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

Untersuchung der Eignung von Deep Learning für die Identifizierung von Erklärungsbedarf aus Gesichtsaufnahmen

Exploring the suitability of deep learning for identifying explanation needs from facial imaging

Bachelorarbeit

im Studiengang Informatik

von

Sophie Seifert

Prüfer: Prof. Dr. rer. nat. Kurt Schneider
Zweitprüfer: Dr. rer. nat. Jil Ann-Christin Klünder
Betreuer: Hannah Luca Deters

Hannover, 30.08.2024

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 30.08.2024

Sophie Seifert

Zusammenfassung

In den letzten Jahren haben Software-Systeme eine immer größer werdende Relevanz in unserem Alltag gewonnen. Jedoch werden diese auch immer komplexer und machen es Nutzenden somit schwer, sie tiefer und nicht nur oberflächlich zu verstehen.
Obwohl es Literatur zur Erklärbarkeit von Software gibt, ist noch kein bestimmtes Vorgehen bekannt, mit dem es möglich ist, den Erklärungsbedarf der Nutzer zur Laufzeit zu erkennen und daraufhin Erklärungen zur Software nur dann zu geben, wenn diese benötigt werden.
Um eine mögliche Lösung für dieses Problem zu finden, wurde in dieser Arbeit die Eignung von Gesichtserkennung getestet, um Erklärungsbedarf zu erkennen. Dazu wurde eine Studie mit 15 Teilnehmern durchgeführt, bei der diese gefilmt sowie der Bildschirm aufgenommen wurde. Innerhalb der Studie haben die Teilnehmer verschiedene Aufgaben durchgeführt, bei denen zum Teil Erklärungsbedarf ausgelöst werden sollte. Dabei wurden *Instruction* und *Algorithm* als Arten von Erklärungsbedarf genutzt. Danach wurde der tatsächliche Erklärungsbedarf, der durch eine anschließende Umfrage notiert wurde, mit den Ergebnissen der Analyse mittels Gesichterkennung verglichen. Dazu wurde zum einen eine Bibliothek mit *Facial Emotion Recognition* genutzt sowie eine manuelle Gesichtserkennungsanalyse durchgeführt.
Die höchste Genauigkeit der Vorhersage wurde bei der *Algorithm*-Aufgabe erzielt. Des Weiteren konnte die Genauigkeit der Vorhersage gesteigert werden, indem lediglich die Emotionen *sad* und *angry* bei der Analyse durch die Bibliothek betrachtet wurden. Eine generell zuverlässige Vorhersage hingegen konnte durch die Gesichtsanalyse nicht erzielt werden. Die Analyse durch die *Facial Emotion Recognition*-Bibliothek ergab einen F1-Score von rund 41%, die manuelle Analyse immerhin von circa 68%.

Abstract

Exploring the suitability of deep learning for identifying explanation needs from facial imaging

In recent years, software systems have become increasingly relevant in our everyday lives. However, they are also becoming more and more complex, making it difficult for users to understand them in depth rather than just superficially.
Although literature on the explainability of software exists, no specific procedure is known that makes it possible to recognize the explanation need of the user at runtime and then only provide explanations about the software when they are needed.
In order to find a possible solution to this problem, this work tested the suitability of facial recognition to identify the need for explanation. A study was conducted with 15 participants, in which they were filmed and the screen was recorded. During the study, the participants carried out various tasks, some of which were intended to trigger a need for explanation. *Instruction* and *Algorithm* were used as types of explanation need. The actual need for explanation, which was noted in a subsequent survey, was then compared with the results of the analysis using facial recognition. A library with facial emotion recognition was used, and a manual facial recognition analysis was carried out.
The highest accuracy of the prediction was achieved with the *Algorithm* task. Furthermore, the accuracy of the prediction could be increased by only considering the emotions *sad* and *angry* in the analysis by the library. A generally reliable prediction, however, could not be achieved by the facial analysis. The analysis by the facial emotion recognition library resulted in an F1 score of around 41%, while the manual analysis resulted in around 68%.

Inhaltsverzeichnis

Kapitel 1

Einleitung

In den letzten Jahren haben Softwaresysteme immer mehr Relevanz in unserem Alltag und generell unserem Leben gewonnen. Diese Softwaresysteme werden immer komplexer. Aus diesem Grund ist es wichtig, dem Nutzer dieses System offener darzulegen, damit er diesem nicht kritisch und mit Angst gegenübersteht[16]. Dabei ist es von besonderer Relevanz, den Nutzern ein Verständnis für die Verhaltensweisen des Systems zu geben und worauf diese basieren. Ebenfalls wichtig ist aber auch zu erklären, wie sie mit dem System interagieren und dieses benutzen können[3][18].
Zu viele nicht benötigte Erklärungen sind möglicherweise störend[4], da sie den Nutzungsablauf der Software unterbrechen. Zu wenige Erklärungen können zu Verwirrung und einem Mangel an Transparenz, Vertrauen sowie Akzeptanz führen[4][2]. Aufgrund dessen ist es wichtig, den Umfang und die Stelle für die benötigten Erklärungen zu ermitteln.

1.1 Problemstellung

Wie oben erklärt ist es besonders wichtig, den Nutzer mit genügend Informationen über das System auszustatten. So ist es notwendig herauszufinden, wann der Nutzer das Bedürfnis für Informationen und auch nach welcher Art von Informationen hat. Dabei ist der Wunsch nach Erklärung von Nutzer zu Nutzer unterschiedlich[3] und eine Einteilung in Nutzergruppen zur Einordnung wäre zu unspezifisch und lediglich eine Hilfe in der anfänglichen Softwareentwicklung[10].
Ein neuer Ansatz, um Erklärungsbedarf zu ermitteln, ist die Nutzung von Gesichtserkennungssoftware, welche in der Lage ist, die Emotionen des Nutzenden zu identifizieren. Somit könnte man einen Anhaltspunkt gewinnen, was den Erklärungsbedarf des Benutzers angeht. Dieses könnte dann individuell für jeden Einzelnen aussagen, ob dieser in dem bestimmten

Moment einer Erklärung bedarf.
Dabei stellt sich die Frage, ob eine bestimmte Veränderung der Mimik generell mit dem Aufkommen des Wunsches nach Erklärung zusammenhängt. Des Weiteren muss herausgefunden werden, ob diese sogenannte FER (Facial Emotion Recognition) für die Erkennung von Erklärungsbedarf präzise genug entwickelt ist.

1.2 Lösungsansatz

In dieser Arbeit soll ermittelt werden, ob und welche Modelle der FER zur gesonderten Ermittlung des Erklärungsbedarfs eines Einzelnen herangezogen werden können. Dazu wird im Folgenden eine Nutzerstudie durchgeführt, bei der konkrete Aufgaben vorgegeben werden, die die Teilnehmer ausführen und durch die verschiedene Arten des Erklärungsbedarfs ausgelöst werden sollen. Diese Durchführung wird anschließend mit einem möglichst präzisen Tool mit den passenden Emotionserkennungen ausgewertet.
Um die Genauigkeit des Tools messen zu können, werden die Teilnehmer am Ende gebeten, einen Fragebogen auszufüllen. Dies macht es möglich, den tatsächlichen Erklärungsbedarf während der Durchführung zu ermitteln.
Abschließend werden Videoaufnahmen der Durchführung mittels des FER-Tools ausgewertet und mit dem Fragebogen verglichen, sodass Rückschlüsse auf die Eignung eines solchen Tools für das Erkennen von Erklärungsbedarf gezogen werden können. Zusätzlich werden die Videoaufnahmen manuell ausgewertet, um feststellen zu können, ob eventuelle schlechte Ergebnisse lediglich mit der Genauigkeit der Facial Emotion Recognition zusammenhängen.

1.3 Struktur der Arbeit

Die Struktur dieser Arbeit ergibt sich wie folgt. Zunächst wird in Kapitel 2 auf wichtige Grundlagen zur Erklärbarkeit und Erklärungsbedarf sowie auf wichtige Elemente und Methoden der Emotionserkennung eingegangen. Dazu gehören Definitionen, verschiedene Arten des Erklärungsbedarfs, aber auch die Vorstellung von zwei Emotionserkennungs-Bibliotheken sowie deren Vorgehen. Des Weiteren gehe ich auf verwandte Arbeiten ein und grenze diese von dieser Arbeit ab.
In Kapitel 3 werden Forschungsfragen formuliert, die in dieser Arbeit beantwortet werden sollen. Anschließend wird in Kapitel 4 auf die Studie eingegangen. Hier werden die Planung der Studie, deren Inhalte sowie Informationen zu den Teilnehmern und des Ablaufs erläutert. In Kapitel 5 wird daraufhin auf die Verarbeitung der gesammelten Daten der Studie eingegangen. Kapitel 6 geht dann auf die Auswertung der Analyse ein und bewertet die Ergebnisse. Anschließend diskutiert Kapitel 7 die zuvor

beschriebenen Ergebnisse. Dort werden ebenefalls die Forschungsfragen beantwortet.
Abschließend wird in Kapitel 8 eine Zusammenfassung der Arbeit sowie ein Ausblick auf mögliche zukünftige Arbeiten gegeben.

Kapitel 2

Grundlagen

In den folgenden Unterkapiteln der Grundlagen wird auf die wichtigsten Definitionen eingegangen und das nötige Grundwissen für die Betrachtung der Arbeit näher erklärt.

2.1 Erklärbarkeit

2.1.1 Definitionen

Explainability Chazette et al. [3] haben folgende allgemeine Definition von *Explainability* formuliert:

> *A system S is explainable with respect to an aspect X of S relative to an addressee A in context C if and only if there is an entity E (the explainer) who, by giving a corpus of information I (the explanation of X), enables A to understand X of S in C.*

Diese Definition zeigt, dass die Erklärbarkeit des Systems sowohl von dem Aspekt X, als auch von dem Nutzer A abhängt. Aus diesem Grund werden in dieser Studie mehrere Nutzer sowie mehrere Aspekte untersucht.
Somit bezeichnet *Explainability* die Fähigkeit von Software, sich mittels Erklärungen dem Nutzer verständlich zu machen. Dadurch kann beispielsweise die Transparenz und das Vertrauen in die Software größer werden[5][2].

Explanation Need Unterbusch et al. [21] formulierten die folgende allgemeine Definition von *Explanation Need*:

> *An addressee A has incomplete knowledge about an aspect X of system S in context C and requests a corpus of information I provided by an entity E that allows A to understand X of S in C.*

Auch diese Definition macht deutlich, dass der Erklärungsbedarf von dem Nutzer A sowie von dem Aspekt X abhängt. Aspekte können die verschiedenen Arten von Erklärungsbedarf sein (siehe 2.1.3).

2.1.2 Nutzen von Erklärungen

Erklärungen sind im Bereich der Software von besonderer Relevanz, da diese zum einen das Verständnis des Systems fördern[13]. Des Weiteren helfen sie dem Nutzer, die Schlussfolgerungen und Ergebnisse einfacher nachzuvollziehen[4]. Die Nutzer erlangen durch die Erklärungen ein generell besseres Verständnis der Software und somit wird die Transparenz des Systems gefördert[4]. Dadurch fällt es ihnen leichter, dem System zu vertrauen[2][8][14]. Zudem stoßen Nutzer aufgrund von Erklärungen und besserem Verständnis auf weniger Frustration[22], was die Nutzung der Software ebenfalls angenehmer gestaltet.

