Michael Gref, **Nike Matthiesen**

Erkennung wahrgenommener Emotionalität mit Künstlicher Intelligenz in audiovisuellen Zeitzeugeninterviews

→ Deep Learning, Emotionserkennung, Zeitzeugeninterviews, Multimodalität

Zeitzeugeninterviews sind ein wichtiger Bestandteil musealer Ausstellungs- und Vermittlungspraxis. Bisher wurden in der Auswertung jedoch primär die Transkriptionen des gesprochenen Wortes und damit die Inhalte der Erzählung berücksichtigt. Für eine vertiefte Erschließung ist allerdings nicht nur interessant was gesagt wird, sondern auch wie es gesagt wird. Die automatisierte Erkennung wahrgenommener Emotionalität kann dabei helfen, Zeitzeugeninterviews auf eine neue Weise zu erschließen. In diesem Aufsatz wird ein Forschungsprojekt zur Entwicklung eines Software-Prototyps für Emotionserkennung vorgestellt. Der Prototyp basiert auf einem multimodalen Ansatz, der sich an der menschlichen Fähigkeit orientiert, den emotionalen Zustand anderer Menschen erkennen zu können (Dekodierungskompetenz). Eine wichtige Modalität hierbei ist die automatisierte Bilderkennung. Neben der Vorstellung der konzeptionellen Überlegungen und ersten Ergebnissen der Experimente werden auch die besonderen Herausforderungen des Projekts erläutert. Menschen nehmen Emotionen subjektiv und oft mehrdeutig wahr. Diese Annahme der Mehrdeutigkeit menschlicher Wahrnehmung von Emotionen zeigt sich bereits in den ersten untersuchten Ergebnissen. Ein Ungleichgewicht der verschiedenen Emotionsklassen beim Training und ein Mangel an repräsentativen Trainingsdaten führen ebenfalls zu Herausforderungen bei der technischen Umsetzung. Gleichzeitig offenbaren die Ergebnisse spannende Beobachtungen und vielversprechende Ideen für die zukünftige Anwendung und Forschung.

↠ Zitierhinweis Early View
Michael Gref, Nike Matthiesen: Erkennung wahrgenommener Emotionalität mit Künstlicher Intelligenz, in: Dieckmann et al. (Hg.): 4D → Dimensionen | Disziplinen | Digitalität | Daten, Heidelberg: arthistoricum.net, Advance online publication, 04.10.2022, **https://doi.org/10.11588/arthistoricum.1100.c15424**.

Die automatische Bilderkennung stellt einen wichtigen und stetig wachsenden Anwendungsbereich der Künstlichen Intelligenz dar. Sie findet neben der vielfachen Nutzung zur Erschließung und Recherche in großen Bilddatenbanken beispielsweise auch Anwendung im E-Commerce wie Online-Verkaufsplattformen, in der medizinischen Radiologie oder im autonomen Fahren. Ein besonderes Forschungsfeld, in dem die Bilderkennung Anwendung findet, ist der Bereich des **Affective Computing** oder auch **Emotional Artifical Intelligence.** 01 Hierbei steht das Messen bzw. Erkennen menschlicher Emotionen im Fokus – aber nicht nur über die Gesichtserkennung, sondern auch über andere Merkmale wie der Körpersprache, Stimme und Gestik oder Biosignalen in Form von Schwitzen oder erhöhtem Puls.

In einem interdisziplinären, kooperativen Forschungsprojekt entwickelt das **Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS)** seit Oktober 2020 gemeinsam mit der **Stiftung Haus der Geschichte der Bundesrepublik Deutschland (HdG)** einen Software-Prototyp, der in der Lage sein soll, wahrgenommene Emotionalität in audiovisuellen Zeitzeugenvideos zu erkennen. In dem Forschungsprojekt **Multimodales Mining von Zeitzeugeninterviews zur Erschließung von audiovisuellem Kulturgut** 02 wird die unterschiedliche menschliche Wahrnehmung von Emotionalität, sowie die Möglichkeiten und Limitationen von unterschiedlichen datengetriebenen Ansätzen untersucht. Die KI-Software soll perspektivisch dabei helfen, eine große Datenmenge zu analysieren, um besser zu verstehen, welche Rolle Emotionen beim historischen Erinnern spielen. Die Entwicklung und bisherigen Forschungsergebnisse werden in diesem Beitrag vorgestellt.

■ 01
Vgl. Rosalind Picard, Affective Computing, Cambridge 1997. Vgl. auch Soujanya Poria et al., A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion, in: Information Fusion, 37 (2017), S. 98–125.

■ 02
Das Forschungsprojekt wird über eine Förderlaufzeit von zwei Jahren im Rahmen der KI-Strategie der Bundesregierung über die Beauftragte für Kultur und Medien gefördert.

Erkennung wahrgenommener Emotionalität

Das Forschungsprojekt stellt aufgrund der interdisziplinären Ausrichtung und der neuartigen Verknüpfung von erzählter Geschichte, Emotionen und der Verwendung von Deep Learning Verfahren einige spezifische methodisch-konzeptionelle Anforderungen. Eine elementare Anforderung beinhaltet die Beantwortung der Fragen, was wir unter Emotionen verstehen und wie sie von einer Software erkannt werden. Ganz grundsätzlich können Emotionen als episodische, psychophysische Reaktionen aus subjektiven und objektiven Faktoren beschrieben werden, denen wir Menschen weitgehend passiv unterliegen. 03 Auch wenn in der emotionswissenschaftlichen Forschung der Untersuchungsgegenstand **Emotion** nicht einheitlich definiert wird, haben viele der unterschiedlichen Ansätze gemeinsam, dass Emotionen aus mehreren Komponenten bestehen: der kognitiven, der psychophysiologischen, der motivationalen, der expressiven Komponente und der Komponente des subjektiven Erlebens. 04 Für die Entwicklung des Software-Prototyps ist die expressive Komponente elementar, denn die Software soll als »maschinelles Gegenüber« erkennen, welche Emotionen von den Personen ausgesendet werden. Ausschlaggebend hierfür ist, die Prozesse der menschlichen Erkennung von Emotionen in ihrem Ablauf zu verstehen.

