Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2022–2031
Marseille, 20-25 June 2022

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), Germany
[2]Haus der Geschichte der Bundesrepublik Deutschland Foundation (HdG), Bonn, Germany
[3]University of Applied Sciences Bonn-Rhein-Sieg, Germany
[4]Autonomous Intelligent Systems (AIS), Computer Science Institute VI, University of Bonn, Germany
{michael.gref, sreenivasa.hikkal.venugopala, shalaka.satheesh, aswinkumar.vijayananth,
duc.bach.ha, sven.behnke, joachim.koehler}@iais.fraunhofer.de, matthiesen@hdg.de

Abstract

For research in audiovisual interview archives often it is not only of interest what is said but also how. Sentiment analysis and emotion recognition can help capture, categorize and make these different facets searchable. In particular, for oral history archives, such indexing technologies can be of great interest. These technologies can help understand the role of emotions in historical remembering. However, humans often perceive sentiments and emotions ambiguously and subjectively. Moreover, oral history interviews have multi-layered levels of complex, sometimes contradictory, sometimes very subtle facets of emotions. Therefore, the question arises of the chance machines and humans have capturing and assigning these into predefined categories. This paper investigates the ambiguity in human perception of emotions and sentiment in German oral history interviews and the impact on machine learning systems. Our experiments reveal substantial differences in human perception for different emotions. Furthermore, we report from ongoing machine learning experiments with different modalities. We show that the human perceptual ambiguity and other challenges, such as class imbalance and lack of training data, currently limit the opportunities of these technologies for oral history archives. Nonetheless, our work uncovers promising observations and possibilities for further research.

Keywords: emotion recognition, sentiment analysis, language, oral history, speech emotion recognition, facial emotion recognition, annotation, ambiguity

1. Introduction

Oral history archives are often large audiovisual data repositories composing numerous interviews of contemporary witnesses to historical events. Deep learning can help to make these archives more accessible quantitatively and qualitatively. A prominent example in recent years is automatic speech recognition for transcription of oral history interviews.

However, many potential deep learning applications for oral history interviews are still untapped. In a recent survey, Pessanha and Salah (2022) discuss potential applications of computational technologies for oral history archives. Among other aspects, the authors point out the potential benefits of these technologies in understanding the changes in emotions during remembering, storytelling, and conversing. In conjunction with the transcriptions, researchers can better infer not only *what* is being said but also *how*. In a recent research project, we study sentiment analysis and emotion recognition for German oral history interviews as the foundation for such complex search and indexing approaches for archives.

Various challenges arise when transferring research results to real-world, "in-the-wild" applications. As with many AI-based approaches, suitable representative training data of adequate scale is one of the key challenges. For natural data sets, the current gold standard for data annotation is to use other people's perception of emotion as the learning target, cf. (Schuller, 2018). However, even the annotation is a significant challenge since it is ambiguous and subjective. Emotions actually felt by the recorded persons and the emotions perceived by annotators may differ—and human recognition rates usually do not exceed 90 %, cf. Akçay and Oğuz (2020).

We assumed this value to be an upper limit. Dupré et al. (2020) compared the emotion recognition performance of humans and eight commercial systems using facial expression videos. The experiments show an overall human recognition accuracy of 72 % for the six basic Ekman emotions classes (Ekman et al., 1980). A 48–62 % range in recognition accuracy was observed for eight different tested commercial systems. The authors found that the machines' accuracy was consistently lower for spontaneous expressions.

Krumhuber et al. (2020) studied human and machine emotion recognition using fourteen different dynamic facial expressions data sets—nine with posed/acted and five with spontaneous emotions. For posed emotions, they report mean human recognition scores of roughly 60–80 %. However, for the spontaneous five corpora, the mean human scores are roughly 35–65 %.