2.1.3 Arten von Erklärungsbedarf

Unterbusch et al. [21] stellten eine Taxonomie[1] von Erklärungsbedarf auf, welche sich in *Primary Concern* und *Secondary Concern* unterteilt. Dabei beschreibt *Primary Concern* den Erklärungsbedarf, bei dem das mangelnde Wissen des Nutzers alleine steht. Bei einem *Secondary Concern* hingegen kommen noch weitere Probleme, wie zum Beispiel eine fehlende Funktion des Softwareprogramms, zu der Wissenslücke hinzu.
Die Kategorie *Primary Concern* unterteilt sich weiter in die Unterkategorien *Training*, *Interaction* und *Business*.
Es handelt sich um *Training*-Erklärungsbedarf, wenn die Benutzer noch nicht genug über das System oder über gewisse Systemteile wissen.
Wenn Endnutzer allerdings mit dem System vertraut sind, dieses aber unerwartetes Verhalten liefert, beschreibt dies die Kategorie *Interaction*.
Kommt hingegen ein generelles Bedürfnis nach Erklärungen auf, welches nicht während der Interaktion mit dem System entsteht, so zählt dieser Erklärungsbedarf zu der Kategorie *Business*.
Da vor allem die Unterkategorien *Training* und *Interaction* wichtig für diese Arbeit sind, werde ich auf diese genauer eingehen.
Die Kategorie *Training* unterteilt sich weiter in *Instruction*, *Features Offered* und *Effect-Of*. Diese beschreiben folgende Situationen:

Um **Instruction**-Erklärungsbedarf handelt es sich, wenn Benutzer Erklärungen fordern, um zu verstehen, wie man das System, eine Funktion oder eine Einstellung verwenden kann.
Die Art **Features Offered** von Erklärungsbedarf beschreibt die Situation, in der der Nutzende ein Verständnis für die Funktionen des Systems erlangen und somit herausfinden möchte, was das System zu bieten hat.
Wenn Anwender Informationen über das Resultat bestimmter Aktionen erhalten möchten, handelt es sich um den Erklärungsbedarf **Effect-Of**.

[1] Einordnung in ein bestimmtes System; https://www.duden.de/rechtschreibung/Taxonomie (04.07.24)

Dabei wissen diese bereits, wie eine Funktion auszuführen ist, aber die Information über das Ergebnis ist nicht bekannt.

Man kann ebenfalls die Kategorie *Interaction* in drei weitere Teile gliedern:
Bei der Kategorie **Algorithm** möchte ein Nutzer nachvollziehen können, warum das System eine gewisse Ausgabe erzeugt hat, verbunden mit den beeinflussenden Faktoren für dieses Ergebnis.
Es handelt sich um **Design Decision**-Erklärungsbedarf, wenn die Benutzer sich über eine bestimmte Ausführung des Softwaresystems wundern. Das bedeutet sie verstehen nicht, warum Dinge auf eine gewisse Art und Weise gemacht wurden.
Möchten Nutzer nun über Definitionen, Symbole oder Ähnliches Bescheid wissen, so kennzeichnet dies den Erklärungsbedarf **Signification**. Sie wollen dabei verstehen, welchen Sinn das System beabsichtigt hat.
Die drei eben erwähnten, für diese Arbeit relevanten Arten des Erklärungsbedarfs (*Instruction, Algorithm, Signification*) werden ebenfalls von Droste et al. [9] genannt. Dabei werden diese als *Interaction, System Behavior* und *Domain Knowledge* bezeichnet. Eine Übersicht dieser ist in Tabelle 2.1 zu sehen. Ein Erklärungsbedarf in der Kategorie *Interaction* bedeutet, dass der Nutzer Erklärungen zur Interaktion zwischen Benutzer und Software benötigt. Dazu gehört beispielsweise, dass der Nutzer wissen möchte, wie ein bestimmter Vorgang mit dem Softwaresystem funktioniert. Dies stellt eine alternative Beschreibung für den Begriff *Instruction* von Unterbusch et al. [21] dar. *System Behavior* erläutern sie als die Art von Erklärungsbedarf, bei der eine Beschreibung, wie die Software funktioniert und warum sie sich auf eine bestimmte Weise verhält, nötig ist. Unterbusch et al. [21] bezeichnen jenes als *Algorithm*. Der Begriff *Domain Knowledge* von Droste et al. [9], der dem *Signification*-Erklärungsbedarf von Unterbusch et al. [21] ähnelt, wird als fehlendes Wissen bezüglich der Domäne der Software beschrieben.

Art des Erklärungsbedarfs		Beschreibung
Unterbusch et al.[21]	**Droste et al.[9]**	
Instruction	Interaction	Nutzer benötigt Informationen, wie das System zu nutzen ist, unter anderem über Funktionen und Einstellungen.
Algorithm	System Behavior	Nutzer möchte verstehen, warum das System bestimmte Ausgaben generiert.
Signification	Domain Knowledge	Nutzer will über Definitionen und Elemente des Systems Bescheid wissen und deren Bedeutung verstehen.

Tabelle 2.1: Arten von Erklärungsbedarf

2.2 Facial Emotion Recognition

Facial Emotion Recognition bezeichnet die Analyse der Mimik einer Person, um deren Emotion zu identifizieren. Dies wird mit Künstlicher Intelligenz, und somit neuronalen Netzen, durchgeführt. Dazu wird zunächst ein Gesicht in einem Foto gesucht und anschließend die Emotionen anhand der Mimik herausgefiltert[1].Die grundlegenden Schritte dieses Vorgehens werden im Folgenden genauer betrachtet.

2.2.1 Schritte der Emotionserkennung

Meistens werden mehrere Stufen bei Facial Emotion Recognition benötigt, um die Emotionen korrekt zu identifizieren und einzuordnen. Die wichtigsten drei Prozeduren laut Rana et al. [17] heißen wie folgt:

1. Gesichtserkennung

Bei dem ersten Schritt werden zunächst die Gesichter herausgesucht, um danach den Bereich, welcher das Gesicht beinhaltet, zu isolieren.

2. Merkmalsextrahierung (Feature Extraction)

Bei dem zweiten Schritt werden die entscheidenden Charakteristiken herausgearbeitet. Für Gesichtsausdrücke werden folgende Charakteristiken herangezogen:

(a) **Geometrische Eigenschaften**
Zu geometrischen Eigenschaften gehören wichtige Gesichtspunkte, wie zum Beispiel die Position der Nase oder der Augen.

(b) **Merkmale der Textur**
Die Textur oder das Aussehen des Gesichts werden herausgefiltert.

(c) **Statistische Eigenschaften**
Dies beschreibt die Pixelintensitäten in gewissen Gesichtsbereichen.

(d) **Temporale Charakteristiken**
Hier werden zeitliche Veränderungen des Gesichtsausdrucks herausgearbeitet.

3. Emotionsklassifizierung

Im dritten und letzten Schritt werden dann die eben genannten Charakteristiken genutzt, um die erkannten Gesichtsausrücke einer oder mehreren Emotionen zuzuordnen. Um dies umzusetzen, werden Machine Learning- und Deep Learning-Ansätze genutzt.

Machine Learning Die Charakteristiken können im Folgenden genutzt werden, um Machine Learning(ML)-Algorithmen zu trainieren. Dazu gehören beispielsweise Support Vector Machines (SVM) oder Decision Trees (Entscheidungsbäume).

Convolutional Neural Networks (CNN) Eine weitere Möglichkeit, die herausgearbeiteten Charakteristiken zu verwenden, stellen CNNs dar, welche mit diesen Charakteristiken gefüllt werden können. Die CNNs lernen somit Muster und Skalen für die Emotionserkennung. Sie sind dafür bekannt, eigenständig komplizierte hierarchische Eigenschaften der Daten zu lernen. CNNs stellen außerdem Gesichtsdetektierung und Identitätsverifikation für Gesichtserkennungssysteme bereit.

Multi-Modale Ansätze Um die Genauigkeit der Emotionserkennung zu steigern, werden oft zusätzliche Bedingungen, wie zum Beispiel die Tonanalyse, herangezogen.

2.2.2 Bibliotheken und Module

Bei meiner Recherche bin ich vor allem auf zwei Bibliotheken gestoßen, die mit Convolutional Neural Networks (CNN) arbeiten. Dabei handelt es sich

um die FER-Bibliothek[2] sowie die DeepFace-Bibliothek[3], welches beides Python-Bibliotheken sind.

DeepFace-Bibliothek Mittels dieser Bibliothek ist es möglich, durch die Benutzung von Deep Learning-Techniken Gesichtsmerkmale in einem Foto zu analysieren und diese im Anschluss mit der Datenbank der für das Modell bekannten Gesichter zu vergleichen, um Personen zu erkennen.
Es werden mehrere Möglichkeiten geboten, darunter Gesichtsverifikation, Gesichtserkennung, Gesichtsgruppierung, aber auch die Analyse von Gesichtsattributen wie Alter, Geschlecht oder Emotion.
DeepFace beinhaltet bereits vortrainierte Modelle, die einen einfachen Gebrauch der Funktionen möglich machen[15].

FER-Bibliothek Die FER-Bibliothek ist entwickelt worden, um die Emotionserkennung zu vereinfachen. Es handelt sich hierbei um eine Bibliothek, welche OpenCV (2.2.2) sowie Keras-Bibliotheken[4] verwendet, um die Benutzung bereits trainierter Deep Learning-Modelle zu ermöglichen. Die Bibliothek ist des außerdem in der Lage, mittels bereitgestellter Funktionen sowohl Fotos, als auch Videos zu analysieren[15].

Die FER-Bibliothek nutzt selbst weitere Bibliotheken. Zudem habe ich zur weiteren Analyse der Ergebnisse durch die FER-Bibliothek noch zusätzliche Bibliotheken genutzt. Aus diesem Grund werde ich ein paar dieser Bibliotheken und deren Inhalte im Folgenden kurz erläutern.

OpenCV (cv2) OpenCV ist eine Bibliothek, die Funktionen wie die Erkennung von Gesichtern oder die Entdeckung von Gesichtspunkten bereitstellt.

Pandas Die Bibliothek *Pandas* wird genutzt, um die Verarbeitung von organisierten Daten zu vereinfachen. Sie bietet verschiedene Datenstrukturen und Funktionen, um numerische Daten zu handhaben[11].

TensorFlow TensorFlow ist ein Open-Source-Modell, welches im Bereich Machine Learning und Deep Learning zum Einsatz kommt. Es werden Datenflussdiagramme genutzt, um Neuronale Netze zu bilden. Es wird vor allem zur Klassifizierung, Erkennung und Vorhersage genutzt[19].

[2] https://pypi.org/project/fer/ (04.07.24)
[3] https://pypi.org/project/deepface/ (04.07.24)
[4] https://pypi.org/project/keras/ (04.07.24)

2.3 Auswertungsmetriken

Um die mit der Studie erhobenen Daten auswerten zu können, werden mittels True Positives, False Positives und False Negatives die Werte für *Precision, Recall* und *F1-Score* berechnet. Dies geschieht wie folgt[6]:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.2}$$

$$\text{F1-Score} = \frac{2PR}{P + R} \tag{2.3}$$

Precision beschreibt dabei den prozentualen Anteil des durch die FER-/manuellen Analyse korrekt erkannten Erklärungsbedarfs. Dies zeigt dabei an, mit welcher Wahrscheinlichkeit es sich tatsächlich um Erklärungsbedarf handelt, beziehungsweise fälschlicherweise Erklärungsbedarf erkennt. Recall hingegen ist der Anteil der gefundenen Erklärungsbedarfe, die insgesamt zu finden waren. Dieser macht deutlich, mit welcher Wahrscheinlichkeit ein tatsächlicher Erklärungsbedarf durch die Analyse gefunden wird. Um diese beiden Werte zu vereinen und gleichermaßen in eine Bewertung einzubeziehen, wird der F1-Score berechnet.