■ 03
Vgl. Paul R. Kleinginna, Anne M. Kleinginna, A Categorized List of Emotion Definitions, with Suggestions for a Consensual Definition, in: Motivation and Emotion, 5 1984, S. 345–379 sowie an Catrin Misselhorn, Künstliche Intelligenz und Empathie. Vom Leben mit Emotionserkennung, Sexrobotern & Co, Ditzingen 2021, S. 13ff.

■ 04
Vgl. Ulrich Mees, Zum Forschungsstand der Emotionspsychologie – eine Skizze, in: Rainer Schützeichel (Hg.), Emotionen und Sozialtheorie. Disziplinäre Ansätze, Frankfurt 2006, S. 106.

Emotionen spielen in der zwischenmenschlichen Beziehung und Kommunikation eine große Rolle, wie bereits in dem klassischen Sender-Empfänger-Modell nach Shannon und Weaver 05 beschrieben wird. In dem Modell kodiert der Sender ein Signal, das der Empfänger dekodiert und dann entsprechend darauf reagiert. Das Signal wird in Form der verbalen Kommunikation (gesprochenes Wort), der paraverbalen Kommunikation (Art der Artikulation, Spektrum der Stimme) und/oder der nonverbalen Kommunikation (Gestik, Mimik, Körperhaltung) gesendet. 06 Die Fähigkeit des Senders, das Signal, also in diesem Fall speziell emotionale Zustände, so auszudrücken, dass sie von anderen Menschen erkannt werden, heißt Enkodierungskompetenz. 07 Die Fähigkeit, den emotionalen Zustand anderer Menschen zu erkennen, nennt man Dekodierungskompetenz. 08

Die automatisierte, maschinelle Erkennung der Emotionen setzt bei der Dekodierungskompetenz, also das war für das Gegenüber erkennbar wird, an. Diesen Dekodierungsprozess versuchen wir mithilfe der KI-basierten Software nachzustellen. Das Wichtige ist, dass nur das bewertet werden kann und soll, was durch Formen des Ausdrucks und der Verhaltensweisen wahrnehmbar ist. Aus diesem Grund verwenden wir anstelle des Begriffs der **Emotionserkennung** die Formulierung **Erkennung wahrgenommener Emotionalität**. Diese präzisere Formulierung soll hervorheben, wie das System trainiert wird und was das System wirklich leisten kann. Anders als der Begriff der **Emotionserkennung** suggeriert, zielt unser System nicht darauf ab, zu erkennen, was eine Person wirklich gefühlt hat, sondern das zu erkennen, was andere Menschen an Emotionen wahrnehmen, wenn sie sich ein Zeitzeugenvideo anschauen.

Die Software wird mit von Menschen gelabelten Daten trainiert. Das heißt, dass Menschen die erkannten, also für sie wahrnehmbaren Emotionen in den Beispielvideos annotiert haben. Folglich soll die KI-Software eigene Muster aus den von Menschen gelabelten Trainingsdaten zur Erkennung wahrgenommener Emotionalität entwickeln. Der multimodale Ansatz der Software spiegelt hierbei auch die menschliche Dekodierungskompetenz auf mehreren Ebenen wider.

Dieses Vorgehen grenzt sich von Forschungsansätzen ab, die auf Grundlage von emotionswissenschaftlichem Wissen, beispielsweise durch Kombination unterschiedlicher **action units** im Gesicht, regelbasiert versuchen, Emotionen zu erkennen. 09 Genauso wenig wie wir Menschen kann die von uns entwickelte Software in den Kopf schauen. Allgemeine Kritik erfährt die Verwendung von automatisierter Emotionserkennung insbesondere von Emotionswissenschaftlerinnen und -wissenschaftlern, die anführen, dass wichtige Bestandteile der menschlichen Dekodierungskompetenz, wie die Empathie, soziale Sensitivität und vor allem die körperliche Komponente des Mitfühlens und Erfahrungsprozesses, nicht von Maschinen reproduziert werden können. 10 Wir entkräftigen durch unseren Ansatz den oft geäußerten Vorwurf des Lügendetektors, da nichts aufgedeckt wird, was nicht offensichtlich erkennbar ist. Diese feine Unterscheidung ist aus unserer Sicht wichtig, um essenzielle ethische und rechtliche Fragen präzise und sachkundig beantworten zu können.

Wir arbeiten in unserem Software-Prototyp nicht ausschließlich an der automatisierten Erkennung von Emotionen, sondern auch an der Erkennung positiver, negativer oder neutraler Meinungsäußerungen auf der Textebene (Sentimentanalyse 11). Aufgrund dessen verwenden wir im Entwicklungsprozess

05
Vgl. Claude E. Shannon, Warren Weaver, The Mathematical Theory of Communication, University of Illinois 1949.

06
Vgl. Jörg Merten, Einführung in die Emotionspsychologie, Stuttgart 2003, S. 145.

07
Vgl. ebd.

08
Vgl. ebd.

09
Vgl. u. a. Marco Del Giudice, Livia Colle, Differences between children and adults in the recognition of enjoyment smiles, in: Developmental Psychology, 43 (3), 2007, S. 796–803.

10
Vgl. u. a. »Can Machines Perceive Emotion?« Dr. Lisa Feldman Barrett, Talks at Google (2018), zu finden unter: https://www.youtube.com/watch?v=HlJQXfL_GeM, zuletzt abgerufen am 31.01.2022.

11
Vgl. Bing Liu, Sentiment Analysis. Mining Opinions, Sentiments, and Emotions, Cambridge 2015.

des multimodalen Software-Prototyps den Begriff der **Emotionalität,** da dieser in unserem Anwendungskontext nicht nur Emotionen, sondern auch weitere emotionale Ausdrucksformen, wie zum Beispiel Meinungsäußerungen widerspiegelt.

Für die »klassische« Erkennung von Emotionen orientierten wir uns am Goldstandard des **Affective Computing**: der biologisch-evolutionär geprägten Emotionstheorie nach Charles Darwin [12] und auf dessen Erkenntnissen aufbauenden Forschungen des Psychologen Paul Ekman. [13] Nach dieser Theorie gibt es sechs Basisemotionen, denen Mechanismen zugrunde liegen, die in der stammesgeschichtlichen Entwicklung entstanden sind, weil sie zur Lösung eines spezifischen Anpassungsproblems beigetragen haben. [14] Diese sechs Basisemotionen sind Freude, Trauer, Ärger, Überraschung, Angst und Ekel / Verachtung. [15]

■ 12
Vgl. Charles Darwin, The Expression of the Emotions in Man and Animals, London 1872.

