A Study on the Ambiguity in Human Annotation of German Oral History Interviews for Perceived Emotion Recognition and Sentiment Analysis

Michael Gref[1], Nike Matthiesen[2], Sreenivasa Hikkal Venugopala[1,3], Shalaka Satheesh[1,3], Aswinkumar Vijayananth[1,3], Duc Bach Ha[1], Sven Behnke[1,4], Joachim Köhler[1]

[1]Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS)
[2]Haus der Geschichte der Bundesrepublik Deutschland Foundation (HdG)
[3]University of Applied Sciences Bonn-Rhein-Sieg
[4]Autonomous Intelligent Systems (AIS), Computer Science Institute VI, University of Bonn
Contact: michael.gref@iais.fraunhofer.de

Introduction

- Deep learning can help to make large audiovisual archives more accessible
- Prominent example in recent years: ASR for transcription of oral history interviews
- Beyond ASR: Sentiment analysis and emotion recognition can help capture, categorize and make different facets of interviews searchable
- However, humans perceive sentiments and emotions ambiguously and subjectively. Moreover, oral history interviews have multi-layered levels of complex facets of emotions
- To what degree humans and machines capture and assign these facets into commonly predefined classes? What are reasons for the ambiguity in human perception?

The HdG *Zeitzeugenportal* Corpus

Zeitzeugenportal (Portal of Oral History)

- `https://www.zeitzeugen-portal.de`
- German online service by *Haus der Geschichte* (House of the History) Foundation (HdG)
- Offers a central collection of contemporary German oral history interviews
- Currently hosts more than 8,000 clips from around 1,000 interviews

Project: Multi-Modal Mining for Oral History

- Goal: Investigation of complex search modalities for indexing oral history interviews
- Focus not only on *what* is being told but also on *how* it is being told
- Automated analysis can help better understand the role emotions play in historical remembering
- Recognition of perceived emotion

Table 1: Overview of HdG oral history data sets after annotation and split into speaker-independent subsets.

HdG Set	Videos	Segments	Hours
Training	104	1,863	6.35
Development	27	430	1.44
Test	33	471	1.74

Annotation of Perceived Emotions and Sentiment in Oral History

- Annotation of perceived emotion and sentiment by three employees at the *HdG*, who have an academic background in history
- The annotators received pre-segmented interview videos based on ASR, i.e., video-stream including audio and a raw ASR transcript

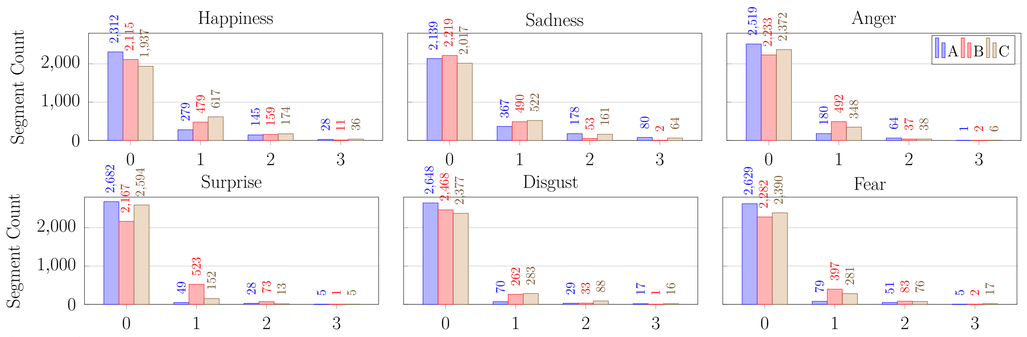


Figure 1: Histograms of the annotation scores for each emotion. Each color bar represents a different annotator.

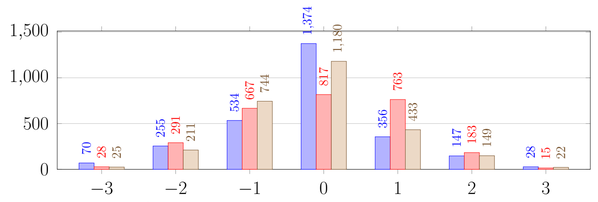


Figure 2: Sentiment annotation score histograms.

Correlation Analysis of Annotation Pairs

Table 2: Spearman rank-order correlation coefficients between the annotated labels of two transcribers.

	Transcriber Pairs			
Class	**A–B**	**A–C**	**B–C**	**Avg.**
sentiment	0.66	0.61	0.61	0.63
happy	0.52	0.52	0.60	0.55
sad	0.45	0.52	0.44	0.47
anger	0.29	0.35	0.36	0.33
surprise	0.14	0.26	0.19	0.20
disgust	0.31	0.32	0.38	0.34
fear	0.36	0.38	0.41	0.38

Inter-Class Correlation Analysis

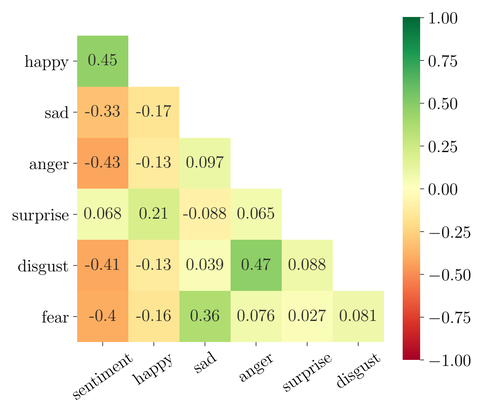


Figure 3: Spearman correlation between the annotated average scores of classes.

