Anteil an neutralen Aussagen liegt hier bei 83,40%. Bei der Betrachtung der Verteilung des englischen Voting-Classifiers, sieht man einen Unterschied in der Verteilung der Daten im Vergleich zu dem deutschen. Dieser hat eine Verteilung von 5,00% positiver, 4,56% negativer und 90,44% neutraler Aussagen. Der deutsche Voting-Classifier hat somit mehr positive als negative Aussagen und der englische Voting-Classifier eine ungefähr gleiche Verteilung an positiven und negativen Aussagen.

Problembenennung:
„*Problembenennung*“ enthält analog zu „*Wissenstransfer*“ und „*Informationsweitergabe*“ Daten aus der act4teams-Kategorie „Organisationales Wissen“. Daher kommt es hier zu einer ähnlichen Verteilung der Daten. Problembenennung enthält 12,94% positive Aussagen und 5,44% negative Aussagen. Der Voting-Classifier auf Basis der englischsprachigen Tools klassifiziert dagegen 5,19% der Aussagen positiv und 5,65% negativ. Bei diesem ist hier eine eher gleiche Verteilung an positiven und negativen Aussagen zu erkennen.

Kollegiales Verhalten:
Beim „*Kollegialen Verhalten*“ handelt es sich um Aussagen, die andere mit Wertschätzungen und Humor einbindet. Bei dieser Kategorie ist zu erwarten, dass die Aussagen mehr positiv als negativ sind. Dieses Ergebnis wird auch durch beide Voting-Classifier bestätigt. Bei Kollegialem Verhalten sind 10,40% der Aussagen positiv und 1,00% der Aussagen negativ.
Somit ist auch hier der größte Teil der Aussagen neutral einzuordnen. Allerdings zeigt sich, dass die Aussagen in diesem Bereich, wenn sie nicht neutral sind, meist positiv einzuordnen sind.

Insgesamt lässt sich sagen, dass die Betrachtung aller Kategorien zeigt, dass fast alle eine leicht unterschiedliche Verteilung der Sentimente aufweisen. Ausgenommen davon sind die Kategorien „*Wissenstransfer*“, „*Informationsweitergabe*“ und „*Problembenennung*“, welche eine ähnliche Verteilung an positiven, neutralen und negativen Aussagen aufweisen. Zudem ist zusätzliche die Betrachtung des englischen Voting-Classifiers sinnvoll, da es Kategorien gibt, bei denen die Voting-Classifier unterschiedliche Ergebnisse liefern. Dadurch ist eine genauere Betrachtung der Unterschiede möglich. Eine genauere Analyse der Daten wird in Kapitel 5.2 diskutiert.

Kapitel 5

Diskussion

In dieser Arbeit wurden Interaktionen und Stimmungen in Meetings von Softwareprojekten untersucht. Dabei wurden für einen vorhandenen Datensatz aus dieser Domäne die einzelnen Interaktionen in Form von act4teams-SHORT-Kategorien in Zusammenhang mit den Stimmungen dieser Aussagen gesetzt. Im Nachfolgenden wird in Abschnitt 5.1 die Forschungsfrage beantwortet. Anschließend werden in Abschnitt 5.2 die Ergebnisse interpretiert und im darauf folgenden Abschnitt wird auf 5.3 mögliche Einflussfaktoren der Validität eingegangen.

5.1 Beantwortung der Forschungsfrage

Um Zusammenhänge zwischen Interaktionen und Stimmungen in Meetings von Softwareprojekten zu untersuchen, wurde in dieser Arbeit Aussagen aus dieser Domäne bezüglich ihrer Interaktionen und Stimmungen untersucht. Dafür wurden diese auf Grundlage von act4teams-SHORT bezüglich ihrer Interaktionen eingeteilt [6]. Auf dieser Basis wurden die Daten gegliedert und auf die Stimmungen innerhalb dieser Kategorien untersucht. Dafür wurden diese mit Hilfe von Sentiment-Analyse-Tools auf ihre Stimmungen klassifiziert und analysiert. Für diese Untersuchung wurde am Anfang dieser Arbeit die folgende Forschungsfrage gestellt.

Forschungsfrage: Gibt es Zusammenhänge zwischen den Interaktionen in einem Softwareprojektmeeting und den Stimmungen dieser Interaktionen und wo liegen die gegebenenfalls?

