HAW HAMBURG

BACHELOR THESIS
Jimmy Cu

Konzeption und Entwicklung einer Affective Computing Anwendung

FAKULTÄT DESIGN, MEDIEN UND INFORMATION
Department Medientechnik

Faculty of Design, Media and Information
Department Media Technology

HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFTEN HAMBURG
Hamburg University of Applied Sciences

Jimmy Cu

Konzeption und Entwicklung einer Affective Computing Anwendung

Bachelorarbeit eingereicht im Rahmen der Bachelorprüfung
im Studiengang *Bachelor of Science Media Systems*
am Department Medientechnik
der Fakultät Design, Medien und Information
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuende Prüferin: Prof. Dr. Larissa Putzar
Zweitgutachter: Thorben Ortmann

Eingereicht am: 20. Juli 2022

Jimmy Cu

Thema der Arbeit

Konzeption und Entwicklung einer Affective Computing Anwendung

Stichworte

Sprachassistent, KI, Emotionserkennung, maschinelles Lernen

Kurzzusammenfassung

Affective Computing beschäftigt sich mit der Fragestellung, wie Technologien zur Emotionserkennung verwendet werden können, um digitale Interaktionen zu humanisieren. So kann das emotionale Verhalten von Menschen gedeutet und analysiert werden und so einen Vorteil in vielen Anwendungsbereichen bieten. Die vorliegende Bachelorarbeit befasst sich mit diesem Forschungsgebiet, der Konzeption einer solchen Anwendung und der prototypischen Implementierung des Konzepts für einen selbst entwickelten Sprachassistenten.

Ausgehend von der Recherche hat sich gezeigt, dass die Emotionserkennung auf Basis von verschiedenen Eingaben erfolgen kann. Eine der attraktivsten Lösungen zur Emotionserkennung erfolgt durch Informationsflüsse der Sprache, der Mimik und der Gestik. Aufgrund dieser Erkenntnisse wurde eine Anwendung entwickelt, welche Emotionen sowohl über die Sprache als auch über die Mimik und Gestik erkennt. Die Anwendung wurde ausschließlich in Python umgesetzt.

Jimmy Cu

Title of Thesis

Conception and development of an Affective Computing application

Keywords

Voice assistent, AI, emotion recognition, machine learning

Abstract

Affective computing deals with the question of how emotion recognition technologies can be used to humanize digital interactions. In this way, the emotional behavior of people can be interpreted and analyzed. Furthermore, affective computing offers an advantage in almost every scope of application. This bachelor thesis deals with the research of emotion recognition, the conception of such an application and the prototypical implementation for a self-developed voice assistant. It has been shown that emotion recognition can be based on various inputs. One of the most attractive solutions for emotion recognition is through the inputs language, the facial expressions and the gestures. Based on these findings an application was developed, which recognizes emotions through language, facial expressions and gestures. The application was implemented exclusively in Python.

Inhaltsverzeichnis

1 Hintergrund

In diesem Kapitel soll verdeutlicht werden, welchen Rahmen diese Thesis umfassen soll. Hierfür wird zunächst ein Einblick in das Forschungsgebiet Affective Computing gewährt. Weiterführend wird die Thematik dieser Thesis konkretisiert, indem die Ziele dieser Thesis definiert werden. Abschließend wird die Struktur der Arbeit vorgestellt.

1.1 Einführung

Der Mensch zeichnet sich als ein emotionales Wesen aus [1] und Computer werden immer besser darin, die Gefühle des Menschen zu verstehen und zu deuten.
Affective Computing oder auch Emotion Artificial Intelligence umfasst die Interaktion zwischen Mensch und Computer. Hierbei werden Daten aus Gesichtern, Stimmen und der Körpersprache gesammelt, um die Gefühle des Menschen zu ordnen und zu messen. Eine Fragestellung, die sich aus diesem Forschungsgebiet entwickelt, ist, wie sich die Mensch-Maschinen-Interaktion durch das Einbeziehen von Emotionen verbessern kann. [2] [3] Denn eines steht fest: Die Kommunikation zwischen Mensch und Maschine erweist sich bisher noch als äußerst kompliziert. Ein Bestandteil dieses Problems kann unter anderem die computerhafte Stimme sein oder auch, dass der Computer noch nicht gut genug die Mimik und Gestik des Menschen deuten kann. Seither forschen viele Start-ups in dem Bereich. Große Firmen wie Amazon, Microsoft, Google und IBM haben mittlerweile Systeme für die Emotionserkennung entwickelt. Während Microsoft mit Microsoft Azure, Google mit Vision AI und Amazon mit Amazon Recognition ihre Emotionserkennung basierend auf den Gesichtszügen einer Person entwickelt hat, verarbeitet das System IBM Watson die natürliche Sprache in Form einer Textanalyse. [2] Doch auch wenn es durch große Firmen wie Amazon, Microsoft und IBM bereits Algorithmen zur Emotionserkennung gibt, gibt es noch keinen konkreten Beweis dafür, dass diese Systeme die „echten“ Gefühle eines Menschen erkennen, so die Neurowissenschaftlerin und Professorin für Psychologie Lisa Feldman Barrett. [4] Feldmann Barrett et al. konnten aufzeigen, dass Emotionen sehr komplex und nicht auf den ersten Blick zu deuten sind. Aus der Studie kam hervor, dass es „unmöglich sei, aus einem einfachen Lächeln mit Sicherheit die Emotion Freude zu identifizieren.“ [4] Das Erleben von Emotionen ist subjektiv, so unterscheiden sich die Empfindungen einer Emotion von Mensch zu Mensch und von Situation zu Situation. [4] [5]

1.2 Zielsetzung

Diese Arbeit befasst sich mit der Affective Computing Forschung und die Entwicklung einer Mensch-Computer-Schnittstelle, die den emotionalen Zustand eines Endbenutzers erkennen und angemessen darauf reagieren kann. Dazu werden mithilfe von Recherchen zu Emotionen und bereits existierenden Algorithmen und Open-Source-Projekten ein Konzept für eine Anwendung zur Emotionserkennung entwickelt. Um einen genauen Vergleich ziehen zu können, soll bedacht werden, dass die Emotionserkennung mehrere Eingabeströme berücksichtigt. Auf dieser Grundlage werden Leitfragen definiert, die diese Thesis zu einem Ergebnis führen sollen. Infolgedessen sollen folgende Fragen im Laufe der Thesis beantwortet werden.

- *Wie werden Emotionen ausgedrückt?*

  Es soll gezeigt werden, auf welche Art und Weise Emotionen geäußert werden können.

- *Wie können Emotionen maschinell wahrgenommen werden?*

  Es soll untersucht werden, über welche Sensoren Computer in der Lage sind, Emotionen erkennen zu können.

- *Wie ist der Stand von nicht kommerziellen Möglichkeiten? Wie einfach ist die Umsetzung einer Emotionserkennung für unerfahrene Personen?*

  Es soll herausgefunden werden, wie anfängerfreundlich und einfach die Nutzung von nicht kommerziellen Möglichkeiten ist. Hierzu soll untersucht werden, ob es möglich ist, mit nicht-kommerziellen Möglichkeiten eine präzise Emotionserkennung zu implementieren.

- *Für welchen Anwendungsfall könnte die Umsetzung einer hybriden Emotionserkennung geeignet sein?*

  Es soll herausgefunden werden, für welchen Anwendungsfall die Verarbeitung von mehreren Eingabeströmen sinnvoll umgesetzt wird.

- *Welche Faktoren können einen positiven bzw. negativen Einfluss auf die Emotionserkennung nehmen?*

  Es soll aufgezeigt werden, welche Faktoren Einfluss auf die Emotionserkennung nehmen können.

1.3 Struktur der Arbeit

Damit das Ziel der Thesis erreicht werden kann, wird im Kapitel 2 zunächst der Forschungsbereich um Emotionen thematisiert und analysiert. Es werden die Grundlagen von Emotionen näher erläutert und folgende Fragen werden in diesem Abschnitt untersucht: Was sind Emotionen? Wie funktionieren sie? Und wie werden Emotionen ausgedrückt?
Weiterführend wird im Kapitel 3 im Bezug auf Affective Computing der Stand der Technik aufgeführt. Dabei werden die Möglichkeiten der maschinellen Wahrnehmung zur Emotionserkennung untersucht. Die Komponenten Bild, Ton und Text, welche Bestandteil der prototypischen Umsetzung sind, werden in diesem Abschnitt näher betrachtet. Um für die folgenden Kapitel ein besseres Verständnis zu entwickeln, wird der Begriff der künstlichen Intelligenz näher erläutert. Passend dazu werden die verschiedenen Techniken untersucht, die in dem Projekt angewendet und umgesetzt worden sind. Ausgehend von diesem Kapitel wird im Kapitel 4 das Konzept für jene Anwendung erstellt. Des Weiteren werden in diesem Kapitel nur die Möglichkeiten der nicht kommerziellen Dienste in Betracht gezogen. Die Umsetzung der Arbeit setzt den Fokus auf jene Möglichkeiten. Im nächsten Kapitel, dem Kapitel 5, wird die Implementierung des Konzepts beschrieben. Bei der Implementation wird auf die verwendete Architektur des Prototypen eingegangen. Das Kapitel umfasst unter anderem die verwendeten Tools sowie die Struktur der Anwendung. Im Anschluss wird im Kapitel 6 der Bereich der Evaluation beleuchtet. Für die Untersuchung wird der Prototyp zum Einsatz gebracht. Das Kapitel beschreibt den durchgeführten Testdurchlauf und die daraus resultierenden Beobachtungen. Abschließend wird im Kapitel 7 eine Zusammenfassung auf den Inhalt der Arbeit und ein Ausblick darüber, wo Affective Computing in Zukunft stehen kann, gegeben.

2 Grundlagen von Emotionen

Das Konstrukt Emotion ist ein Bestandteil unseres Alltags. Es vergeht beinahe kein Tag, an dem der Mensch nicht emotionale Zustände wie Freude, Angst oder Ärger empfindet. Um ein grundsätzliches Verständnis für die darauf folgenden Kapitel zu schaffen, wird zunächst der psychologische Begriff „Emotion“ näher untersucht. Zunächst werden die Grundlagen von Emotionen erläutert, welche das Fundament dieser Arbeit bilden. Hierbei liegt der Fokus vor allem auf der umfangreichen Definition von Emotionen, der Abgrenzung des Emotionsbegriffs und deren Klassifizierung.

2.1 Definition

Der Begriff „Emotion“ stellt Neurowissenschaftler seit Jahrhunderten vor eine große Herausforderung. [6] Der Begriff Emotion wird vielschichtig und unterschiedlich definiert, sodass es Wissenschaftlern bislang nicht gelungen ist, sich auf eine Definition zu einigen. [7] Die Schwierigkeiten liegen in der Abgrenzung zu anderen Komponenten, wie zum Beispiel der Motivation. Fasst man jedoch die Gemeinsamkeiten der Definitionen zusammen, wie Kleinginna et al. [8], so haben Emotionen „subjektive erfahrbare und objektive erfassbare Komponenten, die zielgerichtetes Verhalten begleiten bzw. fördern, das dem Organismus eine Anpassung an seine Lebensbedingungen ermöglicht.“ [7] Im Grunde bedeutet es, dass bestimmte Ereignisse konkrete Emotionen hervorrufen, wie zum Beispiel Angst, Ärger, Freude oder Überraschung. Diese Emotionen können wiederum Gefühlsempfindungen auslösen. Zu diesen Empfindungen zählen Verhaltensreaktionen wie bspw. der Ausdruck einer Emotion wie ein Weinen bei Traurigkeit. Aber auch physiologische Reaktionen, wie z. B. die Veränderung der Herzfrequenz oder die Erweiterung der Blutgefäße zählen zu diesen Gefühlsempfindungen. Diese Zustände sind meist mit Ereignissen verbunden, die für uns von persönlicher Relevanz sind. Dass uns Emotionen in eine bestimmte Richtung bewegen können, kann am Beispiel der Emotion Angst erkannt werden. Die Angst motiviert Menschen zur Vermeidung von bestimmten Situationen. [9] [10]
Emotionen werden als ein natürliches Phänomen empfunden. In Abbildung 2.1 (oben) ist diese Ansicht dargestellt. Anhänger der Gegenposition hingegen betrachten Emotionen als Konstruktionen von Menschen, siehe Abbildung 2.1 (unten). Die Ursache für Emotionen sind Ereignisse, die bestimmte Reaktionen bei dem Menschen auslösen, die er bei sich selbst oder bei anderen wahrnimmt. [10]

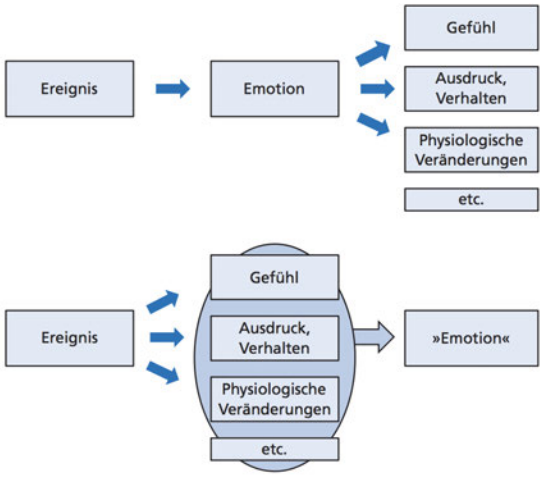


Abbildung 2.1: Emotionen als natürliche Phänomene (oben) oder als soziale Konstruktionen (unten)[10]

Es wird bereits ersichtlich, dass das Konstrukt Emotion viele Facetten aufweist. Abgesehen von den bereits genannten Gefühlsempfindungen werden von Autoren weitere Reaktionen aufgeführt. Demzufolge ist ein weiterer Bestandteil einer Emotion auch das Bewerten des Ereignisses oder eine bestimmte Motivationslage. Diese Veränderungen werden meistens als Komponenten von Emotionen angesehen. [10]
In der Abbildung 2.2 ist zur Veranschaulichung ein Beispiel eines Komponentenmodells dargestellt. Das Modell nach Rothermund und Eder [11] beschränkt sich auf fünf Reaktionen, die auf eine Emotion folgen: Das Erleben, die Kognition, die Physiologie, der Ausdruck und die Motivation. In dem Modell sind drei Aspekte von Emotionen vertreten, die von den meisten Emotionstheoretikern als charakteristisch anerkannt worden sind. [9] Aufgrund der Komplexität des Begriffs Emotion und der Relevanz für die Thesis werden auch nur diese drei Aspekte weiter thematisiert.

Der **Erlebensaspekt** beschreibt, dass Menschen Gefühle subjektiv erleben und entsprechend auch empfinden. Somit empfindet jeder Mensch eine Emotion unterschiedlich. [9]

Die **physiologische Reaktion** zeigt, dass Emotionen körperliche Veränderungen auslösen wie z. B. Schwitzen, Erröten, aber auch zur Erhöhung der Herzfrequenz resultieren können. Diese physiologischen Reaktionen stoßen eine Anpassung an jene Bedingungen an, die diese Erregungen hervorrufen können. Sie bereiten z. B. den Körper bei Gefahr auf eine Flucht vor. [9]

Der **Verhaltensaspekt (Ausdruck)** wird durch die Reaktionen auf eine Emotion definiert. Das bedeutet, dass sich Emotionen im Verhalten ausdrücken lassen, indem sie mit Bewegungen bestimmter Gesichtsmuskeln oder einer bestimmten Körperhaltung einhergehen. So kann der Verhaltensaspekt als eine spezifische Mimik, Gestik, Haltung oder Stimme wahrgenommen werden oder auch als konkreter Handlungsimpuls, wie z. B. die Vermeidung bei Ekel. [9]

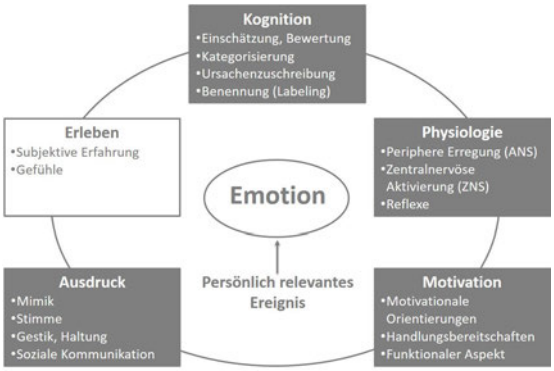


Abbildung 2.2: Das Komponentenmodell der Emotionen nach Rothermund und Eder [11]

2.2 Abgrenzung

Emotionen werden auf verschiedene Arten und Weisen ausgedrückt, so wird der Begriff Emotion oftmals in der Wissenschaft mit ähnlich verwandten Konstrukten verwendet. Um für den vorliegenden Abschnitt den Emotionsbegriff von **Stimmungen**, **Affekten** oder **Gefühlen** abzugrenzen, wurde der Ansatz von Meyer, Reisenzein und Schutzwohl [12] herangezogen. Emotionen sind demnach durch folgende Merkmale charakterisiert.

a) Sie sind aktuelle psychische Zustände von Personen und zeitlich datierte, unwiederholbare Ereignisse:
Das bedeutet, dass eine Emotion in einer zeitlich begrenzten **Dauer** vorkommt und jede Emotion ein unwiederholbares Ereignis ist. Wenn wir zum Beispiel Freude über einen Sieg von unserer Heimmannschaft empfinden, so tritt diese Freude unmittelbar in dieser Situation auf, hält eine Weile an und verflüchtigt sich schließlich wieder. [9]

b) Emotionen haben eine Qualität, Intensität und Dauer:
Emotionen sind durch drei Merkmale zu klassifizieren. Das wichtigste Merkmal stellt die **Qualität** der Emotion dar, womit die spezifischen Basisemotionen gemeint sind, wie zum Beispiel Freude, Angst und Wut. Im herkömmlichen Sprachgebrauch wird die Qualität von Emotionen mit dem Begriff Emotion meist gleichgesetzt. [9]

Die **Intensität** einer Emotion bildet das zweite wichtige Klassifizierungsmerkmal einer Emotion. Der Intensitätsgrad beschreibt die Veränderung der Emotion in ihrer Intensität. Demzufolge werden die Basisemotionen anhand ihrer Stärke abgebildet und erzeugen demnach vielfältige Emotionen, zum Beispiel Schrecken als stärkere Version der Emotion Angst. In der Abbildung 2.3 wird die Intensität von Emotionen visuell dargestellt. [9]

Wie schon in Punkt a) aufgeführt, sind Emotionen temporär und somit nur über eine bestimmte Dauer vorhanden. [9]

c) Emotionen sind in der Regel objektgerichtet:
Die **Objektgerichtetheit** beschreibt, dass Emotionen sich immer auf „etwas“ beziehen, wie zum Beispiel die Freude über ein Geschenk. Dabei müssen die Objekte nicht zwingend existieren, lediglich die Überzeugung und Interpretation der betroffenen Person ist für die Emotionsentstehung ausschlaggebend. [9]