■ 13
Vgl. Paul Ekman, Universals and Cultural Differences in Facial Expressions of Emotion, in: J. Cole (Hg.), Nebraska Symposium on Motivation, 19, 1971, S. 207–282. Vgl. auch Paul Ekman, Erika L. Rosenberg (Hg.), What the Face reveals. Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS), Oxford 2020.

■ 14
Vgl. ebd.

■ 15
An dieser Stelle sei zu erwähnen, dass wir Ekel und Verachtung grundsätzlich als zwei unterschiedliche Emotionen interpretieren. Wir haben uns dennoch in unserem Anwendungsfall dafür entschieden, beide Emotionen in einer Kategorie einzusortieren.

Datensatz

Für die Entwicklung unseres Software-Prototyps haben wir einen eigenen Datensatz mit rund zehn Stunden Interviewmaterial zusammengestellt. Die dort verwendeten audiovisuellen Zeitzeugeninterviews sind zum Großteil bereits auf dem Zeitzeugen-Portal, einer zentralen Interviewsammlung zur deutschen Geschichte von der **Stiftung Haus der Geschichte**, verfügbar. Allein auf der Seite **www.zeitzeugen-portal.de** befinden sich über 8.000 Clips aus über 1.000 Interviews. Der große Datenbestand bietet nicht nur spannende Inhalte, sondern lässt die Betrachterin und den Betrachter auch durch emotional behaftete Geschichten Erlebtes mitfühlen. Unser zehnstündiger HdG-Datensatz umfasst 164 verschiedene audiovisuelle Interviewclips von 147 Interviewpartnerinnen und -partnern. Die ausgewählten Interviews wurden zwischen 2010 und 2020 aufgezeichnet. Unser Ziel bei der Zusammenstellung war es, unterschiedliche Emotionen zu erfassen und einen heterogenen Datensatz in Bezug auf Alter, Professionalität des Sprechenden [16] und Geschlecht zu schaffen, der repräsentativ für den internen Zeitzeugenbestand der **Stiftung Haus der Geschichte** ist.

■ 16
Mit Professionalität meinen wir die Redeerfahrung vor Publikum oder in einem öffentlichen Kontext. Die professionellen Zeitzeuginnen und Zeitzeugen haben bzw. hatten zumeist ein repräsentatives Amt inne.