Mittels Untersuchung der Ergebnisse aus dem vorherigen Kapiteln lässt sich diese Forschungsfrage beantworten. Insgesamt lässt sich feststellen, dass Aussagen über alle Interaktionen hinweg zum größten Teil eine neutrale Stimmung der Teilnehmer zugrunde liegt. In den einzelnen Interaktionskategorien lassen sich verschiedene Verteilungen der Stimmungen erkennen. Daher gibt es zwischen diesen einzelnen Kategorien Unterschiede. Aufgrund dieser Verteilung lassen sich zwischen den Interaktionen und Stimmungen Zusammenhänge feststellen. Zum Beispiel ist bei „*proaktivem Verhalten*“ zu erkennen, dass diese Aussagen einen

größeren Anteil an positiven Aussagen haben als die anderen Kategorien. Auch ist bei „*Kollegialem Verhalten*“ ein höherer Anteil an positiven als negativen Aussagen vorhanden. Damit lässt sich erkennen, dass die Verteilungen der einzelnen Interaktionen mit den jeweiligen Stimmungen in diesen Kategorien zusammen hängen.

5.2 Interpretation der Ergebnisse

Bei der „*Problemvernetzung*“ gibt es Unterschiede bezüglich der Verteilung der Polaritäten zwischen dem deutsche und dem englischen Voting-Classifier. Ein genaueres Betrachten zeigt, dass der deutsche Voting-Classifier mehr Aussagen positiv klassifiziert. Die stichprobenartige Betrachtung der Aussagen wie „Deshalb wären bei uns die Emails auch ganz wichtig.“, welche positiv klassifiziert wurde, zeigt dass es zu positiven Klassifizierungen kommt, welche wahrscheinlich neutral einzuordnen sind. Beim englischen Voting-Classifier wurden mehr Aussagen negativ klassifiziert, welche beim deutschen neutral klassifiziert sind. Auch hier zeigt eine stichprobenartige Betrachtung von Aussagen, wie „That's not really the problem.“, welche negativ klassifiziert ist, dass diese Aussage wahrscheinlich neutral ist. Somit scheint hier eine eher gleiche Verteilung vorzuliegen.
Die Betrachtung von „*Destruktives Verhalten*“ legt die Vermutung nahe, dass mehr negativ klassifizierte Aussagen auftreten als positive. Der deutsche Voting-Classifier klassifiziert allerdings mehr Aussagen positiv als negativ. Das Betrachten einiger dieser Aussagen wie „Okay Jungs lass und mal wieder ein bisschen konzentieren bitte.“, welche positiv klassifiziert wurde, zeigt dass es bei einigen Aussage zu einer positiven Klassifizierung kommt, obwohl diese wahrscheinlich keine positive Stimmung widerspiegeln. Somit scheint hier der englische Voting-Classifier akkurater zu sein bei der Klassifizierung und es mehr negative als positive Aussagen zu geben.
In der Kategorie „*Verknüpfung und Vernetzung*“, labelt der deutsche Voting-Classifier mehr Aussagen positiv als negativ. Dabei werden Aussagen wie „ich finde das immer wichtig“ oder „mir ist jetzt noch wichtig“ als positiv eingeordnet. Diese scheinen aber eine eher neutrale Stimmung zu haben. Daher scheint hier auch die Klassifizierung des englische Voting-Classifiers eher akkurat zu sein, welcher mehr Aussagen negativ klassifiziert als positiv.
Bei „*Wissenstransfer*“ und „*Informationsweitergabe*“ zeigt die Betrachtung einiger positiver Aussagen, dass diese warhscheinlich neutral einzuordnen sind, wie „Ja, es muss auch klar definiert werden.“. Diese Aussagen stammen vor allem aus der act4teams Kategorie „*Organisationales Wissen*“. Der englische Voting-Classifier, hat diese Aussagen öfter in neutral eingeordnet. Diese Aussagen stammen am meisten aus der act4teams-Kategorie „*Organisationales Wissen*“. „*Wissenstransfer*“ scheint daher eine ungefähr gleich viele positive und negative Aussagen zu haben.
„*Problembennenung*“ hat eine Verteilung analog zu „*Wissenstransfer*“ und „*Informationsweitergabe*“. Hier kommen auch die meisten Aussagen aus der act4teams-Kategorie „*Organisationales Wissen*“. Daher scheint auch hier die Verteilung des englischen Voting-Classifiers akkurater zu sein, womit es ungefähr gleich viele positive und negative Aussagen in dieser Kategorie gibt.