Mit diesen Merkmalen lassen sich Emotionen von anderen ähnlichen Konstrukten abgrenzen. Hierbei ist die zeitliche Begrenzung ein wichtiges Unterscheidungsmerkmal, um diese von dem verwandten Konstrukt der **emotionalen Disposition** zu differenzieren. Eine emotionale Disposition beschreibt einen emotionalen Zustand, der sich über einen längeren Zeitraum ereignet. Genauso lässt sich die Emotion von dem Konstrukt Stimmung abgrenzen. Ein wesentlicher Unterschied ist, dass Stimmungen sich anders als Emotionen nicht auf Objekte beziehen. Stimmungen lassen sich durch die folgenden drei Merkmale charakterisieren: Eine längere Dauer, eine geringere Intensität und eine Unbestimmtheit. Demnach sind Stimmungen nur vage zu beschreiben und sind in ihrer Intensität weniger stark ausgeprägt, dafür werden sie in einem längeren Zeitraum empfunden als Emotionen. „In einer ängstlichen Stimmung zu sein bedeutet daher, sich über eine längere Zeitspanne hinweg ängstlich zu fühlen, ohne jedoch konkret Angst vor etwas (hier das dritte Kriterium der Unbestimmtheit) zu haben.“ [9] Anders als bei Stimmungen werden Affekte oftmals als kurzfristig auftretende Reaktionen definiert. Affekte werden durch bestimmte Situationen ausgelöst und können kognitiv wenig kontrolliert werden. Schließlich sind Emotionen noch von Gefühlen abzugrenzen. Ein Gefühl beschränkt sich nur auf das subjektive Erleben der Emotion und ist dementsprechend nur ein Bestandteil der Emotion. [9] [11]

2.3 Klassifizierung

In der Geschichte der Emotionspsychologie haben sich viele verschiedene Ansätze entwickelt, um Emotionen zu kategorisieren. [7] Dabei wird zwischen dimensionalen und kategorialen Konzeptionen unterschieden. Hierbei sollen die kategorialen Konzeptionen als Grundlage für die eigene Entwicklung der Affective Computing Anwendung dienen. Aus diesem Grund wird diese Art der Klassifizierung im Fokus stehen. [7]

In den **dimensionalen Konzeptionen** von Emotionen wird davon ausgegangen, dass sich Emotionen in ihrer Ausprägung auf verschiedenen Dimensionen einordnen lassen. Demzufolge lässt sich eine Emotion zunächst auf der sog. Valenzdimension kategorisieren. Hierbei wird eine Unterscheidung in eine positive oder negative Emotion durchgeführt. Auf der Intensitätsdimension lässt sich dann bestimmen, wie stark die jeweilige Emotion als positiv oder negativ erlebt wird. [7]

Verfechter der **kategorialen Konzeptionen** hingegen gehen davon aus, dass sich die Emotionen aus bestehenden Basisemotionen zusammensetzen. Basisemotionen sind die Grundlage, aus denen sich alle komplexeren Emotionen zusammensetzen. Hierbei geht

es weniger darum, Emotionen nach ihrer Ausprägung zu ordnen, sondern darum, qualitativ verschiedene Emotionen inhaltlich voneinander abzugrenzen. Robert Plutchik, ein Vertreter der evolutionspsychologischen Emotionsforschung, hat hierzu eines der bekanntesten Modelle „Das Rad der Emotionen“ entwickelt, um die Anordnung der Emotionen zu veranschaulichen. [7]

Hierzu extrahierte Plutchik acht Basisemotionen: Groll, Erwartung, Freude, Vertrauen, Angst, Überraschung, Traurigkeit und Abneigung. Dabei werden diese Emotionen nach Intensität und Ähnlichkeit zu anderen Emotionen ringförmig angeordnet. In dem Ring sind die Emotionen so angeordnet, dass ähnliche Emotionen möglichst nah beieinander und gegensätzliche Emotionen weit voneinander entfernt liegen. Die intensivsten Emotionen stehen in der Mitte des Kreises, sodass die Intensität der Emotion vom inneren bis zum äußeren Ring immer weiter abnimmt. Emotionen, die sich aus zwei in diesem Ring direkt benachbarten Emotionen zusammensetzen, bezeichnete Plutchik als Primäremotionen. So bildet z. B. die Kombination aus Freude und Vertrauen die Primäremotion Liebe. Auch wenn in Plutchiks Rad der Emotionen acht Basisemotionen vertreten sind, ist man bis heute uneinig über die Gesamtmenge der Basisemotionen. Demnach haben bis heute verschiedene Forscher unterschiedliche Kriterien dafür, wann eine Emotion als Basisemotion kategorisiert werden kann. [7] Trotz der unterschiedlichen Kriterien werden die Emotionen Freude, Traurigkeit, Furcht und Wut dennoch einstimmig als Basisemotionen angesehen, da diese kulturübergreifend vertreten sind und überall nahezu gleich ausgedrückt werden. [7]

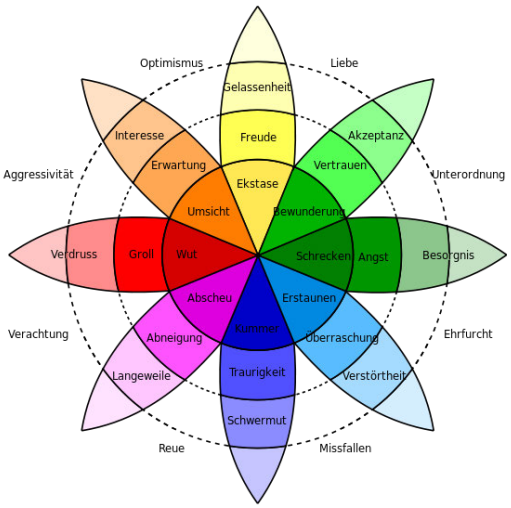


Abbildung 2.3: Das Rad der Emotionen [13]

3 Affective Computing

Affective Computing bildet die Schnittstelle zwischen Mensch und Maschine mit der Besonderheit, dass Emotionen im Mittelpunkt stehen. [14] In diesem Kapitel wird der Forschungsbereich Affective Computing näher beleuchtet. Die Grundlagen des Affective Computing schaffen das Know-how für die nachfolgenden Kapitel. Das Kapitel soll dabei helfen, folgende Frage zu verstehen: Wie sollen Maschinen lernen, den Menschen zu verstehen?
Um der Frage auf den Grund zu gehen, werden zunächst die Möglichkeiten der Emotionserkennung aufgezeigt. Es werden der Stand der Technik und dazugehörige Beispiele vorgestellt. Im Rahmen der Arbeit sollen die Möglichkeiten für jeden greifbar sein, weswegen nicht kommerzielle Möglichkeiten im Fokus stehen. Mithilfe dessen soll das Grundverständnis für eine affektive Anwendung gefördert werden und gilt als Vorbereitung für das nächste Kapitel, der Konzeption. Ebenso wird der Begriff der künstlichen Intelligenz näher untersucht, da diese einen wichtigen Bestandteil der Affective Computing Forschung bildet. Unter anderem werden die Ausdrücke Machine Learning, Deep Learning sowie das künstliche neuronale Netz (KNN) erklärt. Diese Arbeit beschränkt sich auch hier wieder nur auf die Bereiche der künstlichen Intelligenz, die für diese Thesis relevant sind.

3.1 Einführung

„Eine Welt, in der auf Ihre Gefühle besondere Rücksicht genommen wird.“ [14] Eine Vorstellung, die dank Affective Computing nun greifbar ist. Eine Vorstellung, die allgemein Rosalind Picard zugeschrieben wird, einer Professorin und Wissenschaftlerin am Massachusetts Institute of Technology (MIT). Bereits 1995 hat Rosalind Picard einen Fachbericht mit dem Titel „Affective Computing“ veröffentlicht. [15] Basierend auf dem Wissen hat sie 1997 ein in der Forschung oft herangezogenes Buch mit demselben Titel veröffentlicht. [3] Ihr ist es zu verdanken, dass sich Affective Computing seitdem als interdisziplinärer Forschungsbereich etabliert hat. [16]
Wie zuvor erwähnt, sollen dabei affektive Technologien die menschlichen Emotionen wahrnehmen und/oder sogar beeinflussen können. Deutlich wird hierbei, dass genauso wie der Mensch Emotionen auf verschiedene Art und Weise ausdrücken kann, auch affektive Technologien auf verschiedene Art und Weise mit Emotionen umgehen können. [14]

Innerhalb des Affective Computing gibt es entsprechend unterschiedliche Systeme, welche Emotionen unterschiedlich verarbeiten sollen. [17] Es gibt unter anderem ein System „fühlender Mensch“, welches den Fokus auf die Gefühle des Menschen legt. Emotionssensitive Technologien sollen dabei die Gefühle des Menschen verstehen können. Mittels Sensoren wie Kameras, Blutdruckmesser oder Mikrofone sollen Emotionen erkannt und entsprechend darauf reagiert werden. Der Gegenpol dieses Systems ist ein System aus dem Feld „Emotional Robotic“. Innerhalb dieses Forschungsfelds sollen emotionale Technologien verwendet werden, um die Emotionen eines Menschen zu simulieren. Es kann dazu dienen, einen Roboter glaubwürdiger und menschenähnlicher erscheinen zu lassen, sodass dieser bspw. fröhlicher reagiert, nachdem man ihm ein Kompliment gemacht hat. Man spricht hierbei von einem „fühlenden Computer“. Neben diesen beiden Unterscheidungen gibt es eine weitere. Hierzu werden unterschiedliche Systeme zusammengeführt: Das eine funktioniert emotionsbeeinflussend, während das andere emotionsneutral agiert. Ersteres ist ein System, das auf die Gefühle des Menschen einwirkt. Emotionsneutrale Systeme hingegen sollen die Emotionen des Menschen erfassen können, ohne dabei die Gefühle des Menschen manipulieren zu wollen. [14] [17]
Aus diesen Systemen lässt sich ein breites Spektrum an verschiedenen Anwendungsfeldern bilden. Demnach findet Affective Computing unter anderem Anwendung im Bereich Marketing, im Gesundheitswesen oder in der Bildung. Im Bereich Marketing können damit die Effizienz von Produktrezensionen analysiert und die Kundenzufriedenheit verbessert werden. Im Gesundheitswesen können affektive Systeme zur Behandlung von Depressionen genutzt werden. Auch für den Alltag wurden mittlerweile affektive Systeme entwickelt. Der „Buddy“ ist ein Beispiel eines menschenähnlichen und emotionsbeeinflussenden Systems und findet Anwendung im Haushalt. Das System besitzt ein menschliches Gesicht, das die eigenen Emotionen widerspiegelt. Es verfolgt den Zweck, die Stimmung von Menschen positiv zu beeinflussen. [18] [19]

Abbildung 3.1: Das emotionsbeeinflussende System Buddy [20]

3.2 Künstliche Intelligenz

> „Ziel der KI ist es, Maschinen zu entwickeln, die sich verhalten, als verfügten sie über Intelligenz.“ John McCarthy [21]

Ähnlich wie das Konstrukt Emotionen gibt es für den Begriff der künstlichen Intelligenz (KI) diverse Beschreibungen. Das Forschungsgebiet der KI lässt sich in verschiedene Teilgebiete unterteilen. Um ein tiefergehendes Verständnis für die darauffolgenden Abschnitte zu entwickeln, werden bestimmte Gebiete der KI erläutert. [21]

3.2.1 Machine Learning

Machine Learning (Maschinelles Lernen) ist ein Verfahren, welches in Expertenkreisen als Schlüsseltechnologie der KI gehandhabt wird. Ähnlich wie bei der menschlichen Intelligenz wird ein künstliches System über Lernprozesse entwickelt. Das bedeutet, dass ein System aus Erfahrungen Wissen generiert, indem es von Beispielen lernt und so ein komplexes Modell entwickelt. Anschließend können das Modell und die damit erworbenen Wissensrepräsentationen auf neue Eingabewerte angewendet werden, um das Lösen von Problemen kontinuierlich zu verbessern. [22]

3.2.2 Deep Learning und künstliche neuronale Netze

Deep Learning ist ein Teilbereich des Machine Learnings und eine spezielle Methode der Informationsverarbeitung. Deep Learning arbeitet mit künstlichen neuronalen Netzen (KNN). [23] Bei einem KNN handelt es sich um eine Art künstliches Abstraktionsmodell, um das menschliche Gehirn abzubilden. Es besteht aus mehreren Schichten von miteinander verbundenen künstlichen Neuronen und funktioniert ähnlich wie bei einem menschlichen Gehirn. Ein KNN verfügt über eine Eingabe- und eine Ausgabeschicht. Dazwischen befindet sich die verborgene Schicht, bestehend aus beliebig vielen Neuronen. Die zu analysierenden Informationen werden hierzu über Eingangsneuronen in das neuronale Netz eingelesen und das Ergebnis lässt sich an den Ausgangsneuronen ablesen. Ein Neuron besteht aus mehreren Eingängen, einer Aktivierungsfunktion und einem Ausgang. Über sog. Kanten sind die Neuronen miteinander verbunden. Diese sind einer bestimmten Gewichtung zugeordnet, die aussagt, wie viel Einfluss ein Neuron auf ein anderes haben kann. Der Zustand des Neuronenausgangs ist dementsprechend abhängig von der gewichteten Summe aller Eingänge. Über einen gesetzten Schwellwert der Aktivierungsfunktion wird schließlich der Zustand des Neuronenausgangs bestimmt. Abhängig davon, ob der Schwellwert durch die gewichtete Summe aller Eingänge überschritten oder unterschritten wird, werden die Informationen weitergeleitet oder nicht. Das bedeutet, dass Informationen nur dann weitergegeben werden, wenn die gewichtete Summe aller Eingänge den Schwellwert der Aktivierungsfunktion überschreitet, andernfalls wird der Zustand des Neurons auf inaktiv gesetzt. [22] [24]

Mittlerweile gibt es eine Vielzahl an neuronalen Netzen und je nach Art und Schwierigkeit der Aufgabe kommen die entsprechenden neuronalen Netze zum Einsatz, die sich in ihrem Aufbau und Anzahl der Neuronenschichten voneinander unterscheiden. [25]

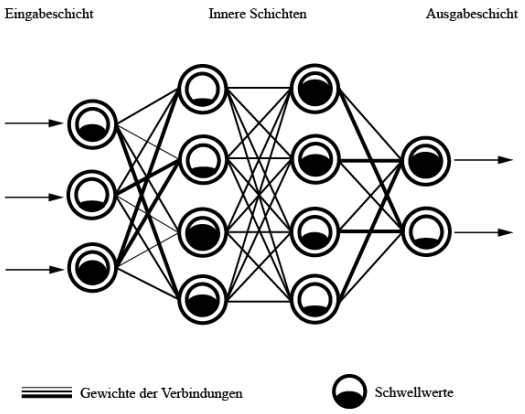


Abbildung 3.2: Neuronales Netz [25]

3.2.3 Mehrlagiges Perzeptron

Das mehrlagige Perzeptron (MLP) ist ein Modell eines KNN, der sowohl für die Klassifikation als auch für die Regression verwendet wird. [26] Für die Thesis wird der Fokus nur auf den Teilbereich der Klassifikation gelegt. Ein MLP besteht in der einfachsten Form aus einem Neuron mit mehreren Eingängen und einem Ausgang. Der Ausgabewert wird hierbei durch die Gewichtung der Eingänge und durch die Schwellwerte bestimmt. Diese Gewichtungen können anhand des Lernprozesses verändert werden, um bestimmte Klassifikationsaufgaben zu erfüllen. Ein mehrlagiges Perzeptron besteht aus mehreren Schichten mit Neuronen, die untereinander vernetzt sind. Die Anzahl an Schichten definiert hierbei die Vielfältigkeit der Klassifizierungsfähigkeiten des neuronalen Netzes. Das bedeutet, je mehr Schichten in dem Perzeptron vorhanden sind, desto mehr lassen sich komplexere Klassifizierungsformen bilden. [27]

3.2.4 Convolutional Neural Network

Das Convolutional Neural Network (CNN) ist eine weitere Form des KNN. Es ist optimal geeignet für die Verarbeitung von Bild- und Audiodaten. Ein CNN ist dem menschlichen Gehirn nachempfunden und soll die Sehrinde eines Gehirns abbilden. Ein CNN ist aus mehreren Schichten aufgebaut, bestehend aus folgenden einzelnen Schichten.

- Convolutional-Layer
- Pooling-Layer
- Fully-Connected Layer

Ein CNN ist so aufgebaut, dass nach einem Convolutional-Layer ein Pooling-Layer folgt. Hierbei kann es sein, dass die Kombination der beiden Schichten mehrfach hintereinander vorhanden ist. Zum Ende hin folgt ein sog. Fully-Connected Layer. [28] Zur Veranschaulichung steht die Abbildung 3.3 zur Verfügung.

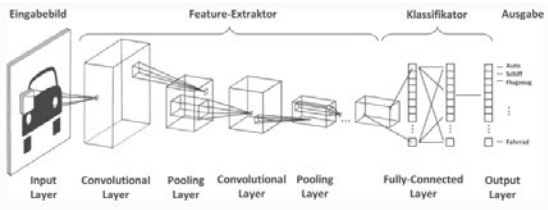


Abbildung 3.3: Aufbau eines CNN [28]

Die Aufgabe eines **Convolutional-Layers** ist es, aus den Eingabedaten Merkmale zu erkennen und zu extrahieren. Demnach können in der Bildverarbeitung Merkmale wie Linien, Kanten oder andere bestimmte Formen extrahiert werden. Hierzu werden die Eingabedaten in Form einer Matrix verarbeitet. Die Matrizen sind in Breite x Höhe x Kanäle definiert. Der **Pooling-Layer** dient dazu, überflüssige Informationen zu verwerfen. Hierzu werden die erkannten Merkmale verdichtet und die Auflösung reduziert. Infolgedessen werden die Datenmengen reduziert und die Berechnungsgeschwindigkeit erhöht sich. Der **Fully-Connected Layer** dient dazu, Merkmale und Elemente der vorherigen Schichten mit jedem Ausgabemerkmal zu verknüpfen. Die Neuronen können in mehreren Ebenen angeordnet sein. Die Anzahl der Neuronen ist abhängig von den Klassen oder Objekten, die das neuronale Netz unterscheiden soll. [28]

3.3 Stand der Technik

Damit ein Computer lernt, mit den Emotionen des Menschen umzugehen, musst er zunächst lernen, diese erfassen zu können. Für die Emotionserkennung existieren die verschiedensten Dienste, Open-Source-Projekte und Algorithmen. Diese basieren auf verschiedenen Eingabeströmen für die Erkennung von Emotionen und je nach Anwendung variieren diese. Zu den am häufigsten genutzten Eingabeströmen gehören die Bilderkennung, die Spracherkennung und die Texterkennung. Eher selten werden Eingabeströme wie *Eye-Tracking, Elektrodermale Aktivität (EDA), Elektrokardiogramm (EKG), Elektroenzephalografie (EEG) oder Elektromyografie (EMG)* angewendet. Hinsichtlich der Relevanz werden nur die am häufigsten genutzten Eingabeströme wie Bild, Text und Audio näher thematisiert. [29] [30]

Auch im Rahmen dieser Thesis werden weitere Faktoren wie das Alter und das Geschlecht außen vor gelassen. Zugleich mit dem Wissen, dass durch die Einbeziehung des Alters und des Geschlechts eine Verbesserung möglich ist. Eine wissenschaftliche Arbeit zu sprachbasierter Emotionserkennung konnte durch die Einbeziehung des Alters eine Verbesserung von 7 % erreichen. [31] Eine weitere wissenschaftliche Arbeit dazu konnte ebenfalls die Genauigkeit der Emotionserkennung erhöhen, aufgrund der Tatsache, dass das Geschlecht einbezogen worden ist. [32]

3.3.1 Bilderkennung

Viele Daten für die gesichtsbasierte Emotionserkennung basieren auf den Arbeiten von Paul Ekman. Paul Ekman ist Professor für Psychologie an der University of California und einer der bekanntesten Emotionswissenschaftler weltweit. Mehr als 40 Jahre widmete sich Paul Ekman dem Konstrukt Emotionen. [33] Dabei interessierte ihn überwiegend deren Ausdruck. Durch das Beobachten von Ur-Völkern kam er zur Annahme, dass Emotionen universell sind. Das bedeutet, dass Emotionen kulturübergreifend gezeigt und auch verstanden werden können. Laut seiner Thesis gibt es sieben sogenannte Basisemotionen, die in allen Kulturen der Welt vertreten sind und nahezu gleich ausgedrückt werden. [7] [34]

- Freude
- Trauer
- Überraschung
- Ekel
- Angst
- Zorn
- Verachtung

Diese Basisemotionen können anhand von Merkmalen im menschlichen Gesicht identifiziert und beispielsweise über Kameras erfasst werden. [35] Bei seiner Forschung kam er zu dem Ergebnis, dass der Mensch über 10.000 Gesichtsausdrücke verfügt, aber nicht alle eine Emotion auch transportieren können. Demnach setzte er seinen Fokus darauf, jene Gesichtsausdrücke zu dechiffrieren, welche auch tatsächlich eine Emotion übermitteln. Basierend auf diesem Wissen entwickelte er 1978 mit Wallace Friesen ein Kodierungsverfahren zur Beschreibung von Emotionen, das Facial Action Coding System (FACS). In der heutigen Zeit ist das FACS die Grundlage vieler Gesichtserkennungssoftwares. [36] [37] Das FACS ist ein anatomisch basiertes System und dient dazu, jegliche beobachtbaren Gesichtsbewegungen zu beschreiben. Das System erkennt dabei nur die emotionalen nonverbalen Gesichtsausdrücke, während die nicht emotionalen Gesichtsausdrücke ausgeblendet werden. Hierbei wird jede beobachtbare Komponente der Gesichtsbewegung

einer Aktionseinheit (Action Unit, AU) zugeordnet. Bei einer Aktionseinheit werden einzelne oder mehrere Muskelbewegungen zusammengefasst. Demnach wird z. B das Heben der Augenbrauen in einer sog. Aktionseinheit zusammengefasst. Insgesamt fasst das FACS 44 Einheiten zusammen, das Obergesicht umfasst 12, das Untergesicht 32 AU. Durch die Kombination der Einheiten können dann die Basisemotionen bestimmt werden. [2] [38] Die folgende Abbildung 3.4 veranschaulicht beschriebenes.