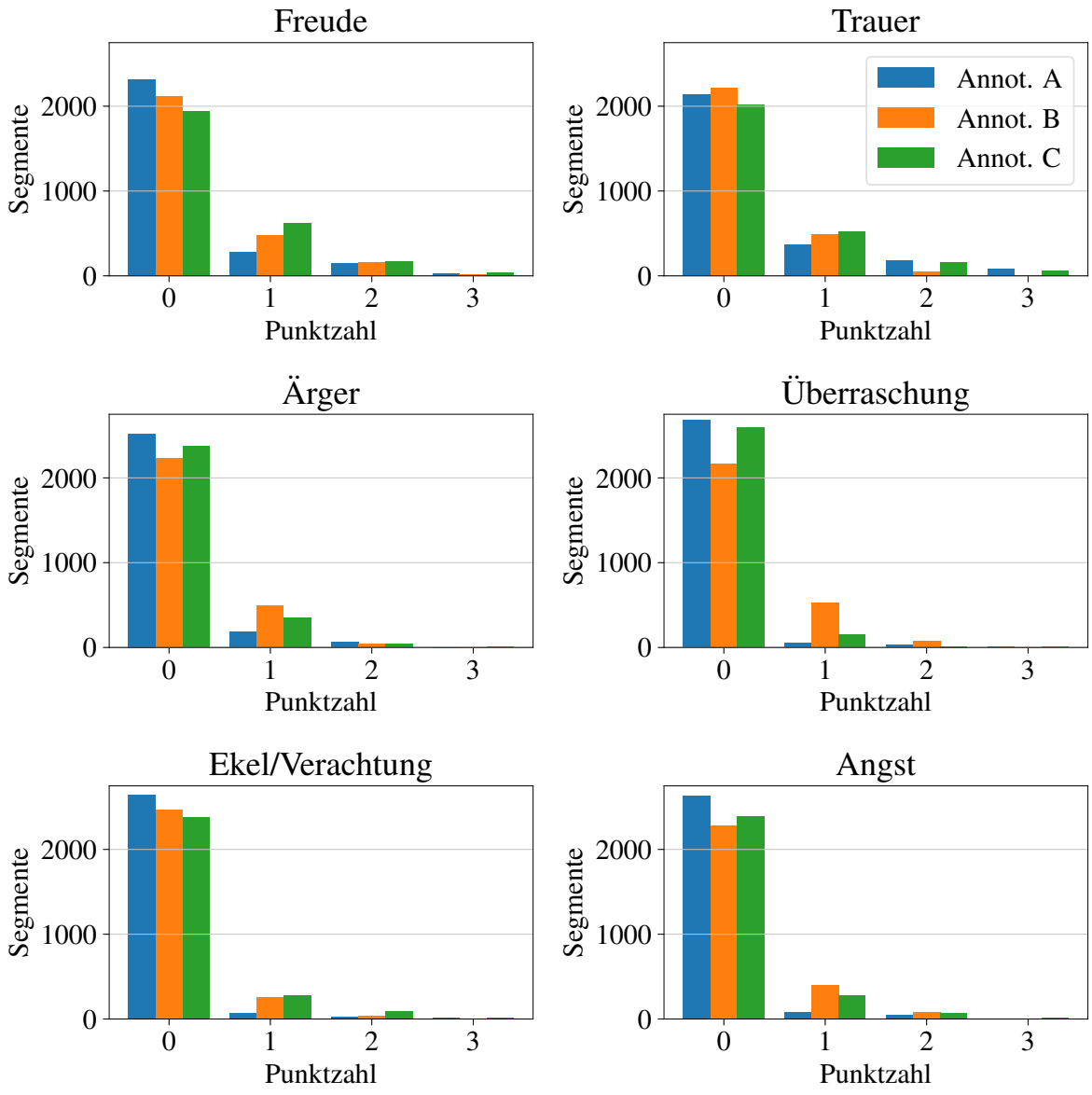
Annotation

Damit der zusammengestellte HdG-Datensatz effektiv zum Training für das KI-basierte System genutzt werden kann, muss dieser mit von Menschen gelabelten Daten angereichert werden. Hierfür wurde im ersten Schritt das Ergebnis einer ASR (Automatic Speech Recognition) verwendet, um die Interviews an den längsten Sprechpausen in kurze Segmente zu zerlegen, bis wir Segmente von 30 Sekunden oder weniger erhalten. Auf der Grundlage dieser Segmentierung haben drei Mitarbeitende der **Stiftung Haus der Geschichte** die Zeitzeugenclips segmentweise annotiert. Gleichzeitig wurde eine Referenztranskription durch Korrektur des ASR-Transkripts erstellt.

Die Annotation der wahrgenommenen Emotionalität erfolgt im selben Durchgang und für dieselben Segmente wie die Korrektur der ASR-Transkripte. Wir haben die zuvor erläuterten sechs Ekman-Klassen für die Emotionsannotation verwendet.

Pro Segment haben die drei Annotierenden eine Punktzahl auf einer 4-Punkte-Likert-Skala von 0 (keine Wahrnehmung) bis 3 (stark) für jede der sechs Emotionsklassen verwendet. Null steht für keine Wahrnehmung der Emotion und ein steigender Zahlenwert für eine stärkere Wahrnehmung der Emotion.

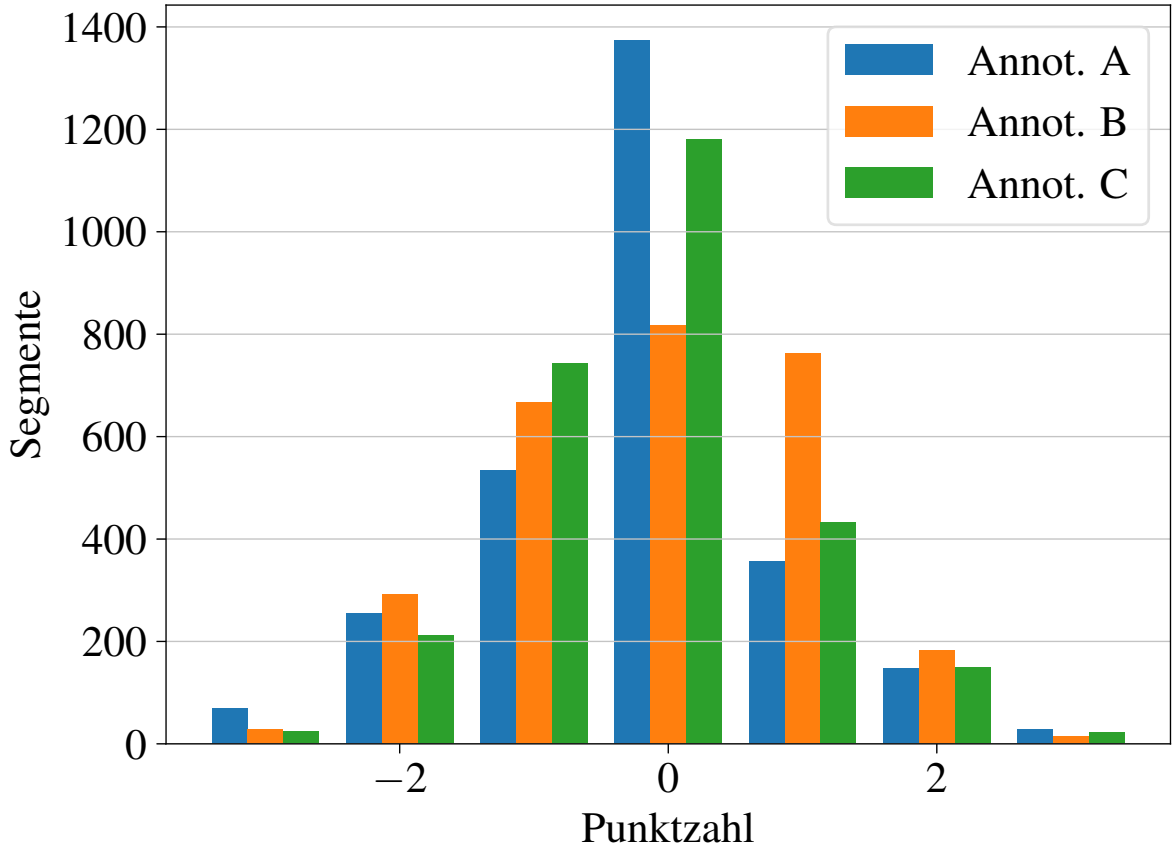
Die Annotation erfolgt unabhängig für jede Emotionsklasse, sodass in jedem Segment mehrere Emotionen in unterschiedlicher Stärke gleichzeitig auftreten können. Ähnlich wie bei den Emotionen, erfolgt die Annotation des Sentiments auf einer Likert-Skala von -3 (sehr negativ) bis 3 (sehr positiv). Negative Werte stehen für ein stark negatives Sentiment, positive Werte für ein stark positives.



☐ 01
Histogramm des Annotationsergebnisses für Emotionen.

01 stellt die Verteilung der annotierten Punktzahlen pro Emotionsklasse für jeden der drei Annotierenden dar. Die Verteilung folgt einem Muster, das für natürliche, reale Daten zu erwarten ist: Bei allen Emotionen dominiert eine neutrale Bewertung. Mit steigender Punktzahl nimmt die Segmentanzahl ab. Obwohl Emotionen beim Erinnern eine entscheidende Rolle spielen, sind die Zeitzeuginnen und Zeitzeugen dieses Datensatzes beim Erzählen oft sehr gefasst. Ein starker Ausdruck von Emotionen ist daher selten.