In den einzelnen Kategorien sieht man, dass es eine verschiedene Verteilung der Sentimente gibt. Somit kann man davon ausgehen, dass die Verteilung der Stimmung mit den Interaktionen zusammenhängt. Die Aussagen der Teilnehmer scheinen entsprechend der Kategorie von act4teams-SHORT eine gewisse Verteilung der Stimmung zu besitzen. Für alle Kategorien gilt zwar, dass die Stimmung in den meisten Fällen neutral ist, allerdings variiert die genaue Verteilung der nicht neutralen Aussagen je nach Kategorie. Zum Beispiel enthalten „*Proaktives Verhalten*“ und „*Kollegiales Verhalten*“, deutlich mehr positive als negative Aussagen. Daher lässt sich bei diesen Kategorien abschätzen, dass die Stimmung der Meetingteilnehmer bei diesen Aussagen allgemein mehr positiv sind. Bei „*Destruktives Verhalten*“ dagegen, lassen sich mehr negative Aussagen erkennen. Dies lässt den Schluss zu, dass die Teilnehmer bei destruktiven Aussagen öfter eine negative als positive Stimmung haben. Somit scheinen die einzelnen Kategorien Auswirkungen auf die Verteilung der Stimmung in diesen zu haben, sodass dort eine Korrelation besteht. Insgesamt scheint es so, dass es einen Zusammenhang zwischen den Interaktionen und den Stimmungen der Meetingteilnehmer gibt und diese miteinander verknüpft sind. Auch wenn anzumerken ist, dass die Stimmungen der Teilnehmer, bei den meisten Aussagen die getroffen werden, neutral ist.
Da es Verbindungen zwischen der Zufriedenheit von Entwicklern und ihrer Produktivität gibt [6], können diese genauer betrachtet werden, um das Verhalten und somit auch die damit verbundene Stimmung der Entwickler zu erfassen. Mit diesen Ergebnissen ist es dann möglich, diese Einflüsse in einem Meeting zu erkennen. Somit können Meetings durch geeignetes Entgegensteuern wieder auf Kurs gebracht werden [6].

5.3 Threats of Validity

In diesem Kapitel wird auf die Einflussfaktoren eingegangen, welche die Validität der Arbeit beschränken können und die Generalisierbarkeit der Arbeit beeinflussen. Dazu werden die einzelnen Threats of Validity aufgezeigt und es wird benannt, welche Maßnahmen gegebenenfalls getroffen wurden, um diesen entgegen zu wirken.

5.3.1 Internal Validity

Bei dieser Arbeit sind Änderungen am ursprünglichen Datensatz vorgenommen worden. Es wurde darauf geachtet, dass vorgenommene Anpassungen am Datensatz möglichst keine Auswirkung auf die Klassifizierung der Sentiment-Analyse-Tools haben und dadurch die Ergebnisse beeinflusst werden. Allerdings ist nicht auszuschließen, dass die vorgenommen Änderungen Auswirkungen auf die Klassifizierung der Daten hatten.
Auch die Übersetzung des Datensatzes stellt eine Bedrohung der internen Validität dar. Dabei kann allerdings nicht ausgeschlossen werden, dass die Übersetzung keinen Einfluss auf die Ergebnisse der Klassifizierung der Aussagen hatte. Um

dieses zu mindern, wurde die Übersetzung wurde durch schnelles Durchgehen und stichprobenhafte Kontrolle einzelner Aussagen überprüft.
Eine weitere Bedrohung der internen Validität ist die zufällige Auswahl von Testdaten zur Bestimmung der Eignung der einzelnen Tools für die vorhandenen Daten. Die Auswahl der Daten und deren manuelle Klassifizierung können die Auswahl der Tools beeinflussen. Der Testdatensatz auf dem die Tools getestet wurden, um die Genauigkeit zu bestimmen, sind abhängig von den Labeln, die an den Testdaten angebracht wurden. Somit kann die Auswahl der Tools von dem Testdatensatz beeinflusst worden sein.
Eine andere mögliche Bedrohung der internen Validität ist das Mapping von act4teams zu act4teams-SHORT. Bei dem Mapping werden nicht alle Kategorien von act4teams-Kategorien direkt zu act4teams-SHORT gemappt. Dies kann dazu führen, dass die Zuordnungen von den Aussagen in die act4teams-SHORT-Kategorien nicht mit denen übereinstimmen, welche bei einem direkten Meeting vorgenommen worden wären. Um das Mapping möglichst genau zu halten, wurde das Mapping auf Basis der Arbeit von Klünder [7] vorgenommen und die notwendigen Erweiterungen in Kapitel 3.3 beschrieben.

