# Chancen und Grenzen der automatisierten Erkennung von Emotionen und Sentiments in Zeitzeugeninterviews

## Ergebnisbericht eines interdisziplinären KI-Forschungsprojekts

Nike Höfer, Michael Gref, Jana Beinlich, Sarah Zimmermann, Markus Würz und Ruth Rosenberger

### 1. Einleitung

Dass Geschichte nicht nur aus Zahlen und Fakten besteht, sondern auch aus Emotionen, bestreitet heute wohl niemand mehr. Nicht zufällig fällt die Entstehung der Oral History[1] als Zweig der Geschichtswissenschaft in das letzte Drittel des 20. Jahrhunderts, das mit seiner Hinwendung zur Subjektivität ebenso der Emotionsgeschichte zu Auftrieb verholfen hat.[2] Systematische Annäherungen an emotionale Zäsuren der (deutschen) Geschichte sowie an das Spektrum der Emotionen, die sie begleiten, könnte die Auswertung größerer Sammlungen erzählter Geschichte in Form von Zeitzeugeninterviews erbringen. So stellen mit der Kamera gefilmte Zeitzeugeninterviews historische Erfahrungsberichte dar, die zugleich die Emotionen dokumentieren, die Erinnerungen begleiten. Sie veranschaulichen die individuelle Verarbeitung von Geschichte, führen Multiperspektivität persönlicher Sichtweisen und Erfahrungen vor Augen und können so ein differenziertes Bild von historischem Erleben zeichnen. Erzählungen von Zeitzeuginnen und Zeitzeugen spielen in der musealen Ausstellungs- und Vermittlungspraxis historischer Häuser eine zentrale Rolle (De Jong 2018), bieten aber auch für die geschichtswissenschaftliche Forschung ganz eigene Potentiale (Wierling 2003; Apel 2015).

Bei der Auswertung von Zeitzeugeninterviews wird zumeist die Textfassung der erzählten Inhalte berücksichtigt. Dabei bleibt die ihnen eingeschriebene Emotionalität als zentrales Moment, das diese Quelle kennzeichnet, oft wenig beachtet. Dies in den Blick zu nehmen war einer der zentralen Ausgangspunkte für das interdisziplinäre Forschungsprojekt „Multimodales Mining von Zeitzeugeninterviews zur Erschließung von audiovisuellem Kulturgut“.[3] In diesem sollten mit Hilfe von Methoden der KI-gestützten Erschließung von Zeitzeugeninterviews neue Arbeitsinstrumente entwickelt und

1 Zur Geschichte und Forschungspraxis der Oral History siehe zuletzt: Althaus/Apel 2023.

2 Als Überblick zur Emotionsgeschichte siehe: Hitzer 2011; Plamper 2012, 2013.

3 Vgl. https://www.hdg.de/stiftung/projekte (9.6.2023). Dank gilt Jonathan Heil, Annika Kreuziger und Marius Engel für die Durchführung der zeitaufwändigen Annotationen und für die Unterstützung bei der Auswertung der Ergebnisse. Dank gilt auch Sreenivasa Hikkal Venugopala, Shalaka Satheesh und Aswin Kumar Vijay Ananth für die Durchführung der Software-Experimente und die Bereitstellung der Ergebnisse.

 https://doi.org/10.3224/bios.v35i2.07

evaluiert werden, die diesem Untersuchungsgegenstand Rechnung tragen. Denn sowohl für die Forschung wie auch für die museale Zeitgeschichte ist nicht nur relevant, was erzählt wird, sondern auch, wie es erzählt wird.

Hinzu kommt, dass die Menge an archivierten Zeitzeugeninterviews und -Beständen, wie sie etwa bei der Stiftung Haus der Geschichte der Bundesrepublik Deutschland, aber auch bei anderen sammelnden Institutionen vorliegen,[4] sich nicht mehr händisch erschließen lassen. Neben den neuen inhaltlichen Fokus tritt somit die alltagspraktische wie forschungsstrategische Aufgabe einer digital unterstützenden Erschließung und entsprechender Instrumente. Denn Algorithmen erleichtern im digitalen Zeitalter die Suche und das Filtern von Informationen in großen Datenbeständen. Sie unterstützen den Menschen bei der Erschließung von Daten und gewinnen daher zunehmend an Bedeutung auch für kulturelle Einrichtungen mit umfangreichen Sammlungen. Vor diesem Hintergrund müssen sich Museen und sammelnde Institutionen fragen, wie auch KI-Technologien genutzt werden können, um die inhaltliche und dokumentarische Erschließung ihrer Sammlungsbestände zu unterstützen oder zu verbessern. Dies setzt jedoch voraus, dass sie die technischen Details von Verfahren der Künstlichen Intelligenz reflektieren und eine kritische Haltung entwickeln, indem sie sich Potenziale und Problematiken der Technologie bewusst machen.[5] Erste Schritte in diese Richtung unternahm die Stiftung Haus der Geschichte der Bundesrepublik Deutschland gemeinsam mit dem Fraunhofer Institut für Intelligente Analyse- und Informationssysteme (IAIS) anhand einer exemplarischen Anwendung im Forschungsprojekt, das von Oktober 2020 bis September 2022 im Rahmen der KI-Strategie der Bundesregierung über die Staatsministerin für Kultur und Medien gefördert wurde. Ziel des Projekts war es, mit Hilfe von KI-Technologien in videografierten Zeitzeugeninterviews ausgedrückte Emotionen zu identifizieren.

In der musealen Praxis der Stiftung wird erzählte Geschichte in Form von Zeitzeugeninterviews genutzt, um Besucherinnen und Besucher subjektiv und emotional anzusprechen, ihnen einen Zugang zur oftmals abstrakten Geschichte zu eröffnen und die Multiperspektivität von historischen Erfahrungen zu vermitteln (Petschow/Würz 2022). Zur Sammlung der Stiftung gehören rund 2.600 Zeitzeugeninterviews, die in der Objektdatenbank erfasst sind – davon sind derzeit gut 1.100 Interviews über das Angebot www.zeitzeugen-portal.de auch online zugänglich. Grundlage der bisherigen dokumentarischen und inhaltlichen Erschließung sind Transkriptionen des gesprochenen

4 Dies sind insbesondere das Archiv „Deutsches Gedächtnis“ des Instituts für Geschichte und Biographie an der Fernuniversität Hagen, die „Werkstatt der Erinnerungen“ der Forschungsstelle für Zeitgeschichte in Hamburg oder die Digitalen Interview-Sammlungen des „Centers für Digitale Systeme“ an der FU Berlin. Es sind aber auch Gedenkstätten zum Unrecht des SED-Staates wie beispielsweise die Gedenkstätte Berlin-Hohenschönhausen oder die Gedenkstätten an ehemalige nationalsozialistische Konzentrationslager wie unter anderem die Gedenkstätte Buchenwald, die Mahn- und Gedenkstätte Ravensbrück oder die Gedenkstätte und Museum Sachsenhausen. Seit Ende 2023 bietet das Interviewportal Oral-History.digital Zugang zu einer Vielzahl von Interviews aus unterschiedlichen Sammlungen und Institutionen. Siehe: https://portal.oral-history.digital/de (26.2.2024).

5 Anfang Dezember 2022 fand im Landesmuseum Karlsruhe die Konferenz „Kulturen der Künstlichen Intelligenz. Neue Perspektiven für Museen“ statt, die erstmals in Deutschland eine kritische Bestandsaufnahme über Chancen und Risiken des Einsatzes von KI-Verfahren in Museen versuchte. Siehe https://www.landesmuseum.de/fileadmin/BLM_-_Cultures_of_Artificial_Intelligence_-_New_Perspectives_for_Museums.pdf (7.2.2024).