Abbildung 3.4: Aktionseinheiten, die in der Serie „Lie to me“ zur Ausdrucksbewertung berücksichtigt worden sind. Ekman (2009) [39]

Der Bereich um die gesichtsbasierte Emotionserkennung hat sich fortgehend immer weiter entwickelt und mittlerweile existieren viele verschiedene Ansätze und Algorithmen dazu. Facial Emotion Recognition, kurz auch FER, ist eine der gängigsten Methoden zur Emotionserkennung. [40] Das liegt unter anderem daran, dass Gesichtsausdrücke zu den wichtigsten Merkmalen der Emotionserkennung gehören. Heutzutage findet man FER Systeme in den verschiedensten Bereichen. Neben der üblichen Anwendung im Bereich KI und der Mensch-Maschinen-Interaktion findet man FER Systeme in Bereichen wie augmented reality, affective gaming oder autonomem Fahren. FER Systeme lassen sich in zwei Arten unterteilen. Zum einen gibt es traditionelle Systeme, die auf Machine Learning basieren und Systeme, die auf Deep Learning basieren. [40]

1.) FER Systeme basierend auf Machine Learning

Die Emotionserkennung basierend auf Machine Learning ist eine traditionelle Methode und erfolgt in der Regel in drei Schritten. [40] Zu Beginn wird aus den eingelesenen Informationen ein Gesichtsbild extrahiert. Demnach werden aus einem bereits existierenden Bild die Informationen zu dem Gesicht entfernt werden. Man spricht hierbei von einem statischen FER (static FER). Ein weiterer Ansatz arbeitet mit Videodateien. Hierzu wird das Video in kurze Frames unterteilt und die Bildsequenzen als Datensatz verwendet. Man spricht von einem dynamischen FER (dynamic FER). Im nächsten Schritt erfolgt die eigentliche Detektion des Gesichts. Hierbei werden Informationen, die nicht zum Gesicht gehören, gefiltert und eliminiert. Anschließend werden die Gesichtsmerkmale wie Augen, Nase und Mund durch Punkte rekonstruiert. Zum Schluss werden diese Merkmale an einem Emotionsklassifikator angewendet, wie z. B. eine **Support Vector Machine** (SVM). SVM beschreibt eine Methode, um Daten zu analysieren und Objekte in bestimmten Klassen unterzuordnen. Ausgangsbasis für eine SVM ist die Trainingsphase. Die Objekte müssen demzufolge bereits Klassen zugeordnet sein. In der gesichtsbasierten Emotionserkennung bedeutet es, dass die Gesichtsmerkmale den

verschiedenen Emotionen zugeordnet werden. Hierbei werden in den meisten Fällen die Basisemotionen nach Paul Ekman verwendet. Die Auswahl der bestimmten Emotionen variiert jedoch von System zu System. Manche Algorithmen sind unter anderem zusätzlich auf einen neutralen Gesichtsausdruck ausgelegt. [40] [41]

2.) FER Systeme basierend auf Deep Learning

Emotionserkennung basierend auf Deep Learning ist eine eher neue Technik. [40] Im Gegensatz zu der traditionellen Methode werden die Ansätze eines Convolutional Neural Network herangezogen. Aus dem Eingabebild wird das Gesichtsbild zunächst extrahiert. Anschließend erfolgt der Einsatz eines CNN. Die Daten werden nach und nach den einzelnen Schichten unterzogen, sodass aus dem Gesichtsbild Merkmale extrahiert und die Daten danach reduziert werden. Zum Schluss wird das Gesichtsbild im Fully-Connected Layer einer Emotion zugeordnet. [40]

Aus einer Studie [40] wurden die Ansätze eines dynamischen FER basierend auf Deep Learning ausgewählt und umgesetzt. Das System wurde auf sechs Basisemotionen ausgelegt: Freude, Trauer, Wut, Ekel, Angst und Neutral. Das Modell diente dazu, aus einem Gesichtsausdruck eine der sechs Emotionen vorauszusagen. Hierzu wurde die Open-Source Bibliothek Dlib benutzt. Es handelt sich um ein Tool, das in C++ geschrieben ist und enthält Algorithmen für maschinelles Lernen, Bildverarbeitung und maschinelles Sehen. Zusätzlich enthält es eine Deep Learning Methode zur gesichtsbasierten Emotionserkennung. Für das resultierende Modell wurden die Daten zunächst aus einer Webcam entnommen. Im nächsten Schritt wurden mithilfe von Orientierungspunkten das Gesicht rekonstruiert und Merkmale extrahiert. Die Orientierungspunkte entsprechen hierbei den Action-Units vom FACS. Anders als das FACS enthält das menschliche Gesicht laut Dlib 68 Orientierungspunkte. Die Abbildung aus 3.5 zeigt ein Beispiel einer solchen Gesichtsrekonstruktion. [40]

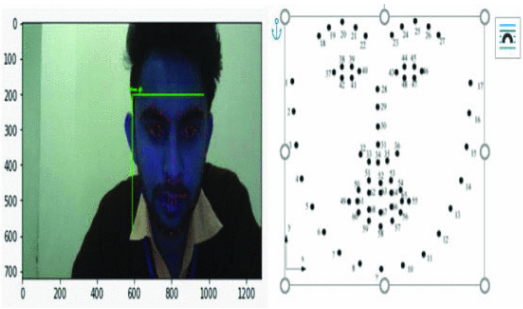


Abbildung 3.5: Orientierungspunkte [40]

Zum Schluss wurden die Merkmale an einem bereits trainierten CNN angewendet, um den Gesichtsausdruck einer Basisemotion zuzuordnen. Aus der Abbildung 3.6 kann der Prozess im Detail entnommen werden. Der Trainingsdatensatz bestand aus drei verschiedenen Datensätzen zur Gesichtserkennung: The labeled faces in the wild, the Yale Face database B und das google facial expression comparison dataset. Insgesamt fassten die Datensätze mehr als 100.000 Bilder zusammen. Das CNN Modell wurde jedoch nur

aus einem zuvor verarbeiteten Datensatz, bestehend aus mehr als 30.000 Bildern, antrainiert. Diese Daten wurden dann entsprechend einer Emotion zugeordnet. Das Modell konnte eine Genauigkeit von etwa 93 % erreichen. [40]

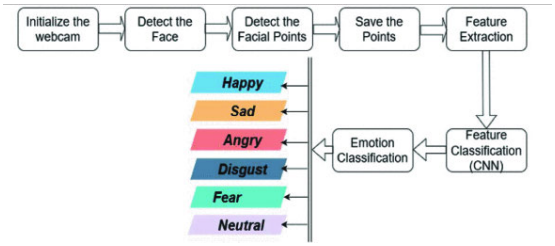


Abbildung 3.6: Prozess eines dynamischen FER mit einem CNN [40]

Die gesichtsbasierte Emotionserkennung wird nicht nur in zahlreichen Studien erforscht. Große Firmen wie Microsoft stellen ihre eigenen KI-Dienste bereit. Mit Microsoft Azure bietet die Firma Microsoft eine Gesichtserkennungssoftware an. Hierfür werden Datensätze aus Bildern entnommen. Neben der Emotionserkennung können mithilfe dieser Software das Alter und das Geschlecht einer Person erkannt werden. Die analysierenden Informationen werden aus dem Gesichtsausdruck entnommen und werden dann einer Emotion zugeordnet. Microsoft Azure ist auf die Emotionen Wut, Verachtung, Ekel, Angst, Freude, Trauer, Überraschung und Neutral ausgelegt. [42]

Auch mit DeepFace ist eine Gesichtserkennung möglich. Die von Facebook entwickelte Software basiert auf den Ansätzen der KI, genauer gesagt des Deep Learning.
Mit DeepFace ist es möglich menschliche Gesichter auf Fotos oder Videoaufzeichnungen zu erkennen und entsprechend einer Emotion zuzuordnen. Laut Facebook erreicht die Software eine Genauigkeit von mehr als 97 %. Mittlerweile ist DeepFace in Python als Bibliothek verfügbar. Mit der Python Bibliothek ist es möglich, zusätzlich zu der Emotion, das Alter, das Geschlecht und die Herkunft einer Person zu bestimmen. [43]

3.3.2 Spracherkennung

Die Sprache ist für den Menschen die natürlichste Form der Kommunikation. [44] Die entsprechenden Emotionen, die dabei vermittelt werden, können auf zwei verschiedene Arten erkannt werden. [45] Man unterscheidet hierbei zwischen der expliziten und der impliziten Analyse der Sprache. Für ein umfangreiches Verständnis werden die zwei Arten der Analyse auf die passenden Unterkapitel zugeordnet. In diesem Kapitel wird die implizite Analyse näher beleuchtet, während die explizite Analyse in der Texterkennung thematisiert wird. Die implizite Analyse der Sprache beschäftigt sich mit dem „wie“. Der Inhalt des Gesprochenen wird außen vor gelassen und vielmehr wird analysiert, wie der Inhalt gesprochen wird. Demnach kann bspw. über die Lautstärke oder die Tonlage eine bestimmte Emotion identifiziert werden. Die Umsetzung einer sprachbasierten Anwendung erfolgt üblicherweise in drei Schritten. Zu Beginn werden die eingelesenen

Sprachsignale vorverarbeitet und basierend darauf werden die Merkmale zur Emotionserkennung extrahiert. Zuletzt werden die verarbeiteten Daten an einem Klassifikator angewendet, um die Emotionen zu bestimmen. Mittlerweile wurden die Ansätze des Deep Learning hinzugezogen, da diese den Vorteil bieten, dass die Erkennung effizienter ist. Demnach werden die extrahierten Merkmale in ein neuronales Netz weitergegeben, um die Emotionen zu bestimmen. [44] [46]

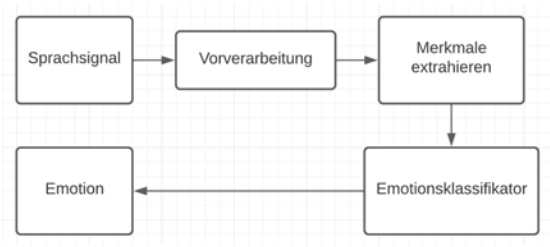


Abbildung 3.7: Traditionelle sprachbasierte Emotionserkennung modifiziert nach [44]

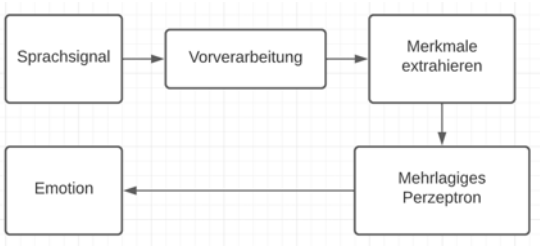


Abbildung 3.8: Sprachbasierte Emotionserkennung mit MLP modifiziert nach [44]

Emotionen durch die Eingabe eines Sprachsignals zu erkennen ist kein einfach zu lösendes Problem. Für die Verarbeitung der eingegebenen Informationen gibt es verschiedene Ansätze, wobei das Grundprinzip jedes Mal dasselbe ist. Datensätze werden vorverarbeitet, dann einer Machine Learning Technik unterzogen und anschließend einer Emotion zugeordnet. Die Genauigkeit des resultierenden Modells hängt von der Wahl der Vorverarbeitung und der Wahl des neuronalen Netzes ab. Je nach Auswahl der entsprechenden Komponenten kann die Genauigkeit unterschiedlich ausfallen. [44]
In einer wissenschaftlichen Arbeit [44] wurden hierzu drei unterschiedliche Prozesse vorgestellt. Bei dem ersten Prozess werden Sprachsignale unverändert in ein neuronales Netz gegeben. Der zweite Prozess beschreibt die Verarbeitung des Sprachsignals in ein **Spektrogramm** und anschließend die Eingabe in ein neuronales Netz. Bei einem Spektrogramm handelt es sich um die bildliche Darstellung eines Frequenzspektrums eines Signals. [47] In dem dritten Prozess ist das besondere Merkmal die direkte Signalverarbeitung. Die eingebenden Signale werden verarbeitet und anschließend in ein neuronales Netz gegeben. Mithilfe von Techniken aus der Audioverarbeitung sollen überflüssige Informationen eliminiert werden. **Mel-Frequenz-Cepstrum-Koeffizienten** (MFCC) beschreibt ein solches Verfahren und wurde ursprünglich zur automatischen Spracherkennung verwendet. Das Ergebnis eines MFCC führt zu einer kompakten Darstellung des Frequenzspektrums. Das entsprechende Resultat wird, wie zuvor erwähnt, zuletzt in ein neuronales Netz gegeben. [48] Bei der Auswahl des neuronalen Netzes ist man sich in der Regel einig, dass im Rahmen der Spracherkennung die Nutzung eines CNN

sehr vorteilhaft ist. Hierzu haben Forschungen bereits positive Ergebnisse gezeigt. [44] In einer Forschungsarbeit [49] zu Emotionserkennungssystemen wurden die Ansätze der letzten beiden Prozesse näher thematisiert und das Ergebnis vorgestellt. Im Rahmen dieser Arbeit wurden verschiedene Modelle aus verschiedenen Ansätzen entwickelt, um durch Sprach- oder Textsignale eine bestimmte Emotion zu erkennen. Zu Beginn des Experiments wurden die Komponenten festgelegt. Bei dem neuronalen Netz entschied man sich für ein CNN. Zusätzlich wurden Spektrogramme und MFCC genutzt, um nur die benötigten Sprachmerkmale zu extrahieren. Um dies umsetzen zu können, wurde das Pythonpaket **librosa** angewendet. Librosa ist ein Paket, um Audioeingaben zu analysieren und bietet Möglichkeiten an, gewünschte Sprachmerkmale daraus zu extrahieren. [49] [50] In der folgenden Tabelle 3.1 sind die Ergebnisse der einzelnen Modelle veranschaulicht.

Tabelle 3.1: Genauigkeitstabelle modifiziert nach [49]

Methode	Eingabe	Gesamtgenauigkeit	Klassifizierungsgenauigkeit
Modell 1	Text	64.4 %	47.9 %
Modell 2A	Spektrogramm	71.2 %	61.9 %
Modell 2B	Spektrogramm	71.3 %	61.6 %
Modell 3	MFCC	71.6 %	59.9 %
Modell 4A	Spektrogramm & MFCC	73.6 %	62.9 %
Modell 4B	Text & Spektrogramm	75.1 %	69.5 %
Modell 4C	Text & MFCC	76.1 %	69.5 %

Die resultierenden Systeme ordnen sich in drei Kategorien unter. Systeme der ersten Kategorie wurden basierend auf Texteingaben umgesetzt. Hierbei gilt es zu erwähnen, dass es sich bei den Texteingaben um transkribierte Audioaufnahmen handelt. Systeme aus der zweiten Kategorie basieren auf den Eingaben von Sprachmerkmalen. Das bedeutet, dass rohe Audiosignale entweder durch ein Spektrogramm oder ein MFCC, umgewandelt bzw. verarbeitet werden. In der letzten Kategorie sind Systeme enthalten, die verschiedene Eingaben miteinander kombinieren. Die Eingabe erfolgt dementsprechend aus Spektrogramm und MFCC, aus Text und Spektrogramm oder aus Text und MFCC. [49]
Für die abschließende Evaluation wurden die verarbeiteten Audiosignale an einem bereits trainierten CNN angewendet. Der Trainingsdatensatz, der hierfür verwendet wurde, war das Interactive Emotional Motion Capture. [51] Der Datensatz enthält Hunderte Videos über eine Konversation zwischen zwei Personen. Dabei wurden über improvisierte und gestellte Szenen neun verschiedene Emotionen dargestellt. Die Modelle aus der Forschungsarbeit sind jedoch nur auf vier von diesen Emotionen ausgelegt: Freude, Trauer, Wut und Neutral. Nach der Evaluation kam als Ergebnis hervor, dass die Nutzung mehrerer Eingabe eine höhere Genauigkeit erzielt. Demnach kann man aus der Tabelle 3.1 entnehmen, dass das Modell 4B durch die Eingabe aus Text und MFCC eine Genauigkeit von 76,1 % erreichen konnte. Ein Modell, welches nur einen Eingabestrom

berücksichtigt hat, wie das Modell 1 konnte nur eine Gesamtgenauigkeit von 64,4 % erzielen. [49]

Mit den Ansätzen des Machine Learning hat sich eine zentrale Komponente in der Spracherkennung etabliert. Heutige Systeme arbeiten ausschließlich mit diesen Methoden. Doch auch wenn es mittlerweile zahlreiche Optionen gibt, Emotionen über die Sprache zu erkennen, birgt diese Technologie noch viele Schwierigkeiten. Besonders anfällig ist diese Technologie bei Hintergrundgeräuschen. Zusätzlich zu den Hintergrundgeräuschen ist die sprachbasierte Emotionserkennung abhängig von der Sprache. Die Sprechweise und die Betonung einer Sprache sind von Kultur zu Kultur unterschiedlich, sodass für jede vorkommende Sprache und dessen Dialekt ein eigenes Modell entwickelt bzw. trainiert werden muss. Durch die verschiedenen Charakteristiken in einer Sprache ist diese Technologie nicht universell einsetzbar. [52] [53]

Viele Unternehmen, darunter auch Start-ups, beschäftigen sich mit dieser Problematik. Mit openSMILE bietet das Start-up-Unternehmen audEERING ein Open-Source-Toolkit an, um solche Charakteristiken zu extrahieren. OpenSMILE ist ein in C++ geschriebenes Programm für die Extraktion von Audiomerkmalen und die Klassifizierung von Sprach- und Musiksignalen. Die Software ist in der Lage, neben Charakteristiken wie den Emotionen auch das Alter und das Geschlecht zu erkennen. Inzwischen wurden mit openSMILE viele Produkte im Rahmen der Affective Computing Forschung entwickelt. Es wurden hierzu unterschiedliche Systeme entworfen, um aus Audio, Text oder aus einer Kombination von beiden Emotionen zu erkennen. Die Software kann bis zu vier verschiedene Emotionen unterscheiden: Freude, Wut, Trauer und Neutral. Für den akademischen Zweck bietet audEERING eine einfach zu bedienende Python-API an. [54]