Human-annotated training data is the crucial building block for emotion recognition and sentiment analysis machine learning systems. However, since the human

perception of spontaneous emotions is ambiguous and subjective, we address this issue for oral history interviews in this paper. We investigate to what extent different persons perceive and annotate the emotions and sentiment of interviewees in oral history recordings differently, comparing annotations of three different persons on a German oral history data set. We believe this contributes to assessing the general capabilities of such approaches for oral history interviews. With initial experiments on three different modalities, we further study the influence of the annotation ambiguity and class imbalance for these tasks. We uncover challenges of sentiment analysis and emotion recognition for the oral history use case that need to be addressed in further research.

2. The HdG *Zeitzeugenportal*

Zeitzeugenportal[1] (Portal of Oral History) is a German online service by *Haus der Geschichte* (House of the History) Foundation (HdG) that offers a central collection of contemporary German oral history interviews. More than 8,000 clips from around 1,000 interviews can currently be found at *Zeitzeugenportal*.

2.1. *Multi-Modal Mining for Oral History*

In the research project *Multi-Modal Mining of German Oral History Interviews for the Indexing of Audiovisual Cultural Heritage*, *Fraunhofer IAIS* cooperates with HdG to investigate complex search modalities for indexing oral history interviews. These audiovisual interviews illustrate the individual processing of history and demonstrate the multiperspectivity of personal views and experiences. Emotions are an important factor in the memory process, so automated analysis can help better understand the role emotions play in historical remembering.

2.2. The *HdG* Oral History Data Set

We selected 10 hours of German oral history interviews from the HdG *Zeitzeugenportal* for our experiments. Our *HdG data set* comprises 164 different interview videos of 147 interviewees. The selected interviews were recorded between 2010 and 2020. Thus, the selection is representative of the more recent videos on the portal—including both professional and non-professional speakers. In addition, we aimed to represent different emotions in the selection and create a heterogeneous data set in terms of age and gender.

For prepossessing the HdG data set for annotation, we apply our current automatic transcription system Fraunhofer IAIS Audio Mining (Schmidt et al., 2016) with our latest robust broadcast ASR model to create a raw ASR transcript, including punctuation. We use the ASR result to chunk the interviews into short segments at the longest speech pauses until we obtain segments of 30 seconds or less.

[1] `https://www.zeitzeugen-portal.de`

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

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

Three employees at the *Haus der Geschichte*, who have an academic background in history, annotated the emotions and sentiment for the pre-segmented interview videos. At the same time, a reference transcription is obtained by correcting the ASR transcript. Details are presented in the following section. After the annotation, the HdG data set was split into speaker-independent training, development, and test subset for model training and evaluation, as presented in Table 1. The data set is not published and is only used in-house due to the General Data Protection Regulation and the personal rights of the interviewees.

3. Annotation of Perceived Emotions and Sentiment in Oral History

As discussed, human perceptions of emotion and sentiment are often subjective, ambiguous and may differ greatly from the speaker-internal state. We use the phrase *recognition of perceived emotion* to emphasize this issue and how machine learning systems are trained on such annotated data. Such systems merely reproduce human decoding competence that works with different levels to decode emotions: the verbal (text), para verbal (voice), and nonverbal (face) level, similar to the Shannon-Weaver Model of Communication (Shannon and Weaver, 1949).

3.1. Emotion and Sentiment Annotation

The annotators received the segmented interview videos, i.e., video-stream including audio and a raw ASR transcript. The annotation of emotion and sentiment is done in the same pass and on the same segments as the correction of ASR transcripts—all using the multi-modal input. We use the six Ekman classes (Ekman et al., 1980) for the raw emotion annotation: happiness, sadness, anger, surprise, disgust, and fear. This choice was made to ensure comparability to established data sets and under the assumption that the annotation for this emotion inventory would be more intuitive for non-experts and thus more reliable to annotate than more complex inventories. Per segment, the three annotators assign a score on a 4-point Likert scale from 0 (no perception) to 3 (strong) for each of the six emotion classes following (Zadeh et al., 2018). The annotation is done independently for each emotion class so that multiple emotions can appear in each segment to different degrees. Similar to the emotions, sentiment annotation is done on a Likert scale from -3 (very negative) to 3 (very positive).