Wortes. Die Vision ist es, mit Hilfe KI-basierter Software die Interviewdaten zu analysieren, um automatisiert Metainformationen, wie etwa genannte Personen, Orte, Institutionen/Körperschaften und auch ausgedrückte Emotionen zu identifizieren, um so die Interviews in der Sammlungsdatenbank mit diesen Informationen anzureichern, die dokumentarische Erfassung zu vertiefen und die kuratorische Arbeit zu erleichtern. Existieren in der KI-Forschung bereits Ansätze für die Extraktion einiger der genannten Informationen,[6] steht die automatisierte Suche nach ausgedrückten Emotionen noch am Anfang. Ziel des interdisziplinären Kooperationsprojektes war es, Erkenntnisse aus dem Bereich des Maschinellen Lernens mit der Arbeit an audiovisuellen Zeitzeugeninterviews zu kombinieren, Methoden aus der bisherigen Forschung hierzu zu reflektieren und explorativ einen Software-Prototypen[7] zu entwickeln, der in audiovisuellen Zeitzeugenvideos ausgedrückte Emotionen erkennen soll. Die Projektpartner betrieben mit dem Vorhaben Grundlagenforschung, die im *Affective Computing* zu verorten ist, das heißt, der Disziplin, die die maschinelle Erkennung, Verarbeitung und Simulation menschlicher Emotionen erforscht (Picard 1997; Poria et al. 2017; Pessanha/Salah 2022). Audiovisuell aufgezeichnete Erzählungen, Emotionen und die Verwendung von Deep-Learning-Verfahren wurden methodisch-konzeptionell zusammengeführt.

## 2. Was sind Emotionen?

Emotionen bestehen aus mehreren zusammenspielenden Komponenten: einer kognitiven, einer psychophysiologischen, einer motivationalen, einer expressiven und dem subjektiven Erleben (Mees 2006: 106). Ein Individuum nimmt ein Objekt – dies kann eine Person, eine Sache, ein Zustand, ein Ort oder ein Ereignis sein – über seine Sinne wahr (Russell/Feldman Barrett 1999). Über die kognitive Komponente werden subjektive Erfahrungswerte angesprochen und Bezüge zur „persönlichen Wichtigkeit oder Bedeutung der aufgefassten Welt" hergestellt (Mees 2006: 106). Durch die psychophysiologische Komponente werden bestimmte Neurotransmitter und Hormone ausgeschüttet, die den physiologischen Zustand des Individuums verändern und eine körperliche Erregung veranlassen. Die motivationale Komponente beschreibt die (verhaltensbasierte) Reaktion auf das Wahrgenommene. Über die expressive Komponente wird schließlich die Emotion ausgedrückt.

In der Emotionsforschung variieren die Theorien darin, welche von diesen Komponenten sie primär in den Fokus nehmen und wie sie das Wechselspiel zwischen den Komponenten beschreiben (Misselhorn 2021: 16). Forschungshypothesen, die auf Erkenntnissen von Charles Darwins *The Expression of the Emotions in Man and Animals* (1872) beruhen, beschreiben sechs, teils acht basale Emotionen, die auf verschiedenen Ebenen (Mimik, Gestik, physiologische Veränderungen) „durch die ganze Welt mit merkwürdiger Gleichförmigkeit" (Euler 2000: 46, zitiert nach Darwin 1872: 17) sichtbar werden. Diesen evolutionsbasierten Ansatz verfolgte der Psychologe Paul Ekman in seinen grundlegenden Studien zu den sechs Basisemotionen Freude, Ärger, Trauer, Ekel, Angst und Überraschung weiter (Ekman 1971). Nach seiner Theorie liegt jeder Basisemotion ein Mechanismus zugrunde, der in der stammesgeschichtlichen Entwick-

6 Zum Einsatz digitaler Methoden in der Geschichtswissenschaft siehe unter anderem König 2017.

7 Die im Projekt entwickelten Prototypen waren nur im Rahmen der Forschungsumgebung des Fraunhofer IAIS nutzbar und nicht öffentlich zugänglich.

lung entstanden ist, weil er zur Lösung eines spezifischen Anpassungsproblems beigetragen hat. Beispielsweise kann Angst Flucht vor Bedrohungen auslösen. Basierend auf dieser Theorie entwickelte Ekman mit Wallace Friesen ein Kodierungsverfahren, das rund 44 sogenannte *action units*, sichtbare Bewegungen der mimischen Muskulatur, zusammenfasst und bis heute weltweit in vielen interdisziplinären Gebieten Anwendung findet (Ekman/Friesen 1978; Ekman/Rosenberg 2020). Der Vorteil des Ekman'schen Ansatzes ist es, dass sich Basisemotionen beschreiben und klassifizieren und damit für intelligente Computerprogramme erkennbar machen lassen, weshalb dieser oftmals als „Goldstandard" des *Affective Computing* bezeichnet wird. Kritik erfährt dieser universell geprägte Ansatz der angeborenen Emotionen von Emotionspsychologen und -psychologinnen wie Lisa Feldman Barrett, die mit ihrer „Theory of constructed emotions" dafür plädiert, dass Emotionen keinesfalls universell seien, sondern von der Person selbst durch das Zusammenspiel individueller physischer Eigenschaften des Körpers, einem flexiblen Gehirn sowie kulturellen Bedingungen kreiert würden (Feldman Barrett 2017a, 2017b, 2018). Auch geistes- und sozialwissenschaftliche Forschungen folgen zumeist einem sozial konstruierten, kulturell bedingten und veränderbaren Konzept von Emotionen (Plamper 2013: 13). Die emotionshistorische Forschung wiederum beschäftigt sich primär mit Fragen zur Geschichtsmächtigkeit von Gefühlen und der Normierung und Variabilität von Emotionen in verschiedenen Gesellschaften zu unterschiedlichen Zeiten (Frevert 2020; Gammerl 2021).

Um eine operationalisierbare Methodik zu entwickeln, positionierte sich das Forschungsprojekt zwischen diesen Ansätzen, indem es seinen Blick sowohl auf die expressive Komponente des Emotionsprozesses richtete als auch auf die menschliche Wahrnehmung von Emotionen. Dies lässt sich mit Hilfe des von Shannon und Weaver beschriebenen Sender-Empfänger-Modells (Shannon/Weaver 1949) fassen. Der Sender kodiert ein Signal, das der Empfänger dekodiert und dann darauf reagiert. Das Signal wird auf Ebenen der verbalen Kommunikation (gesprochenes Wort), der paraverbalen Kommunikation (Art der Artikulation, Spektrum der Stimme) und/oder der nonverbalen Kommunikation (Gestik, Mimik, Körperhaltung) gesendet (Merten 2003: 145). Die Fähigkeit des Senders, das Signal (hier: emotionale Zustände) so auszudrücken, dass es von anderen Menschen erkannt wird, wird als Enkodierungskompetenz beschrieben (ebd.). Die Fähigkeit, den emotionalen Zustand anderer Menschen zu erkennen, ist entsprechend die Dekodierungskompetenz (ebd.). Forschungsprojekte mit Künstlicher Intelligenz versuchen, menschliches Problemlösungsverhalten nachzubilden. Die automatisierte, maschinelle Erkennung von Emotionen setzt daher bei der Dekodierungskompetenz an, um diese mithilfe der KI-basierten Software nachzustellen. Es können dabei nur Emotionen identifiziert und bewertet werden, die auf der Ebene von Ausdruck und Verhalten explizit durch ein Gegenüber (Empfänger) wahrnehmbar sind. Der Emotionsbegriff muss daher präziser als wahrnehmbare oder wahrgenommene Emotion gefasst werden, denn die von Zeitzeuginnen und Zeitzeugen im Gesprächsverlauf intrinsisch erlebten Emotionen lassen sich nicht abbilden.