Abbildung 3.9: Sprachbasierte Emotionserkennung von audEERING [54]

Speech Emotion Analyzer ist ein weiteres Open-Source-Projekt zur Erkennung von Emotionen. Aus Audiosignalen kann die Software unterschiedliche Emotionen erkennen wie Freude, Wut, Ruhe, Angst, Trauer, Ekel, Überraschung und Neutral. Zusätzlich wurden die Ansätze des Geschlechts hinzugezogen, sodass das Programm erkennen kann, ob die Person männlich oder weiblich ist. Das Projekt wurde in Python umgesetzt und basiert auf den Ansätzen des Machine Learning. [55]

3.3.3 Texterkennung

Die explizite Analyse der Sprache untersucht den Inhalt des Gesprochenen. Demnach können über negative Wörter wie schlecht oder aggressiv eine negative Basisemotion wie Trauer oder Wut identifiziert werden. [45]
Natural Language Processing (NLP) beschäftigt sich mit der maschinellen Verarbeitung der natürlichen Sprache. Hierzu soll die natürliche Sprache erfasst werden und mithilfe von Regeln und Algorithmen computerbasierend verarbeitet werden. Um solche Lösungen zu schaffen, muss das NLP in der Lage sein, nicht nur einzelne Wörter und Sätze zu verstehen, sondern vielmehr komplette Textzusammenhänge und Sachverhalte deuten können. Die Komplexität der menschlichen Sprache und ihrer Mehrdeutigkeit bereitet hierbei viele Schwierigkeiten. Mithilfe von Algorithmen und Verfahren der künstlichen Intelligenz und des Machine Learning sollen diese Probleme überwindet werden. [56]
Sentiment Analysis (SA) beschreibt genau jenen Bereich und beschäftigt sich mit der Gewinnung bedeutungsvoller Muster aus Textdaten. Durch den Einsatz von Textanalyseverfahren können aus Textdaten Emotionen interpretiert und klassifiziert werden. Machine Learning spielt dabei eine wichtige Rolle, um Texte ganz einheitlich erkennen zu können. Für die Analyse ist es somit notwendig, im Vorfeld große Datenmenge zu erfassen und bereits erkannte Muster heranzuziehen. Das Potenzial hierzu ist gewaltig. Besonders geeignet ist die SA im Social Media Bereich oder für die Analyse von Rezensionen. [57]
Auch im Bereich der Texterkennung gab es zwischenzeitlich einen Wandel. Mit der Zeit hat sich die textbasierte Emotionserkennung von der groben Erkennung in die feine Erkennung entwickelt. Bei der groben Emotionserkennung werden Textsignale in zwei verschiedene Emotionen eingeordnet: Positiv und negativ. Die feine Emotionserkennung hingegen ist in der Lage, Textsignale in viele verschiedene Emotionen zu klassifizieren. In der Regel werden die Signale einer bestimmten Auswahl von Basisemotionen zugeordnet. Für die textbasierte Emotionserkennung gibt es genauso wie in der Bild- und Spracherkennung bereits viele verschiedene Ansätze. [58] [59] In der Abbildung 3.10 sind die verschiedenen Methoden veranschaulicht.

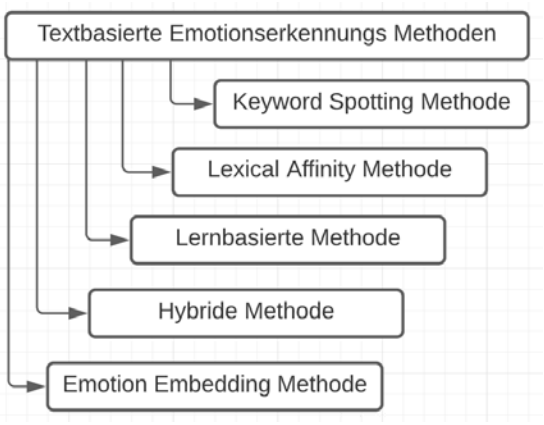


Abbildung 3.10: Verschiedene Methoden zur textbasierten Emotionserkennung. Abbildung modifiziert nach [58]

Die **Keyword-Spotting Methode** ist die am häufigsten genutzte Methode in der textbasierten Emotionserkennung. Aufgrund ihrer leichten Umsetzbarkeit ist diese Methode weitverbreitet. Zur Vorhersage einer Emotion werden zunächst Wörter herausgesucht, die sich direkt auf eine Emotion beziehen, bspw. Freude. Mit einer Liste von den entsprechenden Hinweiswörtern und Regeln werden die Emotionen in einem Satz bestimmt. [58] [59]

Die **Lexical-Affinity Methode** ist eine leichte Erweiterung der Keyword-Spotting Methode. Bei dieser Methode werden nicht nur Wörter betrachtet, die sich direkt auf eine Emotion beziehen, sondern auch Wörter, die einen emotionalen Gehalt bieten. In diesem Zusammenhang wird als Beispiel das Wort „Unfall“ in der Regel in einem negativen Kontext betrachtet, sodass durch die Lexical-Affinity Methode der Satz einer negativen Emotion zugeordnet wird. [58] [59]

Die **lernbasierte Methode** zieht die Ansätze des Machine Learning hinzu und basiert darauf, Emotionen anhand eines vortrainierten Modells zu bestimmen. Für die Klassifikation von Emotionen werden Machine-Learning-Methoden, wie das SVM, genutzt. Um Emotionen innerhalb eines Textes zu erkennen, werden die Ansätze des Deep Learning hinzugezogen. [58] [59]

Die **hybride Methode** kombiniert, einfach gesagt, die Ansätze des Keyword-Spotting und die der lernbasierten Methode. [58]

In einer Forschungsarbeit [59] zur textbasierten Emotionserkennung wurde eine neue Methode zur Klassifizierung von Emotionen vorgestellt: Die **Word Embedding Methode**. Die Word Embedding Methode beschreibt den Grundsatz der modernen Texterkennung. Hierbei werden die zu analysierenden Wörter durch Vektoren ersetzt. Die Zusammenhänge der einzelnen Wörter werden durch die Abstände der Vektoren dargestellt. Das bedeutet, dass Wörter, die eine ähnliche Bedeutung haben, näher zueinander stehen als Wörter, die komplett verschieden sind. Auf diese Weise werden die Textsignale klassifiziert. [59] In der folgenden Abbildung 3.11 ist der Prozess eines Word Embedding Modells visualisiert.

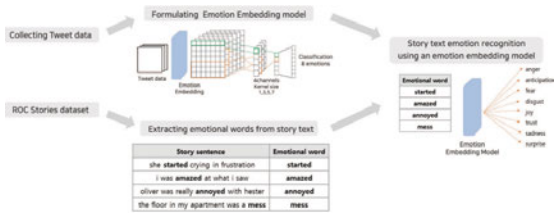


Abbildung 3.11: Prozess des Emotion Embedding Modells [59]

In dieser Forschungsarbeit [59] wurde das Word Embedding Modell am Beispiel einer textbasierten Emotionserkennung umgesetzt. Für die Implementierung dieses Systems wurden zunächst Daten aus Twitter gesammelt. Hierzu wurden emotionale Tweets mit sog. „Emotionshashtags“ gesucht. „Ich habe mit meinem Freund Schluss gemacht. #traurig“ beschreibt ein Beispiel eines solchen Tweets. Es wurden über 140.000 Twitterdaten

gesammelt. Die „Emotionshashtags“ sowie das resultierende System wurden auf die acht Basisemotionen von Plutchiks Rad der Emotionen ausgelegt. Im nächsten Schritt erfolgte die Umsetzung des Emotion Embedding Modells. Um die Emotionen zu klassifizieren, wurde ein CNN verwendet. Das Modell wurde mit den entsprechenden Twitterdaten antrainiert. Während der Lernphase wird eine Embedding-Schicht geschaffen, um die zu analysierenden Texte zu vektorisieren. Mit dieser Schicht sollen die Emotionen erkannt und schließlich klassifiziert werden. Zusammenfassend beutet es, dass die gesammelten Twitterdaten als Eingabe in das CNN gegeben werden. Die Daten werden innerhalb dieses neuronalen Netzes verarbeitet und schließlich einer der acht Basisemotionen zugeordnet. Im dritten Schritt wurden aus Textgeschichten emotionale Wörter extrahiert. Hierfür wurde der Datensatz ROCStories, bestehend aus verschiedenen Geschichten, benutzt. Mithilfe des NLTK Vader Sentiment Analyzer wurden die emotionalen Wörter aus den Textgeschichten erkannt und bewertet. Hierzu werden als Ergebnis vier Werte zurückgegeben: positiv, negativ, neutral und die Gesamtsumme der drei Werte. Der Wert der Gesamtsumme gibt hierbei den Grad der Emotion an und kann einen Wert zwischen -1 und +1 annehmen. Das bedeutet, dass -1 die höchste Negativität und +1 die höchste Positivität darstellt. Entscheidend dabei ist, dass nur der höchste absolute Wert der Gesamtsumme genutzt wird, somit wird auch nur das emotionalste Wort in einem Satz gesucht. In der Abbildung 3.12 ist ein Beispiel veranschaulicht. [59]

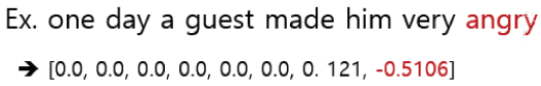


Abbildung 3.12: Beispiel zum Auserwählen eines emotionalen Wortes mit der höchsten Polarität [59]

Als letzten Schritt wurden die emotionalen Wörter anhand des Emotion Embedding Model einer Basisemotion zugeordnet. Um dies gewährleisten zu können, werden ähnliche Wörter im Emotion Embedding Modell gesucht, sodass dem Wort „grief“ bspw. die Emotion Trauer zugeordnet wird. Im Rahmen dieser Arbeit wurden mehr als 130.000 Sätze analysiert. Für die abschließende Evaluation wurden zum Schluss 120 Sätze aus dem Datensatz auserwählt, die von Studierenden einer Emotion zugeordnet werden sollen. Als Ergebnis kam hervor, dass mit einer Wahrscheinlichkeit von etwa 73,3 % die Emotion Freude am genausten vorhergesagt werden konnte. Während Wut mit einer Genauigkeit von etwa 36,7 % am ungenausten abgeschnitten hat. [59]

Genau schon wie die Bild- und Spracherkennung ist die textbasierte Emotionserkennung noch nicht ganz ausgereift. Die Technologie ist durch die Wahl von nur einem emotionalen Wort nicht in der Lage, emotionale Textzusammenhänge zu erkennen und zu klassifizieren. Zusätzlich werden Verneinungen nicht beachtet. Das bedeutet, dass durch den Satz „Ich bin keineswegs wütend auf dich“ das Wort „wütend“ trotz Verneinung als emotionales Wort festgelegt und somit die Emotion Wut erkannt wird. Zuletzt stellt die Sprache eine weitere Herausforderung dar. Das Modell ist nur in der Lage, Wörter zu erkennen, die dieselbe Sprache haben wie der trainierte Datensatz. [58] [59]

Große Firmen wie IBM Watson beschäftigen sich mit dieser Problematik und bieten ihren eigenen Dienst an. Mit Watson Natural Language Understanding von IBM Watson sollen die Emotionen einer Person durch Texte erkannt werden. Der Dienst von IBM Watson bietet nicht nur die Emotionserkennung an. Zusätzlich können bspw. Stimmungen, Konzepte oder Schlüsselwörter extrahiert werden. Watson Natural Language Understanding umfasst ein Sprachpaket von über 20 Sprachen und ist dabei stetig zu wachsen. Für die Programmiersprache Python gibt es inzwischen einen einfach gehaltenen API-Service. [60]

4 Konzeption

In diesem Abschnitt wird ein Konzept für eine prototypische Emotionserkennung entwickelt. Das Wissen aus dem Bereich Affective Computing dient als Grundlage und soll dabei helfen, eine Idee der Umsetzung zu kristallisieren.
Bis zu diesem Zeitpunkt ist die Größe dieses Forschungsbereiches klar und deutlich geworden. Aufgrund der Komplexität der Thematik ist es daher sinnvoll, dass dieses Kapitel in Abschnitte unterteilt wird. Zu Beginn wird die Idee einer Affective Computing Anwendung konkretisiert. Es werden hierzu die Faktoren definiert, die in eine solche Anwendung einfließen sollten. Im Anschluss wird das resultierende Konzept vorgestellt, welches auf den zuvor genannten Anforderungen basiert.

4.1 Abgrenzung

Um ein Konzept für ein komplexes Thema dieser Größe planen und ausarbeiten zu können, müssen die entsprechenden Komponenten festgelegt werden. Aus dem vorherigen Kapitel wurde deutlich, dass eine Anwendung zur Emotionserkennung viele Umsetzungsmöglichkeiten besitzt. Die Menge an verschiedenen Machine Learning Ansätzen führt zu vielen verschiedenen Ergebnissen. Um den Stand der Technik jedoch gerecht zu werden, müssten die Ansätze des Deep Learning hinzugezogen werden. Dementsprechend müsste zunächst ein neuronales Netz aufgesetzt und mit vielen Daten antrainiert werden. Für die ideale Anwendung müssten hier bereits die Anforderung einer solchen Anwendung definiert werden. Das bedeutet, dass beim Trainieren des neuronalen Netzes bereits feststehen sollte, unter welchen Umständen die Emotionserkennung funktionieren soll. Faktoren wie Lichtverhältnisse, unterstützte Sprachen oder Hintergrundgeräusche müssten hierzu alle in das Konzept einfließen. Ergänzend dazu sollte das neuronale Netz mit den eigenen Daten trainiert werden, um das bestmögliche Ergebnis zu erzielen. Zusätzlich ist es sinnvoll zu wissen, welche Eingabeströme einfließen sollen. Zuletzt kann die Programmiersprache passend ausgewählt werden, je nachdem, welches affektive System gewünscht ist bzw. besser in das eigene Konzept passt. Jedes dieser Faktoren hat seine Daseinsberechtigung und sollte im optimalen Fall in ein solches Konzept einfließen. Für die Thesis werden einige Faktoren nicht weiter in Betracht gezogen, da sich diese Thesis auf eine einfache Anwendung stützt und dies den Rahmen der Arbeit übersteigen würde. Folglich werden die Faktoren Lichtverhältnisse, mehrere unterstütze Sprachen, Hintergrundgeräusche, eigene Datensätze, textbasierte

Emotionserkennung, passende Programmiersprache, Alter und Geschlecht nicht weiter berücksichtigt.

4.2 Entwurf

Aus den genannten Abgrenzungen wird schließlich das eigene Konzept entworfen.

4.2.1 Komponenten des neuronalen Netzes

Zu Beginn werden die Komponenten des neuronalen Netzes definiert. Die Wahl des neuronalen Netzes und der Daten kann entscheidend für das resultierende Ergebnis sein. Für die Anwendung werden bereits existierende Datensätze verwendet. Unter anderem sollen Bild- oder Audiodateien genutzt werden, die abhängig von dem Gesichtsausdruck oder der Aussprache eine Emotion abbilden. Ein Beispiel solcher Daten enthält der bereits erwähnte Datensatz IEMOCAP. [51] Für die Klassifikation von Emotionen bedeutet es, dass die Anzahl an Basisemotionen abhängig von den Datensätzen ist. Demnach stehen nur die Basisemotionen zur Auswahl, die im Datensatz vertreten worden sind. Für das eigene Konzept werden aus dem Emotionsset Ruhe, Freude, Trauer, Wut, Angst, Ekel, Überraschung und Neutral nur die Emotionen Wut, Trauer, Freude und Neutral verwendet. Diese sind jedoch frei wählbar.
Für eine möglichst schnelle und benutzerfreundliche Umsetzung soll hierfür ein sog. feedforward neuronales Netzwerk mit einem Backpropagations-Algorithmus verwendet werden. Das Modell gehört zu den meist verwendeten Arten eines KNN und wird oftmals für Klassifizierungsaufgaben eingesetzt. [61]
Der **Backpropagation-Algorithmus** ist ein Algorithmus, der zur Gruppe des überwachten Lernens gehört. Innerhalb dieser Gruppe werden Algorithmen beschrieben, die für den Lernprozess Datenpaare, bestehend aus Eingabe und Ausgabe, benötigen. Mithilfe dieser Datenpaare soll ein bekanntes Problem gelöst und dem Netz die Fähigkeit antrainiert werden, Assoziationen zu erstellen. Konkret bedeutet es, dass bei der Emotionserkennung bestimmte Merkmale bestimmten Emotionen zugeordnet sind. Demnach hängen ein offener Mund und hochgezogene Augenbrauen mit der Emotion Überraschung zusammen. Über diese Datenpaare soll das Netzwerk die Fähigkeit erlangen, Daten, die eine ähnliche Struktur zu den gelernten Datensätzen besitzen, eigenständig zu klassifizieren. Der Backpropagation-Algorithmus ist die weitverbreitetste Lernmethode für neuronale Netze und hilft dabei, die Fehlerfunktion minimal zu halten. [61] Dieser Lernalgorithmus erfolgt in drei Phasen.

1. **Forward-Pass**: Zu Beginn wird das neuronale Netz mit zufälligen Werten initialisiert. Über diese Gewichtungen und den eingehenden Signalen wird die Ausgabe des neuronalen Netzes bestimmt. [61]

2. **Fehlerbestimmung**: In der zweiten Phase wird ein Fehler mithilfe einer Fehlerfunktion bestimmt. Hierfür werden die im Forward-Pass berechneten Ausgaben mit den erwarteten Ausgaben verglichen. Sollte ein Unterschied vorliegen, so wird dies als Fehler des neuronalen Netzes angesehen. [61]

3. **Backward-Pass**: Bei einem berechneten Fehler werden in der dritten Phase, ausgehend von der Ausgabeschicht hin bis zur Eingabeschicht der Fehler zurück propagiert. Hierbei werden die Gewichtungen entsprechend verändert, sodass bei erneuter Eingabe ein besseres Ergebnis resultieren kann. [61]

Für die Umsetzung eines solchen Algorithmus werden zwei bestimmte Datensätze benötigt: Ein Trainingsdatensatz und ein Testdatensatz. Der Trainingsdatensatz wird ausschließlich für das Erlernen des neuronalen Netzes verwendet. Nach abschließendem Training wird mit dem Testdatensatz überprüft, ob das neuronale Netz etwas gelernt hat. [61]

4.2.2 Anwendung

Für die vorliegende Thesis soll ein emotionssensitives System entwickelt werden, bei dem der emotionale Zustand eines Endbenutzers erkannt und anschließend angemessen darauf reagiert wird. Für eine möglichst genaue Emotionserkennung ist das Ziel, eine Grundlage für ein hybrides System zu schaffen. Das bedeutet, dass mehrere Eingabeströme in das Projekt einfließen. Nach abschließender Abgrenzung werden für die prototypische Implementierung nur die Eingabeströme Bild und Sprache berücksichtigt. Damit das Wissen aus der Texterkennung trotzdem einfließt, wird das affektive System auf ein Anwendungsfall beschränkt, der zumindest Methoden der Texterkennung nutzt. Basierend auf diesen Kriterien wird die Affective Computing Anwendung am Beispiel eines Sprachassistenten umgesetzt. Demzufolge soll ein Sprachassistent umgesetzt werden, der über eingehende Bildsignale oder Sprachsignale die Emotion des Benutzers bestimmt. Abhängig von der ausgegebenen Emotion soll das System anschließend Musik empfehlen, um auf diese Weise den Benutzer aufzuheitern bzw. seine Emotion Beeinflussen. Das System soll bei der Ausgabe von „Trauer“ eine fröhliche Musik empfehlen. Bei der Emotion „Wut“ z. B. soll das System eine Playlist empfehlen, die Rock-Musik beinhaltet. Die Wahl der Musik orientiert sich dabei an dem Buch „Macht der Musik“ [62]. Laut der Literatur ist aggressive Musik ein Medium, um Wut abzubauen. Für eine einfache Implementierung wird die Musik, die ausgegeben wird, aus der Webseite YouTube entnommen.