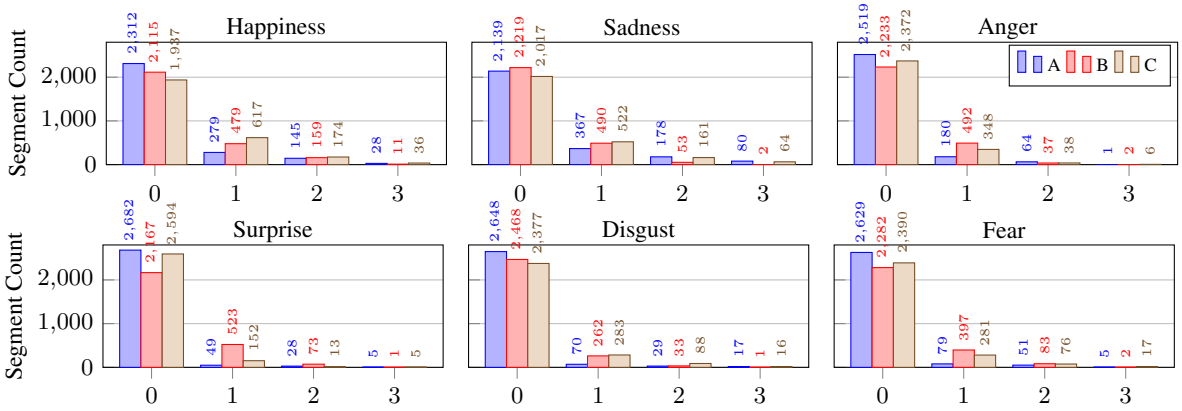

Figure 1: Histograms of the annotation scores for each emotion. Each color bar represents a different annotator.

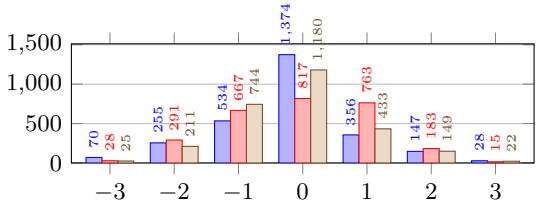
Figure 2: Sentiment annotation score histograms.

Class	Transcriber Pairs 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

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

Figure 1 shows the distribution of annotated scores per emotion class for each of the three annotators. The distribution follows a pattern expected for natural, real-world data: A neutral score dominates for all emotions. With increasing score, the segment-count decreases. Although emotions play a decisive role in remembering, contemporary witnesses are often very composed when narrating. Therefore, a strong expression of emotions is rare. Happiness and sadness are the most represented emotions in our data, surprise and disgust are the weakest.

Figure 2 presents the histogram of for the sentiment scores. As for emotions, the neutral score is most dominant. Unlike many other unstructured, real-world data sets, our data have negative sentiment more pronounced. This is likely due to the nature of the interviews: many interviews cover war and post-war experiences when Germany was divided into two states.

3.2. Correlation Analysis of Annotation Pairs

Table 2 shows the class-wise relationship between the annotation for each pair of annotators in terms of correlation. We use the Spearman rank-order correlation coefficient measuring the monotony instead of a linear relationship between two annotators. Overall, the values for each annotator pair are in a similar range of values with no strong outliers. Therefore, we assume no fundamentally different understanding of the task or approach to annotation for any of the three annotators.

Sentiment has the strongest correlation among all classes. Thus, the annotators seemed to have comparatively the same perceptions regarding the sentiment. However, the correlation is just above moderate, with a mean value of 0.63, indicating some substantial differences between all three annotators.

Emotion is often considered more ambiguous and subjective than sentiment, as evidenced by the systematically lower correlation of these classes. Thus, there seem to be greater differences in perception or interpretation of emotions in our interviews. Happiness and sadness have the highest correlation coefficient with 0.55 and 0.47, respectively. Even if there is no consensus, we assume a fundamental agreement in a sufficient number of segments.