2.4 Verwandte Arbeiten

Deters et al. [7] haben eine Methode getestet, bei der physiologische Daten mit einer Armbanduhr aufgenommen und analysiert wurden, um das Bedürfnis nach Erklärungen der Nutzer während der Systemnutzung zu überprüfen. Dazu wurde eine Studie mit 9 Teilnehmern durchgeführt, bei der diese Aufgaben mit Microsoft Excel bearbeiteten. Von 5 Aufgaben mit jeweils 5 bis 7 Unteraufgaben waren 3 Teilaufgaben so konzipiert, dass sie Erklärungsbedarf auslösen sollten. Dabei wurden drei Arten von Erklärungsbedarf ausgelöst. Dazu gehört der Wunsch nach Erklärung, wie das System korrekt zu nutzen ist, warum das System ein unerwartetes Ergebnis liefert und was ein bestimmter Begriff bedeutet. Die ersten zwei dieser Arten von Erklärungsbedarf wurden ebenfalls in dieser Arbeit ausgelöst. Während der Studie wurden zwei verschiedene Daten gesammelt. Zum einen wurde eine sogenannte *Empatica Watch* getragen, um biometrische Daten, wie elektrodermale Aktivität, Blutvolumenpuls, Herzfrequenz und Temperatur zu messen. Außerdem wurde der Bildschirm während der Durchführung aufgenommen, um später die Zeiten des Erklärungsbedarfs manuell zu notieren und somit zu verifizieren. Des Weiteren wurde vor der Durchführung ein Fragebogen zu Alter, Geschlecht und Vorkenntnissen mit Excel durch die Teilnehmer ausgefüllt. Ergebnis der Studie war, dass lediglich die zwei ersten Arten von Erklärungsbedarf bei den Teilnehmern aufgetreten sind. Dabei ist bei 56% der Teilnehmer der Erklärungsbedarf zu der korrekten Benutzung der Software aufgetreten und bei 78% der Teilnehmer der zu dem unerwarteten Ergebnis. Dabei wurden 20% der Erklärungsbedarfe zur korrekten Benutzung von der elektrodermalen Aktivität sowie von dem Blutvolumenpuls erkannt. Von dem Erklärungsbedarf des unerwarteten Systemverhaltens wurden 57% von der elektrodermalen Aktivität und 85% von dem Blutvolumenpuls erfasst. Herzfrequenz und Temperatur hingegen gaben keinen Aufschluss über den Erklärungsbedarf. Die Präzision des Blutvolumenpulses hingegen lag bei 28% korrekter Vorhersage und die der elektrodermalen Aktivität bei 68%. Diese Arbeit untersucht ebenfalls die Erkennung von Erklärungsbedarf der Nutzer in Echtzeit. Allerdings betrachtet diese Arbeit dazu die Emotionen in den Gesichtern mit Facial Emotion Recognition und nicht eine biometrische Uhr.
Des Weiteren haben Deters et al. [7] dabei festgestellt, dass Erklärungsbedarf dann auftritt, wenn Frustration oder Verwirrung durch die Software auftritt.

Menzil et al. [15] haben sich damit beschäftigt, inwiefern Facial Emotion Recognition (Emotionserkennung) die mögliche Unzufriedenheit während Videokonferenzen erkennen lässt. Dies stellt ein ähnliches Ziel zu dieser Arbeit dar, die ebenfalls die Kundenzufriedenheit mittels FER analysiert, jedoch bezogen auf Erklärbarkeit. In der Arbeit von Menzil et al. [15] wurden

anschließend drei mögliche Modelle zur Emotionserkennung genutzt. Dazu gehörten DeepFace, das Vision Transformer Model und die FER Bibliothek, die ebenfalls in dieser Arbeit genutzt wird.
Um diese Modelle zu evaluieren, wurden mehrere Bilder aus einem Video herausgearbeitet sowie Beispielfotos genutzt. Es sollte herausgefunden werden, ob es damit möglich ist, genauestens positive beziehungsweise negative Emotionen in einem Video zu detektieren.
Dabei ergaben sich F1-Scores von 0.78 für das Vision Transformer Model, 0.72 für die FER-Bibliothek und 0.63 für DeepFace.

Eine ähnliche Feststellung für das Potential des Erkennens der Nutzerzufriedenheit mittels Facial Emotion Recognition haben Singh et al. [19] getroffen. Ihr Ziel war es, eine Desktop-Anwendung auf Kundenseite zu entwickeln, die Emotionsdaten der Nutzenden verwendet, um Videoempfehlungen in Echtzeit zu geben. Damit wird gewährleistet, dass Nutzende nur Videos empfohlen bekommen, die zu der aktuellen Stimmung passen. Auch dieses Ziel weist Parallelen zu dieser Arbeit auf, da beide nach einem personalisierten und optimalen Benutzererlebnis streben. Die Arbeit von Singh et al. [19] bezogen auf Videoempfehlungen und diese Arbeit bezogen auf Erklärungen zur Software.
Singh et al. [19] nutzten dazu DeepFace und außerdem die OpenSource Bibliothek OpenCV. Zudem wurde ein sogenannter *Request Executor* verwendet. Dieser sollte kurzzeitige, spontane Bildveränderungen normalisieren und herausfiltern. Dabei wurden die Häufigkeiten der Emotionen gezählt und schlussendlich diejenige ausgewählt, die am meisten vorkam. Eine neue Anfrage kommt dann nur bei veränderter Emotion zum vorherigen Teil auf. Die Studie ergab, dass 65% der aktuellen Emotionen während des Ansehens von Videos richtig identifiziert wurden. Dies zeigt Potential für die Nutzung von Facial Emotion Recognition bei ähnlichen Absichten. Es zeigt aber auch, dass der Zusammenhang zwischen erkannter Emotion und durchschnittlicher Emotion während des gesamten Videos wichtig ist.

Kapitel 3

Forschungsfragen

Die vorherigen Kapitel zeigen bereits auf, dass Erklärbarkeit ein wichtiges Thema ist, um die Software für die Nutzer transparent zu machen und somit die Akzeptanz der Benutzer zu fördern. Dabei ist es jedoch schwierig, die Stellen und die notwendige Ausführlichkeit der Erklärungen abzuschätzen, sodass die Nutzer nicht überladen werden.
Um Nutzern an den richtigen Stellen passende Erklärungen zu liefern, müssen Auslöser gesucht werden, die Erklärungsbedarf erkennen können. Aus diesem Grund ist es Ziel dieser Arbeit, herauszufinden, ob sich Facial Emotion Recognition für das Erkennen von Erklärungsbedarf eignet. Falls dies eine zuverlässige Erkennung zulassen würde, wäre es Softwareentwicklern möglich, den Erklärungsbedarf in Echtzeit an jeden Nutzer angepasst zu erkennen. So bekäme ein Nutzer nur dann Erklärungen, sofern er diese auch benötigt.
Um dies zu analysieren, ist es zunächst notwendig, eine passende Bibliothek für Facial Emotion Recognition zu finden. Mithilfe dieser Bibliothek soll anschließend überprüft werden, ob sich FER-Bibliotheken zur Erkennung von Erklärungsbedarf eignen, was Forschungsfrage RQ 2 ergibt. Da die einzelnen erkannten Emotionen der Bibliothek unterschiedlich gute Ergebnisse liefern könnten, muss jeweils die Eignung einer Emotion überprüft werden. Dies resultiert in Forschungsfrage RQ 1. Die Forschungsfragen sind im Folgenden aufgelistet:

> **RQ 1:** Welche Emotionen der Bibliothek eignen sich am besten, um Erklärungsbedarf zu identifizieren?

> **RQ 2:** Ist Gesichtserkennung im Allgemeinen in der Lage, das Aufkommen von Erklärungsbedarf zu erkennen?

Kapitel 4

Nutzerstudie

Dieses Kapitel beschreibt die Planung und die Durchführung der Studie. Dabei wird auf die Teilnehmer, den Ablauf sowie einzelne zu bearbeitende Aufgaben eingegangen.

4.1 Planung

Um die Studie durchführen zu können, musste zunächst ein Softwareprogramm ausgewählt werden, welches sich für die Nutzung durch die Teilnehmer eignet. Dabei ist es notwendig, eine Software auszuwählen, mit der die Nutzer ein wenig vertraut sind. Außerdem soll es möglich sein, bei den Nutzern Erklärungsbedarf auszulösen.
Die Microsoft Office Programme sind komplex, sodass Erklärungsbedarf ausgelöst werden kann. Allerdings sind viele Personen mit Microsoft PowerPoint[1] bereits zu vertraut. Da es aber möglich sein sollte, die Teilnehmer zu verwirren, um Erklärungsbedarf auszulösen, habe ich mich im Rahmen der Studie für das OpenOffice Programm Libre Office Impress[2] entschieden.
In einem nächsten Schritt sollten Aufgaben ausgewählt werden, die Erklärungsbedarf auslösen können, jedoch auch zwischendurch Zeit zur Erholung bieten. Dabei wurden die Aufgaben so konzipiert, dass sich die Durchführung auf etwa 10-15 Minuten beläuft. Dazu habe ich mir 5 verschiedene Aufgaben überlegt, von denen 3 dazu gedacht waren, Erklärungsbedarf auszulösen. Diese wurden so sortiert, dass zu Beginn eine funktionierende Aufgabe gegeben wurde, um den Einstieg zu erleichtern. Des Weiteren wurde versucht, nach jeder Aufgabe, die Erklärungsbedarf auslösen soll, eine funktionierende oder zumindest eine Aufgabe mit leichten Unterpunkten am Anfang einzubinden. Dies soll verhindern, dass die Teilnehmer dauerhafte Frustration verspüren. Ziel war es, den Teilnehmern

[1] https://www.microsoft.com/de-de/microsoft-365/powerpoint (11.07.24)
[2] https://de.libreoffice.org/discover/impress/ (11.07.24)

nach jeder schwierigen Aufgabe ein Erfolgserlebnis zu verschaffen. So sollte stetige Unzufriedenheit und Enttäuschung umgangen werden.
In Kapitel 2.1.3 wurde bereits etwas genauer auf die verschiedenen Arten von Erklärungsbedarf, die Unterbusch et al. [21] herausgearbeitet haben, eingegangen. In einer anderen Studie [7] wurden ebenfalls ähnliche Aufgaben mit beabsichtigter Auslösung von Erklärungsbedarf durchgeführt. Dabei hat der *Signification*-Erklärungsbedarf nie ausgelöst. Aufgrund dessen habe ich diese Art des Erklärungsbedarfs nicht in die Aufgaben dieser Studie eingebunden.
Bei Aufgabe 2 sowie Aufgabe 5 der Studie war es vorgesehen, *Algorithm*-Erklärungsbedarf auszulösen. Aufgabe 4 hingegen sollte *Instruction*-Erklärungsbedarf auslösen.
Da vor allem die Aufgaben von Relevanz sind, die Erklärungsbedarf auslösen sollten, werde ich diese im Folgenden kurz beschreiben. Alle Aufgaben sind in Anhang A zu finden.

Aufgabe 2 (*Algorithm*) Ziel von Aufgabe 2 war es, das Haupttextfeld einer Präsentationsfolie rot auszufüllen. Dabei wurde das Vorgehen in den Unteraufgaben beschrieben. Allerdings war es nach diesem beschriebenen Vorgehen nicht möglich, das Textfeld rot auszufüllen. Die Teilnehmer sollten sich hierbei darüber wundern, dass das System nicht das erwartete Ergebnis liefert.