Für die Umsetzung einer sprachbasierten Emotionserkennung ist es möglich, die analysierenden Sprachsignale aufzunehmen und zu speichern. Anschließend werden die Signale verarbeitet und relevante Informationen für die Emotionserkennung extrahiert. Zum Schluss werden diese vorverarbeiteten Signale an einem neuronalen Netz angewendet. Anhand dessen werden die Sprachsignale einer Emotion zugeordnet. Im Zuge der gesichtsbasierten Emotionserkennung werden aus einer Kamera Live-Daten aus dem Gesicht entnommen. Über ein Framework wie DeepFace kann das Gesicht anhand von Gesichtsmerkmalen identifiziert und schließlich über ein neuronales Netz einer Emotion zugeordnet werden. Hierzu benötigen beide Modelle eine Menge an emotionsbezogenen Trainingsdaten, um ein Muster zu erkennen. Auf Grundlage dieser Möglichkeiten werden die Sprachbefehle zu den Emotionserkennungen definiert.

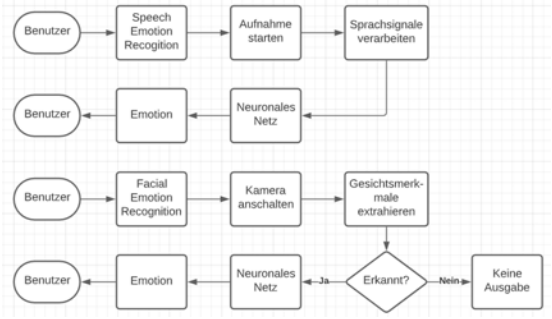


Abbildung 4.1: Flussdiagramm der Emotionserkennung

4.2.3 Sprachbefehle

Neben den zwei Emotionserkennungen soll der Assistent in der Lage sein, typische Aufgaben eines Sprachassistenten zu erledigen. In der folgenden Liste sind einige Sprachbefehle mit ihren Funktionen dargestellt.

- **Open [Name der Webseite]**: Über diesen Befehl wird die entsprechende Webseite im Browser aufgerufen. Das System unterstützt die Webseiten Wikipedia, Google, YouTube und Gmail.
- **Tell me the Time**: Über diesen Befehl wird die Zeit angesagt.
- **Tell me the Weather**: Über diesen Befehl wird das Wetter vorhergesagt. Über die API von OpenWeatherMap [63] werden wetterbezogene Daten zu einer Stadt einer Wahl angefragt und ausgegeben.
- **Tell me the News**: Über diesen Befehl werden die Top 5 neuesten News aus der Seite BBC ausgegeben. Die Daten kommen von der API NewsAPI [64].
- **Lights on/off**: Über diesen Befehl können Lampen von PhillipsHue ein- bzw. ausgeschaltet werden. Über die API von PhillipsHue [65] können diese Lampen angesteuert werden.

- **Open DeepFace**: Über diesen Befehl wird die gesichtsbasierte Emotionserkennung aufgerufen. Hierzu schaltet sich die Kamera an und anhand von Gesichtsausdrücken wird die Emotion des Benutzers erkannt und schließlich eine bestimmte Musikrichtung als Empfehlung zurückgegeben.
- **Open Emotion Recognition**: Über diesen Befehl wird die sprachbasierte Emotionserkennung aufgerufen. Hierzu wird eine Aufnahme gestartet, in die der Benutzer einsprechen kann. Über die enthaltenden Sprachmerkmale wird die Emotion erkannt und schließlich eine bestimmte Musikrichtung als Empfehlung zurückgegeben.

4.2.4 Resultat

Zusammenfassend zeigt die folgende Abbildung 4.2 das resultierende Konzept.

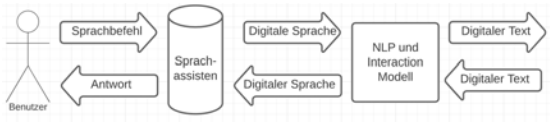


Abbildung 4.2: Prozess eines Sprachassistenten

Nach Starten der Anwendung soll der Benutzer zunächst begrüßt werden und anschließend gefragt werden, was der Assistent für einen tun kann. Daraufhin kann der Benutzer über die Sprachbefehle bestimmte Funktionen auslösen. Hierbei wird die menschliche Sprache über den Assistenten in digitale Sprache umgewandelt. Das System wandelt die digitale Sprache wiederum in digitalen Text um und wird dem Interaction Modell zugeordnet. Über dieses Modell wird die ausgewählte Interaktion angesprochen und ausgeführt. Zuletzt wird der digitale Text zurück in die digitale Sprache umgewandelt und über den Lautsprecher eine Antwort zurückgegeben. Nach diesem Prinzip ist der Benutzer in der Lage, bewusst zu entscheiden, ob seine Emotionen erkannt werden soll oder nicht.

Mit diesem Konzept ist es möglich, ein affektives System zu implementieren, welches drei unterschiedliche Eingabeströme verarbeitet und eine Grundlage bietet, um eine hybride Emotionserkennung umzusetzen.

5 Implementierung

Nachdem im vorherigen Kapitel ein Konzept für eine Affective Computing Anwendung entworfen wurde, wird in diesem Abschnitt die Umsetzung dieses Konzepts vorgestellt. Beginnend mit der Wahl der Werkzeuge wird die Programmiersprache festgelegt sowie verwendete Bibliotheken vorgestellt. Anschließend wird die Struktur der Anwendung thematisiert. Hierbei werden die Bereiche vorgestellt, die für die prototypische Umsetzung im Fokus stehen. Die Struktur unterteilt sich in drei Teile: Den Sprachassistenten, die gesichtsbasierte Emotionserkennung und die sprachbasierte Emotionserkennung. Zuletzt werden auf aufgetretene Probleme eingegangen.

5.1 Toolselection

Für die Implementierung wurden nur anfängerfreundliche und nicht kommerzielle Möglichkeiten in Betracht gezogen. Basierend auf dieser Voraussetzung und der vorherigen Abgrenzung wurde letztendlich die Wahl der Werkzeuge getätigt.

5.1.1 Programmiersprache

Die gesamte Anwendung wurde in der Programmiersprache Python umgesetzt. Python verfolgt den Anspruch, gut lesbar zu sein und wird oftmals im Zusammenhang der KI-Forschung angewendet. [66] Durch die große Auswahl an bestehender und weiterführender Bibliotheken im Bereich Machine Learning eignet sich Python für die Implementierung einer simplen Anwendung. Die einfache Bedienung ist hierbei sehr hilfreich. Die Anwendung wurde in der Entwicklungsumgebung PyCharm umgesetzt.

5.1.2 Bibliotheken

Für die Implementierung der Anwendung wurden Bibliotheken für einen „normalen“ Sprachassistenten, für die sprachbasierte Emotionserkennung sowie für die gesichtsbasierte Emotionserkennung genutzt. Hieraus ergeben sich etwas mehr als 20 unterschiedliche Bibliotheken. Für die Relevanz der Arbeit werden nur die Bibliotheken weiter thematisiert, die auch im Fokus stehen.

- speech_recognition
- pyttsx3
- deepface
- OpenCV
- sklearn
- librosa

Beginnend mit dem Sprachassistenten sind die Pakete „Speech recognition“ [67] und „pyttsx3“ [68] unvermeidbar. Speech recognition ist eine Bibliothek zur automatischen Spracherkennung. Die Hauptfunktion dieser Bibliothek besteht darin, die menschliche Sprache in Textsignale zu konvertieren. Pyttsx3 ist das entsprechende Gegenstück zu der Bibliothek und dient dazu, Textsignale in Sprachsignale umzuwandeln. Diese Bibliotheken bilden den typischen Ablauf einer Sprachanfrage.
Für die Umsetzung der gesichtsbasierten Emotionserkennung wurden die Bibliotheken „DeepFace“ [69] und „OpenCV“ [70] verwendet. OpenCV ist ein Framework, um Probleme innerhalb des Bereichs Computer Vision und Machine Learning zu lösen. Die Bibliothek bietet eine Vielzahl an Algorithmen an, mit denen es möglich ist, Gesichter in Echtzeit zu erkennen. DeepFace ist ein Open-Souce-Framework, um Gesichtsmerkmale zu klassifizieren. DeepFace ist in der Lage, anhand von Gesichtsmerkmalen das Alter, das Geschlecht, die Emotionen oder die Herkunft zu bestimmen. Für die vorliegende Thesis wurden nur die Emotionen berücksichtigt.
Die Implementierung der sprachbasierten Emotionserkennung ist der aufwendigste Teil dieser Anwendung. Für diesen Bereich stehen die Bibliotheken „librosa“ [50] und „Scikit-learn“ [71] im Vordergrund. Librosa ist die bereits erwähnte Bibliothek, um bestimmte Sprachmerkmale zu extrahieren. Scikit-Learn ist eine Bibliothek mit umfangreichen Methoden zu dem Bereich Machine Learning. Neben Regressionsalgorithmen stellt diese Bibliothek verschiedene Klassifikationsmöglichkeiten zur Verfügung, die zu dem gewählten Konzept passen. Unter anderem bietet Scikit-Learn eine Methode an, die die Ansätze des Backpropagation-Algorithmus nutzt.

5.2 Struktur der Anwendung

Die Anwendung besteht aus mehreren Dateien und setzt sich im Grunde aus sechs Python Dateien und einer Textdatei zusammen. Im folgenden Diagramm 5.1 werden die Verhältnisse der Dateien bildlich dargestellt. Für ein näheres Verständnis werden im nächsten Abschnitt die Dateien vorgestellt.

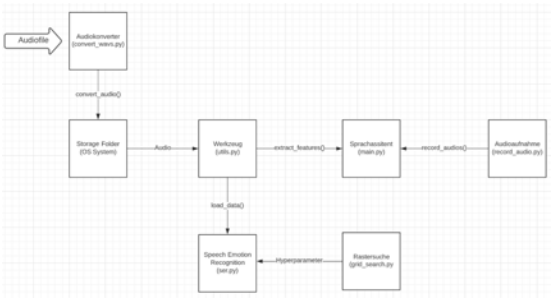


Abbildung 5.1: Blockdiagramm zur Darstellung der Verhältnisse

Um den Programmablauf möglichst genau nachzustellen, wurde das Flussdiagramm 5.2 als visueller Leitfaden erstellt.

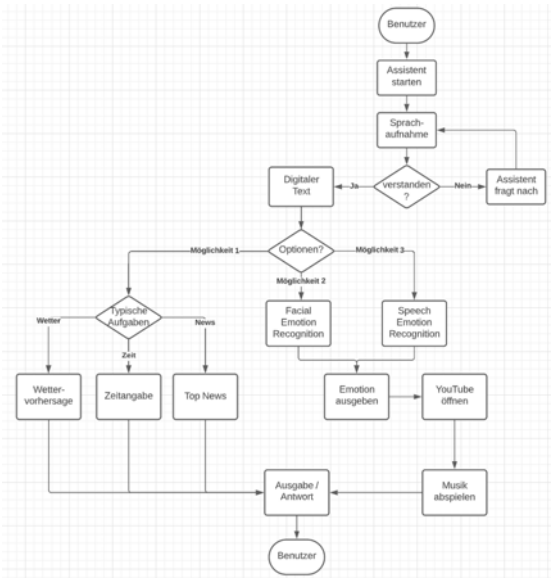


Abbildung 5.2: Flussdiagramm des Sprachassistenten

5.2.1 Sprachassistent

Beginnend mit der „main.py“ wird in diesem Programm der Sprachassistent erzeugt. Diese Datei ist dafür zuständig, die Sprachbefehle des Benutzers entgegen zu nehmen, dem Interaction Modell zuzuordnen und anschließend diese auszuführen. Um die Interaktionen ausführen zu können, müssen zunächst die Prozesse der automatischen Spracherkennung umgesetzt werden.
Entsprechend dafür wurde die Funktion **command_me** erstellt. Über das Mikrofon werden die Stimmen des Benutzers aufgenommen und anschließend in Textsignale verarbeitet. Für die Verarbeitung der Sprachsignale in Textsignale wurde die Speech-To-Text API von Google benutzt, da diese API es ermöglicht, einfach und in Echtzeit Audioeingaben in Textsignale umzuwandeln. In der Abbildung 5.3 kann die Funktion im Detail entnommen werden. Die Erkennung der Stimme wird innerhalb eines Exceptions Blocks umgesetzt, um entsprechende Fehler zu behandeln. Im Falle eines Fehlers wird

```python
def command_me():
    listener = sr.Recognizer()
    with sr.Microphone() as source:
        print("Listening to...")
        listener.adjust_for_ambient_noise(source, duration=1)
        voice = listener.listen(source)

        try:
            voice_command = listener.recognize_google(voice, language='en-us')
            print(f"User said: {voice_command}\n")

        except Exception:
            speak("Excuse me, could you please repeat yourself")
            return "None"
        return voice_command
```

Abbildung 5.3: Speech-to-Text

der Benutzer aufgefordert, seinen Sprachbefehl zu wiederholen. Um die entsprechenden Textsignale anschließend wieder in Sprachsignale umzuwandeln, wurden die Methoden „init“ und „say“ aus der Bibliothek „pyttsx3“ definiert und innerhalb einer Funktion umgesetzt.

Weitergehend im Flussdiagramm wird dargestellt, wie die Hauptfunktion, das Interaction Modell, implementiert wurde. Hierzu wurde eine While-Schleife aufgesetzt, die die Textsignale zunächst in die Variable „voice_command“ speichert. Innerhalb dieser While-Schleife wurde für jede Interaktionsmöglichkeit eine If-Anweisung implementiert. Um die entsprechenden Interaktionen anzusprechen und auszulösen, wurden hierzu Schlüsselwörter definiert, die innerhalb der Variable „voice_command“ auftreten müssen. Diese werden in der Abbildung 5.4 veranschaulicht.

```python
if __name__ == '__main__':

    speak("Is there anything i can do for you?")
    while True:
        voice_command = command_me().lower()

        # Open Browser
        if "open wikipedia" in voice_command:
            webbrowser.open_new_tab("https://www.wikipedia.de")
            speak("Wikipedia is open now.")
            time.sleep(4)
```

Abbildung 5.4: Hauptfunktion

Am Beispiel einer Browseranfrage wurde das Schlüsselwort „open Wikipedia“ definiert. Wenn vom Benutzer das Schlüsselwort entnommen wird, springt die Anwendung in diese If-Bedingung und löst die Anweisung innerhalb des Blocks aus. Als Ergebnis öffnet der Assistent die Webseite Wikipedia auf. Über diese einfache Implementierung ist es möglich, alle definierten Sprachbefehle abzubilden.

5.2.2 Gesichtsbasierte Emotionserkennung

Für die gesichtsbasierte Emotionserkennung wurden die Open-Source-Frameworks OpenCV und DeepFace benutzt. Aufgrund von mangelnder Zeit wurden diese Frameworks mit ihren hergehenden Methoden implementiert, anstelle des Aufsetzens eines eigenen neuronalen Netzes. Diese Art der Implementierung wird erst im nächsten Abschnitt

innerhalb der sprachbasierten Emotionserkennung umgesetzt. Für die Umsetzung einer gesichtsbasierten Emotionserkennung mit DeepFace und OpenCV ist der einfachste Weg, die Installation über einen pip-Befehl.

```
pip install deepface
pip install opencv-python
```

Basierend auf der wissenschaftlichen Arbeit [40] wurde dieser Bereich der Anwendung umgesetzt. Beim Starten der Anwendung wird die Webcam geöffnet. Über diese Live-Daten wird das Gesicht des Benutzers erkannt und abhängig vom Gesichtsausdruck wird schließlich eine Emotion zugeordnet. Hierfür muss zunächst die Gesichtserkennung implementiert werden. Für diesen Abschnitt ist das Framework OpenCV zuständig. OpenCV stellt XML-Dateien zur Verfügung, die sog. Haar-Kaskaden enthalten, um Instanzen von Objekten in Bildern zu erkennen. Das Erstellen eines Haar-Kaskaden-Klassifikators ist eine Methode zur Objekterkennung und basiert auf den Ansätzen des Machine Learnings. Innerhalb der Gesichtserkennung wird eine Haar-Kaskade dafür genutzt, um Gesichter in einem Foto zu erkennen. Hierzu werden viele positive und negative Bilder zum Trainieren des Klassifikators verwendet. Bilder mit Objekten, die der Klassifikator identifizieren soll, nennt man dabei positive Bilder. Bilder von allem anderen sind negative Bilder. Bezogen auf die Gesichtserkennung werden Bilder mit Gesichtsmerkmalen als positive Bilder und Bilder ohne als negative Bilder angesehen. Der bereitgestellten Methode **CascadeClassifier()** wird demzufolge die XML-Datei „haarcascade_frontalface_alt2.xml“ als Parameter übergeben. Über diese Methode wird die Gesichtserkennung schließlich initiiert. Damit der Benutzer genau verfolgen kann, welcher Bereich des Gesichts erfasst wird, wurde zusätzlich ein Rahmen gezeichnet, welches in der Abbildung 5.5 visualisiert wird.

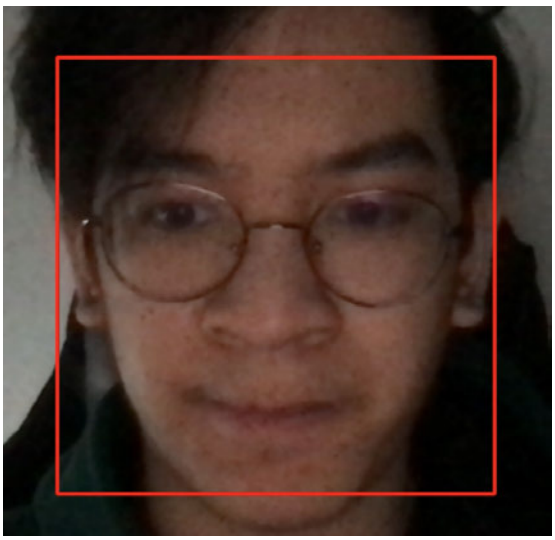


Abbildung 5.5: Die Gesichtserkennung

Nach erfolgreicher Identifikation eines Gesichtes wird die Emotionserkennung angewendet. Mithilfe der Methode **analyze()** von DeepFace können die erkannten Gesichter

einem Attribut zugeordnet werden. Über den Parameter Action kann entschieden werden, welches Gesichtsattribut analysiert werden soll. Hierzu kann das von DeepFace vortrainierte CNN-Modell zwischen sieben Emotionen unterscheiden: Wut, Angst, Trauer, Ekel, Freude, Überraschung und Neutral.

5.2.3 Sprachbasierte Emotionserkennung

Die sprachbasierte Emotionserkennung bildet den letzten und aufwendigsten Teil der Anwendung und besteht aus drei Dateien. Die Emotionserkennung wurde vergleichsweise wie in einer wissenschaftlichen Arbeit [49] zur SER umgesetzt.