The annotators seemed to have severely different perceptions for the other four emotion classes—with *surprise* having the lowest correlation. Since even two humans seem to have only a conditionally identical perception for these emotions, we hypothesize that this ambiguity in annotation severely limits the recognition performance for oral history interviews—at least using these predefined classes.

3.3. Inter-Class Correlation Analysis

In a further correlation analysis, we investigate the co-occurrence of different emotions and sentiment. We combine the three annotations for this analysis by taking the arithmetic mean for each segment. A correlation matrix for the different classes is shown in Figure 3. A moderate correlation exists between emotions and sentiment. In particular, happiness and a positive sentiment have a moderate correlation. An analogous correlation exists between negative sentiment and anger,

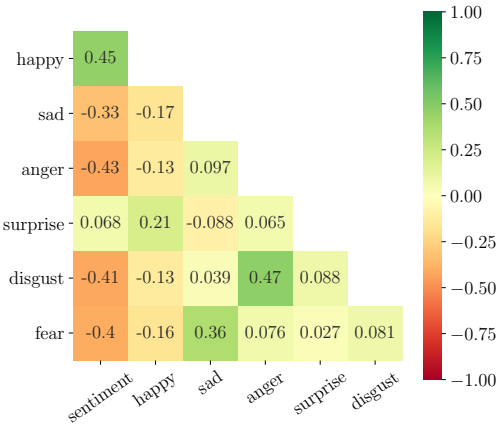

Figure 3: Spearman correlation between the annotated average scores of the different classes.

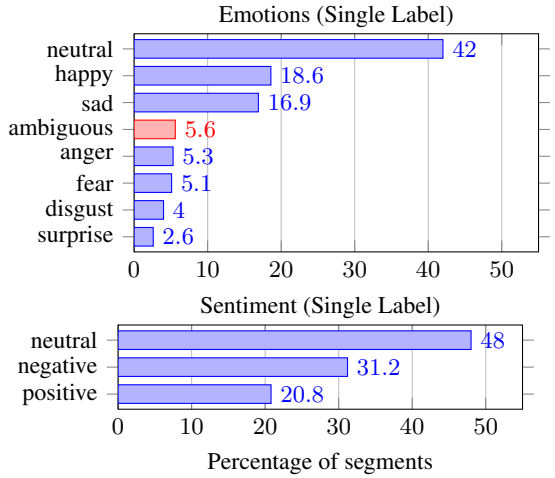

Figure 4: Percentages of emotion class and sentiment classes of whole data set after mapping to single-label.

disgust, and fear.

In most cases, the correlation between the emotion classes is in the range of coincidence. Exceptions are disgust with anger (0.47) and fear with sadness (0.36). Thus, these class pairs seem to occur together more than just by coincidence and could be of interest for a detailed, qualitative analysis of the oral history interviews. We illuminate possible causes for these correlations by surveying the annotators.

3.4. Qualitative Survey of Annotators

In a qualitative survey, the annotators reported various challenges. One challenge is that the 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. In terms of the different emotions, the annotators agreed that the given Ekman classes are insufficient to reflect the complexity of emotions in oral history interviews. Nuances of emotions do not fit into the six categories. Therefore, the persons 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. For example, overwhelm was identified as an important emotion in some interviews in which interviewees have talked about the Fall of the Berlin Wall. In combination, disgust and anger occurred more frequently in narratives reporting oppression or persecution.