Weil die menschliche Dekodierungskompetenz, wie geschildert, auf mehreren Ebenen (Wort, Art des Sprechens und Mimik/Gestik) stattfindet, sollte die im Projekt entstehende Software in einem multimodalen Ansatz nach Vorbild des Menschen als „Gegenüber" erkennen lernen, welche Emotionen von den Personen ausgesendet werden. Orientiert an den Ekman'schen Basisemotionen sollten dies Freude, Trauer, Ärger, Überraschung, Angst und Ekel/Verachtung sein. Zudem sollte auf Textebene, das heißt,

der gesagten Wörter, die Ausprägung von wertenden Meinungsäußerungen (Sentiment) erhoben werden, die ein Indiz für Emotionen sein können. Die Sentimentanalyse ist ein Bereich der Computerlinguistik, der die Meinungen von Menschen zu verschiedenen Objekten oder Themen analysiert und vielfach im Marketing, in der Social-Media- oder in der Kundenkommunikation zum Einsatz kommt. Dabei soll festgestellt werden, ob die Autorinnen und Autoren in einem Text eine positive, negative oder neutrale Haltung (Polarität) äußern (siehe Liu 2015).

## 3. Prototypenentwicklung Emotionserkennung[8]

Die Umsetzung der methodischen Überlegungen in einen Prototyp der automatischen Emotionserkennung erfolgte in drei Phasen. In der ersten Projektphase stand die Bereitstellung, Generierung und Analyse von Trainings- und Testdaten im Vordergrund. In der zweiten Phase wurden für die drei einzelnen Modalitäten „Audio" (das heißt paralinguistische Signale wie Sprechpausen, Sprechrhythmus, Intonation, Tonhöhe und Lautstärke), „Video" (Gesichtsausdruck) und „Text" (welche Worte wurden gesprochen) jeweils unimodale Prototypen trainiert. Die abschließende Projektphase hatte zum Ziel, die vorherigen unimodalen Prototypen in einen multimodalen Prototyp zu überführen, der alle drei Modalitäten gleichzeitig berücksichtigt. Zudem wurde ein Prototyp für die Erkennung von „Sentiment" trainiert und getestet.

### *3.1 Daten und Annotation*

Daten sind die Grundlage, um KI-Systeme zu trainieren. Zwar findet automatisierte Emotionserkennung heute schon vielfach Anwendung, etwa in der Personalisierung von Werbung und Marketing, der Identifikation bei Strafverfolgung oder im Gesundheitswesen (Misselhorn 2021: 37 ff.). Trainingsdaten, die mit (deutschsprachigen) Zeitzeugeninterviews arbeiten, existieren aber nicht. Darin besteht eine große Herausforderung. Daher wurde in der ersten Projektphase ein eigener Datensatz mit rund zehn Stunden Interviewmaterial aus den Beständen der Stiftung Haus der Geschichte zusammengestellt und durch Mitarbeitende der Stiftung annotiert. Der Datensatz umfasste 164 verschiedene audiovisuelle Interviewclips von 147 Personen, die zwischen 2010 und 2020 interviewt wurden. Ziel bei der Zusammenstellung war es, unterschiedliche Emotionen zu erfassen und einen möglichst heterogenen Datensatz in Bezug auf Alter, Geschlecht und Professionalität der Sprechenden[9] zu schaffen, der repräsentativ für den Sammlungsbestand der Stiftung ist.

Für die Annotation wurden die Interviewclips automatisiert (anhand der längsten Sprechpause) in kurze Segmente mit einer maximalen Länge von 30 Sekunden zerlegt. Diese Segmente wurden dann von drei Mitarbeitenden der Stiftung Haus der Geschichte einzeln gesichtet und mit Werten für die Basisemotionen und das Sentiment versehen. Pro Segment haben die drei Annotierenden eine Punktzahl auf einer Likert-Skala von 0 (keine Wahrnehmung) bis 3 (stark) für jede der sechs Emotionsklassen

8 Ausführlich zu diesem Abschnitt siehe: Gref et al. 2022a und Gref und Matthiesen 2022.

9 Professionalität meint, über welche Redeerfahrung vor Publikum oder in einem öffentlichen Kontext Zeitzeuginnen und Zeitzeugen verfügen. Diese ist bei „professionellen" Interviewpartnern, die oftmals ein repräsentatives Amt innehatten, ausgeprägter.

verwendet. Null steht für keine Wahrnehmung der Emotion und ein steigender Zahlenwert für eine stärkere Wahrnehmung der Emotion. Die Annotation erfolgte unabhängig für jede Emotionsklasse, sodass in jedem Segment mehrere Emotionen in unterschiedlicher Stärke gleichzeitig auftreten können. Ähnlich wie bei den Emotionen erfolgte 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.

Eine Erkenntnis dieses Arbeitsschrittes ist, dass sich die vielschichtigen Emotionen, die in Zeitzeugeninterviews wahrzunehmen sind, mit den sechs Ekman'schen Basisemotionen nur unpräzise abbilden lassen. Die Gründe hierfür liegen in der Besonderheit von Zeitzeugeninterviews als Quellengattung. Zeitzeugeninterviews sind retroperspektive Erzählungen, die auf subjektiven Erinnerungen beruhen. Daher können Emotionen auf mehreren Ebenen sichtbar werden. Die Interviewsituation stellt für zahlreiche Interviewte eine ungewohnte Situation dar und ruft Nervosität hervor, die sich in schnellem, aufgeregtem Sprechen zeigen kann. Zudem lässt der Erinnerungsprozesses während des Erzählens im Interview Emotionen entstehen. Zuletzt beschreibt die Erzählung (verbal) emotionale Zustände der Vergangenheit. Die Mitarbeitenden haben bei der Annotation der Interviewclips intuitiv mehrere Emotionsklassen kombiniert, um komplexere Emotionen wie Hass (Verachtung und Wut), Verzweiflung/Hilflosigkeit (Angst und Trauer), Ironie und Sarkasmus (Freude und Verachtung) oder Überwältigung (Freude und Überraschung) in der Annotation darzustellen. Auf welchen Ebenen diese Emotionen lagen, konnte jedoch nicht abgebildet werden.

Insgesamt trat während der Annotation deutlich hervor, dass die menschliche Wahrnehmung von Emotionen subjektiv und oftmals uneindeutig ist. Dies ließ bereits zu diesem Zeitpunkt Auswirkungen auf die Erkennungsgenauigkeit der Software vermuten, die mit diesen Daten trainiert wird. Als wesentlich eindeutiger stellte sich die menschliche Wahrnehmung von Sentiment dar, sodass die Erkennungsleistung der mit diesen Daten trainierten Software höher vermutet wurde.

### *3.2 Unimodale Emotionserkennung Prototypen*

Für das Training der Prototypen wurden jedem Videosegment Klassen zugewiesen. Bei den initialen Experimenten waren dies pro Segment lediglich die am eindeutigsten wahrgenommene Emotion sowie eine Meinungspolarität (positiv, negativ oder neutral). Dazu wurde das arithmetische Mittel der drei Annotationen genutzt, vergleichsweise uneindeutige Wahrnehmungen unter einem Schwellenwert wurden als neutral eingestuft. Mit welcher Stärke die Emotion oder das Sentiment annotiert wurden, wurde nicht berücksichtigt.

Für jede Modalität (Video, Audio und Text) wurde unabhängig voneinander ein Prototyp entworfen und trainiert. Konkret wurden zunächst für jede Modalität vielversprechende Machine-Learning-Ansätze aus aktuellen Veröffentlichungen ausgewählt und diese empirisch miteinander verglichen. Für das Training der Modelle wurden sowohl Teile des hauseigenen Datensatzes als auch andere öffentlich verfügbare deutschsprachigen Datensätze der jeweiligen Modalität verwendet, um eine möglichst hohe Robustheit und Generalisierbarkeit der Erkennung zu erreichen. Es wurde sowohl die Kombination der Datensätze als auch ein mehrstufiges Training mit den unterschiedli-

chen Datensätzen erprobt. Eine beispielhafte Erläuterung des Vorgehens beim Sentiment-Prototypen wird in Kapitel 3.3 gegeben.[10] In einem iterativen Prozess, bei dem wiederholt einzelne Hyperparameter angepasst wurden, wurden die Prototypen einzeln trainiert. Ein Teil des selbst erstellten Datensatzes, der nicht für das Training und die Anpassung der Hyperparameter verwendet wurde (Testdatensatz), wurde schließlich zum Testen der Prototypen verwendet. Es wurde sichergestellt, dass in dem Testdatensatz keine Zeitzeugen vorkamen, die in dem Trainingsdatensatz verwendet wurden.