5.3.2 External Validity

Auch ist anzumerken, dass es sich bei dem genutzten Datensatz um Meetings studentischer Projekte handelt. Das bedeutet, dass es sich bei den Teilnehmern um Studenten handelt, welche nicht alle die Erfahrungen haben, die ein beruflich professionelles Entwicklerteam hat. Somit kann nicht generell davon ausgegangen werden, dass diese Daten sich auch für eine professionelle Umgebung verallgemeinern lassen. Eine Möglichkeit dem vorzubeugen wäre es, einen Datensatz aus einem professionelleren oder beruflichen Umfeld zu nutzen, um so sicher zu stellen, dass die Teilnehmer des Meetings mehr Erfahrungen haben.

5.3.3 Conclusion Validity

Bei der Klassifizierung der einzelnen Aussagen aus dem Datensatz, kann es dazu kommen, dass einige von diesen zufällig gelabelt sind. Jenes kann aufgrund des Voting Classifiers geschehen, da dieser bei einer gleichen Verteilung der Stimmen zufällig entscheidet, welches von den gewählten Sentimenten genommen wird. Diese Zufallsentscheidung sorgt dafür, dass die Ergebnisse bei wiederholter Ausführung unterschiedlich ausfallen können. Die Anzahl der Zufallsentscheidungen liegen hierbei bei dem Voting-Classifer mit den deutschsprachigen Tools bei 2,04% und bei dem Voting-Classifier mit den englischsprachigen Tools bei 1,25%. Um dies zu minimieren, wurde mit Hilfe von Fleiss' Kappa überprüft, wie einig sich die Tools bei der Klassifizierung sind. Für den deutschprachigen Voting-Classifier liegt dieser Wert bei 0,27 und bei dem englischprachigen Voting-Classifier liegt dieser Wert bei 0,21. Nach Landis und Koch [9] handelt es sich für beide Voting-Classifier um einen angemessenen Wert.

5.3.4 Construct Validity

Die meisten genutzten Tools sind nicht für die Domäne des Software Engineering entwickelt worden. Das kann dazu führen, dass das Einfluss auf die Klassifizierung des Datensatzes hatte. Um sicher zu stellen, dass die Auswahl der Tools best möglich getroffen wurde, wurden bei der Betrachtung mehrere Metriken hinzugezogen. Als relevanteste Metrik zur Auswahl des Tools wurde die Accuracy gewählt. Um eine best mögliche Auswahl unter den Tools zu treffen, wurden als Erweiterungen die Metriken Precision, Recall und F1-Measurement hinzugezogen. Somit werden mehrere Aspekte der Tools gegenüber gestellt und betrachtet, um einem Threat of Construct Validity entgegen zu wirken.

Kapitel 6

Verwandte Arbeiten

Die Arbeit beschäftigt sich mit den Bereichen der Interaktionsanalyse und der Stimmungsanalyse und deren Zusammenhänge. Dabei werden andere Arbeiten vorgestellt, welche sich mit ähnlichen Themen oder mit Teilen dieser Themen beschäftigen. Im ersten Abschnitt 6.1 werden zuerst Arbeiten vorgestellt, welche sich mit der Interaktionsanalyse in Meetings von Softwareprojekten beschäftigen. Im zweiten Abschnitt 6.2 wird auf Arbeiten eingegangen, welche sich mit der Thematik der Stimmungsanalyse in Meetings von Softwareprojekten auseinander setzen. Anschließend wird in Abschnitt 6.3 aufgezeigt, wie sich diese Arbeiten von der hier genannten Arbeiten unterscheidet.

6.1 Interaktionsanalyse

Das Verfahren, welches für die Untersuchung der Interaktionen verwendet wurde, basiert auf der Arbeit von Kauffeld et al. [5]. In ihrer Arbeit entwickelten die Autoren ein Verfahren, mit welchem diese Interaktionen untersucht werden können. Dieses Verfahren nennen sie *act4teams*, mit welchem es möglich ist die Interaktionen in einem Meeting in verschiedene Kategorien aufzuteilen und diese dann analysieren zu können.
Eine weitere relevante Arbeit ist von Klünder et al. [6]. Bei dieser geht es um die Analyse von Teaminteraktionen in Teammeetings. Dafür wurde untersucht, wie die einzelnen Mitglieder am Anfang eines Projektes interagieren und wie sich dies über die Zeit ändert. Auch wurde betrachtet, ob in den Meetings genannte Probleme gelöst werden. Um das zu untersuchen nutzten die Autoren das Verfahren von *act4teams-SHORT*, welches auf *act4teams* basiert. Dabei stellten sie fest, dass ein Großteil der Zeit zum Austausch von Informationen genutzt wurde. Zudem wurden genannte Probleme in Zusammenhang mit Lösungen gesetzt. Somit lies sich am Ende eine gute Interaktionsverteilung erkennen [6].
Im Rahmen der Dissertation von Künder [7], wurde unter anderem ein Verfahren gezeigt, um die Kategorisierung von Interaktionen, welche in act4teams kategorisiert sind, zu act4teams-SHORT zu überführen. Dabei wird erklärt, was die Ziele von act4teams-SHORT sind und wo dessen Vorteile liegen. Im Verlauf der Dissertation wird darauf eingegangen, wie ein Mapping zwischen act4teams und act4teams-SHORT vorgenommen werden kann [7].
In ihrer Arbeit untersuchten Schneider et al. [15] die Interaktionen in 32 Meetings