1. Datensätze vorbereiten
2. Datensätze laden
3. Modell trainieren
4. Modell testen

Im ersten Schritt wurden die Datensätze vorbereitet. Es wurden hierzu bereits existierende Datensätze genutzt. Zum einen wurde der Datensatz „Ryson Audio-Visual Database of Emotional Speech“ (RAVDESS) [72] und zum anderen der Datensatz „Toronto emotional speech set“ (TESS) [73] verwendet. RAVDESS enthält einen Datensatz von über 7000 Dateien. Der Datensatz beinhaltet gesprochene oder gesungene Sätze von 24 verschiedenen schauspielenden Personen. Die Sätze wurden dabei in den Emotionen Ruhe, Freude, Trauer, Angst, Wut, Überraschung, Ekel und Neutral ausgedrückt. TESS ist ein Datensatz, der von zwei Schauspielern gesprochen wurde. Dieser Datensatz enthält 200 gesprochene Wörter, welche in sieben unterschiedlichen Emotionen ausgedrückt wurden. Bei den sieben Emotionen handelt es sich um Wut, Ekel, Angst, Freude, Überraschung, Trauer und Neutral. Innerhalb der Datei „convert_wavs.py“ wurden die entsprechenden Datensätze für die Extraktion vorbereitet. Die Datei enthält die Funktion, um die Samplerate einer Audiodatei zu reduzieren und den Audiokanal auf Mono umzustellen, um eine optimale Verarbeitung zu gewährleisten. Die Samplerate beschreibt in der Signalverarbeitung die Häufigkeit, mit der ein Signal in einer vorgegebenen Zeit abgetastet wird. Auf diese Weise wurden die Daten aus den Datensätzen TESS und RAVDESS neu konvertiert und gespeichert. Für die Verarbeitung der Audiodateien wurde das Open-Source-Projekt „FFmpeg“ genutzt. Das FFmpeg-Projekt enthält eine Reihe von Computerprogrammen und Programmbibliotheken zur Verarbeitung von Audio- oder Videomaterial. Unter anderem ist es möglich, mit FFmpeg Audiodateien nach den eigenen Wünschen anzupassen und zu konvertieren.
Der nächste Schritt wird in der Datei „utils.py“ umgesetzt und beinhaltet die Extraktion der Sprachsignale sowie das Laden der Datensätze. Demnach wurde die Funktion **extract_feature** erstellt. Die Funktion enthält die Verarbeitung der eingehenden Sprachsignale. Mithilfe von Techniken wie dass MFCC werden aus einer rohen Audiodatei die

für die Emotionserkennung relevanten Informationen extrahiert. Als Nächstes wird die Funktion **load_data** implementiert. An dieser Stelle werden bereits die Vorbereitungen für den Lernalgorithmus umgesetzt. Hierzu werden beim Laden der Datensätze die Daten zunächst in zwei Arrays eingeordnet: X und y. Dabei enthält das Array X die extrahierten Sprachsignale und das Array y die Emotionen, die ausgedrückt werden. Über die von Scikit-learn bereitgestellte Train-Split Methode werden anschließend die Arrays in zufällig aufgeteilte Trainings- und Testdatensets aufgeteilt. Wichtig zu wissen ist, dass beim Laden der Daten nicht alle Emotionen berücksichtigt werden. Das System beschränkt sich auf die ausgewählten Emotionen Wut, Trauer, Freude und Neutral.

```
def load_data(test_size=0.25):
    X, y = [], []
    for file in glob.glob("data/Actor_*/*.wav"):
        filename = os.path.basename(file)
        emotion = emotion_set[filename.split("-")[2]]

        if emotion not in needed_emotions:
            continue

        speech_features = extract_feature(file, mfcc=True, mel=True, chroma=True)
        X.append(speech_features)
        y.append(emotion)

    return train_test_split(np.array(X), y, test_size=test_size, random_state=5)
```

Abbildung 5.6: Funktion load_data

Abschließend wurde das Klassifizierungsmodell in der Datei „ser.py“ aufgesetzt. Mithilfe der Bibliothek Scikit-Learn wurde der komplette Prozess umgesetzt. Dies beinhaltet das Aufsetzen des Modells, dem Trainieren und der abschließenden Evaluation. Über Scikit-Learn ist die Implementierung eines solchen Modells einfach gelöst. Mit dem MLPClassifier von Scikit-Learn kann ohne viel Rechenleistung ein Modell in Echtzeit aufgesetzt werden. Neben diesem Vorteil nutzt die Klasse ein feedforward neuronales Netzwerk mit Backpropagation-Algorithmus und ist daher passend zum Konzept angeschnitten. Mithilfe der bereitgestellten Methode wird das Modell initiiert und über die verschiedenen Hyperparameter kann die Komplexität des neuronalen Netzes definiert werden. Ein Hyperparameter ist ein Parameter, der zur Steuerung eines Trainingsalgorithmus verwendet wird. Sie definieren die Eigenschaften des trainierten Modells. Über diese Hyperparameter kann z. B. eine unterschiedliche Aktivierungsfunktion oder die Anzahl der Hidden-Layers in einem neuronalen Netz definiert werden. Dadurch können komplexere Klassifizierungsmodelle trainiert und so eine Steigerung der Genauigkeit erzielt werden. [74] Über die bereitgestellte Methode von Scikit-Learn wurde für die eigene Implementation innerhalb der Datei „grid_search.py“ eine sog. Rastersuche umgesetzt. Die Rastersuche beschreibt die Suche nach den optimalen Hyperparameter. Hierfür wird für jeden Parameter ein Bereich definiert, der innerhalb der Rastersuche untersucht wird. Während des Trainings werden alle möglichen Kombinationen getestet und als Ergebnis wird das Modell mit der höchsten Bewertung ausgegeben. Nach abschließender Rastersuche haben sich folgende Hyperparameter ergeben.

```
best_params = {
    'activation': 'tanh',
    'alpha': 0.001,
    'batch_size': 256,
    'hidden_layer_sizes': (300,),
    'learning_rate': 'constant',
    'max_iter': 500,
    'solver': 'adam',
    'verbose': 'true',
}
```

Abbildung 5.7: Hyperparameter nach einer Rastersuche

Das bedeutet, dass das neuronale Netz genau eine verborgene Schicht, bestehend aus 300 Neuronen, umfasst. Während des Trainings werden 256 Datensätze gleichzeitig ins neuronale Netz gegeben und verarbeitet. Die Anzahl eines kompletten Durchlaufs wurde auf 500 Iterationen festgelegt. Für ein tiefergehendes Verständnis steht die Dokumentation von Scikit-Learn [71] zur Verfügung, in der alle Parameter erklärt werden.

Damit das eigene Modell letztendlich Emotionen klassifizieren kann, muss das neuronale Netz mit den vorbereiteten Daten trainiert werden, um entsprechende Muster zu erkennen. Über den folgenden Befehl werden zunächst die Daten geladen und in verschiedene Trainings- und Testdatensets gespeichert.

X_train, X_test, y_train, y_test = load_data(test_size=0.25)

Der Parameter test_size definiert dabei das Verhältnis der Trainings- und Testdaten. So werden 75 % der Daten für das Trainieren des Modells genutzt und 25 % der Daten für die Evaluation. Anschließend kann das Modell über die Methode **fit** trainiert werden. Die Variable X_train definiert hierbei die Eingabe und die Variable y_train die Ausgabe. Demzufolge werden die extrahierten Sprachsignale als Eingabe und die entsprechenden Emotionen als Ausgabe definiert.

model.fit(X_train, y_train)

Zuletzt kann über die Methode **predict** das Modell getestet werden. Für die Evaluation werden die restlichen 25 % genutzt. Nach 296 Iterationen ergibt sich eine Genauigkeit von 93,47 %.

```
Iteration 296, loss = 0.08719615
Training loss did not improve more than tol=0.000100 for 10 consecutive epochs. Stopping.
Accuracy: 93.47%
```

Abbildung 5.8: Genauigkeit des Modells

Für ein näheres Verständnis der Fehlerbestimmung, die innerhalb der Anwendung vollzogen wurde, steht das folgende Kurvendiagramm 5.9 zur Verfügung.

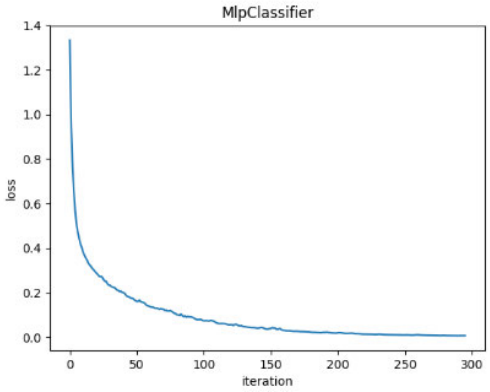


Abbildung 5.9: Kurvendiagramm der Fehlerbestimmung

Laut dem Diagramm ist deutlich zu sehen, dass nach fast jeder Iteration sich der Ausgabewert der Fehlerfunktion immer weiter minimiert. Nach der ersten Iteration ergab sich ein Fehlerwert von etwa 1,33 und nach der letzten Iteration ergab sich nur noch ein Fehlerwert von etwa 0,007. Demzufolge konnte sich der Ausgabewert der Fehlerfunktion nach 296 Iterationen um 1,323 verbessern. Auffällig jedoch ist, dass zwischen der 50. und 150. Iteration die Kurve oftmals angeschlagen ist. Das bedeutet, dass innerhalb dem Bereich der Fehlerwert zwischenzeitlich sich erhöht hat. Zuletzt wurde zur Auswertung der Klassifizierung eine sog. Konfusionsmatrix implementiert. Für die Erstellung der Matrix wurde die bereitgestellte Methode **confusion_matrix** von Scikit-Learn verwendet. Eine Konfusionsmatrix dient zur Visualisierung der Leistung eines Algorithmus. Im folgenden Diagramm 5.10 ist die Konfusionsmatrix für das eigene Modell dargestellt.

Abbildung 5.10: Konfusionsmatrix (Y-Achse: True label, X-Achse: Predicted label)

Laut der Matrix gehören von den 536 genutzten Testdaten 141 Daten der Emotion Freude an. Von diesen 141 Daten wurden 127 Testexemplare erfolgreich der Emotion Freude zugeordnet. Für die anderen Emotionen wurden ähnliche Erfolgsquoten ausgewertet. Demzufolge wurden die Emotionen Trauer, Wut und Neutral ebenso über 90 %

richtig zugeordnet. Dennoch gibt es wenige Ausnahmen, die sich in den jeweiligen Emotionen eingeschlichen haben. Unter anderem wurden bspw. neun Testdaten der Emotion Freude zugeordnet, obwohl diese eigentlich die Emotion Wut widerspiegeln.

Damit die Anwendung auch die Stimme des Benutzers erfassen kann, muss das Modell in den Sprachassistenten implementiert werden. Hierzu wird innerhalb der Hauptdatei eine weitere If-Anweisung definiert und innerhalb der Datei „record_audio“ wird eine Funktion definiert, um Mikrofoneingaben aufzunehmen. Über das Schlüsselwort „emotion recognition“ wird die Anweisung angesprochen und ausgelöst. Mithilfe der Bibliothek „pickle“ wird das Modell geladen und der Benutzer wird aufgefordert, etwas zu erzählen. Über die importierte Funktion **record_audios** wird schließlich die Eingabe gespeichert und über die Funktion **extract_feature** werden daraus die relevanten Informationen extrahiert. Die verarbeiteten Informationen werden anschließend dem Klassifizierungsmodell als Parameter übergeben. Über die Methode predict wird zuletzt die entsprechende Emotion vorhergesagt. Abhängig von der Emotion wird dann als Ausgabe eine Musikplaylist zurückgegeben, um die Emotion des Benutzers zu beeinflussen.

5.3 Probleme

Die Implementierung der Anwendung verlief nicht einwandfrei. Während der Umsetzung sind verschiedene Probleme entstanden. Die ersten Probleme sind während der Installation der Bibliotheken aufgetreten. Einige Bibliotheken werden von einer höheren Python-Version nicht unterstützt. Aus diesem Grund wurde für diese Anwendung die Python-Version 3.6 verwendet, da innerhalb dieser Version alle verwendeten Bibliotheken funktionieren. Hierbei ist es wichtig zu wissen, dass einige Bibliotheken auf eine höhere Python-Version aufgebaut ist. Daher muss für die Installation der Pakete eine genaue Version angegeben werden. Auch die Umsetzung der gesichtsbasierten Emotionserkennung verlief nicht einwandfrei. Während der Testphase hat sich ergeben, dass der Einsatz der Spracherkennung einen großen Aufwand an Rechenleistung benötigt. Die Auswertung von Live-Daten aus einer Kamera benötigen als einzelne Komponente bereits viel Rechenleistung. Innerhalb des Sprachassistenten ist diese Leistung stark erhöht. Aufgrund dieser hohen Anforderungen kam es öfter dazu, dass nach einem langen Testdurchlauf Veränderungen in Echtzeit sich stark verzögerten. Um dieses Problem zu lösen, wäre ein möglicher Ansatz die Verarbeitung der Daten anhand eines Bildes anstatt einer laufenden Kamera. Demzufolge wird nach Aufruf der gesichtsbasierten Emotionserkennung ein Foto von dem Benutzer geschossen. Dieses Foto wird anschließend in das Klassifizierungsmodell eingelesen und einer Emotion zugeordnet. Auf diese Weise wird die Rechenleistung reduziert und trotzdem Live-Daten aus dem Benutzer entnommen. Der Ansatz wurde nachträglich implementiert. Jedoch wurde der Ansatz aufgrund von mangelnder Zeit nicht für die Testphase berücksichtigt.

6 Benutzertest

Nachdem die Anwendung prototypisch implementiert wurde, wird in diesem Abschnitt der Bereich der Evaluation gedeckt. Hierzu soll die Anwendung zum Einsatz gebracht und anhand von ausgewählten Probanden getestet werden. Weiterführend werden in diesem Kapitel das Ziel der Anwendung und die Testbeschreibung aufgeführt. Abschließend werden die Beobachtungen und die Ergebnisse vorgestellt.

6.1 Hintergrund

Damit das System schließlich auf seine Funktionalitäten getestet werden konnte, wurden acht Probanden aus dem Bekanntenkreis angefragt, die den Prototypen ausgiebig testen konnten. Das Alter der Probanden variierte zwischen 25 und 30 Jahren. Um den Prototypen als Ganzes auswerten zu können, sollten hierfür die Testpersonen sowohl normale Sprachbefehle wie eine Wettervorhersage oder einer Googlesuche als auch Sprachbefehle zur Emotionserkennung testen. Für das Testen der normalen Aufgaben waren keine Anforderungen gegeben. Für die Emotionserkennung hingegen wurde den Testpersonen eine besondere Aufgabe gestellt. Die Aufgabe der Probanden war es, die Emotionen Wut, Trauer, Freude und Neutral nach eigener Interpretation nachzustellen. Demzufolge sollten die Testpersonen bei der gesichtsbasierten Emotionserkennung mithilfe ihrer Gesichtsausdrücke die vorgegebenen Emotionen nachstellen. Bei der sprachbasierten Emotionserkennung sollten die Testpersonen einen vorher beliebig bestimmten Satz in der jeweiligen Emotion nachstellen. Ziel dieser Übung war es, mit Live-Daten die Genauigkeit der Emotionserkennungssysteme zu testen und zu dokumentieren. Abschließend wurde dem Probanden drei Fragen gestellt, um feststellen zu können, ob der Prototyp zusätzlich zu den Sprachbefehlen und der Emotionserkennung auch die Emotion des Benutzers beeinflussen kann.

- Funktioniert der Sprachassistent?
- Kann Musik deine Emotion beeinflussen?
- War deine Emotion beeinflussbar?

6.2 Auswertung

Aus den durchgeführten Testdurchläufen lässt sich aussagen, dass der Prototyp trotz Probleme als Ganzes funktioniert. Von den Probanden konnten bis auf die gesichtsbasierte Emotionserkennung alle Sprachbefehle ohne weitere Schwierigkeiten getestet werden. Für eine ausführliche Auswertung wird dieser Abschnitt in vier Teile aufgeteilt: Der Sprachassistent, die Gesichtserkennung, die Spracherkennung, das Ergebnis.

6.2.1 Leistung des Sprachassistenten

Ein wichtiger Bestandteil eines Sprachassistenten ist die Aufnahme der menschlichen Sprache und die Umwandlung dieser in digitale Sprache bzw. digitalen Text. Nach zahlreichen Tests konnte von den Probanden bestätigt werden, dass in den meisten Fällen die Aufnahme der Sprache keine Probleme bereitet hat. Der Assistent war in der Lage, die gesprochenen Wörter direkt abzubilden, um so die entsprechenden Sprachbefehle aufzurufen. Auffällig war jedoch, dass Hintergrundgeräusche große Schwierigkeiten bereiten. Eine laute Umgebung erschwert dem System, die Wörter richtig zu deuten.

6.2.2 Leistung der Gesichtserkennung

Von allen Sprachbefehlen konnte der Befehl zu der gesichtsbasierten Emotionserkennung nicht gründlich getestet werden. Zwar konnte von DeepFace jeder Gesichtsausdruck richtig zugeordnet werden, jedoch hatte der Aufruf der Gesichtserkennung für eine Beeinträchtigung der Rechenleistung gesorgt. Aus diesem Grund wurde die Auswertung eines emotionsbeinflussenden Systems letztendlich nur für die sprachbasierte Emotionserkennung berücksichtigt.

6.2.3 Leistung der Spracherkennung

Die sprachbasierte Emotionserkennung spiegelt den spannendsten und wichtigsten Abschnitt der Evaluierung wider. Von den Probanden wurde die Aufgabe gefordert, die Genauigkeit des Klassifizierungsmodells zu testen. Hierzu sollte folgender Satz in den vier genannten Emotionen ausgesprochen werden.

> „We know each other for a long time.“

Während der Testphase gab es bis auf Weiteres keine großen Schwierigkeiten. Ein einziger Störfaktor waren zudem die Hintergrundgeräusche. Nach zahlreichen Testdurchläufen wurden die absoluten Werte einer Ausgabe dokumentiert. Wenn der Proband also von den zehn Versuchen die Emotion Wut abzubilden versucht, neunmal die Emotion

Freude zurückgegeben bekommt und nur einmal die Emotion Wut, so wird als Ergebnis die Emotion Freude festgehalten. Aus diesen absoluten Werten ergibt sich folgende Tabelle 6.1.

Tabelle 6.1: Emotionsausgabe (Anzahl eigentlicher Emotion/Versuchsanzahl)

	Wut	Neutral	Freude	Trauer
Proband 1	Freude (4/10)	Freude (1/10)	Freude (10/10)	Trauer (10/10)
Proband 2	Wut (10/10)	Freude (2/10)	Freude (10/10)	Freude (4/10)
Proband 3	Wut (10/10)	Freude (4/10)	Wut (4/10)	Trauer (9/10)
Proband 4	Wut (9/10)	Wut (2/10)	Freude (9/10)	Wut (4/10)
Proband 5	Freude (4/10)	Freude (4/10)	Freude (10/10)	Freude (3/10)
Proband 6	Wut (10/10)	Freude (2/10)	Freude (10/10)	Trauer (10/10)
Proband 7	Wut (10/10)	Neutral (9/10)	Freude (10/10)	Trauer (10/10)
Proband 8	Freude (3/10)	Wut (2/10)	Freude (10/10)	Freude (4/10)

Aus der entsprechenden Tabelle wurden anschließend die Genauigkeiten der einzelnen Emotionen berechnet. Hieraus ergibt sich folgende Auswertung, die in der Abbildung 6.2 dargestellt werden.