Die Erkennungsraten der Prototypen zu den einzelnen Modalitäten waren allerdings wenig zufriedenstellend. Keiner der drei Prototypen erreichte bei der Erkennung der sieben Klassen (sechs Ekman'sche Basisemotionen und „neutral") eine Genauigkeit von über 40 Prozent. Insgesamt war der Prototyp für die Modalität „Video/Gesichtsausdruck" am vielversprechendsten. Die Erkennung für die Emotionsklassen „Freude" funktionierte mit 60 Prozent, für „Trauer" mit 44 Prozent. „Angst" wurde noch mit 52 Prozent erkannt.[11] Allerdings kam es häufig zu Verwechslungen mit der neutralen Klasse. Weiterhin wurde häufig die Klasse „Ärger" mit „Trauer" vertauscht. Bei den anderen Klassen kam es hingegen fast immer zu nicht-systematischen bzw. zufällig wirkenden Verwechslungen. Die Erkennungsleistung für einzelne Klassen konnte verbessert werden, wenn bei der Klassifikation weniger Emotionsklassen zugelassen bzw. einzelne Emotionsklassen ausgelassen wurden (vgl. Gref et al. 2022a).

Die im Projekt trainierten sprach- und textbasierten Prototypen blieben wesentlich hinter dem Prototyp für die Modalität „Video/Gesichtsausdruck" zurück. Selbst die Differenzierung zwischen „Freude" und „Trauer" war für die Prototypen dieser beiden Modalitäten oft nicht möglich. Um die Ursachen hierfür zu untersuchen, wurde die Anzahl der Emotionsklassen sukzessive reduziert. Auch dann blieb jedoch die Unterscheidung zwischen „Freude" und „Trauer" für das sprachbasierte System eine große Herausforderung. Eine stichprobenartige Erhebung von Audiosegmenten zeigte, dass für viele dieser Segmente eine Differenzierung allein anhand der Audiospur auch für Menschen kaum möglich ist. Nur in wenigen Segmenten gibt es hinreichend eindeutige Hinweise auf den emotionalen Zustand, die sich beispielsweise durch Lachen, Weinen, eine erhöhte Sprechgeschwindigkeit oder in der Tonlage äußern.

Die insgesamt zu geringe Erkennungsgenauigkeit aller drei Prototypen bestätigte die Vermutung, dass sich die Uneindeutigkeit in der menschlichen Wahrnehmung von Emotionen, die bei der Annotation der Daten festgestellt wurde, auf die Leistungsfähigkeit der Software auswirkte. Für das Training problematisch war zudem, dass die Emotionsklasse „neutral" gegenüber den anderen Klassen überrepräsentiert war.

### *3.3 Ausblick: Multimodale Emotionserkennung*

In der abschließenden Projektphase ging es darum, Lösungsansätze für die zuvor erkannten Herausforderungen zu erarbeiten, um eine multimodale Emotionserkennung in einem Softwareprototypen zu ermöglichen. Gezielt wurden daher die Uneindeutigkeit bzw. Subjektivität der menschlichen Wahrnehmung bei der Annotation von Emotionen

10 Für eine detaillierte technische Beschreibung und Auswertung der Experimente zum Training der unimodalen Prototypen (sowie weiterführender Forschungsliteratur) siehe Gref et al. 2022a und Gref/Matthiesen 2022.

11 Die Prozentzahlen entsprechen dem so genannten „F1-Score", einer Metrik, um die Effektivität von Machine-Learning-Modellen zu bewerten. Er ist ein gewichtetes Mittel aus Genauigkeit und Sensitivität.

mithilfe der Ekman'schen Klassen noch einmal in den Blick genommen. Statt, wie in der Forschung üblich, durch Mittelwerte und Verrechnen der unterschiedlichen Annotationen eine „gemeinsame Wahrheit" zu finden, sollte das System lernen, dass es Uneindeutigkeiten in der Annotation gibt. Die Software wurde dabei gezielt parallel auf die einzelnen Annotationen hin trainiert.

Diesen Ansatz legt Viswanath (2023) in einer eigenen Studie dar. So wurde zunächst das Training eines multimodalen Prototypen zur Emotionserkennung mit einem „Multi-Label"-Klassifikationsansatz und unter Einsatz eines Multimodal-Transformer (Tsai et al. 2019) untersucht. Bei der „Multi-Label"-Klassifikation wird jede Emotionsklasse unabhängig von den anderen klassifiziert, sodass gleichzeitig beliebige Kombinationen auftreten können – in Einklang mit der Art, wie die Daten annotiert wurden. Die „Multimodal-Transformer"-Architektur ist darauf ausgelegt, gleichzeitig unterschiedliche Modalitäten verarbeiten zu können, wobei automatisch gelernt werden kann, dass es in den verschiedenen Modalitäten zu unterschiedlichen Zeitpunkten Merkmale für das Klassifikationsergebnis gibt. Mit Einsatz dieser Architektur und der Multi-Label-Klassifikation gelang es, erhebliche Verbesserungen in der Emotionserkennung zu erzielen, wobei eine durchschnittliche Genauigkeit von über 70 Prozent erreicht wurde.

Um die Uneindeutigkeiten in der Wahrnehmung der Annotatoren bzw. die Unterschiede in den Annotationen zu untersuchen und die Auswirkungen beim Modelltraining zu verstehen, wurde ein neuer Trainingsansatz, das „Multi-Annotator-Learning", untersucht. Die Annotation jedes Annotators wurde hierbei als eigene Klasse behandelt und das Modell auf die gleichzeitige Vorhersage der Annotation aller drei Annotatoren für jede Emotionsklasse trainiert. Dabei liegt die Annahme zugrunde, dass die annotierenden Personen zwar unterschiedliche Emotionsklassen für den gleichen Videoclip vergeben haben, dies aber innerhalb ihrer eigenen Annotation konsistent getan haben, sodass das Modell für jeden Annotator eine präzisere Vorhersage als für den Mittelwert treffen kann.

Weiterführende Arbeiten, um aus der Erkennung der individuellen Annotationen ein indizierbares, annotatorunabhängiges Erkennungsergebnis zu erhalten, stehen noch aus. Statt einer „Emotionserkennung" soll die Software perspektivisch beispielsweise eine „Emotionsgruppierung" bzw. ein „Emotionsclustering" realisieren. Sie soll aus den unterschiedlichen Daten eine abstrakte Repräsentation für die Emotionen in den Interviewclips erlernen, mit denen ähnliche Emotionen in anderen Videos verglichen und gefunden werden können. Die Arbeiten an diesem innovativen Ansatz dauern an, so dass an dieser Stelle noch kein finaler Ergebnisbericht erfolgen kann.

## 4. Prototypentwicklung Sentiment-Klassifikation

Neben der Emotionserkennung wurde „Sentiment" als relevante und für weitere Untersuchungen interessante Analysekategorie für Zeitzeugeninterviews identifiziert. Mit Hilfe einer maschinellen Sentimentanalyse soll dabei die positive, negative oder neutrale Polarität einer Meinungsäußerung in einem Text identifiziert werden. Eindeutige Identifikatoren sind sprachspezifische Ausdrücke oder auch Satzkonstruktionen, die aufgrund ihrer (Wort-)Bedeutung positiv oder negativ bewertet sind.