von Studentenprojekten auf Basis des act4teams-Schemas. Die Ergebnisse der Analyse zeigte, das konstruktive Aussagen einen positiven Einfluss auf den Gruppenaffekt haben, wenn auf diese unterstützenden Aussagen folgen [15]. Dabei stellten sie fest, dass subtile Verhaltensmuster den Gruppenaffekt beeinflussen können. Softwareprojekte könnten daher von unterstützendem Verhalten profitieren [15].

6.2 Stimmungsanalyse

Ein wichtiger Bereich dieser Arbeit ist der der Stimmungsanalyse. Dafür relevant ist die Arbeit von Islam et al. [4]. In ihrer Arbeit haben die Autoren ein Sentiment-Analyse-Tool entwickelt, basierend auf einem bereits existierenden Tool[4]. Das Ziel war es dabei, dass Tool auf die Domäne des Software Engineering arbeitet. Das Ergebnis ihrer Arbeit zeigte, dass die Entwicklung eines domänenspezifischen Tools, die Genauigkeit der Klassifizierung innerhalb dieser Domäne verbessert. Somit arbeitet das erstellte Tool auf der Domäne besser, als das Tool, welches nicht für diese gedacht ist[4].
Eine weitere relevante Arbeit ist die Arbeit von Tymann et al. [21]. In dieser untersuchen die Autoren, ob die Übersetzung eines Datensatzes in die englische Sprache sinnvoll ist, wenn anschließend ein englischsprachiges Sentiment-Analyse-Tool genutzt wird, statt eines deutschen [21]. Um dies zu untersuchen haben die Autoren ihr entwickeltes Tool GerVADER mit dem Tool VADER verglichen. Dabei stellten sie fest, dass das englischsprachige Tool VADER, auf dem gleichen Datensatz in übersetzter Form, besser abschneidet, als das deutschsprachige Tool GerVADER [21].

6.3 Abgrenzung von anderen Arbeiten

Die hier geschriebene Arbeit teilt einige gemeinsame Aspekte mit den hier genannten Arbeiten. Es gibt Verbindungen in den Bereichen der Untersuchung der Interaktionen von Meetings in Softwareprojekten und auch Überschneidungen bei der Klassifizierung und Analyse von Stimmungen in dieser Domäne. Diese Arbeit beruht auf den hier vorgestellten Arbeiten und baut auf deren Ergebnissen auf. Diese Arbeiten beschäftigten sich mit einzelnen Aspekten dieser Arbeit, wie die Interaktionsanalyse oder der Stimmungsanalyse. Diese Arbeit hat allerdings als Ziel, Interaktionen und Stimmungen auf Zusammenhänge zu untersuchen. Somit werden in dieser Arbeit die beiden Punkte Interaktionsanalyse und Stimmungsanalyse miteinander verbunden und im Zusammenhang untersucht und nicht einzeln.

Kapitel 7

Zusammenfassung und Ausblick

Dieses Kapitel dient dazu, einen abschließenden Überblick über die Inhalte und die wichtigsten Ergebnisse der Arbeit zu geben. Dafür wird in Abschnitt 7.1 die Problemstellung noch einmal aufgegriffen und die gewonnenen Erkenntnisse der Arbeit zusammengefasst. Anschließend wird in Abschnitt 7.3 ein Ausblick gezeigt, in welchem weitere Möglichkeiten genannt werden, um auf Basis dieser Arbeit weiter aufzubauen.