Freude und Trauer sind am stärksten vertreten, mit etwa 23% der Segmente mit einer Annotation über 0. Überraschung und Ekel sind mit etwa 10% die schwächsten Klassen. Unterschiede in der Annotation zwischen den drei Annotierenden lassen sich bereits anhand der Histogramme erahnen. Diese Unterschiede werden im nächsten Abschnitt näher untersucht.



□ 02
Histogramm des Annotationsergebnisses für Sentiment.

02 zeigt das Histogramm des Sentiments. Wie bei den Emotionen ist die neutrale Bewertung am dominantesten. Es ist jedoch bemerkenswert, dass eine Annotation stark von den beiden anderen abweicht. Diese Person hat häufiger einen Wert ungleich neutral vergeben als die beiden anderen.

Im Gegensatz zu vielen anderen unstrukturierten, realen Datensätzen sind bei unseren Daten negative Sentiments stärker ausgeprägt. Dies ist wahrscheinlich auf die Art der Interviews zurückzuführen: In vielen Interviews berichten die Zeitzeuginnen und Zeitzeugen von Erfahrungen aus dem Krieg oder der Nachkriegszeit, der Zeit des geteilten Deutschlands.

Klasse	Annotation A	Annotation B	Annotation C	Mittelwert
Sentiment	0.66	0.61	0.61	0.63
Freude	0.52	0.52	0.60	0.55
Trauer	0.45	0.52	0.44	0.47
Ärger	0.29	0.35	0.36	0.33
Überraschung	0.14	0.26	0.19	0.20
Ekel/Verachtung	0.31	0.32	0.38	0.34
Angst	0.36	0.38	0.41	0.38

☐ T1
Korrelation zwischen den Annotationen

T1 zeigt für jede Klasse die Korrelation zwischen den verschiedenen Annotationen. Wir verwenden den Korrelationskoeffizienten nach Spearman 17 zur Messung einer monotonen anstelle einer linearen Beziehung zwischen den Daten zweier Annotierenden. Insgesamt liegen die Werte für jedes Annotationspaar in einem ähnlichen Wertebereich und weisen keine starken Ausreißer auf. Daher gehen wir davon aus, dass keiner der drei Annotierenden ein grundlegend anderes Verständnis der Aufgabe oder einen anderen Ansatz für die Annotation hat.

Die stärkste Korrelation unter allen Klassen weist das Sentiment auf. Die Annotierenden scheinen in weiten Teilen die gleichen Wahrnehmungen bezüglich des Sentiments zu haben. Allerdings ist die Korrelation mit einem Mittelwert von lediglich 0,63 nicht als »stark« zu interpretieren, was auf in Teilen erhebliche Unterschiede zwischen allen drei Annotationen hindeutet.

Wir gehen davon aus, dass ein System, das auf diesen Daten trainiert wurde, das Sentiment grundsätzlich erlernen kann, so dass Nutzerinnen und Nutzer des Systems in den meisten Fällen eine ähnliche Meinung haben oder zumindest die Klassifikation als sinnvoll erachten würden.

Emotionen werden oft als mehrdeutiger und subjektiver angesehen als das Sentiment, was sich in der systematisch geringeren Korrelation dieser Klassen zeigt. Auch wenn im Vorfeld Annotationsrichtlinien erarbeitet wurden, in denen grob die sechs Emotionsklassen beschrieben wurden, scheint es größere Unterschiede in der Wahrnehmung oder Interpretation von Emotionen in unseren Interviews zu geben. Freude und Trauer haben unter den Emotionen die höchsten Korrelationskoeffizienten. Auch wenn es unter den Annotierenden keinen echten Konsens gibt, gehen wir von einer grundsätzlichen Übereinstimmung in einer ausreichenden Anzahl von Segmenten aus. Wir stellen die Hypothese auf, dass das Lernen der Wahrnehmung dieser Emotionen für unsere Daten prinzipiell funktionieren kann, wenn die drei verschiedenen Annotationen sinnvoll zusammengeführt werden. Es ist jedoch wahrscheinlich nicht davon auszugehen, dass beliebige Nutzerinnen und Nutzer die Einschätzung des Systems in allen Fällen teilen würden.

Die Annotierenden schienen für die verbleibenden vier Emotionsklassen sehr unterschiedliche Wahrnehmungen zu haben, wobei die Emotion Überraschung die geringste Übereinstimmung aufwies. Wenn sogar Menschen nur

■ **17**
Charles Spearman, The Proof and Measurement of Association between Two Things, in: The American Journal of Psychology, 15 (1), 1904, S. 72–101.

eine bedingt identische Wahrnehmung für diese Emotionen zu haben scheinen, vermuten wir, dass diese Mehrdeutigkeit in der Annotation die Erkennungsleistung von Emotionen in Zeitzeugeninterviews stark einschränkt – zumindest bei der Verwendung der vordefinierten Ekman-Klassen.

In einer weiteren Korrelationsanalyse untersuchen wir das gemeinsame Auftreten von verschiedenen Emotionen und des Sentiments. Wir kombinieren die drei Annotationen für diese Analyse, indem wir das arithmetische Mittel für jedes Segment bilden. Eine Korrelationsmatrix für die verschiedenen Klassen ist in T2 dargestellt. Es besteht eine moderate Korrelation zwischen Emotionen und dem Sentiment. Insbesondere besteht eine mäßige Korrelation zwischen dem Sentiment und einer positiven Gefühlslage sowie analog zwischen einem negativen Sentiment und Ärger, Ekel/Verachtung und Angst.

☐ T2
Korrelation zwischen den unterschiedlichen Klassen anhand der Mittelwerte der Annotationen pro Segment

Freude	0.45					
Trauer	-0.33	-0.17				
Ärger	-0.43	-0.13	0.10			
Überraschung	0.07	0.21	-0.09	0.07		
Ekel/ Verachtung	-0.41	-0.13	0.04	0.47	0.09	
Angst	-0.40	-0.16	0.36	0.08	0.03	0.08
	Sentiment	**Freude**	**Trauer**	**Ärger**	**Überraschung**	**Ekel/ Verachtung**

In den meisten Fällen liegt die Korrelation zwischen den Emotionsklassen im Bereich des Zufalls. Ausnahmen sind die Emotionspaare Verachtung mit Ärger und Angst mit Trauer. Diese Emotionspaare scheinen mehr als nur zufällig häufig zusammen aufzutreten und könnten für eine detaillierte, qualitative Analyse der Zeitzeugeninterviews von Interesse sein. Mögliche Ursachen für diese Emotionspaare beleuchten wir mittels einer qualitativen Befragung der Annotierenden.