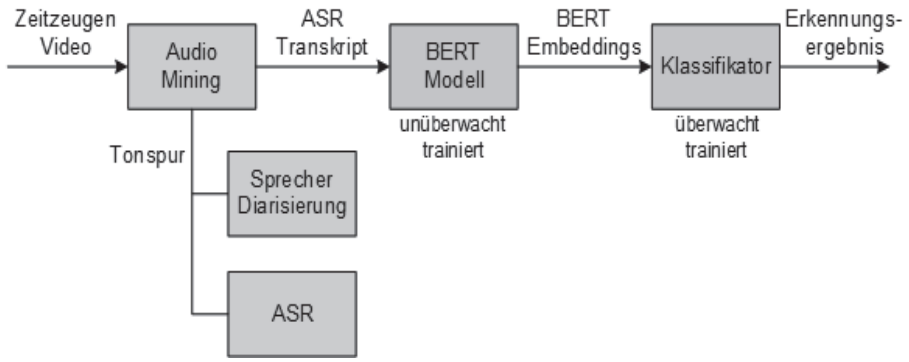


*Abbildung 1: Vollautomatische, textbasierte Sentimentanalyse*

Grundlage für die Sentimentanalyse sind Transkriptionen des gesprochenen Wortes, die durch automatische Spracherkennung erstellt wurden. Das Konzept des verwendeten Systems ist in Abbildung 1 schematisch dargestellt. Genutzt wurde die Software „Audio Mining" (Köhler et al. 2017), die in der Audiospur erkennt, wer wann spricht (Diarisierung), diese Audiospur segmentiert, um anschließend die gesprochene Sprache zu erkennen (ASR, automatic speech recognition). Dieses Verfahren unterstützt die Erschließung und Analyse großer Interviewbestände, für die keine Transkriptionen vorliegen (siehe beispielsweise Leh 2021). Gleichwohl arbeiten ASR-Systeme nie fehlerfrei.

Das im Projekt für die Analyse verwendete Sentiment-System basiert auf dem vortrainierten Sprachmodell BERT (Devlin et al. 2019) zur Extraktion von „Embeddings" („Einbettungen") aus geschriebenen Texten. Die identifizierten „Embeddings" sind dabei eine abstrakte mathematische Vektorraumrepräsentation der Bedeutung von Sätzen, die es computerbasierten Systemen erlaubt, Ähnlichkeiten und Unterschiede zwischen Sätzen mathematisch zu untersuchen. Als Basismodell wurde ein für Deutsch vortrainiertes Modell der Plattform *Huggingface*[12] verwendet. Dieses Modell wurde mit einer großen Datenmenge aus verschiedenen Open-Source-Quellen wie Wikipedia Untertiteln von Filmen und Serien, öffentlich zugänglichen Buch-Corpora sowie Internet- und Newsseiten unüberwacht, das heißt ohne annotierte Daten, vortrainiert. Die „Embeddings" des BERT-Modells dienen als Eingabewerte für einen Klassifikator bestehend aus einem neuronalen Netz. Das System wurde mit einem mehrstufigen Trainingsansatz für die Sentiment-Klassifikation überwacht, also mit Sentiment-annotierten Daten, trainiert. Hierfür wurde zunächst der deutsche Teil des sogenannten CMU-MOSEAS-Datensatzes (Zadeh et al. 2020) verwendet, bestehend aus ca. 18,6 Stunden und 10.000 Sätzen, der Annotationen von wahrgenommenem Sentiment und entsprechenden Emotionen enthält. Dieser Datensatz ist aktuell einer der größten, natürlich-sprachlichen mit dieser Art der Annotation. Danach erfolgte ein „Fine-Tuning" mit einem Teil des selbst erstellten Datensatzes. Die verbleibenden Interviews des Datensatzes wurden für Tests und Evaluationen des trainierten Systems verwendet. Um die Anwendung „robust" zu machen, wurden für das Training der Software unbearbeitete ASR-Transkriptionen verwendet, in denen Fehler, das heißt Worte, die von der Software nicht richtig erkannt

12 https://huggingface.co/dbmdz/bert-base-german-cased (05.0.2024).

wurden, nicht nachträglich von Menschen korrigiert wurden. Im Durchschnitt lag die Wortfehlerrate in den Transkriptionen bei 16 bis 17 Prozent (Gref et al. 2022b).

Die Sentimentanalyse erreichte eine für natürlich-sprachliche und unstrukturierte Daten annehmbare Genauigkeit von 64,1 Prozent verglichen mit den in den menschlichen Annotationen wahrgenommenen Sentiments. Als häufigster Fehler des Systems tritt die Schätzung eines negativen oder positiven Sentiments auf, während die Daten tatsächlich mit „neutral" ausgezeichnet sind.[13]

### *4.1 Sentimentanalyse: Zwei Fallstudien*

Um die Qualität der automatisierten Sentimentanalyse besser einschätzen zu können, wurden zusätzlich zwei qualitative Fallstudien durchgeführt. Bei beiden erfolgten Auswertung und Bewertung durch jeweils eine wissenschaftliche Mitarbeiterin der Stiftung Haus der Geschichte, die bis dahin nicht in das Projekt eingebunden war.

Wie präzise eine automatisch generierte Sentimentanalyse im Einzelfall funktioniert, wurde in Fallstudie 1 beispielhaft an sechs deutschsprachigen Videointerviews des Regisseurs Hermann Vaske untersucht. Über einen Zeitraum von mehr als 30 Jahren hat Vaske Menschen weltweit nach den Beweggründen für ihr kreatives Tun gefragt („Why are you creative?"). Die Stiftung Haus der Geschichte hat den Bestand als Teil ihrer Sammlung übernommen. In Vorbereitung auf die Sentimentanalyse wurden zunächst Rohtranskriptionen der Interviews erstellt, die aus insgesamt 568 gesprochenen Sätzen bestanden. Verglichen wurde dabei das subjektive Rating der annotierenden Person (positiv, neutral, negativ) mit dem im Zuge der automatisierten Sentimentanalyse generierten Rating. Die Annotation erfolgte „blind", das heißt, ohne Kenntnis der Ergebnisse der Sentimentanalyse. Anschließend wurde verglichen, ob die Einschätzungen von Mensch und Maschine a) „konvergent" (von annotierender Person und Prototyp übereinstimmend als positiv-positiv; neutral-neutral; negativ-negativ bewertet), b) „nahe konvergent" (leicht voneinander abweichend, zum Beispiel neutral-positiv; neutral-negativ) oder c) „divergent" (negativ-positiv; positiv-negativ) sind.

Das Ergebnis: Die computergenerierte Einschätzung wich nur in 5 Prozent der untersuchten Sätze divergent von der subjektiven Einschätzung der annotierenden Person ab. Dagegen lag in 57 Prozent der Fälle eine konvergente, in 38 Prozent der Fälle eine nahe konvergente Einschätzung vor. Insbesondere wenn die Interviewten in Metaphern, Bildern, abstrakten Umschreibungen sprechen, scheint die automatisierte Sprachanalyse an Grenzen zu stoßen. Zur korrekten Einschätzung ist hier entweder der Bezug zum Kontext des gleichen oder der umgebenden Sätze relevant oder die bildhafte, nicht wortwörtliche Interpretation des Gesagten. Hierzu wiederum müssen Metaphern und Redewendungen als solche erkannt werden, für deren Identifikation im Einzelfall soziokulturelles Vorwissen relevant ist, das dem Prototyp (noch) fehlt. So kann der Satzabschnitt „etwas Verrücktes" Ausdruck positiver Meinung im Sinne von Faszination oder positiver Überraschung sein. Etwas, das als „nicht erklärbar" beschrieben wird, kann durchaus positiv gemeint sein im Sinne eines geheimnisvollen, interessanten, Neugierde weckenden Ereignisses, das gerade im künstlerischen Kontext nicht negativ,

13 Die Genauigkeit der Sentimentanalyse verbesserte sich nur leicht auf 64,8 bis 66,0 Prozent, wenn von Menschen korrigierte Transkripte verwendet wurden – abhängig davon, wie die Korrekturen vorgenommen wurden. Von der ASR-Software produzierte Fehler betrafen somit nur geringfügig die für Sentimentanalyse relevanten Wörter.

sondern erstrebenswert ist. Dieser Umstand führt unter anderem dazu, dass subtil positives Sentiment durch den annotierenden Mensch erkannt werden, während die Sentimentanalyse-Software das Prädikat „neutral“ vergibt, oder gar divergente (negative) Einschätzungen vornimmt.