7.1 Zusammenfassung

Bei dieser Arbeit wurde sich mit den Zusammenhängen von Interaktionen und Stimmungen von Meetings in Softwareprojekten beschäftigt. Dabei wurde untersucht, ob Zusammenhänge zwischen den Interaktionen und den Stimmungen existieren und was für Zusammenhänge es sich gegebenenfalls handelt. Um dies zu erreichen, wurde ein Datensatz aus dieser Domäne auf die Interaktionen und die jeweiligen Stimmungen der Aussagen innerhalb dieser Interaktionsgruppen untersucht.
Im ersten Schritt wurde dafür der Datensatz für die weitere Verarbeitung vorbereitet. Um das zu erreichen, wurden am Datensatz Änderungen vorgenommen, sowohl bei der Formatierung als auch bei der inhaltlichen Form. Das war nötig, um eine eventuelle Beeinflussung der Klassifizierung der Aussagen durch die Sentiment-Analyse-Tools zu verhindern.
Im nächsten Schritt wurde der vorhandene Datensatz auf die enthaltenen Interaktionen untersucht, mit Hilfe der Einteilung in act4teams-Kategorien [5], welche bei dem Datensatz bereits vorhanden waren. Diese Kategorien wurden auf Grundlage der Arbeit von Klünder [7] zu act4teams-SHORT-Kategorien
gemappt[6][7]. Mit diesen ist es möglich, die Interaktionsverteilung des Meetings gut erkennen und zu untersuchen[6]. Im Anschluss wurde ermittelt, welche Sentiment-Analyse-Tools sich gut für die Benutzung auf dem vorhandenen Datensatz eignen. Dafür wurden mehrere Tool getestet und mit geeigneten Metriken auf ihre Performance auf dem vorhandenen Datensatz untersucht. Zusätzlich wurde dies mit englischsprachigen Tools auf Basis einer Übersetzung des Datensatzes getestet. Durch die Verwendung eines Voting-Classifiers konnte die Performance für die deutschsprachigen und die englischsprachigen Tools, nochmal verbessert werden. Die Voting-Classifier wurden genutzt, um die einzelnen Aussagen des

Datensatzes, geordnet nach den act4teams-SHORT-Kategorien, auf die einzelnen Polaritäten positiv, neutral und negativ zu untersuchen. Somit sind die Verteilungen der Polaritäten in den einzelnen act4teams-SHORT-Kategorien dargestellt. Die daraus erhaltenen Ergebnisse wurden anschließend untersucht und unter Berücksichtigung der Forschungsfrage ausgewertet .
Dabei zeigte sich, dass die einzelnen Interaktionskategorien verschiedene Verteilungen an Stimmungen aufweisen. Bei näherem Betrachten lassen sich Beziehungen in Form von Verteilung der Sentimente feststellen. Des weiteren zeigte sich bei allen untersuchten Kategorien, dass der größte Anteil an Aussagen eine neutrale Polarität besitzt und somit eine neutrale Stimmung der Teilnehmer bei diesen Aussagen widerspiegelt. Somit sind Zusammenhänge zwischen den Interaktionen und die Stimmungen bei Softwareprojektmeetings innerhalb dieser Kategorien zu sehen. Da es Zusammenhänge gibt zwischen der Produktivität von Entwicklern und deren Zufriedenheit, können mit diesen Ergebnissen die Meetings genauer betrachtet werden. Somit können eventuelle Probleme frühzeitig erkannt werden und diesen kann dann entsprechend entgegengesteuert werden.

7.2 Verbesserungsmöglichkeiten

Eine Möglichkeit eine Verbesserung wäre es eine größere Anzahl an Sentiment-Analyse-Tools zu nutzen. Durch die Betrachtung weiterer bestehender Tools, könnten welche gefunden werden, welche auf dem vorhandenen Datensatz eine bessere Performance aufweisen. Dies würde dabei helfen, die Genauigkeit bei der Klassifizierung zu erhöhen und dadurch gegebenenfalls zu einem genaueren Ergebnis der Kategorisierung führen.
Um die Genauigkeit der Tools auf dem Datensatz zu verbessern, könnte ein deutschsprachiges Sentiment-Analyse-Tool gewählt werden, welches für diese Domäne entwickelt wurde. Auch die Nutzung eines Wörterbuches für ein Sentiment-Analyse-Tool, welches für diese Domäne entwickelt wird, wäre eine mögliche Verbesserung. Diese müssten extra für diese Domäne auf deutsch entwickelt werden. Wenn ein solches Tool oder Wörterbuch verwendet werden würde, sollte eine Verbesserung bei Klassifikation bei Betrachtung der gleichen Metriken vorliegen, da diese genauer auf die vorhandene Domäne angepasst wären [12]. Ein solches Tool könnte für die Klassifizierung der Aussagen benutzt werden oder auch ein solches Wörterbuch könnte in bestehende Tools eingebunden werden.
Bei der Kategorisierung der Interaktionen könnte zukünftig für den Datensatz direkt act4teams-SHORT genutzt werden. Wenn dieses direkt beim Meeting gelabelt wird, wäre kein Mapping von act4teams [5] zu act4teams-SHORT [6] notwendig. Somit könnte sichergestellt werden, dass die Daten direkt mit act4teams-SHORT-Kategorien gelabelt sind und es zu keinen Abweichungen bei der Zuordnung vom act4teams zu act4teams-SHORT kommen kann.