Aufgabe 4 (*Instruction*) Bei Aufgabe 4 sollte zu einer beliebigen Folie mit Beispielüberschrift eine Notiz verfasst werden, die man bei Präsentationen als Karteikarte nutzen kann. Dabei wurde jedoch nicht genauer darauf eingegangen, wo man diese Notizübersicht finden kann. So sollte bei den Teilnehmern das Bedürfnis nach einer Erklärung aufkommen, die beschreibt, wo man die Notizübersicht findet und wie man diese verwendet.

Aufgabe 5 (*Algorithm*) Abschließend wurde mit Aufgabe 5 ein prozentules Flächendiagramm mit zwei Flächen (Gruppen) generiert. Auch hier wurden wieder die einzelnen Schritte des Vorgehens genau erklärt. Hat man diese Schritte so ausgeführt, sah das Endergebnis allerdings nicht so aus wie auf dem Aufgabenzettel vorgegeben. Diese unerwartete Ausgabe sollte bei den Teilnehmern Verwunderung und somit daraus resultierenden Erklärungsbedarf hervorrufen.

4.2 Durchführung

4.2.1 Teilnehmer

Zur Durchführung der Studie wurden 15 Teilnehmer herangezogen, wovon 8 männlich und 7 weiblich waren. Dabei lag das durchschnittliche Alter der Teilnehmer bei 22,07 Jahren (min: 19 Jahre, max: 29 Jahre, Standardabweichung ∼2,37) und es handelt sich somit um *Digital Natives*[20]. 60% der Teilnehmer waren Studenten, ∼7% Schüler, ∼7% Auszubildende und die restlichen ∼26% waren in einem Beruf tätig, der nicht im Bereich der Informatik liegt.
Zur allgemeinen Nutzung von Microsoft PowerPoint gaben ∼27% an, PowerPoint ca. 1 Mal pro Woche zu verwenden, ∼13% gebrauchen es ca. 1 Mal im Monat, ∼47% ca. 1 Mal pro Jahr und die restlichen ∼13% weniger als 1 Mal im Jahr. LibreOffice Impress hingegen wurde zuvor noch von keinem der Teilnehmer verwendet.

4.2.2 Ablauf

Eine Übersicht des gesamten Ablaufs der Studie ist bei Abbildung 4.1 zu finden.
Zum Start der Studiendurchführung wurde den Teilnehmern zunächst kurz der Ablauf und das grundsätzliche Thema *Erklärbarkeit* genannt. Dabei wurde das richtige Ziel der Studie nicht erwähnt, um das Verhalten während der Studiendurchführung nicht zu beeinflussen. Die Teilnehmer könnten mit diesem Wissen beispielsweise darauf achten, an den Stellen des Erklärungsbedarfs diesen bewusst durch Emotionen deutlich zu machen oder im Gegenteil, die Emotionen zu unterdrücken.
Anschließend wurde eine Einverständniserklärung unterzeichnet. Danach wurde das Setup für die Aufnahme der Studie aufgebaut. Dieses beinhaltete eine Lampe für ausreichende und gleichwertige Lichtverhältnisse, ein neutraler, möglichst weißer Hintergrund sowie das Starten der Bildschirmaufnahme und der Aufnahme mittels einer Webcam. Um die Teilnehmer nicht zu beeinflussen oder zu stören, wurden diese während der Bearbeitung der Aufgaben in dem Raum alleine gelassen. Nach Durchführung der Aufgaben wurden die Aufnahmen beendet und den Teilnehmern ein Umfragebogen ausgehändigt. In diesem wurden die Stellen abgefragt, an denen sie Erklärungsbedarf hatten und persönliche Fragen gestellt. Die Teilnehmer füllten diesen Fragebogen in meinem Beisein selbst aus. Dies war notwendig, um bei der Analyse einen Vergleichswert für tatsächlichen Erklärungsbedarf zu haben. Anschließend wurde den Teilnehmern das Ziel der Studie ausführlich erläutert. Zum Schluss wurden die Teilnehmer nochmal gebeten, eine abschließende Einverständniserklärung zu unterzeichnen. Dabei wurden diese darauf hingewiesen, dass es ebenfalls möglich ist, diese nicht zu unterzeichnen und die Zustimmung ohne Angabe eines Grundes zurückzuziehen. Wäre dies

der Fall, würden die Aufnahmen und alle zugehörigen Materialien vernichtet werden.

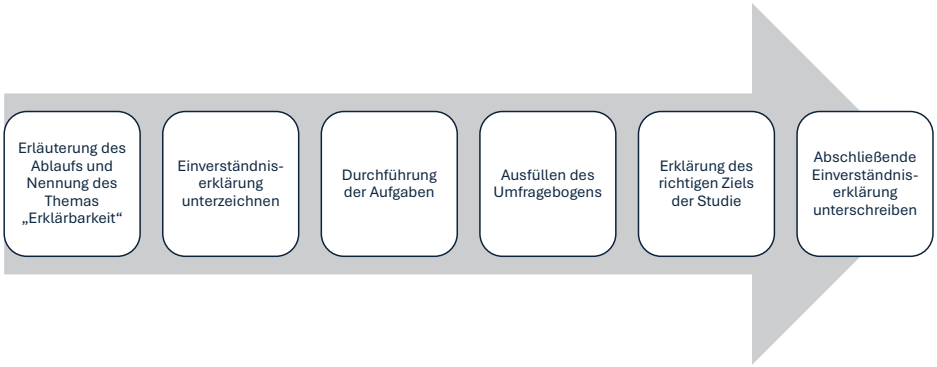


Abbildung 4.1: Ablauf der Studie

4.3 Datenerhebung

Um eine anschließende Analyse durchführen zu können, wurden mehrere Arten von Daten erhoben.
Dazu gehörte die Aufzeichnung des Bildschirms sowie die Aufzeichnung der Teilnehmer mithilfe einer Webcam. Aus der Bildschirmaufnahme sollte herausgearbeitet werden, wann die Teilnehmer die einzelnen Aufgabenteile behandelt haben. Durch die Gesichtsaufnahmen sollten manuell sowie mittels FER-Bibliothek negative Emotionen beobachtet und so mögliche Erklärungsbedarfe detektiert werden.
Des Weiteren wurden Informationen durch die Umfrage gesammelt. Diese sollten genutzt werden, um die Teilaufgaben mit tatsächlichem Erklärungsbedarf zu ermitteln und persönliche Daten, wie Alter und Beruf beziehungsweise Studium der Teilnehmer, zu erhalten.

Kapitel 5

Durchführung der Analyse

Nachdem die gesammelten Daten (Webcam-Aufzeichnung, Bildschirmaufnahme, Umfrage) erhoben wurden, mussten diese im weiteren Verlauf sowohl manuell, als auch mit einer FER-Bibliothek analysiert werden.

5.1 Auswahl der Bibliothek

Mencil et al. [15] evaluierten die zwei Bibliotheken, die bereits in Kapitel 2.2.2 kurz vorgestellt wurden. Dabei ergab sich, dass die FER-Bibliothek mit dem CNN-Modell die DeepFace-Bibliothek im Bezug auf Genauigkeit übertrifft. Überdies kommt hinzu, dass die Genauigkeit der Analyse mittels FER-Bibliothek gesteigert werden kann, indem das MTCNN-Netzwerk statt des OpenCV-Klassifizierers genutzt wird[1]. Des Weiteren stellt die FER-Bibliothek die Analyse von Videos bereit, während die DeepFace-Bibliothek nur Bilder analysieren kann. Aus diesen Gründen wurde sich in dieser Arbeit für die FER-Bibliothek entschieden.

5.2 Analyse mittels FER-Bibliothek

Eine Übersicht der gesamten Verarbeitung durch die FER-Bibliothek ist in Abbildung 5.1 zu finden.
Zur Analyse durch die FER-Bibliothek wurde zunächst der Python-Code für die Analyse mittels FER-Bibliothek geschrieben und mehrfach getestet. Dafür wurden die Bibliotheken OpenCV (2.2.2), CSV, FER (2.2.2) sowie Pandas (2.2.2) genutzt. Der geschriebene Code nimmt das Webcam-Video und analysiert jedes Bild des Videos einzeln. Dabei wird eine Sekunde des Videos in 30 Bilder unterteilt. Es war nötig, die Analyse selbst Bild

[1] https://pypi.org/project/fer/ (04.07.24)

für Bild zu programmieren, da die FER-Bibliothek Bilder ohne erkanntes Gesicht nicht in die resultierende CSV-Datei aufnimmt. Dies war bei manchen Bildern der Fall. Es ist jedoch notwendig, dass jedes Bild auftaucht, unabhängig davon, ob mit erkannten Emotionen oder nicht, da im Anschluss bei der Analyse die Bilder den Sekunden zugeordnet werden müssen.
Nachdem die FER-Bibliothek jedes Bild einzeln analysiert hat, wurden diese prozentualen Ergebnisse der Emotionen für jedes Bild zeilenweise in einer CSV-Datei gesichert.
Zur weiteren Verarbeitung wurde dann mittels Code die dominante Emotion, also die Emotion mit der höchsten Wahrscheinlichkeit, pro Zeile und somit pro Bild ermittelt und in eine neue Spalte eingefügt. Dies ermöglichte späteres Einlesen und Auswerten dieser Emotion.
Zur Vereinfachung wurden anschließend jeweils 30 Bilder zusammen in eine neue CSV-Datei gruppiert, um eine Sekunde des Videos widerzuspiegeln. Alle diese CSV-Dateien, die jeweils eine Sekunde des Videos darstellten, wurden daraufhin genutzt, um jeweils die am meisten auftretende Emotion pro Sekunde herauszufiltern. So ergab sich schlussendlich eine endgültige CSV-Datei für jeden Teilnehmer mit den Spalten *Second* und *Emotion*, die alle Sekunden des gesamten Videos mit zugehörigen Emotionen zeigt.

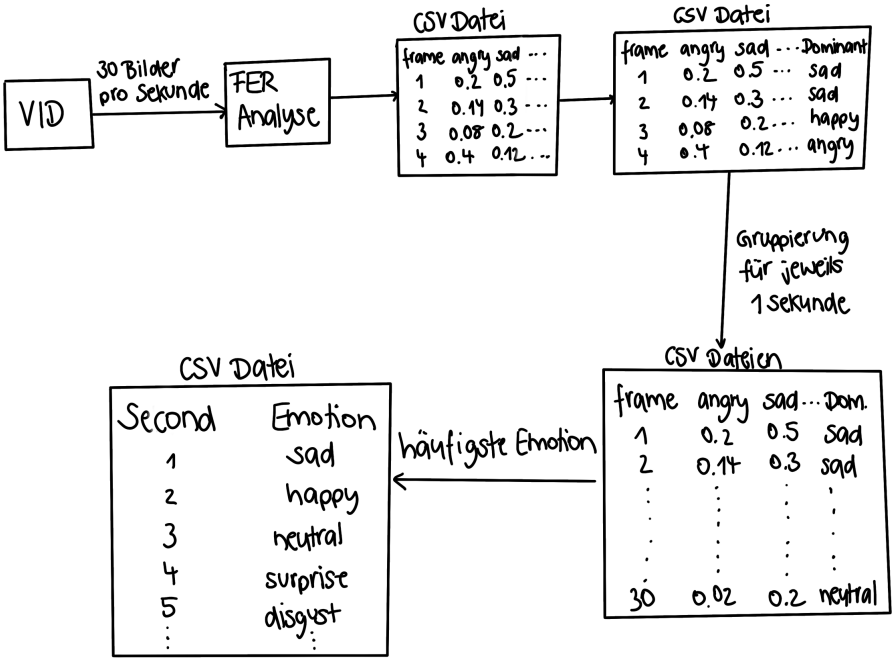


Abbildung 5.1: Graphische Darstellung der Verarbeitungsschritte - Analyse mit FER-Bibliothek

5.3 Manuelle Analyse

Um Forschungsfrage RQ 2 (3) zu beantworten, ist es wichtig zu differenzieren, ob nur die ausgewählte Bibliothek mit Gesichtserkennung geeignet beziehungsweise ungeeignet ist, oder ob das Ergebnis auf die generelle Gesichtserkennung übertragen werden kann. Aus diesem Grund müssen die erhobenen Daten ebenfalls manuell analysiert werden.
Bei der manuellen Analyse wurden die Webcam-Videos ohne zugehörige Bildschirmaufnahme einzeln genau angeschaut. Dabei wurde jedes Mal, wenn eine negative Emotion wahrnehmbar war, die Sekunde des Auftretens notiert. Da eine negative Emotion über mehrere Sekunden anhalten kann, wurden lediglich die Sekunden notiert, bei denen die Emotion das erste Mal auftrat. Also nur, wenn der Gesichtsausdruck zuvor neutral oder positiv war.
Diese Beobachtungen wurden ebenfalls in einer CSV-Datei pro Teilnehmer mit den Spalten *Second* und *Emotion* gespeichert.