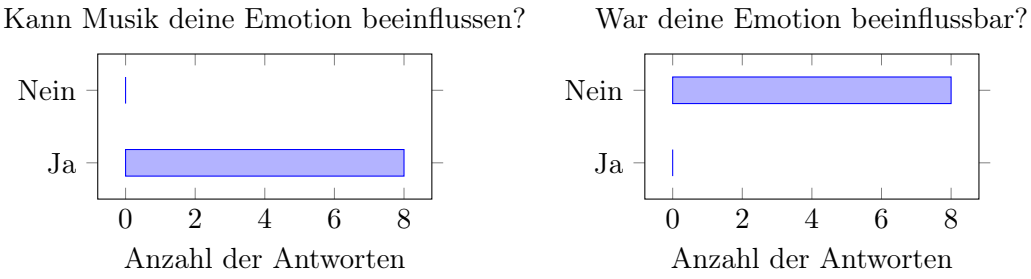
Tabelle 6.2: Genauigkeiten der Emotionen

	Wut	Neutral	Freude	Trauer
Genauigkeit	75 %	32,50 %	91,25 %	67,50 %

Laut der Tabelle konnte die Emotion Freude mit einer Wahrscheinlichkeit von 91,25 % am genausten bestimmt werden. Im Gegensatz dazu war die Emotion Neutral schwierig festzustellen. Hierzu war nur ein einziger Proband in der Lage, konstant die Emotion Neutral abzubilden. Mit einer Wahrscheinlichkeit von 32,50 % war diese Emotion am schwersten zu erkennen. Zwischen diesen beiden Wahrscheinlichkeiten befinden sich die Emotionen Wut und Trauer. Während die Emotion Wut mit einer Wahrscheinlichkeit von 75 % erkannt werden konnte, konnte die Emotion Trauer mit einer Wahrscheinlichkeit von 67,50 % bestimmt werden. Auffällig ist, dass diese Auswertung sich stark von der Konfusionsmatrix des Modells unterscheidet. Das kann unter anderem daran liegen, dass die Emotion von den Probanden nicht richtig ausgesprochen wurde, da diese keine professionellen Schauspieler sind. Weiterhin ist es möglich, dass beim Einsprechen das System Hintergrundgeräusche wahrgenommen hat, sodass die Tonqualität darunter gelitten hat. An dieser Stelle gilt es zu erwähnen, dass diese Daten nur auf den Testdurchläufen der acht Probanden basieren. Für eine genauere Beobachtung müssten weitere Benutzertests durchgeführt werden.

6.2.4 Ergebnis

Für die abschließende Auswertung wurden den Probanden die zwei emotionsbezogenen Fragen gestellt. Das Ergebnis der Fragen ist in den folgenden Diagrammen dargestellt.



Auf die Frage hin, ob Musik die Emotion des Probanden beeinflusst, haben acht von acht der Frage zugestimmt. Die Voraussetzung, dass das System also emotionsbeeinflussend reagieren kann, war demzufolge gegeben. Das Ergebnis der zweiten Frage entsprach genau dem Gegenteil der ersten Frage. Auf die Frage hin, ob die Emotion des Probanden durch das System beeinflussbar war, haben acht von acht die Frage verneint. Laut den Probanden fehlte ihnen die zwischenmenschliche Interaktion. Die Empfehlung von einer computerbasierten Stimme löste in ihnen nichts aus. Zuletzt wurde angemerkt, dass die inszenierte Situation ebenfalls entscheidend gewesen ist, dass das System nicht auf die Emotion des Nutzers einwirken konnte, da sie diese Emotionen nur gespielt und nicht wirklich gefühlt haben.
Weiterhin ist es anzumerken, dass Musik subjektiv wahrgenommen wird. Eine wissenschaftliche Studie [75] hat bereits gezeigt, dass der Geschmack von Musik abhängig von den Charakterzügen ist. Demzufolge wäre es möglich gewesen, dass die Emotion des Probanden bei einer Musikauswahl, die auf seinen Geschmack zugeschnitten ist, eher beinflussbar wäre. Zusammenfassend bedeutet es, dass diese Anwendung zwar die Emotion des Benutzers bestimmen kann, jedoch noch nicht in der Lage ist, die Emotion des Benutzers zu beeinflussen.

7 Fazit

Das Ergebnis dieser Arbeit kann als Grundlage für weitere Projekte genutzt werden. In diesem Kapitel wird der Inhalt der Thesis zusammengefasst und ein Ausblick auf mögliche Weiterentwicklungen vorgestellt.

7.1 Zusammenfassung

Im Rahmen dieser Arbeit wurde ein Überblick über den aktuellen Stand der Technik im Bereich Affective Computing verschafft. Das Ziel war es, den Stand von nicht kommerziellen Möglichkeiten zu überprüfen. Hierfür sollte aus dem erforschten Wissen eine emotionserkennende Anwendung entwickelt werden, basierend auf jenen Möglichkeiten. Zu Beginn dieser Arbeit wurde ein Einblick in die Thematik verschafft und das Ziel dieser Arbeit definiert. Weiterführend wurden die Grundlagen von Emotionen erläutert. Hierzu wurde der Begriff Emotion erklärt, klassifiziert und von anderen ähnlichen Begriffen abgegrenzt. Es wurde verdeutlicht, wie Emotionen ausgedrückt werden können und welche Art der Kategorisierung für eine Affective Computing Anwendung geeignet ist. Anschließend wurde das Forschungsgebiet Affective Computing tiefer untersucht. Auf den Grundlagen der Emotionen wurden verschiedene Systeme vorgestellt, die Emotionen auf verschiedene Art und Weise verarbeiten. Basierend auf diesem Wissen wurden mögliche Anwendungsfelder für diese Systeme vorgestellt. Für ein tiefergehendes Verständnis wurden danach die benötigten Teilgebiete der künstlichen Intelligenz erläutert. Anschließend wurde der Stand der Technik untersucht, indem die verschiedenen Möglichkeiten der maschinellen Wahrnehmung zur Emotionserkennung vorgestellt worden sind. Entsprechend dazu wurden verschiedene Open-Source-Projekte, Bibliotheken, Dienste und vergleichbare Arbeiten präsentiert, die Emotionserkennungen basierend auf verschiedenen Eingabeströmen, umgesetzt haben. Weiterhin wurden mögliche Probleme bei der Entwicklung einer Emotionserkennung dargestellt. Das erlangte Wissen wurde im Anschluss dafür genutzt, um ein Konzept für eine entsprechende Emotionserkennung zu entwickeln. Es wurden dazu die Anforderungen für die Implementierung und ein konkreter Anwendungsfall definiert. Im nächsten Abschnitt wurde der Entwurf realisiert. Hierbei wurden die eingesetzten Tools vorgestellt und die Struktur der Anwendung sowie aufgetretene Probleme dargestellt. Im Anschluss wurde die Anwendung zum Einsatz gebracht und die resultierenden Ergebnisse präsentiert. Hierzu durfte eine ausgewählte Gruppe an Probanden die Anwendung testen. Zuletzt wurden die gewonnenen

Erkenntnisse dargestellt, die Leitfragen beantwortet und Ideen zur Weiterentwicklung der Anwendung präsentiert.

7.2 Diskussion

In dem Kapitel 1 wurden Leitfragen gestellt, die im folgenden Abschnitt beantwortet werden.

- *Wie werden Emotionen ausgedrückt?*

  Emotionen können auf verschiedene Art und Weise ausgedrückt werden und spiegeln sich in einer Reaktion nach einem bestimmten Ereignis wider. Sie zeigen sich durch Mimik, Gestik, Haltung, Stimme, Handlungsimpulse wie die Vermeidung bei Ekel oder durch physiologische Reaktionen wie Schwitzen, Erröten oder Erhöhung der Herzfrequenz.

- *Wie können Emotionen maschinell wahrgenommen werden?*

  Für die Emotionserkennung müssen Computer lernen, die Emotionen des Benutzers erfassen zu können. Es existiert mittlerweile eine Vielzahl an maschinellen Wahrnehmungsmöglichkeiten, um Emotionen zu erkennen. Für die Erfassung von Emotionen können Bild-, Video-, Audio-, Textdateien, Eye-Tracking, ein EDA, ein EKG, ein EEG oder ein EMG genutzt werden.

- *Wie ist der Stand von nicht kommerziellen Möglichkeiten? Wie einfach ist die Umsetzung einer Emotionserkennung für unerfahrene Personen?*

  Um zu bestimmen, wie der aktuelle Stand von nicht kommerziellen Möglichkeiten ist, wurde im Laufe der Thesis ein Konzept für eine Emotionserkennung entwickelt. Die Anwendung stützt sich nur auf Methoden, die freizugänglich sind und wurde diesbezüglich in Python umgesetzt. Python bietet eine große Auswahl an bestehender und weiterführender Bibliotheken im Bereich Machine Learning an und eignet sich, um unerfahrenen Personen einen Einblick in die Emotionserkennung zu verschaffen. Hierbei hat der Benutzer die Wahl vortrainierte Modelle zu nutzen oder ein eigenes Modell selber zu trainieren. Viele Bibliotheken wie DeepFace bieten die Möglichkeit an, bereits trainierte Modelle und Methoden zur Klassifizierung zu nutzen.
  Sollte ein eigenes Modell trainiert werden, kann über Bibliotheken wie Scikit-Learn der Prozess einfach umgesetzt werden. Ein wichtiges Merkmal ist die gute Dokumentation der Bibliothek und ihrer Funktionen. Das Ziel bei der Entwicklung von Scikit-Learn bestand darin, eine solide

Implementierung für Machine Learning bereitzustellen. [76] Die Bibliothek zeichnet sich durch eine saubere, einheitliche und einfache API aus und bietet ein breites Spektrum an Machine Learning Methoden an, die anhand von User-Guides, Tutorien oder ausführlichen Beispielen einfach erklärt werden. [76] Über Scikit-Learn ist es möglich, mit der ausführlichen Dokumentation schnell und einfach ein Modell aufzusetzen, zu trainieren und zu evaluieren. Zusätzlich lässt sich Scikit-Learn gut in viele andere Python-Bibliotheken integrieren wie Matplotlib. Auf diese Weise können bspw. Fehlerfunktionen oder Konfusionsmatrizen visuell dargestellt werden.
Für das eigene Konzept wurden beide Methoden getestet und umgesetzt. Mit DeepFace konnte schnell und einfach ein vortrainiertes Modell integriert werden. Das bereitgestellte CNN-Modell von DeepFace bietet eine Gesichtserkennung mit einer Wahrscheinlichkeit von etwa 97 % an. Diese Wahrscheinlichkeit konnte im eigenen System bestätigt werden. Das Modell war in der Lage, in Echtzeit die Emotion des Benutzers richtig zu bestimmen. Hierbei ist es wichtig zu wissen, dass die Ausgabe in Echtzeit nur bei Bildern vorhanden war und nicht bei Live-Daten. Die Auswertung von Live-Daten aus einer laufenden Kamera benötigen viel Rechenleistung. Aufgrund dieser hohen Anforderungen kam es öfter dazu, dass nach einem langen Testdurchlauf Veränderungen in Echtzeit sich stark verzögerten. Aus diesem Grund ist die Nutzung von Live-Daten nicht empfehlenswert. Im Gegensatz dazu hat die Nutzung von bereits existierenden Bildern einwandfrei funktioniert.
Weiterführend ist die Nutzung von Scikit-Learn empfehlenswert. Ohne viele Vorkenntnisse und ohne viel Rechenleistung konnte mithilfe der Klasse MLPClassifier und den daraus resultierenden Methoden der Prozess eines Klassifizierungsmodells zeitnah erlernt und umgesetzt werden. Nur mit den hergehenden Methoden war es möglich, ohne viele Codezeilen eine Rastersuche zu starten, das Modell zu initiieren und das Modell zu trainieren. Anschließend konnte man über die bereitgestellte Konfusionsmatrix visuell überprüfen, ob das Modell etwas gelernt hat. Auf diese Weise konnte das eigene System überarbeitet werden, um das bestmögliche Ergebnis zu erzielen. Nach Nutzung von Scikit-Learn konnte ein Klassifizierungsmodell umgesetzt werden, dass eine Genauigkeit von 93,47 % erreichen konnte. Demnach ist es möglich, mithilfe von nicht kommerziellen Möglichkeiten eine präzise Emotionserkennung zu implementieren.

- *Für welchen Anwendungsfall könnte die Umsetzung einer hybriden Emotionserkennung geeignet sein?*

In dem Kapitel 1 wurde die Voraussetzung definiert, dass das entwickelte System mehrere Eingabeströme verarbeiten soll. Auf dieser Voraussetzung wurde die Emotionserkennung in einen Sprachassistenten integriert. In einem „normalen“ Sprachassistenten werden bereits Verfahren der Texterkennung umgesetzt. Demzufolge werden Textsignale durch die Texterkennung in Sprachsignale umgewandelt. Auf dieser Grundlage kann mithilfe von Bild- und Audiodateien die Drei meistverwendeten Eingabeströmen in der Emotionserkennung verarbeitet werden. Das System nutzt hierbei die Texterkennung als passives Verfahren, während Bild- und Audiodateien als aktive Eingabeströme für die Emotionserkennung verwendet werden können. Über die Sprachbefehle kann der Benutzer bewusst entscheiden, ob seine Emotion erfasst werden soll oder nicht.

- *Welche Faktoren können einen positiven bzw. negativen Einfluss auf die Emotionserkennung nehmen?*

  Für die Entwicklung einer Emotionserkennung gibt es viele Faktoren, die es zu beachten gilt. Hierbei können die Faktoren einen positiven als auch einen negativen Einfluss auf die Emotionserkennung nehmen. In dem folgenden Abschnitt werden die Faktoren dargestellt, die während der Implementierung nicht weiter berücksichtigt worden sind.

  - **Lichtverhältnisse** ermöglichen es, dass die Anwendung sowohl in hellen als auch in dunklen Lichtverhältnisse funktioniert.
  - Die Sprechweise und die Betonung einer **Sprache** sind von Kultur zu Kultur unterschiedlich, sodass für jede vorkommende Sprache und dessen Dialekt ein eigenes Modell trainiert werden muss.
  - **Eigene Daten** ermöglichen es, präzisere Klassifizierungsmodelle zu trainieren und die Auswahl an Emotionen zur Emotionserkennung zu erweitern.
  - **Hintergrundgeräusche** können bei der Aufnahme der Sprache stören und Probleme verursachen, sodass das System die richtigen Wörter nicht wahrnehmen kann.
  - Durch das Einbeziehen des **Alters** kann die Genauigkeit erhöht werden. [31]
  - Durch das Einbeziehen des **Geschlechts** kann die Genauigkeit erhöht werden. [32]

7.3 Ausblick

Das folgende Kapitel gibt einen Ausblick, wie diese Arbeit erweitert und zukünftig eingesetzt werden kann. Die in diesem Rahmen entwickelte Anwendung bietet eine Grundlage für eine hybride Emotionserkennungsoftware und hat Potenzial zur Weiterentwicklung.

7.3.1 Hyperparameter

Zum einen können im Zuge eines neuronalen Netzes größere Bereiche für Hyperparameter festgelegt werden. Demzufolge könnte für die Rastersuche weitere Parameter berücksichtigt werden, um so eine Vielzahl an Kombinationen zu erschaffen, über die es möglich ist, ein besseres Modell zu erzeugen. Auch der Einsatz von eigenen Daten kann dabei helfen, die Genauigkeit eines Modells zu verbessern. Im Rahmen dieser Arbeit wurden nur einfache Möglichkeiten zur Implementierung gezeigt. Demzufolge gibt es noch viele Möglichkeiten, um ein solches Klassifizierungsmodell komplexer zu gestalten.

7.3.2 Emotionserkennung

Weiterhin ist es möglich, die Wahrnehmungsmöglichkeiten der Emotionserkennung zu vereinen. Im Rahmen dieser Arbeit wurden zwar mehrere Eingabeströme für die Erkennung von Emotionen implementiert, jedoch agieren diese separat voneinander. Für eine weitere Optimierung ist es sinnvoll, die Eingabeströme miteinander zu verbinden. Demzufolge sollte bspw. bei dem Sprachbefehl „Open Emotion Recognition“ sowohl die Bilderkennung als auch die Spracherkennung angewendet werden, um so eine hybride Emotionserkennung zu entwickeln. Eine Forschungsarbeit [49] hat bereits gezeigt, dass der Einsatz von mehreren Eingabeströmen die Genauigkeit erhöht. Unter anderem können so die definierten Herausforderungen einer Emotionserkennungssoftware von Lisa Feldmann Barett überwunden werden. Wie im Kapitel 1 bereits erwähnt, sind nach der Studie [4] von Feldmann Barett et al. „echte“ Gefühle eines Menschen eine große Herausforderung, da diese schwer zu erkennen sind.

7.3.3 Sprachassistent

Auch im Bereich der Anwendung können Verbesserungen umgesetzt werden. Für die Implementierung des Prototyps wurden nur einfache und anfängerfreundliche Methoden genutzt. Auch hier könnte tiefer in das Gebiet eingestiegen werden. Gerade die automatische Spracherkennung ist ein entscheidender Faktor, der unbedingt für einen

Sprachassistenten funktionieren sollte. Aus den geführten Testdurchläufen konnte entnommen werden, dass dieses Verfahren nicht immer einwandfrei funktioniert hat. Besonders wenn der Benutzer schnell spricht oder sich in einer lauteren Umgebung befindet, hat das System Schwierigkeiten, die richtigen Wörter zu erfassen. Neben der einfach genutzten Bibliothek „Speech Recognition“ gibt es weitere Ansätze in diese Richtung, die erforscht werden könnten, um so eine Verbesserung zu erstreben. Z. B. bietet das System von IBM Möglichkeiten zur Verbesserung der Speech-To-Text Verfahren an. Auf deren Webseite [60] können weitere Details entnommen werden. Zusätzlich spielt die Stimme eine wichtige Rolle bei der Entwicklung eines emotionssensitiven Systems. Auch hier konnte aus den Testdurchläufen die Erkenntnis extrahiert werden, dass die computerbasierte Stimme nicht gerade förderlich ist, um eine Person aufzuheitern. Demzufolge könnten auch hier weitere Ansätze erforscht werden. Die genutzte Bibliothek „Pyttsx3“ bietet unter anderem die Möglichkeiten an, die Stimme des Systems nach den eigenen Wünschen anzupassen. Über die entsprechenden Parameter ist es möglich, Eigenschaften wie das Alter, das Geschlecht oder die Lautstärke anzupassen.

7.3.4 Abschluss

Abschließend lässt sich sagen, dass die Entwicklung einer Affective Computing Anwendung ein großes Potenzial verspricht. Die Entwicklung wird durch die Nutzung von Machine Learning Bibliotheken immer einfacher und anfängerfreundlicher. Aufgrund ihrer detaillierten Dokumentationen und Beispielen lässt sich das Konzept einer Emotionserkennung schnell erlernen. Emotionen sind ein Bestandteil des Alltags, wodurch die verschiedenen Einsatzmöglichkeiten entstehen. Die Probleme, die dabei auftreten, können durch das breite Spektrum an Machine Learning Bibliotheken gelöst werden. Neben den vorgestellten Bibliotheken gibt es weitere Bibliotheken wie TensorFlow oder Keras, mit denen es möglich ist, komplexere Modelle zu erstellen.
Ein solches Emotionserkennungssystem kann z. B. im Bereich des E-Learning [77] genutzt werden. Emotionen haben großen Einfluss auf die Motivation und die Entscheidungsfindung. Durch das Erfassen der Emotionen der Schüler kann den Lehrern geholfen werden, den Unterricht interessanter zu gestalten. Auch im Gesundheitswesen wäre solch ein Einsatz vorteilhaft. Über diese Systeme können Krankheiten wie Depression früh erfasst werden und somit dem Patienten frühzeitig beigestanden werden. Hierzu gibt es bereits wissenschaftliche Arbeiten, die sich mit diesem Forschungsgebiet befassen. [78] Zuletzt kann der Anwendungsfall eines Sprachassistenten mit Emotionserkennung für Autisten ideal genutzt werden. [79] Autisten fehlt die Fähigkeit, Emotionen zu erkennen oder diese auszudrücken. Mithilfe des Sprachassistenten wäre es möglich, dass ein Autist lernt, seine Emotionen auszudrücken, indem er mit dem Sprachassistenten übt, diese zu äußern. Über den Sprachassistenten kann der Benutzer überprüfen, ob die Emotion, die er dargestellt hat, die richtige war.