Die Annotierenden waren sich einig, dass die vorgegebenen Ekman-Klassen nicht ausreichen, um die Komplexität der vielschichtigen Emotionen in Zeitzeugeninterviews wiederzugeben. Die Zuordnung in eine der sechs Klassen gestaltete sich in vielen Fällen schwierig. Daher haben die Personen intuitiv mehrere der Emotionsklassen kombiniert, um komplexere Emotionen wie Hass (Verachtung und Wut), Verzweiflung/Hilflosigkeit (Angst und Trauer), Hohn (Freude und Verachtung) und Überwältigung (Freude und Überraschung) in der Annotation darzustellen.

Überwältigung wurde zum Beispiel in einigen Interviews, in denen die Zeitzeuginnen und Zeitzeugen über den Mauerfall sprachen, als wichtige Emotion

identifiziert. Verachtung und Wut traten in Kombination häufiger in Erzählungen auf, die von Unterdrückung oder Verfolgung berichteten. Neben der Schwierigkeit, dass die Auswahlmöglichkeiten der Basisemotionen als zu undifferenziert festgestellt wurden, kommt noch die Besonderheit von Zeitzeugeninterviews als Quellengattung hinzu. [18] Zeitzeugeninterviews sind retroperspektive Erzählungen, die auf subjektiven Erinnerungen beruhen. In einem geführten Zeitzeugengespräch können mehrere (Erzähl-)ebenen differenziert werden, sodass Emotionen auf verschiedenen Ebenen sichtbar werden. Zunächst gibt es die Ebene der Interviewsituation, die bereits für einige Zeitzeuginnen und Zeitzeugen eine ungewohnte Situation darstellt. Hierdurch kann emotionales Verhalten wie Nervosität durch beispielsweise schnelles und aufgeregtes Sprechen sichtbar werden. Auf der Ebene der gegenwärtigen Erzählung werden Emotionen erkennbar, die während des Erinnerungsprozesses entstehen. Das können Emotionen sein, die die Zeitzeuginnen und Zeitzeugen nachempfinden, wenn sie von Vergangenem berichten, wie beispielsweise wieder auftretende Trauer und Wut. Zur Situation der gegenwärtigen Erzählung gehören auch reflektierte Emotionen, wenn zum Beispiel über das Erzählte geschmunzelt und gesagt wird: »Heute kann ich über die Situation lachen.« Als dritte Ebene kann die Ebene der reinen Erzählung aufgeführt werden, auf der von damaligen Emotionen oder Emotionen anderer Personen konkret berichtet wird, wodurch die Emotionen primär über das Gesagte vermittelt werden. Die verschiedenen Ebenen sind aus analytischer Sicht wichtig, können im Annotationsprozess sowie im späteren Training jedoch nicht differenziert betrachtet werden. In unserem Annotationsprozess wurden alle vorkommenden und für den Annotierenden wahrnehmbaren Emotionen annotiert, ganz gleich auf welcher Ebene sie vorkommen. Diese annotierten Emotionen können, wie bereits mehrfach erwähnt, keine Aussagen über die wirklich gefühlten Emotionen der Zeitzeuginnen und Zeitzeugen treffen, sondern geben die subjektiv wahrgenommenen Emotionen der Annotierenden wieder.

■ 18
Vgl. Dorothee Wierling, Oral History, in: Michael Maurer (Hg.), Aufriß der historischen Wissenschaften, Bd. 7: Neue Themen und Methoden der Geschichtswissenschaft, Stuttgart 2003, S. 81–151. Vgl. auch Linde Apel, Erinnern, erzählen, deuten. Oral History in der universitären Lehre, in: BIOS. Zeitschrift für Biographieforschung, Oral History und Lebensverlaufsanalysen, 31 (1), 2018, S. 23–34.

Prototypentwicklung

Für die Entwicklung des Software-Prototyps wurde bewusst ein multimodaler Ansatz ausgewählt. Wie bereits erwähnt, basiert die automatisierte Erkennung wahrgenommener Emotionalität auf der menschlichen Dekodierungskompetenz. Nachfolgend geben wir eine kurze Übersicht über die untersuchten Ansätze und initialen Ergebnisse. Wir richten hierbei einen besonderen Fokus auf das bild-basierte Verfahren.

Für das Training der Prototypen musste anhand der Roh-Annotationen im ersten Schritt jedes Segment einer Klasse zugewiesen werden. Für eine einfachere Umsetzung beim Training beschränken wir uns bei den initialen Experimenten, in Anlehnung an andere gängige Datensätze, darauf, pro Segment lediglich die am eindeutigsten wahrgenommene Emotion zu klassifizieren. Dazu verwenden wir das arithmetische Mittel der drei Annotationen und verwenden einen Schwellenwert, sodass vergleichsweise uneindeutige Wahrnehmungen

als Neutral eingestuft werden. Ähnlich gehen wir beim Sentiment vor, um die Roh-Annotationen den Meinungspolaritäten Positiv, Neutral oder Negativ zuzuordnen. Im aktuellen Ansatz versuchen wir ebenfalls nicht die Stärke der Emotion oder des Sentiments zu klassifizieren.

Methoden zur Emotionserkennung anhand von Gesichtsausdrücken können auf unterschiedliche Arten kategorisiert werden. Eine gängige Methode ist die Unterscheidung zwischen traditionellen, auf von Menschen definierten Merkmalen basierten Verfahren gegenüber Verfahren des Repräsentationslernens. Letztere Verfahren verwenden Systeme, wie zum Beispiel ein neuronales Netz, um abstrakte relevante Merkmale für die Erkennung automatisiert aus Trainingsdaten zu lernen und extrahieren. [19]

Weiterhin erfolgt eine Unterscheidung zwischen statischen und dynamischen Methoden. Statische Verfahren sind mit herkömmlichen Verfahren für andere Bildverarbeitungsaufgaben vergleichbar und berücksichtigen aus einem Video stets nur ein Einzelbild (**Frames**). Dynamische Verfahren hingegen berücksichtigen die zeitlichen Beziehungen zwischen den Einzelbildern eines Videos. Auf Grund der ausgeprägten Lernfähigkeit tiefer neuronaler Netze selbstständig aus Rohdaten relevante Informationen für die Aufgabe zu extrahieren, scheinen dynamische Repräsentationslern-Verfahren einen inhärenten Vorteil gegenüber den anderen Ansätzen zu besitzen. [20] Aus diesem Grund setzen wir diese für die Erforschung unseres Prototyps ein.