5.4 Analyse der Umfrage

Im weiteren Verlauf der Analyse mussten die Ergebnisse durch die FER-Bibliothek sowie die der manuellen Analyse mit dem tatsächlichen Erklärungsbedarf verglichen werden. Dazu wurden die Videos der Webcam- und Bildschirmaufnahmen kombiniert angesehen und die Sekunden-Bereiche notiert, bei denen die Teilnehmer jeweils angegeben haben, dass sie den Wunsch nach zusätzlichen Erklärungen verspürten. Diese Zeitstempel wurden ebenfalls in einer CSV-Datei mit den Spalten *Start* und *End* für jeden Teilnehmer fixiert.

5.5 Vergleich mit tatsächlichem Erklärungsbedarf

Um ein abschließendes Endergebnis zu erhalten, wurden im letzten Schritt die CSV-Dateien der FER-Analyse sowie die der manuellen Analyse mit den Intervallen des tatsächlichen Erklärungsbedarfs verglichen. Aufgrund des hohen Zeitaufwands wurde dies ebenfalls mittels Code automatisiert durchgeführt. Dabei wurde auch die Pandas-Bibliothek (2.2.2) verwendet.
Hierbei wurden mögliche negative Emotionen betrachtet, die bei Erklärungsbedarf auftreten könnten. Diese waren *sad, angry, disgust, fear* und *surprise*, da Nutzer wütend, traurig oder gar ängstlich werden können, wenn sie bestimmte Teile der Software nicht verstehen. Außerdem könnte man einen Gesichtsausdruck des Ekels oder der Überraschung haben, wenn die Software unerwartet handelt.
Zur Auswertung und abschließenden Bewertung habe ich True Positives,

False Positives sowie False Negatives ermittelt. Diese ergeben sich folgendermaßen und sind in der Tabelle 5.1 zusammengetragen:

True Positive: Die Analyse mittels FER, beziehungsweise manuell, ergab eine negative Emotion und der Teilnehmer hat zu diesem Zeitpunkt tatsächlich Erklärungsbedarf geäußert.

False Positive: Die FER-/manuelle Analyse erkennt eine negative Emotion, obwohl der Teilnehmer zu diesem Zeitpunkt keinen Erklärungsbedarf hatte.

False Negative: Es handelt sich um tatsächlichen Erklärungsbedarf des Teilnehmers, die Analyse mittels FER, beziehungsweise manuell, erkennt jedoch keine negative Emotion.

		tatsächlicher Erklärungsbedarf	
		Positive	Negative
Vorhersage	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Tabelle 5.1: Einteilung in TP, FP, TN, FN

Um True Negatives hätte es sich gehandelt, wenn die FER-Analyse keine negative Emotion erkennt und der Teilnehmer tatsächlich keinen Erklärungsbedarf geäußert hat. Da diese zur Ermittlung von Precision, Recall und F1-Score nicht relevant und damit zu vernachlässigen sind, wurden sie bei der Analyse nicht berücksichtigt.

Um die Forschungsfrage RQ 1 nach der am besten geeignetsten Emotion zur Erkennung von Erklärungsbedarf zu beantworten (3), ist es hilfreich, die True Positives und weitere Werte für jede negative Emotion einzeln zu erarbeiten. Somit extrahiert der Code zunächst nur die Zeilen mit der jeweiligen negativen Emotion. Dabei werden nur Einträge herangezogen, bei denen die Emotion mindestens drei aufeinander folgende Sekunden anhält. Dies macht es möglich, fehlerhafte kurzzeitige Emotionsänderungen nicht zu berücksichtigen. Etwas in der Art könnte der Fall sein, wenn eine Person durch äußere Einflüsse kurzzeitig abgelenkt wird und diese spontanen, aber nur sehr kurzen Reaktionen, nichts mit der Software oder Erklärungsbedarf bezüglich dieser zu tun haben. Es wurden 3 Sekunden gewählt, da dies als Zeit einer begründeten Emotion geschätzt wurde. In einer resultierenden CSV-Datei werden schließlich die Sekunden und Emotionen notiert, die der geforderten Emotion entsprechen und mindestens drei Sekunden anhalten.
Danach werden diese CSV-Dateien (jeweils eine für angry, disgust, fear, sad,

surprise sowie eine mit allen negativen Emotionen zusammen) genutzt, um sie mit den Zeitstempeln des tatsächlichen Erklärungsbedarfs zu vergleichen. Dazu wird für jede Zeile der FER-/manuellen Analyse geschaut, ob diese Startzeit der Emotion in einem beliebigen Intervall der Zeitstempel-CSV-Datei des tatsächlichen Erklärungsbedarfs liegt. Ist dies der Fall, so wird dieser Eintrag als True Positive gewertet. Des Weiteren wird das Intervall, in dem die Emotion lag, als erkannt notiert, um später die False Negatives zu ermitteln. Wurde kein Intervall gefunden, indem die Sekunde des Zeileneintrags liegt, so wird dieser als False Positive vermerkt. Abschließend wird für alle Intervalle des Erklärungsbedarfs geschaut, ob mindestens ein Eintrag gefunden wurde, bei dem die Sekunde innerhalb des Intervalls liegt, also ob der Erklärungsbedarf mindestens einmal erkannt wurde. Wurde ein Intervall nicht mindestens einmal erkannt, so handelt es sich um ein False Negative.

Die Ergebnisse der Vergleiche sind in einer gemeinsamen Tabelle für alle Teilnehmer und für jede Emotion einzeln sowie gesamt aufgetragen worden. Diese Tabelle ist in Anhang B zu finden.

Kapitel 6

Auswertung

In diesem Kapitel werden die durch die Studie erhobenen Daten (4) und dessen weitere Verarbeitung (5) ausgewertet und in Kontext gesetzt.
Nach dem Herausarbeiten der True Positives, False Positives und False Negatives (5.5) wurden mittels dieser Precision, Recall sowie F1-Score ermittelt.

6.1 Auswertung einzelner Emotionen

Um Forschungsfrage RQ 1 (3) beantworten zu können, werden im Folgenden zunächst die Ergebnisse der Emotionen einzeln betrachtet.

6.1.1 Emotion *disgust*

Da die Emotion *disgust* lediglich einmal bei Person F erkannt wurde und es sich dabei um einen False Positive Wert handelt, ergibt sich für Precision, Recall und F1-Score jeweils der Wert 0.

6.1.2 Emotion *surprise*

Die Emotion *surprise* wurde bei keinem der Teilnehmer erkannt.

6.1.3 Emotion *sad*

Vergleicht man die Werte von Tabelle 6.1 nun mit den Gesamtwerten (siehe Anhang B), so fällt auf, dass allein durch die Emotion *sad* der Großteil des erkannten Erklärungsbedarfs abgedeckt wird. Auffällig ist auch die hohe Zahl der False Positives, die zu einer Precision von ∼0,25 führen. Die False Negatives hingegen sind sehr gering. Dies bedeutet, dass die meisten Erklärungsbedarfe auch als diese erkannt wurden. Deswegen liegt der Recall bei ∼0,93. Insgesamt führt dies zu einem F1-Score von ∼0,39.

Teilnehmer	**True Positives**	**False Positives**	**False Negatives**
A	0	2	5
B	3	19	1
C	16	34	0
D	4	30	0
E	7	19	0
F	13	49	1
G	13	45	0
H	4	16	0
I	11	51	0
J	23	31	1
K	7	44	1
L	1	4	3
M	20	36	0
N	23	43	0
O	8	39	0
Gesamt:	153	462	12

Tabelle 6.1: Ergebnisse der Emotion *sad*

6.1.4 Emotion *angry*

Betrachtet man die Werte aus Tabelle 6.2, so fällt auf, dass das Verhältnis von True Positives zu False Positives in etwa ausgeglichen ist, wodurch eine Precision von ∼0,44 erzielt wird. Aufgrund der verhältnismäßig hohen Zahl der False Negatives wird ein Recall von ∼0,35 und ein F1-Score von ∼0,39 erreicht.

Teilnehmer	True Positives	False Positives	False Negatives
A	0	0	5
B	0	0	4
C	6	2	2
D	0	0	3
E	0	1	2
F	0	0	5
G	0	0	3
H	5	5	0
I	2	1	1
J	0	0	4
K	0	1	3
L	8	19	2
M	2	0	3
N	0	0	3
O	0	0	3
Gesamt:	23	29	43

Tabelle 6.2: Ergebnisse der Emotion *angry*

6.1.5 Emotion *fear*

Auffallend in Tabelle 6.3 ist, dass im Verhältnis deutlich mehr False Positives erkannt werden als True Positives. So ergibt sich für *fear* eine Precision von ∼0,18. Auch der Wert der False Negatives ist vergleichsweise hoch und resultiert in einem Recall von ∼0,24 und schließlich in einem F1-Score von ∼0,2.

Teilnehmer	True Positives	False Positives	False Negatives
A	0	0	5
B	1	0	3
C	0	0	4
D	0	0	3
E	2	29	0
F	0	0	5
G	1	4	2
H	1	2	0
I	4	15	1
J	0	2	4
K	0	0	3
L	3	7	2
M	1	0	3
N	0	0	3
O	0	1	3
Gesamt:	13	60	41

Tabelle 6.3: Ergebnisse der Emotion *fear*

6.2 Vergleich FER-Bibliothek und manuelle Analyse

Durch die Auswertung der einzelnen Emotionen lässt sich entnehmen, dass die Emotionen *sad* und *angry* besonders relevant sind, um die Erkennung des Erklärungsbedarfs zu ermöglichen. Mit diesen zwei Emotionen allein ist es bereits möglich, alle erkannten Erklärungsbedarfe zu erhalten. Jedoch wird dieser teilweise auch durch die Emotion *fear* erkannt.
Berücksichtigt man nun lediglich *sad*, *angry* und *fear*, so ergibt sich ein F1-Score von ∼0,4. Aufgrund des niedrigen F1-Scores von *fear* ergibt sich bei der weiter eingeschränkten Betrachtung von *sad* und *angry* allerdings ein gering verbesserter F1-Score von ∼0,41. Aus diesem Grund betrachtet die weitere Analyse ausschließlich diese zwei Emotionen.