7.3 Ausblick

Die Untersuchung von Zusammenhängen von Interaktionen und Stimmungen von Meetings in Softwareprojekten kann später für eine erweiterte Betrachtung von Interaktionsverteilungen genutzt werden. Bei einer Untersuchung von Interaktionen auf Grundlage von act4teams-SHORT könnten somit Rückschlüsse auf die Stimmungen der Teilnehmer innerhalb der Kategorien gezogen werden. Dabei sollte allerdings auf die Domäne geachtet werden.
Bei dem hier untersuchten Datensatz, handelte es sich um Meetings studentischer Softwareprojekte. Es kann dadurch nicht ausgeschlossen werden, dass die Erkenntnisse nur auf diese spezielle Domäne mit studentischen Teilnehmern eingeschränkt werden. Somit wäre dies ein guter Ansatzpunkt in Zukunft auch die Verwendung von anderen Datensätzen zum erweitern dieser Arbeit in Betracht zu ziehen. Es könnten Datensätze verwendet werden, welche aus beruflichen Meetings von Softwareprojekten stammen, um eine bessere Generalisierbarkeit der Ergebnisse zu erhalten. Damit können die Ergebnisse auf verschiedene Arten von Softwareprojektmeetings angewendet werden. Somit könnte eine Erweiterung der Betrachtung von Softwareprojektteams zu Arbeitsteams vorgenommen werden.
Teammeating sind in verschiedenen beruflichen Feldern relevant [8] und nicht nur in der Softwareentwicklung. Daher ist es möglich, die Betrachtung über den Bereich der Meeting in der Softwareentwicklung hinaus zu machen. Es wäre möglich, Meetings in anderen beruflichen Bereichen zu betrachten und somit die Domäne nicht nur auf Software Engineering zu beschränken. Diese verschiedenen Ansätze bieten gute Möglichkeiten für weitere Forschungen in dieser Richtung.

Anhang A

Anhang

A.1 Tabellen zu act4teams und act4teams-SHORT

Tabelle A.1: Professionelle Kompetenzen mit Codes nach Klünder [7]

Professionelle Kompetenz		
Problem **P**	Sollentwurf **SL**	Organisationales Wissen **WO**
Problemerläuterung **PE**	Lösungsvorschlag **L**	
	Lösungserläuterung **LE**	
Verknüpfung Problemanalyse **V**	Problem zu Lösung **PL**	Wissen wer **WW**
	Verknüpfung mit Lösung **VL**	Frage **F**

Tabelle A.2: Methodenkompetenz mit Codes nach Klünder [7]

Methodenkompetenz		
Zielorientierung **Z**	Zeitmanagement **ZT**	Verlieren in Details und Beispielen **DB**
Klärung/Konkretisierung **K**	Aufgabenverteilung **A**	
Verfahrensvorschlag **VV**	Visualisierung **VIS**	
Verfahrensfrage **VF**	Kosten-Nutzen-Abwägung **KN**	
Priorisieren **PRIO**	Zusammenfassungtextbf **ZSF**	

Tabelle A.3: Sozialkompetenz mit Codes nach Klünder [7]

Sozialkompetenz		
Ermunternde Ansprache **EA**	Atmosphärische Auflockerung **ATM**	Tadel/Abwertung **TD**
Unterstützung **ZUST**	Ich-Botschaft: Trennung von Meinung und Tatsache **IB**	Unterbrechung **Unt**
Aktives Zuhören **AZ**	Gefühle **G**	Seitengespräche **Seit**
Ablehnung **ABL**	Lob **Lob**	Reputation **R**
Rückmeldung **RM**		

Tabelle A.4: Selbstkompetenz mit Codes nach Klünder [7]

Selbstkompetenz		
Interesse an Veränderung **IN**	Kein Interesse an Veränderungen **KI**	Schuldigensuche **S**
Eigenverantwortung **EV**	Jammern **J**	Betonung autoritärer Elemente **AE**
Maßnahmenplanung **MP**	Abbruch **E**	Phrase **AL**