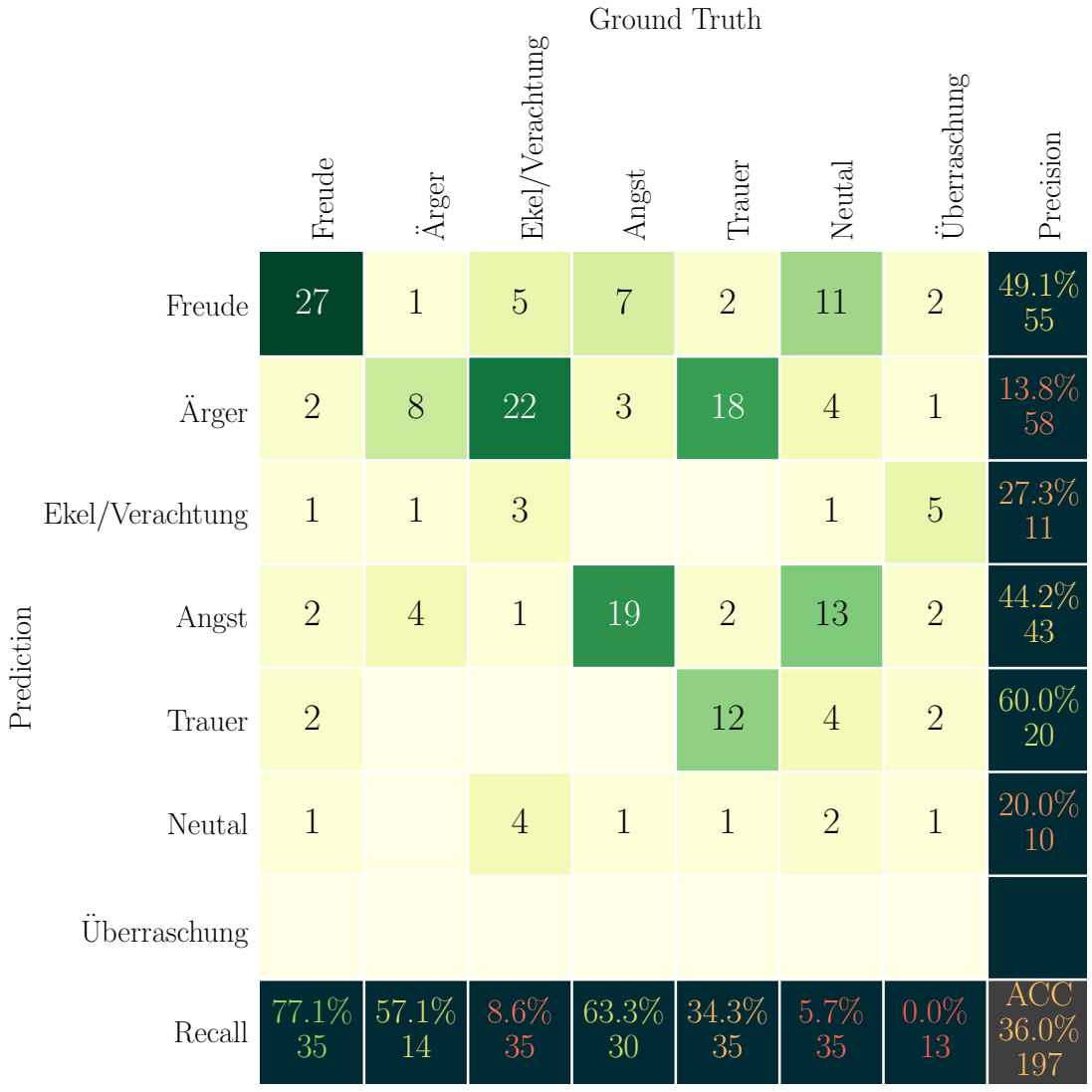
19
Vgl. Wafa Mellouk und Wahida Handouzi, Facial emotion recognition using deep learning: review and insights, in: Procedia Computer Science, 175, 2020, S. 689–694.

20
Vgl. ebd. für eine Übersicht und Vergleich relevanter Arbeiten auf diesem Gebiet.

■ 21
Vgl. Debin Meng et. al., Frame attention networks for facial expression recognition in videos, IEEE International Conference on Image Processing (ICIP), 2019, S. 3866–3870.

Im Detail untersuchen wir den Ansatz der **Frame Attention Networks** 21, welches den bekannten **Attention** Mechanismus einsetzt, um dem neuronalen Netz zu ermöglichen, selbstständig zu erlernen, welche Frames in einem Video für die Erkennung relevant sind. Ergebnisse eines initialen Experimentes mit allen Emotionsklassen sind in 03 in Form einer Konfusionsmatrix dargestellt. Die Auswertung führt zu Beobachtungen, wie sie anhand der Datenanalyse zu erwarten gewesen wären. Für die am eindeutigsten wahrgenommene und häufigste Emotion Freude funktioniert die Erkennung vergleichsweise präzise, wenngleich eine erhöhte Anzahl Segmente mit dem Label »Neutral« fälschlicherweise als »Freude« klassifiziert werden. Ähnliches gilt für Angst und zum Teil für Trauer. Für die anderen Klassen funktioniert die Erkennung kaum. Wir beobachten jedoch systematische Verwechslungen von Ärger und Verachtung sowie Ärger und Trauer. Ähnliche Vertauschungen treten zum Teil auch bei den anderen Modalitäten auf.

□ 03
Ergebnisse gesichtsausdruck-basierter Emotionserkennung auf allen Klassen in initialen Experimenten.