Im nächsten Schritt ist es notwendig, die Ergebnisse der FER-Bibliothek mit denen der manuellen Analyse zu vergleichen, sodass Forschungsfrage RQ 2 (3) beantwortet werden kann. Eine niedrige Precision oder Recall bei bestimmten Teilnehmern könnte aussagen, dass die Emotionen und somit der Erklärungsbedarf lediglich bei diesem Teilnehmer nicht korrekt erkannt werden.

6.2.1 Ergebnisse der FER-Bibliothek

Im Folgenden werden Precision, Recall sowie F1-Score jeder einzelnen Person betrachtet. Daraus ergibt sich folgende Darstellung von Abbildung 6.1:

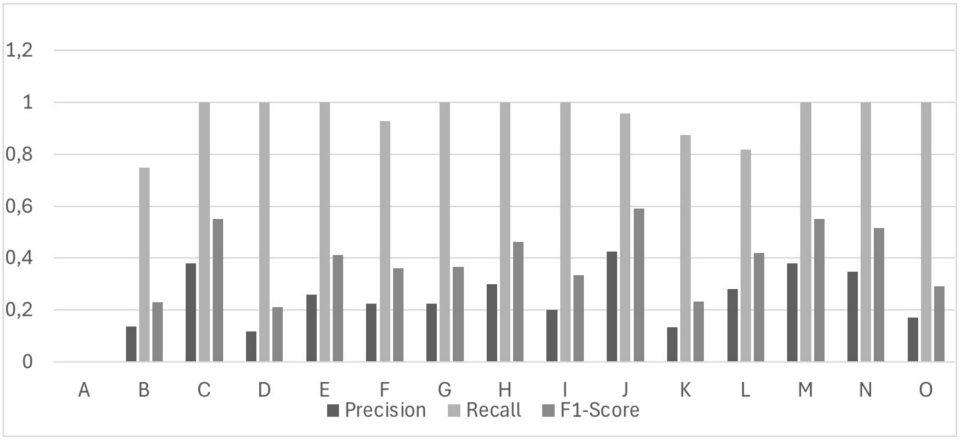


Abbildung 6.1: Genauigkeit der FER-Bibliothek zur Erkennung von Erklärungsbedarf der einzelnen Teilnehmer

Dem Diagramm ist zu entnehmen, dass der Recall bei jedem Teilnehmer bei mindestens 0,75 liegt. Die Precision hingegen bewegt sich zwischen ∼0,11 und ∼0,43, was deutlich geringer ist. Aufgrund dieser zwei Werte liegt der F1-Score durchschnittlich bei ∼0,35. Nimmt man True Positives, False Positives und False Negatives aller Personen zusammen, erhält man einen gesamten F1-Score von ∼0,41.

6.2.2 Ergebnisse der manuellen Analyse

Untersucht man zusätzlich die Werte der manuellen Analyse, resultiert dies in Abbildung 6.2:

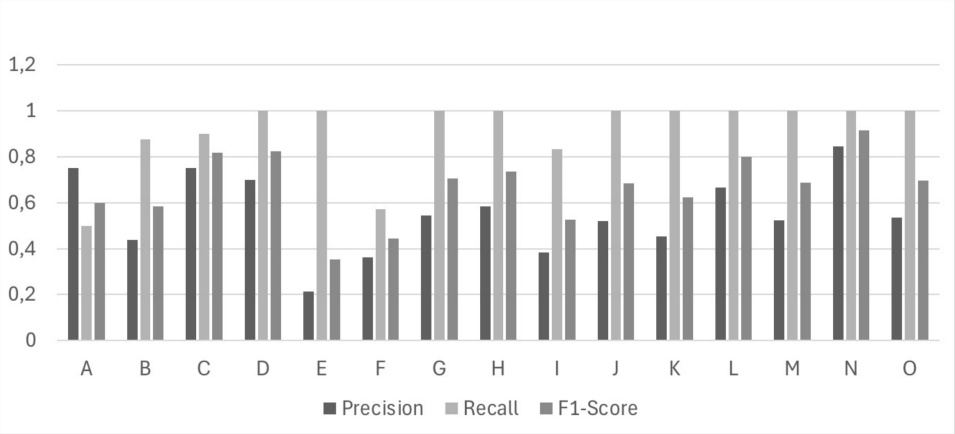


Abbildung 6.2: Genauigkeit der manuellen Analyse zur Erkennung von Erklärungsbedarf der einzelnen Teilnehmer

Auffallend ist, dass sich der Recall in dem Bereich von 0,5 bis 1 und die Precision zwischen ∼0,21 und ∼0,85 befindet. Der F1-Score liegt in einem Bereich von ∼0,35 bis ∼0,92. Dies ergibt einen durchschnittlichen F1-Score von ∼0,67 und einem gesamten F1-Score von ∼0,68.

6.3 Auswertung der einzelnen Aufgabenstellungen

Aufgabe 1 war so konzipiert, dass kein Erklärungsbedarf bei den Teilnehmern aufkommen sollte. Bei 8 der 15 Teilnehmer trat dieser jedoch auf. Die True Positives, False Positives sowie die Precision für Aufgabe 1 sind wie in Tabelle 6.4 verteilt:

Teilnehmer	**TP**	**FP**	**Precision**
A	0	0	NaN
F	0	8	0
G	4	4	0,5
J	6	6	0,5
L	1	0	1
M	8	4	0,66
N	4	12	0,25
O	1	10	0,09
Gesamt	24	44	0,35

Tabelle 6.4: Auswertung Aufgabe 1

Der Wert der False Negatives von Aufgabe 1 liegt bei 2 und ergibt somit einen Recall von ∼0,92.

Bei Aufgabe 2 handelte es sich um eine Aufgabe, die Erklärungsbedarf auslösen sollte. Bei 12 Teilnehmern trat dieser auch auf. Daraus ergeben sich folgende Werte:

Teilnehmer	**TP**	**FP**	**Precision**
A	0	0	NaN
B	1	4	0,2
C	8	4	0,66
D	2	2	0,5
E	4	1	0,8
F	2	1	0,66
G	1	8	0,11
J	0	4	0
K	3	6	0,33
L	0	0	NaN
M	1	4	0,2
N	4	1	0,8
Gesamt	26	35	0,43

Tabelle 6.5: Auswertung Aufgabe 2

Der Tabelle 6.5 ist zu entnehmen, dass der Recall für Aufgabe 2 ∼0,90 beträgt, da es sich um 3 False Negatives handelt.

Obwohl es bei Aufgabe 3 nicht beabsichtigt war, Erklärungsbedarf auszulösen, hatten 4 Teilnehmer der Studie den Wunsch nach Erklärungen. Die Ergebnisse für diese Teilnehmer sind in Tabelle 6.6 zu sehen:

Teilnehmer	**TP**	**FP**	**Precision**
A	0	0	NaN
I	1	13	0,07
L	8	2	0,8
O	5	8	0,38
Gesamt	14	23	0,38

Tabelle 6.6: Auswertung Aufgabe 3

Durch 1 False Negative ergibt sich hierbei ein Recall von ∼0,93.

Mittels Aufgabe 4 sollte Erklärungsbedarf verursacht werden. Dieser wurde auch bei allen 15 Teilnehmern ausgelöst und die Precision ist Tabelle 6.7 zu entnehmen:

Teilnehmer	**TP**	**FP**	**Precision**
A	0	0	NaN
B	1	2	0,33
C	11	0	1
D	1	6	0,14
E	3	0	1
F	5	4	0,55
G	8	3	0,72
H	9	0	1
I	12	4	0,75
J	11	0	1
K	4	2	0,66
L	0	7	0
M	9	5	0,64
N	15	0	1
O	2	5	0,28
Gesamt	91	38	0,70

Tabelle 6.7: Auswertung Aufgabe 4

Der Recall liegt bei ∼0,98, aufgrund von 2 False Negatives.

Aufgabe 5 sollte Erklärungsbedarf auslösen. Aufgetreten ist dieser allerdings nur bei 7 der 15 Teilnehmer. Die Verteilung der True Positives, False Positives und Precision wird in Tabelle 6.8 dargestellt:

Teilnehmer	**TP**	**FP**	**Precision**
A	0	1	0
B	1	7	0,125
C	3	11	0,21
D	1	6	0,14
F	6	21	0,22
J	6	9	0,4
M	4	11	0,26
Gesamt	21	66	0,24

Tabelle 6.8: Auswertung Aufgabe 5

Dies ergibt bei 1 False Negative einen Recall von ∼0,95.

Schlussendlich ist auffallend, dass bei Aufgabe 4 der *Instruction*-Erklärungsbedarf am besten korrekt erkannt wurde. Des Weiteren löst besonders Aufgabe 5 oft aus, obwohl es sich nicht um Erklärungsbedarf handelt (False Positives). Bei Aufgabe 5 handelte es sich um eine Aufgabe, die den Erklärungsbedarf *Algorithm* auslösen sollte. Eine hohe Anzahl an False Positives könnte zeigen, dass die Aufgabe viele Emotionen auslöst, die jedoch keinen Erklärungsbedarf darstellen.

Kapitel 7

Diskussion

In diesem Kapitel werde ich auf die Ergebnisse der Studie genauer eingehen und diese beurteilen. Außerdem werde ich im Folgenden die Forschungsfragen beantworten und mögliche Einschränkungen der Studie erläutern.

7.1 Beantwortung von RQ 1

RQ 1 Welche Emotionen der Bibliothek eignen sich am besten, um Erklärungsbedarf zu identifizieren?

In Kapitel 6.1 konnte man bereits erkennen, dass die Emotionen *surprise* und *disgust* fast gar nicht vorkommen. *Surprise* tritt gar nicht auf, *disgust* hingegen nur ein Mal. Dabei handelt es sich um eine falsche Vorhersage. Das weist darauf hin, dass diese beiden Emotionen nicht relevant für die Erkennung von Erklärungsbedarf sind.
Betrachtet man darauf folgend die Ergebnisse der Emotionen *angry, sad* und *fear*, erkennt man, dass der F1-Score für *fear* sehr gering ist. In Kapitel 6.2 bin ich bereits kurz darauf eingegangen, dass sich der insgesamte F1-Score verbessern lässt, wenn man nur *angry* und *sad* in die Wertung mit einbezieht und *fear* ausschließt. Dabei handelt es sich allerdings nur um eine geringfügige Verbesserung.
Aufgrund der eben genannten Punkte stellen die Emotionen *sad* und *angry* die relevantesten Emotionen zur Erkennung von Erklärungsbedarf dar.

7.2 Beantwortung von RQ 2

RQ 2 Ist Gesichtserkennung im Allgemeinen in der Lage, das Aufkommen von Erklärungsbedarf zu erkennen?

Aus Kapitel 6.2.1 ist zu entnehmen, dass die Werte des *Recalls* bereits sehr hoch und immer mindestens 75% sind. Das bedeutet, dass mindestens 75% des tatsächlichen Erklärungsbedarfs auch durch die Emotionen *sad* und *angry* erkannt worden ist. Allerdings bewegt sich die *Precision* bei allen Teilnehmern zwischen ungefähr 11% und 43%. Somit handelt es sich bei nur 11-43% des erkannten Erklärungsbedarfs auch tatsächlich um diesen. Dies stellt keinen guten Wert zur Ermittlung von Erklärungsbedarf dar, weil über die Hälfte des erkannten angeblichen Erklärungsbedarfs falsch erkannt werden würde und der Nutzer dadurch zu viele nicht benötigte Erklärungen erhalten würde. Auch die daraus resultierende gesamte Genauigkeit von rund 41% (F1-Score) zeigt kein gutes Ergebnis.
Allerdings lässt sich den Ergebnissen entnehmen, dass der *Algorithm*-Erklärungsbedarf mit einer Precision von 70% deutlich besser erkannt wurde. Dies lässt vermuten, dass diese Art von Erklärungsbedarf mithilfe von Gesichtserkennung leichter zu identifizieren ist. Es ist auch zu erwähnen, dass der Erklärungsbedarf bei dieser Art fast die gesamte Bearbeitungszeit der Aufgabe anhält und dadurch auch nur wenige False Positives auftreten können.
Nun stellt sich die Frage, ob die ungenauen Vorhersagen an der möglicherweise zu wenig entwickelten *Facial Emotion Recognition* liegen oder ob Gesichtserkennung im Allgemeinen nicht aussagekräftig genug ist, um Erklärungsbedarf zu detektieren.