3.5. Mapping to Single-Label Data

The mean scores of the raw annotations were to corresponding classes for classification-system training. Our goal for the initial experiments was to keep it simple and compatible with common data sets to better understand the effects of ambiguity during training. Therefore, we aim to classify only the most prevalent emotion (single-label). For this, we use the arithmetic mean of the three annotations and proceed as follows:

If the mean scores of all emotion classes are below 0.5, we assign this segment to *neutral*. This aims to consider only emotion classes with trustworthy annotation, where at least two of three annotators have given a score of 1 ($0.\overline{6}$ on average), or at least one person has given a score of 2 and above. For non-neutral segments, we choose the class with the highest score. In exceptional cases, if two or more classes have the same score above the threshold, we mark them as *ambiguous* and do not use them in the current training.

For the sentiment score s, we apply a similar threshold and mapping: *negative*, if $s \in [-3, -0.5]$; *positive*, if $s \in [0.5, 3]$; *neutral*, if $s \in (-0.5, 0.5)$.

Figure 4 shows the shares of each class in the entire HdG data set for both emotion and sentiment. Overall, the HdG data set is heavily imbalanced—both for the emotion and sentiment tasks.

In the following three sections, we report results of initial, ongoing experiments and analysis with machine learning systems on our HdG data set. Experiments are performed on three different modalities: Speech, facial expressions, and text.

4. Text-Based Sentiment and Emotion Recognition

Sentiment analysis deals with mining opinions from an observer's point of view on content such as videos, speech, text, images. Opinion mining can be classified as polarity mining (sentiment polarity mining) and emotion mining.

4.1. Related Work

Sentiment and emotion analysis have been explored in various fields ranging from neuroscience, psychology, to human-machine interface. Comprehensive survey papers (Pang and Lee, 2007; Kao et al., 2009; Salah et al., 2019; Norambuena et al., 2019; Hemmatian and Sohrabi, 2019; Yadollahi et al., 2017) cover various approaches and methods on sentiment

and emotion analysis tasks. Some of the comprehensive works on emotion analysis on text data are Mishne and de Rijke (2006; Rao et al. (2014; Alm et al. (2005; Neviarouskaya et al. (2007), and Gupta et al. (2013). These works mainly consider that a document or a sentence consists of a single emotion. Only a few approaches deal with multi-label emotion analysis tasks Bhowmick (2009; Liew and Turtle (2016; Khanpour and Caragea (2018), and Schoene (2020).

4.2. Methodology and Implementation

The pipeline used in our approach for sentiment analysis and emotion recognition starts with a BERT model to extract the embeddings from the tokenized text segments. We feed them into a classifier head consisting of two ReLU layers with a dropout layer in-between. We use the bert-base-german-cased pre-trained model[2] as our base model.

We apply a multi-stage training approach using the German part of the CMU-MOSEAS (Zadeh et al., 2020) data set, mapped to single-label, and the HdG data set for fine-tuning in subsequent stages. In the first stage, we use the German CMU-MOSEAS subset comprising 1,000 videos with 480 distinct speakers distributed across 250 topics and 10,000 annotated sentences. In the second stage, we fine-tune the model using the HdG data set. We use the raw ASR transcriptions from the pre-processing and not the human-corrected transcripts as inputs in the second stage. We aim to use the model as a subsequent analysis system after automatic transcription of large oral history data collections. On average, the HdG data set has a 16 to 17 % word error rate with our ASR system, cf. Gref et al. (2022). To handle the class imbalance issue, we estimate the class weights using the *compute class weights* function from the sklearn library that uses a heuristic inspired by logistic regression.

4.3. Results and Inference

The results of the sentiment and emotion classification are presented in Figure 5. Our approach achieves a decent accuracy on the HdG sets for sentiment, however, only a low accuracy for the emotion recognition. For the sentiment model, we see only a few segments confused between all polarities. However, the emotion recognition model has only learned to distinguish between neutral and happiness.

Interestingly, we observe a slightly increased misclassification of the actual class sadness with fear. As our previous correlation analysis of the human annotations in Figure 3 has shown, these two classes have an increased correlation. The reason is that the annotators often intuitively combine these two emotions to express other emotions such as despair or helplessness. Therefore, this misclassification may not be a system failure but a limitation of the single-label approach.