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

[1] Matthias Huber and Sabine Krause. *Bildung und Emotion.* Springer Fachmedien Wiesbaden, 2018. URL https://doi.org/10.1007/978-3-658-18589-3_1.

[2] Alina Beliba. *Challenges of Emotion Recognition in Images and Video.* apriorit, 2019. URL https://www.apriorit.com/dev-blog/642-ai-emotion-recognition. [Zugriff am: 20. Mai 2022].

[3] Rosalind W Picard. *Affective computing.* MIT press, 1997. URL https://mitpress.mit.edu/books/affective-computing. [Zugriff am: 22. Mai 2022].

[4] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. *Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements.* Psychological Science in the Public Interest, 2019. URL https://doi.org/10.1177/1529100619832930.

[5] Kate Crawford. *ARTIFICIAL INTELLIGENCE IS MISREADING HUMAN EMOTION.* The Atlantic, 2021. URL https://www.theatlantic.com/technology/archive/2021/04/artificial-intelligence-misreading-human-emotion/618696/. [Zugriff am: 20. Mai 2022].

[6] Sandra Schlegl. *Nonverbale Einstellungsmessung.* Gabler Verlag, 2011. URL https://doi.org/10.1007/978-3-8349-6192-1. Seite: 31f.

[7] Veronika Brandstätter, Julia Schüler, Rosa Maria Puca, and Ljubica Lozo. *Motivation und Emotion.* Springer Berlin Heidelberg, 2018. URL http://link.springer.com/10.1007/978-3-662-56685-5. Seite: 164ff.

[8] Paul R. Kleinginna and Anne M. Kleinginna. *A categorized list of emotion definitions, with suggestions for a consensual definition.* Motivation and Emotion, 1981. URL https://doi.org/10.1007/BF00992553. Seite: 355.

[9] Rita Faullant. *Psychologische Determinanten der Kundenzufriedenheit: der Einfluss von Emotionen und Persönlichkeit.* 2007. URL https://link.springer.com/book/10.1007/978-3-8350-9506-9. Seite: 37ff.

[10] Lothar Schmidt Atzert, Martin Peper, and Gerhard Stemmler. *Emotionspsychologie.* Kohlhammer, 2014. URL https://download.e-bookshelf.

de/download/0000/8277/53/L-G-0000827753-0014132121.pdf. Seite: 18ff; [Zugriff am: 01. Juni 2022].

[11] Klaus Rothermund and Andreas Eder. *Motivation und Emotion*. Springer-Verlag, 2011. URL https://link.springer.com/book/10.1007/978-3-531-93420-4. Seite: 165ff.

[12] Rainer Reisenzein, Wulf-Uwe Meyer, and Schützwohl Achim. *Einführung in die Emotionspsychologie*. Verlag Hans Huber, 2001. URL https://www.researchgate.net/publication/28356306_Einfuhrung_in_die_Emotionspsychologie_1. Seite: 27ff.

[13] Robert Plutchik. *Rad der Emotion*. wikipedia, 2012. URL https://de.wikipedia.org/wiki/Robert_Plutchik#/media/Datei:Plutchik-wheel_de.svg. [Zugriff am: 04. Juli 2022].

[14] Henrik Kampling, Oliver Heger, and Björn Niehaves. *Computer, die Gefühle verstehen. Zur Akzeptanz affektiver Technologien*. Forschungskolleg Siegen (Hrsg.), 2017. URL https://www.wiwi.uni-siegen.de/is/studien/studien/2017-01_wissen_computer_die_gefuehle_verstehen_-_digital-1.pdf. [Zugriff am: 28. Mai 2022].

[15] Rosalind Picard. *Affective Computing*. MIT Media Laboratory, 1995. URL https://affect.media.mit.edu/pdfs/95.picard.pdf. [Zugriff am: 12. Juni 2022].

[16] Rafael Calvo, Sidney D'Mello, Jonathan Gratch, and Arvid Kappas. *The Promise of Affective Computing*. Oxford University Press, 2015. URL https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199942237.001.0001/oxfordhb-9780199942237-e-013. [Zugriff am: 12. Juni 2022].

[17] Marc Schröder and Roddy Cowie. *Developing a Consistent View on Emotion-Oriented Computing*. Springer Berlin Heidelberg, 2005. URL https://doi.org/10.1007/11677482_17.

[18] Kiel Mark Gilleade, Alan Dix, and Jen Allanson. *Affective Videogames and Modes of Affective Gaming: Assist Me, Challenge Me, Emote Me*. Paper presented at DiGRA 2005: Changing Views – Worlds in Play, 2005. URL http://www.digra.org/wp-content/uploads/digital-library/06278.55257.pdf. [Zugriff am: 12. Juni 2022].

[19] Jianhua Tao and Tieniu Tan. *Affective Computing: A Review*. Springer Berlin Heidelberg, 2005. URL https://doi.org/10.1007/11573548_125.

[20] Rodolphe Hasselvander. *Buddy-Pro*. BLUE FROG ROBOTICS, 2022. URL https://buddytherobot.com/en/buddy-pro/. [Zugriff am: 04. Juli 2022].

[21] Wolfgang Ertel. *Grundkurs Künstliche Intelligenz*. Springer Vieweg Wiesbaden, 2021. URL https://doi.org/10.1007/978-3-658-32075-1. Seite: 1f.

[22] Inga Döbel, Miriam Dr. Leis, Manuel Molina Vogelsang, Dmitry Neustroev, Henning Dr. Petzka, Annamaria Riemer, Stefan Dr. Rüping, Angelika Dr. Voss, Martin Wegele, and Juliane Dr. Welz. *MASCHINELLES LERNEN*. Fraunhofer, 2018. URL https://www.bigdata-ai.fraunhofer.de/content/dam/bigdata/de/documents/Publikationen/Fraunhofer_Studie_ML_201809.pdf. [Zugriff am: 12. Juni 2022].

[23] Stefan Dipl.-Ing. (FH) Luber and Nico Litzel. *Was ist Deep Learning?* BigData Insider, 2017. URL https://www.bigdata-insider.de/was-ist-deep-learning-a-603129/. [Zugriff am: 12. Juni 2022].

[24] Zbigniew A. Styczynski, Krzysztof Rudion, and André Naumann. *Künstliche Neuronale Netzwerke*. Springer Berlin Heidelberg, 2017. URL https://doi.org/10.1007/978-3-662-53172-3_7.

[25] Dorian Grosch. *Neuronale Netze*. Kompetenzzentrum Öffentliche IT Fraunhofer-Institut FOKUS, 2016. URL https://www.oeffentliche-it.de/-/neuronale-netze. [Zugriff am: 12. Juni 2022].

[26] Ethem Alpaydin. *11 Mehrlagige Perzeptronen*. De Gruyter Oldenbourg, 2022. URL https://doi.org/10.1515/9783110740196-011.

[27] Stefan Dipl.-Ing. (FH) Luber and Nico Litzel. *Was ist ein Perzeptron?* BigData Insider, 2019. URL https://www.bigdata-insider.de/was-ist-ein-perzeptron-a-798367/. [Zugriff am: 12. Juni 2022].

[28] Kai Heinrich, Patrick Zschech, Björn Möller, Lukas Breithaupt, and Johannes Maresch. *Objekterkennung im Weinanbau – Eine Fallstudie zur Unterstützung von Winzertätigkeiten mithilfe von Deep Learning*. Springer Berlin Heidelberg, 2019. URL https://doi.org/10.1365/s40702-019-00514-9.

[29] Robert Peters. *Emotionserkennung mittels künstlicher Intelligenz*. Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag (TAB), 2021. URL https://www.bundestag.de/resource/blob/848996/b0a0e4dc737c35ee2626cdf2ffc8d31d/Themenkurzprofil-048-data.pdf. [Zugriff am: 13. Juni 2022].

[30] Marcel Brand, Florian Klompmaker, Peter Schleining, and Fabian Weiß. *Automatische Emotionserkennung – Technologien, Deutung und Anwendungen*. Springer Berlin Heidelberg, 2012. URL https://doi.org/10.1007/s00287-012-0618-3.

[31] Devika Verma and Debajyoti Mukhopadhyay. *Age driven automatic speech emotion recognition system*. IEEE, 2016. URL https://doi.org/10.1109/CCAA.2016.7813862.

[32] Igor Bisio, Alessandro Delfino, Fabio Lavagetto, Mario Marchese, and Andrea Sciarrone. *Gender-Driven Emotion Recognition Through Speech Signals For Ambient Intelligence Applications*. IEEE, 2013. URL https://doi.org/10.1109/TETC.2013.2274797.

[33] Paul Ekman. *Gefühle lesen*. Springer, Berlin, Heidelberg, 2010. URL https://link.springer.com/de/book/9783662532386.

[34] Paul Ekman. *Gefühle lesen*. Springer Berlin Heidelberg, 2010. URL https://link.springer.com/de/book/9783662532386. Seite: 82f.

[35] Paul Ekman. *An argument for basic emotions*. University of California, 1992. URL https://www.paulekman.com/wp-content/uploads/2013/07/An-Argument-For-Basic-Emotions.pdf. [Zugriff am: 12. Juni 2022].

[36] *Gesehen, vermessen, erkannt*. Deutschlandfunk, 2022. URL https://www.deutschlandfunk.de/dokumentarfilm-zu-gesichtserkennung-gesehen-vermessen-100.html. [Zugriff am: 12. Juni 2022].

[37] Paul Ekman. *Gefühle lesen*. Springer Berlin Heidelberg, 2010. URL https://link.springer.com/de/book/9783662532386. Seite: 19f.

[38] *Paul Ekman*. Wikipedia, 2020. URL https://wikigerman.edu.vn/wiki7/2020/11/30/paul-ekman-wikipedia/. [Zugriff am: 13. Juni 2022].

[39] Nazil Perveen and Chalavadi Mohan. *Configural Representation of Facial Action Units for Spontaneous Facial Expression Recognition in the Wild*. Department of Computer Science and Engineering, IIT Hyderabad, 2020. URL https://doi.org/10.5220/0009099700930102.

[40] Rupali Gill and Jaiteg Singh. *A Deep Learning Approach for Real Time Facial Emotion Recognition*. IEEE, 2021. URL https://doi.org/10.1109/SMART52563.2021.9676202.

[41] Stefan Dipl.-Ing. (FH) Luber and Nico Litzel. *Was ist eine Support Vector Machine?* BigData Insider, 2019. URL https://www.bigdata-insider.de/was-ist-eine-support-vector-machine-a-880134/. [Zugriff am: 17. Juni 2022].

[42] Microsoft. *Microsoft Azure*. Microsoft. URL https://azure.microsoft.com/en-us/services/cognitive-services/face/. [Zugriff am: 22. Mai 2022].

[43] Sefik Ilkin Serengil and Alper Ozpinar. *HyperExtended LightFace: A Facial Attribute Analysis Framework*. IEEE, 2021. URL https://doi.org/10.1109/ICEET53442.2021.9659697.

[44] Huijuan Zhao, Ning Ye, and Ruchuan Wang. *A Survey on Automatic Emotion Recognition Using Audio Big Data and Deep Learning Architectures.* IEEE, 2018. URL https://doi.org/10.1109/BDS/HPSC/IDS18.2018.00039.

[45] Eiman Kanjo, Luluah Al-Husain, and Alan Chamberlain. *Emotions in context: examining pervasive affective sensing systems, applications, and analyses.* https://doi.org/10.1007/s00779-015-0842-3, 2015. URL https://doi.org/10.1007/s00779-015-0842-3.

[46] Girija Deshmukh, Apurva Gaonkar, Gauri Golwalkar, and Sukanya Kulkarni. *Speech based Emotion Recognition using Machine Learning.* IEEE, 2019. URL https://doi.org/10.1109/ICCMC.2019.8819858.

[47] Wolfram Research. *Spectrogram.* Wolfram, 2017. URL https://reference.wolfram.com/language/ref/Spectrogram.html. [Zugriff am: 22. Juni 2022].

[48] Kharibam Jilenkumari Devi and Khelchandra Thongam. *Automatic speaker recognition with enhanced swallow swarm optimization and ensemble classification model from speech signals.* Springer Berlin Heidelberg, 2019. URL https://doi.org/10.1007/s12652-019-01414-y.

[49] Suraj Tripathi, Abhiram Ramesh, and Promod Yenigalla. *Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions.* Samsung RD Institute India, 2019. URL https://arxiv.org/pdf/1906.05681v1.pdf. [Zugriff am: 21. Juni 2022].

[50] Brian McFee, Raffel Colin, Liang Dawen, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. *librosa: Audio and music signal analysis in python.* librosa, 2015. URL https://librosa.org/doc/latest/index.html. [Zugriff am: 22. Juni 2022].

[51] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. *IEMOCAP: interactive emotional dyadic motion capture database.* Springer Berlin Heidelberg, 2008. URL https://doi.org/10.1007/s10579-008-9076-6.

[52] Björn Prof. Schuller, Maximilian Schmitt, and Shahin Amiriparian. *Wie Computer uns verstehen lernen.* Entwickler Magazin, 2018. URL https://entwickler.de/machine-learning/wie-computer-uns-verstehen-lernen. [Zugriff am: 21. Juni 2022].

[53] Zhang Wanli, Li Guoxin, and Wang Lirong. *Application of Improved Spectral Subtraction Algorithm for Speech Emotion Recognition.* IEEE, 2015. URL https://doi.org/10.1109/BDCloud.2015.77.

[54] Dagmar Schuller, Björn Prof. Schuller, and Florian Dr. Eyben. *openSMILE*. audEERING, 2008. URL https://www.audeering.com/de/research/opensmile/. [Zugriff am: 22. Juni 2022].

[55] Mitesh Puthran. *Speech-Emotion-Analyzer*. GitHub, 2017. URL https://github.com/MiteshPuthran/Speech-Emotion-Analyzer.

[56] Stefan Dipl.-Ing. (FH) Luber and Nico Litzel. *Was ist Natural Language Processing?* BigData Insider, 2016. URL https://www.bigdata-insider.de/was-ist-natural-language-processing-a-590102/. [Zugriff am: 13. Juni 2022].

[57] Rashed Sabra. *Sentiment Analysis – Emotionen auslesen*. L-One Systems GmbH. URL https://l-one.de/sentiment-analysis-effekte-und-grundmechanismen-teil-i/?lang=de. [Zugriff am: 13. Juni 2022].

[58] Khodijah Hulliyah, Normi Sham Awang Abu Bakar, and Amelia Ritahani Ismail. *Emotion recognition and brain mapping for sentiment analysis: A review*. IEEE, 2017. URL https://doi.org/10.1109/IAC.2017.8280568.

[59] Seo-Hui Park, Byung-Chull Bae, and Yun-Gyung Cheong. *Emotion Recognition from Text Stories Using an Emotion Embedding Model*. IEEE, 2020. URL https://doi.org/10.1109/BigComp48618.2020.00014.

[60] IBM. *IBM Watson*. IBM. URL https://www.ibm.com/cloud/watson-natural-language-understanding. [Zugriff am: 22. Mai 2022].

[61] Nadja Kurz. *Handschrifterkennung mittels neuronaler Netze mit Backpropagation*. Hochschule RheinMain, 2015. URL https://www.cs.hs-rm.de/~ulges/teaching/15MLSEM/files/ausarbeitungen/kurz.pdf.

[62] Sabrina Zehentmair. *Die Macht der Musik: Die Bedeutung von Musik fr Jugendliche und die soziale Arbeit mit Jugendlichen*. Diplomica Verlag, 2013.

[63] OpenWeatherMapAPI. URL https://openweathermap.org. [Zugriff am: 25. Juni 2022].

[64] NewsAPI. URL https://newsapi.org. [Zugriff am: 25. Juni 2022].

[65] PhillipsHueAPI. URL https://developers.meethue.com/develop/get-started-2/. [Zugriff am: 25. Juni 2022].

[66] Uwe Lorenz. *Bestärkendes Lernen als Teilgebiet des Maschinellen Lernens*. Springer Berlin Heidelberg, 2020. URL https://doi.org/10.1007/978-3-662-61651-2_1.

[67] Anthony Zhang. *Speech Recognition*. Github, 2017. URL https://github.com/Uberi/speech_recognition/blob/master/README.rst. [Zugriff am: 03. Juli 2022].

[68] Natesh M Bhat. *Pyttsx3.* Read the Docs, 2017. URL https://pyttsx3.readthedocs.io/en/latest/. [Zugriff am: 03. Juli 2022].

[69] Sefik Ilkin Serengil and Alper Ozpinar. *LightFace: A Hybrid Deep Face Recognition Framework.* IEEE, 2020. URL https://doi.org/10.1109/ASYU50717.2020.9259802.

[70] G. Bradski. *The OpenCV Library.* OpenCV, 2000. [Zugriff am: 03. Juli 2022].

[71] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. *Scikit-learn: Machine learning in Python.* Scikit-Learn, 2011. [Zugriff am: 03. Juli 2022].

[72] Steven R. Livingstone and Frank A. Russo. *The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS).* Zenodo, 2018. URL https://doi.org/10.5281/zenodo.1188976.

[73] M. Kathleen Pichora-Fuller and Kate Dupuis. *Toronto emotional speech set (TESS).* Borealis, 2020. URL https://doi.org/10.5683/SP2/E8H2MF.

[74] Gang Luo. *A review of automatic selection methods for machine learning algorithms and hyper-parameter values.* Springer Berlin Heidelberg, 2016. URL https://doi.org/10.1007/s13721-016-0125-6.

[75] David Greenberg, Sebastian Wride, Daniel Snowden, Dimitris Spathis, Jeff Potter, and Peter Rentfrow. *Universals and variations in musical preferences: A study of preferential reactions to Western music in 53 countries.* Journal of Personality and Social Psychology, 2022. URL https://doi.org/10.1037/pspp0000397.

[76] Stefan Dipl.-Ing. (FH) Luber and Nico Litzel. *Was ist ein Scikit-Learn?* BigData Insider, 2018. URL https://www.bigdata-insider.de/was-ist-scikit-learn-a-756150/. [Zugriff am: 14. Juli 2022].

[77] Nesreen Mejbri, Fathi Essalmi, Mohamed Jemni, and Bader A. Alyoubi. *Trends in the use of affective computing in e-learning environments.* Education and Information Technologies volume, 2022. URL https://doi.org/10.1007/s10639-021-10769-9.

[78] Chiara Zucco, Barbara Calabrese, and Mario Cannataro. *Sentiment analysis and affective computing for depression monitoring.* IEEE, 2017. URL https://10.1109/BIBM.2017.8217966.

[79] Bernd Bösel. *Mensch-Maschine-Interaktion.* J.B. Metzler, 2019. URL https://doi.org/10.1007/978-3-476-05604-7_30. Seite: 223f.

Erklärung zur selbstständigen Bearbeitung

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbständig verfasst und nur die angegebenen Hilfsmittel benutzt habe. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quellen kenntlich gemacht.

Ort	Datum	Unterschrift im Original