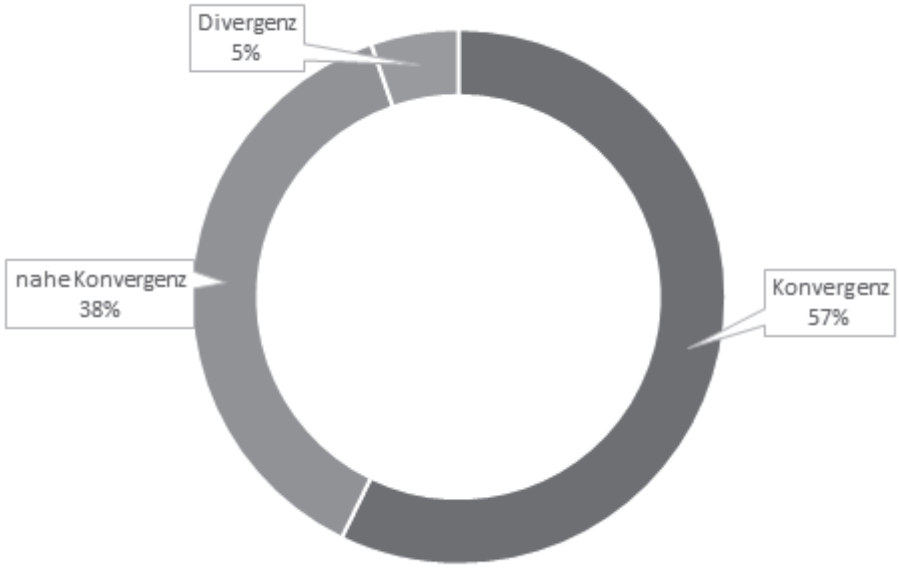


*Abbildung 2: Ergebnisse Fallstudie 1, Sentimentanalyse „Why are you creative?“*

Die zweite Fallstudie nahm in den Blick, wie sich die (grammatikalische) Genauigkeit des gesprochenen Wortes auf die Qualität der Ergebnisse der Sentimentanalyse auswirkt. Als Testbestand dienten Transkriptionen von sechs audiovisuellen Zeitzeugeninterviews von Menschen mit Migrationshintergrund, deren Muttersprache nicht Deutsch ist und deren deutscher Sprachgebrauch sich durch grammatikalische Ungenauigkeiten auszeichnet. Im Ergebnis wich die computergenerierte Einschätzung erneut lediglich in 5 Prozent der untersuchten Sätze divergent von der subjektiven Einschätzung der annotierenden Person ab. Dagegen lag in 66 Prozent der Fälle eine konvergente, in 29 Prozent der Fälle eine nahe konvergente Einschätzung vor.

Eine Ursache für die Konvergenzen war insbesondere der Mangel an repräsentativen Trainingsdaten mit gleichwertiger sprachlicher Fehlerquote. Ausgehend von einer negativen Einschätzung der positiv gemeinten Aussage „Alle Leute haben ganz viel Geduld“ zeigte sich beispielsweise, dass die Phrase „ganz viel“ wie auch die Phrase „sehr viel“ in der Systematik der Trainingsdaten überdurchschnittlich häufig negativ (50 Prozent) konnotiert sind.

Negative oder positive Meinung kann auch über die Zeichensetzung ausgedrückt und interpretiert werden.[14] Ein Ausrufungszeichen wurde beispielsweise durch die Software vermehrt als quantitativer Messwert für positive Meinungsäußerung erkannt, wohingegen die Analyse von Anführungszeichen ausschließlich zu einem negativen Label geführt hat. Probleme bei der Analyse können darüber hinaus auch bei mehrdeutigen Satzzeichen auftreten, zum Beispiel Punkt nach einer Abkürzung. Auch die Satzlänge hatte erheblichen Einfluss auf die Einschätzung. Insbesondere die Ergebnisse stark verschachtelter Sätze mit mehr als einem Satzzeichen sind fehleranfällig und weisen eine hohe Divergenz auf. Divergente Ergebnisse zeigten sich auch, wenn zur Einschätzung der Meinungspolarität historische Kontexte relevant waren. Den Satz „Aber mit dem Vater hat sich das dann schon so ergeben, dass wir zum Teil auch Polnisch gesprochen haben, also schon längere Zeit", zeichnete die Sentimentanalyse als klar „positiv" (68 Prozent) aus. Tatsächlich sprach der Zeitzeuge aber über die traumatische Zeit nach dem Zweiten Weltkrieg in Oberschlesien, in der schlagartig nicht mehr Deutsch, sondern nur noch Polnisch gesprochen werden durfte. Verändert man den Satzbau und ersetzt die Wörter „zum Teil" durch „nur noch", führte dies zu einer korrekten Einschätzung.

In der Summe zeigen beide Fallstudien, dass Prototyp und annotierende Person in der überwiegenden Mehrzahl der Fälle zu annähernd gleichen Einschätzungen der in den Videos ausgedrückten Meinungspolaritäten kamen. Konvergenz zwischen annotierender Person und Prototyp war am häufigsten im Falle neutraler Bedeutungszuschreibungen zu finden. Auch dort, wo Sätze Schlagwörter enthielten, die eindeutig einer Polarität zuzuordnen waren (positiv: zum Beispiel „glücklich", „gut", „wunderbar"; negativ: zum Beispiel „verzweifelt", „Wut", „Ärger"), war die Konvergenz des Sentiment-Prototyps hoch. Weniger präzise fiel die automatisierte Analyse dann aus, wenn positives Sentiment subtiler, differenzierter oder abstrakter formuliert waren und von der automatisierten Sentimentanalyse eher als „neutral" bewertet wurden. Auch im Falle metaphorischer Sprache (zum Beispiel „das war meine Melodie" im Sinne von „das passte gut zu mir") fällt diese Tendenz auf. Es verwundert dementsprechend nicht, dass der dominierende Fehler die Klassifizierung eines neutralen Sentiments war, wohingegen die annotierende Person die gleiche Aussage negativ oder positiv bewertet hat (vgl. nahe Divergenz). Dennoch bestätigt die mit maximal 5 Prozent auftretende starke Divergenz die Robustheit des Prototypen. Der Fokus auf das Sentiment als Annäherung an eine emotionale Äußerung zeigt sich so als mögliche Alternative zur umfassenden Emotionsdetektion und es lohnt sich, weitere Forschungen anzuschließen.

## 5. Fazit

Das interdisziplinäre Forschungsprojekt hat in einer Pilotstudie ein Forschungsdesign für die automatisierte Erkennung von ausgedrückten Emotionen und Sentiments in

14 Es ist anzumerken, dass das der Prototyp während des Trainings nur Satzzeichen berücksichtigt hat, die automatisiert auf Basis des ASR-Transkripts erzeugt wurden. Das „Satzzeichenmodell" wurde auf die Rekonstruktion von Satzzeichen mit Textdaten trainiert, die aus dem Internet gecrawlt wurden. Diese sind nicht immer fehlerfrei. Die Auswirkungen hiervon zu untersuchen, ist ein Aspekt dieser Fallstudie gewesen.