Wir vermuten, dass diese den Ursprung in der im Vorfeld erwähnten verstärkten Korrelation bestimmter Emotionspaare haben. Während der Annotation kombinierten die Personen intuitiv mehrere der Emotionsklassen, um komplexere Emotionen darzustellen, was sich in diesem Ergebnis widerspiegelt. Wir schlussfolgern, dass der Versuch, ausschließlich die am eindeutigsten wahrgenommene Emotion zu klassifizieren, in Zeitzeugeninterviews aufgrund der Vielschichtigkeit der Emotionen nicht zielführend ist. Stattdessen sollte das Ziel sein, auch die Kombination und das gemeinsame Auftreten von Emotionen zu erkennen, um der Komplexität von Emotionen in Zeitzeugeninterviews gerecht zu werden.

Weitere und detaillierte Auswertungen zu allen Modalitäten und die Auswirkungen unterschiedlicher Klassen geben wir in Gref et al. (2022). [22] In der dortigen Untersuchung haben wir darüber hinaus festgestellt, dass sich bei der rein textbasierten Emotionserkennung ein verstärktes Vertauschen von Angst und Trauer zeigt – ebenfalls ein von den Annotierenden häufig kombiniertes Emotionspaar. Die beobachteten Vertauschungen der gesichtsbasierten Emotionserkennung lassen sich bei der Modalität Text jedoch nicht erkennen. Dies lässt die Hypothese zu, dass bestimmte Emotionen oder Emotionspaare durch bestimmte Modalitäten verstärkt transportiert werden. Diese Überlegung bedarf jedoch weiterer Untersuchungen.

■ 22
Vgl. Michael Gref et. al., A Study on the Ambiguity in Human Annotation of German Oral History Interviews for Perceived Emotion Recognition and Sentiment Analysis, In: 13th International Conference on Language Resources and Evaluation (LREC), 2022, S. 2022–2031.

Ethische und rechtliche Fragestellungen

Das interdisziplinäre Forschungsprojekt schlägt eine Brücke zwischen Themenfeldern, die in dieser Form zum ersten Mal im musealen und geschichtswissenschaftlichen Bereich miteinander verknüpft werden. In diesem Kontext ergeben sich nicht nur die bereits ausgeführten konzeptionellen und technologischen Herausforderungen, sondern auch ethische und rechtliche Fragestellungen. Im Zentrum des Projekts stehen die persönlichen Erfahrungsberichte von Zeitzeuginnen und Zeitzeugen. Aus diesem Grund ist es für uns wichtig, dass wir uns an einen sensiblen und bedachten Umgang halten und wir rechtliche und ethische Fragen aus dem öffentlichen Diskurs anwendungsbezogen intensiv diskutieren. Wir orientieren uns hierbei primär an dem **KI-Prüfkatalog** des Fraunhofer IAIS. [23]

Darüber hinaus haben wir uns dafür entschieden, unseren selbst erstellten HdG-Datensatz nicht zu veröffentlichen und ihn ausschließlich intern zu verwenden.

■ 23
Vgl. Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS (Hg.), Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz. KI-Prüfkatalog, 2021, https://www.iais.fraunhofer.de/de/forschung/kuenstliche-intelligenz/ki-pruefkatalog.html. Zuletzt abgerufen am 31.01.2022.

Zusammenfassung

Die Erkennung wahrgenommener Emotionalität mit Künstlicher Intelligenz ist ein interdisziplinäres, komplexes aber auch zukunftsträchtiges Forschungsgebiet für viele verschiedene Anwendungsbereiche. Für unseren konkreten Use Case, der Software-Entwicklung für einen deutschsprachigen Zeitzeugenbestand, konnten wir einige wichtige Ergebnisse sowie Chancen vorstellen, mussten aber auch Grenzen erkennen.

Bereits bei der konzeptionellen Erarbeitung des Forschungsprojekts wurde deutlich, wie wichtig der präzise sowie ethisch und rechtlich korrekte Umgang mit der Erkennung von Emotionen bzw. Emotionalität ist. Auch das Medium Zeitzeugeninterview stellt mit seinen Besonderheiten, zum Beispiel mehreren Erzählebenen, einige Herausforderungen an die Entwicklung des Software-Prototyps.

Die Annotationsergebnisse unseres HdG-Datensatzes zeigen sowohl auf der quantitativ-technischen, als auch auf qualitativ-inhaltlichen Ebene einerseits zu erwartende, aber andererseits auch überraschende Ergebnisse. Auch wenn Emotionen eine wichtige Rolle im Erinnerungsprozess spielen, konnte festgestellt werden, dass Emotionen in der Menge und der Intensität, wie wir sie im Vorfeld erwartet haben, nicht in unserem zusammengestellten Zeitzeugen-Datensatz vorkommen. Hinzu kommt, dass aus inhaltlicher Perspektive durch die notwendige Kategorisierung der Emotionen für das Training des Software-Prototyps die Auswahl der Basisemotionen zu undifferenziert und beschränkt für eine vertiefte Emotionsanalyse ist. Dennoch konnten die ausgewerteten Korrelationen und die von den Annotierenden intuitiv verwendeten Kombinationen zweier Basisemotionen zeigen, dass es trotz der Beschränkung auf die Basisemotionen Möglichkeiten gibt, weitere Emotionsformen darzustellen, die auch zusätzlich durch die Verwendung der Meinungsäußerungen (Sentiment) erweitert werden. Es bleibt eine spannende, interdisziplinäre Forschungsfrage, welche Art der automatisierten Erschließung der Emotionalität in Zeitzeugeninterviews zu einer zielführenden Bereicherung von Forscherinnen und Forschern sowie Nutzerinnen und Nutzern führen kann. Die Ergebnisse und Beobachtungen unserer Arbeit können einen ersten Anstoß für diese weitere Forschung geben.

Danksagung

Wir danken Jonathan Heil, Annika Kreuziger und Marius Engel für die Durchführung der zeitaufwändigen Annotationen und Unterstützung der Auswertung der Ergebnisse.

Wir danken Sreenivasa Hikkal Venugopala, Shalaka Satheesh und Aswinkumar Vijayananth für die Durchführung der Experimente und Bereitstellung der Ergebnisse.