Bei den Ergebnissen der manuellen Analyse (6.2.2) fällt auf, dass die *Precision* bei den Teilnehmern bereits viel höher liegt. Hier ergeben sich teilweise maximale Werte von circa 85%. Dabei ist der maximale Wert der manuellen Analyse ungefähr 42% über dem der Analyse mittels FER-Bibliothek. Der *Recall* liegt bei mindestens 50% bei jedem Teilnehmer. Der gesamte F1-Score stellt mit circa 68% ein deutlich besseres, aber immer noch kein sehr gutes Ergebnis dar. Es deutet aber darauf hin, dass eine prinzipielle Erkennung von Erklärungsbedarf mittels Gesichtserkennung möglich ist.

Insgesamt kann man sagen, dass möglicherweise noch andere Faktoren mit einbezogen werden müssten, wie zum Beispiel der Kontext, die Umgebung der Person sowie spezifische Mimiken, die die Person hat. Das liegt daran, dass verschiedene Personen eigene oder auch mehrere Mimiken und Reaktionen auf bestimmte Probleme haben und diese ebenfalls unterschiedlich stark ausdrücken[11][12]. Zudem könnte man einen zusätzlichen Faktor zu dem Bild hinzunehmen, wie zum Beispiel den Ton. Dies könnte die Genauigkeit der Erkennung von Erklärungsbedarf erhöhen.

7.3 Mögliche Einschränkungen

Im Folgenden soll auf mögliche Einschränkungen oder Faktoren eingegangen werden, die diese Arbeit beeinflusst haben könnten.
Mit 15 Personen stellt die Teilnehmeranzahl einen recht geringen Wert dar, was die Generalisierbarkeit der Studie verringern könnte. Des Weiteren waren die Teilnehmer zwischen 19 und 29 Jahre alt und repräsentieren damit lediglich die Generation der *Digital Natives*[20]. Inwiefern eine Aussage über *Digital Immigrants*[20] im Bezug auf Erkennung von Erklärungsbedarf getroffen werden kann, ist in dieser Arbeit nicht genau spezifiziert.
Zudem könnten die Ergebnisse der manuellen Analyse besser ausgefallen sein, da dort das bewegte Bild betrachtet wurde und nicht 30 Bilder pro Sekunde, wie bei der Analyse mittels FER-Bibliothek. Aufgrund dessen und der menschlichen Fähigkeit, das bewegte Bild interpretieren zu können, konnte man auch Veränderungen der Mimik und Kontexte dieser besser wahrnehmen. Außerdem wäre es möglich, dass ich die Emotionen der Teilnehmer besser einschätzen konnte, da ich diese zum Großteil persönlich kenne. Ein weiterer Faktor ist, dass Personen unterschiedlich mit Frustration und Ähnlichem umgehen. So war bei der abschließenden Betrachtung der Bildschirm- und Webcam-Aufnahme ersichtlich, dass manche Personen sich beim Aufkommen von Erklärungsbedarf amüsiert zeigen. Jedoch wurden lediglich negative Emotionen (*sad, angry, disgust, surprise, fear*) in die Auswertung einbezogen.
Hinzu kommt, dass Personen nicht nur negative Emotionen im Bezug auf die Software zeigen, sondern diese auch durch andere äußere Faktoren verursacht werden können wie zum Beispiel Juckreiz, Niesreiz, generelles Nachdenken beziehungsweise Konzentration oder Gespräche, die nichts mit der Nutzung der Software zu tun haben. Dies macht es schwer, negative Emotionen in Verbindung mit der Software von denen durch äußere Einflüsse zu unterscheiden, was wiederum eine hohe Fehlerrate (FP) verursacht.
Ein weiterer Faktor ist, dass sich bei zwischenzeitlichem Suchen nach beispielsweise bestimmten Funktionen der Software eine erkennbare Emotion im Gesicht abspielt. Dieser kurze Suchprozess oder ähnliche Situationen resultieren jedoch nicht immer direkt in Erklärungsbedarf und wurden aus diesem Grund möglicherweise nicht in der Umfrage angegeben. Somit könnte auch dies das Ergebnis der Studie beeinflussen.
Des Weiteren hat ein Teilnehmer im abschließenden Gespräch erwähnt, sich durch die aufzeichnende Webcam beobachtet gefühlt und deswegen Mimiken reduziert zu haben. Es kann davon ausgegangen werden, dass dies bei mehreren Teilnehmern der Fall war.
Überdies habe ich bei der weiteren Verarbeitung der erhobenen Daten die dominante Emotion für ein Bild gewählt, die am wahrscheinlichsten war. Handelte es sich bei den Werten aber um beispielsweise 51% und 49%, so könnte auch hierdurch das Ergebnis beeinflusst worden sein, indem die 49%

vernachlässigt worden wären. Auch bei der Bestimmung der Emotion für eine Sekunde, die aus 30 Bildern besteht, habe ich die Emotion gewählt, die am häufigsten vorkam. Dort könnte sich die Anzahl ebenfalls nur sehr gering unterschieden haben.
Zusätzlich habe ich den tatsächlichen Erklärungsbedarf ermittelt, indem ich die Umfrageergebnisse mit den Aufnahmen von Webcam und Bildschirm verglichen und mir die Zeitstempel notiert habe. Allerdings ist es schwierig abzuschätzen, bis zu welcher Sekunde sich der Teilnehmer bei einer Aufgabe gedanklich befindet, wenn nicht gerade eine Interaktion am Bildschirm zu sehen ist.

Kapitel 8

Zusammenfassung und Ausblick

8.1 Zusammenfassung

Zur Untersuchung, ob sich Gesichtserkennung zur Identifizierung von Erklärungsbedarf eignet, wurde im Rahmen dieser Arbeit eine Studie durchgeführt. Hierbei haben die Teilnehmer Aufgaben bearbeitet, die dazu gedacht waren, Erklärungsbedarf auszulösen. Dabei wurden die Teilnehmer sowie deren Bildschirm gefilmt, sodass abschließend ein möglicher Zusammenhang erkannt werden könnte.
Dabei wurde festgestellt, dass sich die Emotionen *sad* und *angry* besser als die restlichen eignen, um Erklärungsbedarf zu detektieren. Emotionen wie *disgust* oder *surprise* erzielten kein korrektes Erkennen von Erklärunsbedarf. Mithilfe einer *Facial Emotion Recognition*-Bibliothek konnte nur ein F1-Score von circa 41% erreicht werden. Durch eine manuelle Analyse, bei der die Videos einzeln genau betrachtet wurden, ergab sich ein F1-Score von ungefähr 68%. Dies zeigt, dass die manuelle Analyse bessere Ergebnisse erzielt, als die FER-Bibliothek. Ein zuverlässiges Ergebnis stellt diese aber nicht bereit.
Die Resultate zeigen auf, dass Gesichtserkennung im Allgemeinen Potential zur Erkennung von Erklärungsbedarf zeigt, dies allein jedoch nicht genügt. Das liegt daran, dass auch negative Emotionen durch andere Auslöser, zum Beispiel Konzentration, äußere Einflüsse oder Juckreiz, ausgelöst werden können. Des Weiteren kommt es auch zu kurzen Verwirrungen, die jedoch nicht zwingend zu Erklärungsbedarf führen müssen. Dabei können allerdings negative Emotionen aufkommen, die dann erkannt werden und fälschlicherweise Erklärungsbedarf angeben.
Abschließend kann man feststellen, dass zur Erkennung von Erklärungsbedarf die Nutzung der Emotionen durch Gesichtserkennung nicht genügt, da zu oft Erklärungsbedarf angegeben wird, wenn dieser nicht vorhanden

ist. Um zuverlässige Ergebnisse zu erzielen, ist es möglicherweise nötig, zusätzliche Elemente, wie beispielsweise Audioaufnahmen, zu verwenden. Somit könnte der durch die Emotionserkennung erkannte Erklärungsbedarf verifiziert werden.

8.2 Ausblick

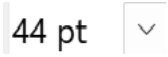
Die Studie macht deutlich, dass ein zuverlässiges Erkennen von Erklärungsbedarf durch Gesichtserkennung allein nicht möglich ist.
In zukünftigen Arbeiten könnte man die Analyse über die Gesichtserkennung hinaus mit weiteren Faktoren, wie zum Beispiel dem aufgenommenen Ton, erweitern. Dies könnte den Kontext mit in Betracht ziehen und so ein besseres Bild über die Situation geben. Zudem könnte man eine Erkennung von Erklärungsbedarf in Echtzeit testen, anders als bei dieser Arbeit im Nachhinein mittels einer Aufnahme.
Des Weiteren ist es möglich, dass sich die Genauigkeit der KI in Zukunft verbessert, sodass die Erkennung dieser Emotionen, die auf Erklärungsbedarf hindeuten, möglicherweise besser als das menschliche Erkennen von Erklärungsbedarf anhand der Emotionen ist. So könnte man diese Studie zu einem späteren Zeitpunkt noch einmal durchführen.
Man könnte jedoch auch andere Erkennungsmerkmale betrachten, um die Emotionen und somit den Erklärungsbedarf zu ermitteln. Man könnte die Emotionen beispielsweise mittels Eyetracking oder anhand biometrischer Daten mittels einer Empatica Watch genauer analysieren.

Anhang A

Nutzerstudie Aufgaben

1 Aufgabe

Bei dieser Aufgabe werden Sie eine Beispielüberschrift generieren und diese mittels Schriftart, Farbe und Größe designen.

1. Klicken Sie zunächst auf das Titel-Textfeld und geben Sie eine beliebige Beispielüberschrift ein.
2. Klicken Sie nun auf den Rahmen des Textfeldes, um dieses auszuwählen.
3. Drücken Sie danach auf den Eigenschaften-Button  , um das Menü auszuklappen, falls dieses noch nicht bereits ausgeklappt ist.
4. Klicken Sie nun im Bereich "Zeichen" auf den Pfeil neben "Liberation Sans" und wählen Sie die Schriftart "Javanese Text" aus der Liste aus.
5. Wählen Sie dann über den Pfeil des Schriftfarbe-Buttons  eine beliebige rote Farbe aus.
6. Ändern Sie nun die Größe der Schrift über das Eingabefeld oben rechts  auf 30 pt.

2 Aufgabe

In dieser Aufgabe sollen Sie den Hintergrund des Textfeldes rot ausfüllen.

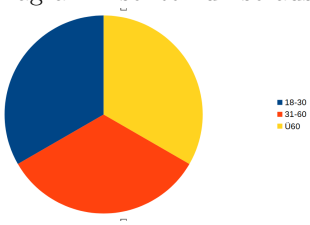
1. Klicken Sie zunächst auf das Textfeld (unteres Feld der Folie).
2. Wählen Sie dann bei der Füllfarbe mithilfe des Pfeiles  eine beliebige rote Farbe aus.
3. Drücken Sie anschließend das Ausfüllsymbol  .