[2] https://huggingface.co/dbmdz/bert-base-german-cased

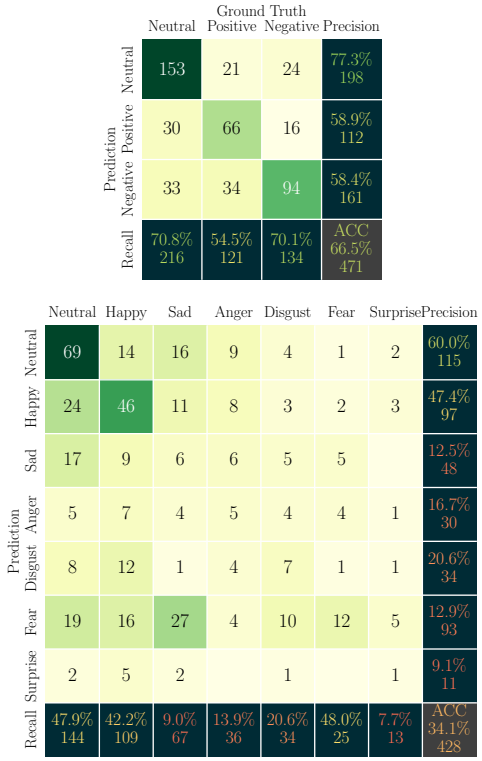

Figure 5: Confusion matrix of the text-based sentiment analysis (top) and emotion recognition model (bottom) on HdG test.

Overall, recognizing emotions from oral history interviews on text alone seems very limited. Nevertheless, interesting observations emerge that deserve further research. This research should reveal whether the poor performance is due to the character of the interviews, the heavy class imbalance in training, or the modality text not conveying emotions appropriately without additional modalities. It might also be that the main reason is the ambiguity of the human annotation we observed on our data. We observe very similar results with the other modalities recognizing seven emotion classes. Therefore, we investigate these modalities with a subset of the classes in the following sections to uncover fundamental problems.

The text-based sentiment analysis works well on our unstructured, imbalanced oral history data. As the data analysis in Section 3.3 indicated, there appeared to be a greater consensus among the three annotators on sentiment than emotion. This tendency seems to be confirmed by the experiment. In particular, we find it noteworthy that the classification works well given that we use raw, erroneous ASR transcripts as input to the model.

5. Speech Emotion Recognition

Speech emotion recognition (SER) is a branch of affective computing that deals with identifying and recognizing the emotional content in speech. One of the significant challenges in this field is identifying appropriate features in the speech signal that best represent

	No. of Samples	
Emotion	**Original**	**Balanced**
Neutral	9843	3039
Happy	3039	3039
Sad	800	2799

Table 3: Combined SER train data set before and after balancing with downsampling and data augmentation.

its emotional content.

5.1. Related Work

A detailed overview of SER is given, for example, by Ayadi et al. (2011), (Schuller, 2018), and (Akçay and Oğuz, 2020). Current approaches utilize convolutional neural network (CNN) or bidirectional recurrent neural network (biRNN) layers for SER—or combining both, such as Dai et al. (2019). The proposed method represents emotion in speech in an end-to-end manner (Zhang et al., 2017). Furthermore, this method focuses on only four categories of emotion: *anger, happy, sad*, and *neutral*, which are identified as the most discriminatory ones.

Further, (Li et al., 2019) propose a Diluted Residual Network (DRN) with multi-head self-attention. The authors employ Low-Level Descriptors (LLDs) extracted from the audio signal as input. (Wang et al., 2020) propose a model consisting of two jointly trained LSTMs: each of these models is separately used to process MFCC features and Mel-spectrograms. Both models predict an output class (emotion) that is averaged to arrive at the result. Some of the currently used techniques also use transfer learning to boost the performance (Akçay and Oğuz, 2020).