Zeitzeugeninterviews entworfen und operationalisiert. Abschließend sind folgende Herausforderungen und Chancen zu bilanzieren:

Sehr deutlich wurde, dass Daten der „Rohstoff" sind, mit dem Systeme Künstlicher Intelligenz trainiert und verbessert werden, um so Anwendungen entwickeln zu können, die die Arbeit oder den Alltag von Menschen erleichtern. Je qualifizierter und breiter die Datengrundlage, desto ausgereifter kann eine KI-basierte Software funktionieren,[15] weshalb die Entwicklung maßgeblich durch US-amerikanische Konzerne wie Alphabet, Amazon, Meta oder Microsoft vorangetrieben wird, die im vergangenen Jahrzehnt große Datenmengen – zum Teil auch ihrer Nutzerinnen und Nutzer – gesammelt haben (Ramge 2019: 87 f.). Der Mangel an qualifizierten Daten, die auf Zeitzeugeninterviews beruhen, führte im Projekt dazu, einen eigenen (deutschsprachigen) Datensatz erstellen zu müssen, um überhaupt eine Datengrundlage zu haben. Der große Mangel an qualifizierten (deutschsprachigen) Daten, die benötigt werden, um eine leistungsfähige Software zu trainieren, wird eine der größten Herausforderungen für die zukünftige Forschung bleiben. Ohne diese wird die Forschung leistungsfähige Erkennungssysteme nicht entwickeln können.[16]

Das Erfassen, Kategorisieren sowie eindeutige Zuordnen von Emotionen zu vordefinierten Klassen ist eine der Grundlagen dafür, dass KI-Systeme lernen können, diese zu identifizieren. Emotionen werden jedoch ein schwer zu definierender Untersuchungsgegenstand bleiben. Anhand der vorgenommenen Daten-Annotationen wurde offensichtlich, wie unterschiedlich bereits die menschliche Wahrnehmung ausgedrückter Emotionen in Zeitzeugeninterviews durch verschiedene Personen ist. Die dem Ansatz zugrundeliegenden Ekman'schen Basisemotionen müssen mit Blick auf das Training der Software während der Annotation sehr eindeutig klassifizierbar sein. Jedoch sind Zeitzeugeninterviews ein hierfür schwieriger Gegenstand, da vielfach eher neutrale Erzählungen vorlagen, während eindeutige Emotionsklassen deutlich unterrepräsentiert waren. Die Annotatoren nutzen Kombinationen aus Emotionsklassen, um die von ihnen wahrgenommenen Emotionen zu beschreiben, da ihnen die Operationalisierung nach Ekman unterkomplex erschien. Zukünftige Forschungsprojekte sollten hieran anknüpfen und das Datenset durch die Auswahl der Segmente aus Zeitzeugeninterviews sowie deren Annotation eindeutiger gestalten.

Als hilfreiches Instrument hat sich die Sentimentanalyse erwiesen, die Wissenschaftlerinnen und Wissenschaftlern Hinweise liefern kann, an welchen Stellen in Zeitzeugeninterviews emotionale Äußerungen zu finden sein können. Das Sentiment hat das Potential als zusätzliche (Such-)Facette vor allem in Kombination mit anderen (inhaltlichen) Erschließungskriterien einen neuen Zugang zu erzählter Geschichte zu bieten. Notwendig sind hierzu weitere Forschungs- und Entwicklungsschritte. Insbesondere die Leistungsfähigkeit der zugrunde liegenden (deutschen) Sprachmodelle (etwa im Hinblick auf Kontextwissen, Interpretation von Metaphern, Bildern und abstrakten Umschreibungen) sowie mögliche Verbindungen mit anderen Erschließungskategorien sind zielführende Ansätze möglicher nächster Schritte, die im Rahmen der Pilotstudie deutlich wurden.

15 Siehe hierzu: KI-Prüfkatalog des Fraunhofer IAIS (Fraunhofer IAIS 2021). KI-basierte Systeme können ungerechtfertigte, problematische oder sogar diskriminierende Entscheidungen treffen – was vielfach mit unzureichenden Trainingsdaten zusammenhängt. Siehe hierzu Beck (2019).

16 Zum Thema Forschungsdaten und Oral History siehe Apel et al. 2022: 213 ff.

Das interdisziplinäre Projekt zwischen Historikerinnen, Historikern und Softwareentwicklungsingenieuren hat gezeigt, dass die Nutzung automatisierter Verfahren die klassischen Regeln der historischen Quellen- und Methodenkritik nicht außer Kraft setzt. Wer softwaregestützt erschließt oder auswertet, sollte Vorteile und Fußangeln dieser Verfahren kennen, um gegebenenfalls ergänzende Methoden heranzuziehen. Angesichts der riesigen Datenmengen und Zukunftsfragen, die sich uns stellen, müssen auch Geisteswissenschaftler es wagen, (software-)technische Aspekte vertieft in den Blick zu nehmen und sich an den Debatten und Entwicklungen neuer Technologien beteiligen. Für die Stiftung Haus der Geschichte der Bundesrepublik war das Projekt ein Ausgangspunkt, um sich grundlegend die Frage zu stellen, wie Künstliche Intelligenz die Erschließung der Sammlungen verändern wird.

## Literatur

Althaus, Andrea und Linde Apel (2023): Oral History. Version: 1.0, in: Docupedia-Zeitgeschichte, 28.03.2023. https://dx.doi.org/10.14765/zzf.dok-2478

Apel, Linde (2015): Oral History reloaded. Zur Zweitauswertung von mündlichen Quellen, in: Westfälische Forschungen, Zeitschrift des LWL-Instituts für westfälische Regionalgeschichte, 65, 243-254.

Apel, Linde, Almut Leh und Cord Pagenstecher (2022): Oral History im digitalen Wandel. Interviews als Forschungsdaten, in: Linde Apel (Hg.): Erinnern, erzählen, Geschichte schreiben. Oral History im 21. Jahrhundert, Forum Zeitgeschichte, Bd. 29, Berlin: Metropol, 193-222.

Beck, Susanne et al. (2019): Künstliche Intelligenz und Diskriminierung: Herausforderungen und Lösungsansätze. Whitepaper aus der Plattform Lernende Systeme. Online: https://www.plattform-lernende-systeme.de/publikationen-details/kuenstliche-intelligenz-und-diskriminierung-herausforderungen-und-loesungsansaetze.html (5.2.2024).

Darwin, Charles (1872): The Expression of the Emotions in Man and Animals. London: John Murray. https://doi.org/10.1037/10001-000

De Jong, Steffi (2018): The Witness As Object. Video Testimony in Memorial Museums, Museums and Collections, Bd. 10, New York: Berghahn. https://doi.org/10.2307/j.ctv3znzsd

Devlin, Jacob, Ming-Wei Chang, Kenton Lee und Kristina Toutanova (2019): BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies, Volume 1 (Long and Short Papers), 4171-4186.

Ekman, Paul (1971): Universals and Cultural Differences in Facial Expressions of Emotion. In: James Cole (Hg.): Nebraska Symposium on Motivation. Cultural Psychology, Bd. 19, Lincoln: University of Nebraska Press, 207-282.

Ekman, Paul und Erika L. Rosenberg (Hg.) (2020): What the Face reveals. Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS), Oxford: Oxford University Press.

Ekman, Paul und Wallace Friesen (1978): Facial Action Coding System. A Technique for the Measurement of Facial Movement, Palo Alto: Consulting Psychologists Press California. https://doi.org/10.1037/t27734-000

Euler, Harald A. (2000): Evolutionstheoretische Ansätze. In: Jürgen H. Otto, Harald A. Euler, Heinz Mandl (Hg.): Emotionspsychologie. Ein Handbuch, Weinheim: Beltz, Psychologie-Verlags-Union, 45-63.

Feldman Barrett, Lisa (2017a): How Emotions are made. The secret life of the brain, London: MacMillan.