Qualitative Survey of Annotators

- Narrative structure of oral history interviews has different levels. Accordingly, emotions become visible in different ways, such as those that arise during remembering or reported emotional situations.
- Annotators agreed: Ekman classes are insufficient to reflect the complexity of emotions in oral history interviews. Nuances of emotions do not fit into the six categories.
- Annotators intuitively combined multiple emotions to represent more complex emotions, such as hate (disgust + anger), despair/helplessness (fear + sadness), scorn (happiness + disgust), and overwhelm (happiness + surprise) in the annotation.
- Examples: *Overwhelm* was identified in some interviews about the Fall of the Berlin Wall. Disgust + anger occurred more frequently in narratives reporting oppression or persecution.

Text-Based Sentiment and Emotion Recognition

- BERT embeddings + DNN classifier
- Multi-staged training with German CMU-MOSEAS and HdG data set
- Input: Raw ASR transcripts
- Class weights are estimated using the compute-class-weights (from sklearn) to handle class imbalance

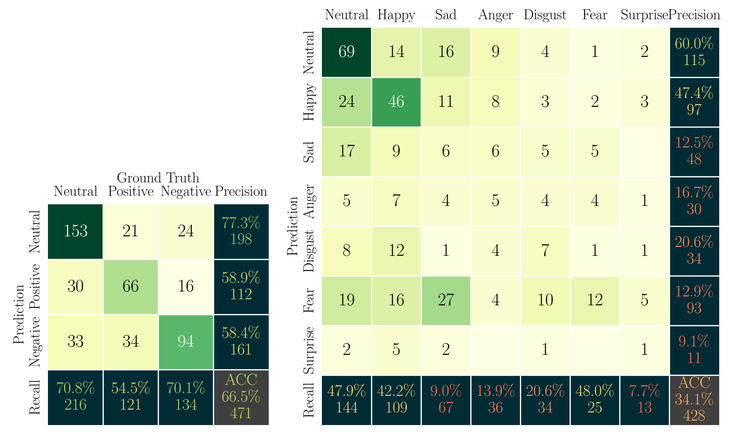
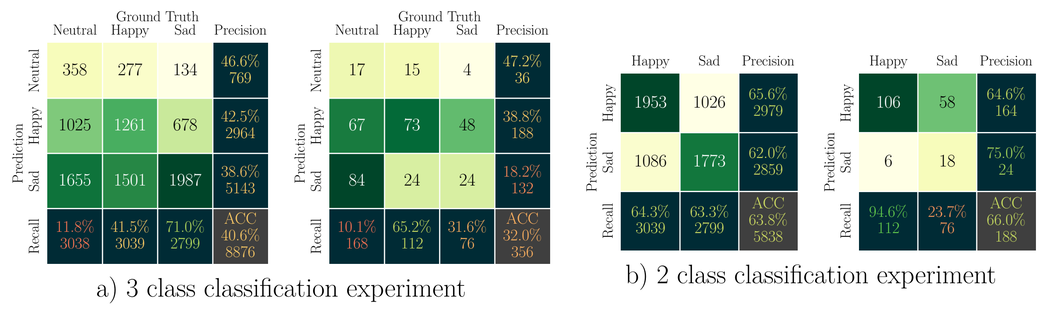


Figure 4: Confusion matrix of the text-based sentiment analysis (left) and emotion recognition model (right) on HdG test.

Speech Emotion Recognition

- Hybrid model: SVM + embeddings from a pretrained VGG-19 model (ImageNet)
- Input: Grayscale log-melspectrograms
- Combined training set: HdG, CMU-MOSEAS (German part), CMU-MOSEI, Berlin EmoDB



a) 3 class classification experiment

b) 2 class classification experiment

Figure 5: Different classification experiments for SER models. For each experiment, the training results are on the left. The results on the HdG test set are on the right.

Facial Emotion Recognition

- Frame Attention Networks (FAN) model (ResNet-18 feature embedding module and a subsequent frame attention network)
- Pre-training with Microsoft FER+ and AFEW data sets; fine-tuning with HdG

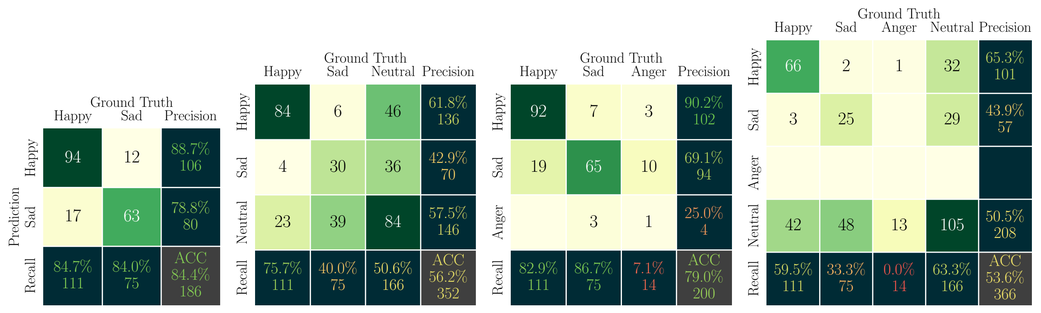


Figure 6: Confusion matrices from results of the FER experiments.

Summary and Conclusion

- Comparing the annotations of three persons using Ekman classes commonly used in emotion recognition revealed substantial differences in human perception for oral history interviews.
- Annotators combine different emotion classes in the annotation to represent the complex emotions of oral history interviews.
- Perceptual ambiguity and other challenges, such as class imbalance and lack of training data, currently limit the opportunities of emotion recognition for oral history archives. The systems can only differentiate fundamental emotions in our interviews, such as happiness and sadness.
- Nonetheless, our work uncovers promising observations and possibilities for further research.

13th International Conference on Language Resources and Evaluation (LREC), 20–25 June 2022, Marseille, France