3 Aufgabe

In dieser Aufgabe sollen sie ein Kreisdiagramm auf der Folie hinzufügen, welches drei Gruppen mit jeweils $\frac{1}{3}$ beinhaltet.

1. Klicken Sie auf das "Diagramm einfügen"-Symbol  in der oberen Reihe.
2. Drücken Sie auf die "Diagrammtyp"-Schaltfläche  und wählen Sie anschließend in dem linken Reiter "Kreisdiagramm" und drücken Sie anschließend am unteren rechten Fensterrand auf "OK".

3. Öffnen Sie nun die Datentabelle  und klicken Sie nun auf die Zelle "Zeile 4" und wählen in dem oberen Reiter "Zeile löschen"  .
4. Drücken Sie nun jeweils auf eine beliebige Zelle der Y-Werte in "Spalte 2" sowie "Spalte 3" und anschließend auf "Datenreihe löschen"  .
5. Klicken Sie nun jeweils für alle drei Gruppen (Zeile 1, Zeile 2, Zeile 3) auf die Zellen der Spalte "Y-Werte" der "Spalte 1" und tragen sie "0,3" ein.
6. Drücken Sie abschließend beim rechten unteren Fensterrand auf "Schließen".
7. Das Diagramm sollte nun so aussehen:



4 Aufgabe

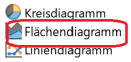
Bei dieser Aufgabe sollen Sie eine Beispielfolie erstellen, der Sie Notizen hinzufügen, sodass man diese während einer Präsentation als eine Art Karteikarte nutzen kann.

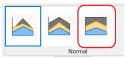
1. Klicken Sie zunächst auf das Titel-Textfeld und geben Sie einen Beispieltitel ein.
2. Öffnen Sie dann die Notizübersicht und geben Sie eine beliebige Beispielnotiz ein.

5 Aufgabe

In dieser Aufgabe sollen Sie ein prozentuales Flächendiagramm auf der Folie abbilden.

1. Klicken Sie dafür zunächst in der obere Zeile auf das Symbol für Diagramme  .
2. Klicken Sie als nächstes auf das Symbol für die Auswahl des Diagrammtyps  .

3. Wählen Sie hier "Flächendiagramm" 

und dann "Prozentual gestapelt" 

aus und drücken Sie auf "OK".

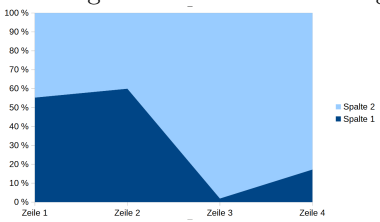
4. Drücken Sie nun oben in der Funktionszeile auf das Symbol für die Datentabelle  .

5. Falls es sich zu Beginn um 3 Spalten handelt, drücken Sie auf eine beliebige Zelle der Spalte 3 und anschließend auf "Datenreihe löschen"  .

6. Tragen Sie nun die Werte für Spalte 1 ein mit 470 für Zeile 1, den Wert 75 für Zeile 2, 5 für Zeile 3 und 104 für Zeile 4 ein.

7. Füllen Sie nun die Zellen der Spalte 2 mit den Werten 380 für Zeile 1, 50 für Zeile 2, 260 für Zeile 3 und 50 für Zeile 4 aus und drücken Sie auf "Schließen".

8. Ihr Diagramm sollte nun wie folgt aussehen:



Anhang B

Auswertungstabelle

Name	**Emotion**	**TP**		**FP**		**FN**	
		manuell	**FER**	**manuell**	**FER**	**manuell**	**FER**
Person A	sad		0		2		5
	angry		0		0		5
	disgust		0		0		5
	fear		0		0		5
	surprise		0		0		5
	Insgesamt	3	0	1	2	3	5
Person B	sad		3		19		1
	angry		0		0		4
	disgust		0		0		4
	fear		1		0		3
	surprise		0		0		4
	Insgesamt	7	4	9	19	1	1
Person C	sad		16		34		0
	angry		6		2		2
	disgust		0		0		4
	fear		0		0		4
	surprise		0		0		4
	Insgesamt	9	22	3	36	1	0
Person D	sad		4		30		0
	angry		0		0		3
	disgust		0		0		3
	fear		0		0		3
	surprise		0		0		3
	Insgesamt	7	4	3	30	0	0
Person E	sad		7		19		0

Name	Emotion	TP		FP		FN	
		manuell	FER	manuell	FER	manuell	FER
	angry		0		1		2
	disgust		0		0		2
	fear		2		29		0
	surprise		0		0		2
	Insgesamt	3	9	11	49	0	0
Person F	sad		13		49		1
	angry		0		0		5
	disgust		0		1		5
	fear		0		0		5
	surprise		0		0		5
	Insgesamt	4	13	7	50	3	1
Person G	sad		13		45		0
	angry		0		0		3
	disgust		0		0		3
	fear		1		4		2
	surprise		0		0		3
	Insgesamt	6	14	5	49	0	0
Person H	sad		4		16		0
	angry		5		5		0
	disgust		0		0		1
	fear		1		2		0
	surprise		0		0		1
	Insgesamt	7	10	5	23	0	0
Person I	sad		11		51		0
	angry		2		1		1
	disgust		0		0		2
	fear		4		15		1
	surprise		0		0		2
	Insgesamt	5	17	8	67	1	0
Person J	sad		23		31		1
	angry		0		0		4
	disgust		0		0		4
	fear		0		2		4
	surprise		0		0		4
	Insgesamt	13	23	12	33	0	1
Person K	sad		7		44		1
	angry		0		1		3
	disgust		0		0		3
	fear		0		0		3

Name	Emotion	TP		FP		FN	
		manuell	FER	manuell	FER	manuell	FER
	surprise		0		0		3
	Insgesamt	5	7	6	45	0	1
Person L	sad		1		4		3
	angry		8		19		2
	disgust		0		0		4
	fear		3		7		2
	surprise		0		0		4
	Insgesamt	8	12	4	30	0	2
Person M	sad		20		36		0
	angry		2		0		3
	disgust		0		0		4
	fear		1		0		3
	surprise		0		0		4
	Insgesamt	11	23	10	36	0	0
Person N	sad		23		43		0
	angry		0		0		3
	disgust		0		0		3
	fear		0		0		3
	surprise		0		0		3
	Insgesamt	11	23	2	43	0	0
Person O	sad		8		39		0
	angry		0		0		3
	disgust		0		0		3
	fear		0		1		3
	surprise		0		0		3
	Insgesamt	15	8	13	40	0	0
Gesamt:	sad		153		462		12
	angry		23		29		43
	disgust		0		1		50
	fear		13		60		41
	surprise		0		0		50
	Insgesamt	114	189	99	552	9	11

Literaturverzeichnis

[1] B. Bakariya, A. Singh, H. Singh, P. Raju, R. Rajpoot, and K. K. Mohbey. Facial emotion recognition and music recommendation system using cnn-based deep learning techniques. *Evolving Systems*, 15(2):641–658, 2024.

[2] A. Bussone, S. Stumpf, and D. O'Sullivan. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics*, pages 160–169. IEEE, 2015.

[3] L. Chazette, W. Brunotte, and T. Speith. Exploring explainability: A definition, a model, and a knowledge catalogue. In *2021 IEEE 29th International Requirements Engineering Conference (RE)*, pages 197–208. IEEE, 2021.

[4] L. Chazette, O. Karras, and K. Schneider. Do end-users want explanations? analyzing the role of explainability as an emerging aspect of non-functional requirements. In *2019 IEEE 27th International Requirements Engineering Conference (RE)*, pages 223–233. IEEE, 2019.

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

[6] L. Derczynski. Complementarity, F-score, and NLP evaluation. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 261–266, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[7] H. Deters, J. Droste, and K. Schneider. On the pulse of requirements elicitation: Physiological triggers and explainability needs. In *REFSQ Workshops*. CEUR Workshop Proceedings, 2024.

[8] F. K. Dosilovic, M. Brcic, and N. Hlupic. Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 0210–0215. IEEE, 2018.

[9] J. Droste, H. Deters, M. Obaidi, and K. Schneider. Explanations in everyday software systems: Towards a taxonomy for explainability needs. In *2024 IEEE 32nd International Requirements Engineering*. IEEE, 2024.

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

[11] M. K. Islam Zim. Opencv and python for emotion analysis of face expressions. In *2023 3rd International Conference on Innovative Practices in Technology and Management (ICIPTM)*, pages 1–7. IEEE, 2023.

[12] A. Kandeel, M. Rahmanian, F. Zulkernine, H. M. Abbas, and H. Hassanein. Facial expression recognition using a simplified convolutional neural network model. In *2020 International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, pages 1–6. IEEE, 2021.

[13] B. Y. Lim and A. K. Dey. Assessing demand for intelligibility in context-aware applications. In S. Helal, H. Gellersen, and S. Consolvo, editors, *Proceedings of the 11th international conference on Ubiquitous computing*, pages 195–204, New York, NY, USA, 2009. ACM.

[14] J. B. Lyons, G. G. Sadler, K. Koltai, H. Battiste, N. T. Ho, L. C. Hoffmann, D. Smith, W. Johnson, and R. Shively. Shaping trust through transparent design: Theoretical and experimental guidelines. In P. Savage-Knepshield and J. Chen, editors, *Advances in Human Factors in Robots and Unmanned Systems*, volume 499 of *Advances in Intelligent Systems and Computing*, pages 127–136. Springer International Publishing, Cham, 2017.

[15] Ç. Menzil, U. E. Sarıkaya, M. S. Aktas, E. Yahsi, M. Keles, and M. Sungur. A business process for detecting facial movements and emotions using deep learning techniques. In *2023 International Conference on Electrical, Communication and Computer Engineering (ICECCE)*, pages 1–8. IEEE, 2023.

[16] A. Pawlicka, M. Pawlicki, R. Kozik, W. Kurek, and M. Choraś. How explainable is explainability? towards better metrics for explainable

ai. In A. Visvizi, O. Troisi, and V. Corvello, editors, *Research and Innovation Forum 2023*, Springer Proceedings in Complexity, pages 685–695. Springer International Publishing, Cham, 2024.

[17] S. Rana, R. Chaudhary, M. Gupta, and P. Garg. Exploring different techniques for emotion detection through face recognition. In *2023 International Conference on Advanced Computing & Communication Technologies (ICACCTech)*, pages 779–786. IEEE, 2023.

[18] M. Sadeghi, V. Klos, and A. Vogelsang. Cases for explainable software systems: Characteristics and examples. In *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, pages 181–187. IEEE, 2021.

[19] A. Singh, S. Gupta, H. Satyawali, V. Sharma, S. Awasthi, and S. Vats. Moodsync: Personalized video recommendation based on user face emotion. In *2024 2nd International Conference on Disruptive Technologies (ICDT)*, pages 975–980. IEEE, 2024.

[20] D. Tapscott. *Grown up digital: How the net generation is changing your world*. McGraw-Hill, New York, 2009.

[21] M. Unterbusch, M. Sadeghi, J. Fischbach, M. Obaidi, and A. Vogelsang. Explanation needs in app reviews: Taxonomy and automated detection. In *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, pages 102–111. IEEE, 2023.

[22] J. P. Winkler and A. Vogelsang. "what does my classifier learn?" a visual approach to understanding natural language text classifiers. In F. Frasincar, A. Ittoo, L. M. Nguyen, and E. Métais, editors, *Natural Language Processing and Information Systems*, volume 10260 of *Lecture Notes in Computer Science*, pages 468–479. Springer International Publishing, Cham, 2017.