Feldman Barrett, Lisa (2017b): The theory of constructed emotion: an active inference account of interoception and categorization. In: Social Cognitive and Affective Neuroscience, 12, Heft 1, 1-23. https://doi.org/10.1093/scan/nsx060

Feldman Barrett, Lisa (2018): Can Machines Perceive Emotion?, Dr. Lisa Feldmann Barrett, Talks at Google. Online: https://www.youtube.com/watch?v=HlJQXfL_GeM (6.2.2024).

Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS (Hg.) (2021): Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz. KI-Prüfkatalog. Online: https://www.iais.fraunhofer.de/de/forschung/kuenstliche-intelligenz/ki-pruefkatalog.html (6.2.2024).

Frevert, Ute (2020): Mächtige Gefühle. Von A wie Angst bis Z wie Zuneigung, Deutsche Geschichte seit 1900, Frankfurt am Main: S. Fischer.

Gammerl, Benno (2021): Anders fühlen. Schwules und lesbisches Leben in der Bundesrepublik. Eine Emotionsgeschichte, München: Carl Hanser Verlag.

Gref, Michael, Nike Matthiesen, Sreenivasa Hikkal Venugopala, Shalaka Satheesh, Aswinkumar Vijayananth, Duc Bach Ha, Sven Behnke und Joachim Köhler (2022a): 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), Marseille: European Language Resources Association, 2022-2031.

Gref, Michael, Nike Matthiesen, Christoph Schmidt, Sven Behnke und Joachim Köhler (2022b): Human and Automatic Speech Recognition Performance on German Oral History Interviews. In: ArXiv PrePrint arXiv:2201.06841 (eess.AS). Online: https://arxiv.org/abs/2201.06841 (6.2.2024).

Gref, Michael und Nike Matthiesen (2022): Erkennung wahrgenommener Emotionalität mit Künstlicher Intelligenz in audiovisuellen Zeitzeugeninterviews. In: Lisa Dieckmann, Bettina Pfleging, Georg Schelbert und Thorsten Wübbena (Hg.): 4D, Dimensionen, Disziplinen, Digitalität, Daten. Tagungsband zur prometheus-Jubiläumstagung 2021, Computing in Art and Architecture, Bd. 6, Heidelberg: Universität Heidelberg, Universitätsbibliothek.

Hitzer, Bettina (2011): Emotionsgeschichte – ein Anfang mit Folgen. In: H-Soz-Kult, 23.11.2011. Online: https://www.hsozkult.de/literaturereview/id/fdl-136824 (6.02.2024).

Köhler, Joachim, Michael Gref und Almut Leh (2017): KA³. Weiterentwicklung von Sprachtechnologien im Kontext der Oral History, in: BIOS – Zeitschrift für Biographieforschung, Oral History und Lebensverlaufsanalysen, 30, Heft 1-2, 43-59. https://doi.org/10.3224/bios.v30i1-2.05

König, Mareike (2017): Digitale Methoden in der Geschichtswissenschaft. Definitionen, Anwendungen, Herausforderungen, in: BIOS – Zeitschrift für Biographieforschung, Oral History und Lebensverlaufsanalysen, 30, Heft 1-2, 7-21. https://doi.org/10.3224/bios.v30i1-2.02

Leh, Almut (2021): Digitale Zeitzeugenschaft – Wenn Algorithmen das digitale Gedächtnis übernehmen. Erfahrungen mit künstlicher Intelligenz im Archiv „Deutsches Gedächtnis“, in: Archivpflege in Westfalen-Lippe, 95, 7-12.

Liu, Bing (2015): Sentiment Analysis. Mining Opinions, Sentiments, and Emotions, Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139084789

Mees, Ulrich (2006): Zum Forschungsstand der Emotionspsychologie – eine Skizze. In: Rainer Schützeichel (Hg.): Emotionen und Sozialtheorie. Disziplinäre Ansätze, Frankfurt am Man, New York: Campus, 104-124.

Merten, Jörg (2003): Einführung in die Emotionspsychologie. Stuttgart: Kohlhammer.

Misselhorn, Catrin (2021): Künstliche Intelligenz und Empathie. Vom Leben mit Emotionserkennung, Sexrobotern & Co., Reclams Universal-Bibliothek, Nr. 14047, Ditzingen: Reclam.

Pessanha, Francisca und Almila Akdag Salah (2022): A Computational Look at Oral History Archives. In: Journal on Computing and Cultural Heritage, 15, Heft 1, 1-16. https://doi.org/10.1145/3477605

Petschow, Annabelle und Markus Würz (2022): Zeitzeugen in zeithistorischen Ausstellungen. In: Stiftung Haus der Geschichte der Bundesrepublik Deutschland (Hg.): Zeithistorische Ausstellungen. Rück- und Ausblick, Bielefeld, Berlin: Kerber, 288-297.

Picard, Rosalind (1997): Affective Computing. Cambridge: The MIT Press. https://doi.org/10.1037/e526112012-054

Plamper, Jan (2012): Geschichte und Gefühl. Grundlagen der Emotionsgeschichte, München: Siedler.

Plamper, Jan (2013): Vergangene Gefühle. Emotionen als historische Quelle, in: Aus Politik- und Zeitgeschichte, 63, Heft 32/33, 12-19.

Poria, Soujanya, Erik Cambria, Rajiv Bajpai und Amir Hussain (2017): A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion. In: Information Fusion, 37, 98-125. https://doi.org/10.1016/j.inffus.2017.02.003

Range, Thomas (2019): Mensch und Maschine. Wie künstliche Intelligenz und Roboter unser Leben verändern, Reclams Universal-Bibliothek, Nr. 14495, Ditzingen: Reclam.

Russell, James und Lisa Feldmann Barrett (1999): Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. In: Journal of personality and social psychology, 76, Heft 5, 805-19. https://doi.org/10.1037//0022-3514.76.5.805

Shannon, Claude E. und Warren Weaver (1949): The Mathematical Theory of Communication. Urbana: University of Illinois.

Tsai, Yao-Hung Hubert, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency und Ruslan Salakhutdinov (2019): Multimodal Transformer for Unaligned Multimodal Language Sequences. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 6558-6569. https://doi.org/10.18653/v1/P19-1656

Viswanath, Anargh (2023): Multi-modal Emotion Categorization in Oral History Interviews. Masterthesis, Hochschule Bonn-Rhein-Sieg, Fachbereich Informatik. https://doi.org/10.24406/publica-1893

Wierling, Dorothee (2003): Oral History. In: Michael Maurer (Hg.): Aufriß der historischen Wissenschaften. Bd. 7: Neue Themen und Methoden der Geschichtswissenschaft, Reclams Universal-Bibliothek, Nr. 17033, Stuttgart: Reclam, 81-151.

Zadeh, Amir, Yansheng Cao, Simon Hessner, Paul Pu Liang, Soujanya Poriaund und Louis-Philippe Morency (2020): CMU-MOSEAS: A Multimodal Language Dataset for Spanish, Portuguese, German and French. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1801-1812. https://doi.org/10.18653/v1/2020.emnlp-main.141

## Zusammenfassung

Historikerinnen und Historiker der Stiftung Haus der Geschichte der Bundesrepublik Deutschland und Informatiker des Fraunhofer Instituts für Intelligente Analyse- und Informationssysteme haben zwei Jahre lang in einem interdisziplinären Forschungsprojekt einen Softwareprototypen entwickelt, der mit Hilfe von Verfahren Künstlicher Intelligenz in audiovisuellen Zeitzeugeninterviews wahrnehmbare Emotionen und wertende Meinungsäußerungen (Sentiment) erkennt. Das Projekt wurde im Zeitraum 2020 bis 2022 finanziell gefördert im Rahmen der KI-Strategie der Bundesregierung. Der Beitrag thematisiert das Forschungsdesign, dessen Operationalisierung sowie daraus ersichtliche Chancen und (derzeitige) Grenzen der automatisierten Analyse von Zeitzeugeninterviews im Kontext der softwareunterstützten Erschließung von (musealen) Sammlungen.