5.2. Methodology and Implementation

In this experiment, we train a hybrid model for SER, which combines traditional machine learning with deep learning. As for the text-based model, we utilize pretrained models and multiple data sets to cope with the lack of training data. We apply a VGG-19 model pretrained on the ImageNet data set and use log-Mel spectrograms treated as grayscale images as input features. The pre-trained VGG-19 model is first fine-tuned on the HdG training set. Then we use a combined data set to extract the embeddings from the fine-tuned VGG-19. These embeddings are used as input for the SVM model. The combined data set contains the HdG train set, the German part of CMU-MOSEAS (Zadeh et al., 2020), CMU-MOSEI (Zadeh et al., 2018), and *Berlin Emotional Database* (Berlin EmoDB) (Burkhardt et al., 2005). Except for CMU-MOSEI, an English data set, all other sets are German. Berlin EmoDB is the only set with acted emotions, whereas all other data sets have natural emotions. For data balancing, we apply data augmentation with 10 dB SNR additive white noise and downsampling the overrepresented. The distribution of the emotional classes in the combined train data set is presented in Table 3.

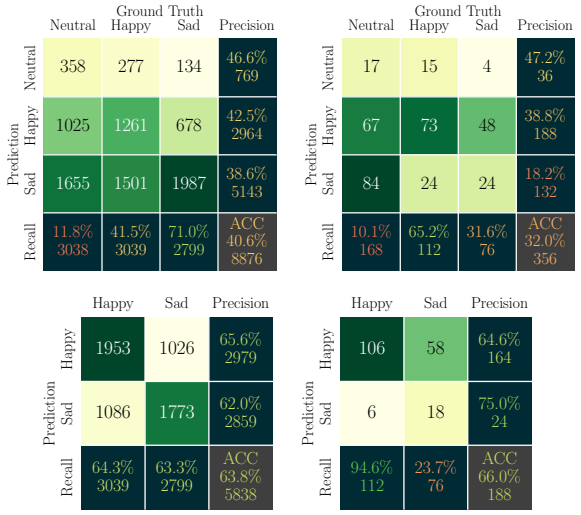

Figure 6: Confusion matrices for 3 class (top row) and 2 class (bottom row) classification of the SER models on the combined training set (left) the HdG test set (right).

As already shown for the text modality, we have not achieved satisfactory recognition performance on our data set so far with seven classes. Since the class *anger* consists of relatively fewer samples in the HdG data set, in this experiment, we only present results considering *happiness, sadness*, and *neutral*. This aims to assess the problems of training better.

5.3. Results and Inference

The model yields a training accuracy of 40.6 % and 32.0 % on the HdG test set for the three-class classification. The results are presented as a confusion matrix in the top row of Figure 6. We observe that the neutral class is often confused with other emotions for both the training and test set. The results of the text modality already indicated this. However, it becomes more substantial in this experiment for the audio modality with three classes. We hypothesize that the neutral class cannot be sufficiently differentiated from subtle emotions in natural speech, leading to confusion in training.

We conducted another experiment in which the neutral class is removed to investigate this issue further. For two-class classification (*happy* and *sad*) the accuracy improves to 63.8 % and 66.0 % for the training and test set, respectively. As shown at the bottom of Figure 6, removing the neutral class results in a structural improvement. However, this does not lead to happiness and sadness being distinguished substantially better for the test set. The high accuracy is mainly attributed to the class imbalance towards happiness. Still, the system favors the happiness class over sadness.

The same model was also tested on the Berlin EmoDB and returned 90.5 % accuracy. We attribute this high accuracy to the fact that this data set consists of acted emotions, unlike the HdG data set, which consists of

naturally occurring emotions. A subjective evaluation of the HdG samples shows that it is challenging to differentiate emotions based on audio samples alone. Thus, we hypothesize that particular attention might have been paid to other modalities, presumably facial expressions, while annotating the interviews.

6. Facial Emotion Recognition

Facial emotion recognition (FER) is the task of recognizing human emotions from facial expressions. The immense number of visual clues present in the human face to identify underlying emotions makes FER an integral part of many emotion recognition systems.