Tabelle A.5: act4teams zu act4teams-SHORT nach Klünder [7]

Kategorie bei act4teams-SHORT	**Kategorie bei act4teams**
Problembennenung	Problem, Problemerläuterung
Problemvernetzung	Verknüpfung bei der Problemanalyse
Lösungsbenennung	Lösungsvorschlag, Lösungserläuterung
Lösungsvernetzung	Verknüpfung mit Lösung, Sollentwurf
Verknüpfung und Vernetzung	Problem zur Lösung
Destruktives Verhalten	Tadel/Abwertung, Schuldigensuche, Jammern, Kein Interesse an Veränderungen, Seitengespräche
Methodisch-strukturierendes Verhalten	Zielorientierung, Priorisieren, Verfahrensvorschlag, Verfahrensfrage, Zusammenfassung, Klärung/Konkretisierung, Visualisierung
Proaktives Verhalten	Interesse an Veränderungen, Eigenverantwortung, Maßnahmenplanung
Wissensaustausch	OrganisationalesWissen,Wissen Wer
Informationsweitergabe	OrganisationalesWissen,Wissen Wer
Kollegiales Verhalten	Atmosphärische Auflockerung, Ermunternde Ansprache, Lob

Literaturverzeichnis

[1] textblob-de. `https://github.com/markuskiller/textblob-de`. Accessed: 2020-07-07.

[2] textblob-de. `https://github.com/sloria/textblob`. Accessed: 2020-07-07.

[3] C. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. 01 2015.

[4] 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.

[5] S. Kauffeld and N. Lehmann-Willenbrock. Meetings matter effects of team meetings on team and organizational success. *Small Group Research*, 43:130–158, 04 2012.

[6] J. Klünder, N. Prenner, A.-K. Windmann, M. Stess, M. Nolting, F. Kortum, L. Handke, K. Schneider, and S. Kauffeld. Do you just discuss or do you solve? meeting analysis in a software project at early stages. ICSEW'20, page 557–562, New York, NY, USA, 2020. Association for Computing Machinery.

[7] J. A.-C. Klünder. *Analyse der Zusammenarbeit in Softwareprojekten mittels Informationsflüssen und Interaktionen in Meetings*. dissertation, Gottfried Wilhelm Leibniz Universität Hannover, 2018.

[8] S. W. Kozlowski and D. R. Ilgen. Enhancing the effectiveness of work groups and teams. *Psychological Science in the Public Interest*, 7(3):77–124, 2006. PMID: 26158912.

[9] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159, mar 1977.

[10] H. Le and N. Trong. A sentiment analyzer for informal text in social media. *Journal of Science and Technology*, 131:6–12, 11 2018.

[11] F. Leon, S.-A. Floria, and C. Badica. Evaluating the effect of voting methods on ensemble-based classification. pages 1–6, 07 2017.

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

[13] R. Remus, U. Quasthoff, and G. Heyer. Sentiws - a publicly available german-language resource for sentiment analysis. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

[14] P. Schaer, P. Mayr, and P. Mutschke. Implications of inter-rater agreement on a student information retrieval evaluation. In M. Atzmüller, D. Benz, A. Hotho, and G. Stumme, editors, *LWA 2010 : Lernen, Wissen Adaptivität ; workshop proceedings*, volume 2010/5 of *Kasseler Informatikschriften (KIS)*, pages 1–7. Universität Kassel, Knowledge and Data Engineering Group, Kassel, 2010.

[15] K. Schneider, J. Klünder, F. Kortum, L. Handke, J. Straube, and S. Kauffeld. Positive affect through interactions in meetings: The role of proactive and supportive statements. *Journal of Systems and Software*, 143, 05 2018.

[16] M. Taboada. Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*, 2, 02 2016.

[17] A. Tharwat. Classification assessment methods. *Applied Computing and Informatics*, 17(1):168–192, 2021.

[18] M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63:163–173, 01 2012.

[19] 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.

[20] K. Tymann, M. Lutz, P. Palsbröker, and C. Gips. Gervader -a german adaptation of the vader sentiment analysis tool for social media texts. 09 2019.

[21] K. Tymann, L. Steinkamp, O. Zhurakovskaya, and C. Gips. Native sentiment analysis tools vs. translation services - comparing gervader and vader. 09 2020.

[22] A. Weeks, H. Swerissen, and J. Belfrage. Issues, challenges, and solutions in translating study instruments. *Evaluation review*, 31:153–65, 05 2007.