6.1. Related Work

FER methods are categorized based on two dimensions: traditional feature- vs. representation-learning-based, and static vs. dynamic methods. Traditional feature learning-based FER methods rely on handcrafted features. In contrast, representation-learning-based methods use a system such as a neural network to learn features from training data automatically. Dynamic methods utilize temporal relationships between frames of an input video while static ones treat every frame independently. Dynamic representation-learning approaches possess an inherent advantage and become potential candidates for further consideration.

To perform the task at hand, we shortlisted Meng et al. (2019; Kuo et al. (2018; Gera and Balasubramanian (2020; Savchenko (2021), and Kuhnke et al. (2020) based on factors such as performance on open-source FER data sets like CK+ (Lucey et al., 2010) and AFEW (Kossaifi et al., 2017), depth of the neural network used (determines the minimum amount of data required for training), and reproducibility of results claimed by authors. Out of the five, Frame Attention Networks (FAN) (Meng et al., 2019) is chosen for its state-of-the-art accuracy on CK+ (99 %) and AFEW (51.18 %) data sets, and its simple yet effective construction.

6.2. Methodology

The HdG videos are pre-processed using the d-lib based face detection and alignment modules to crop out and align faces. These sequences of images are used as inputs for the FAN. The FAN network architecture consists of a CNN-based feature embedding module (ResNet-18) and a subsequent frame attention network. Meng et al. (2019) offer three variants of FAN: baseline, self-attention, and the combination of self and relation attention. The authors report a slightly superior performance of the self-relation-attention variant over the other two. However, we currently use the baseline and self-attention variants due to their simple design, enabling us to better understand their work.

6.3. Design of Experiments

In addition to the challenges of emotion recognition in general and our HdG data set in particular, we hypothesize an additional, specific challenge for FER. Most of the frames in a typical interview video carry faces with neutral or a subtle version of a particular emotion. This adds additional difficulty for any classifier to assign the correct label to the video—especially when *neutral* is one of the possible target classes.

	HdG Training Samples				Test
Exp	**Happy**	**Sad**	**Anger**	**Neutral**	**Acc.**
1	318	316	-	-	84.4%
2	318	316	-	316	56.2%
3	318	316	107	-	79.0%
4	318	316	107	787	53.6%

Table 4: Train data split for the four different FER experiments. The numbers for the emotion classes refer to the total number of segments of the applied HdG train set. For Experiment 2, neutral was reduced from 787 to 316 samples for class balancing.

We conducted four different experiments by training and evaluating the classifiers with different numbers and choices for the target emotion classes to study the effects of class-wise data set imbalance on the model's performance. Table 4 summarizes the experimental setup and the results.

The first experiment was conducted with the already balanced pair of happy and sad classes, with an intent to study these classes' effect on the classifier's predictive performance. In the next experiment, the neutral class was included after under-sampling it to match the other two sizes. Whereas, for the third experiment, the under-represented anger class was added along with the "happy-sad" pair to understand the bias induced from the class imbalance. The final experiment was conducted with unchanged training data to evaluate the current state of the classifier's performance.

All experiments were conducted with both the baseline and self-attention variants of FAN. However, the results presented in the next section are limited to the baseline variant, which performed better in all of the conducted experiments. Models of both variants were pretrained with Microsoft FER+ (Barsoum et al., 2016) and AFEW (Kossaifi et al., 2017) data sets using transfer learning.

6.4. Results and Inference

Figure 7 shows the confusion matrices of the baseline variant of FAN in the four different experiments. The classifier exhibits a decent performance on the balanced pair of happy and sad classes in Experiment 1 with an overall accuracy of 84.4 %, proving its learning capacity. The high class-wise precision value indicates the model's discrimination capability on oral history interviews.

However, the overall accuracy of the classifier drops significantly to 56.3 % with the addition of the neutral class in Experiment 2. This strongly indicates the

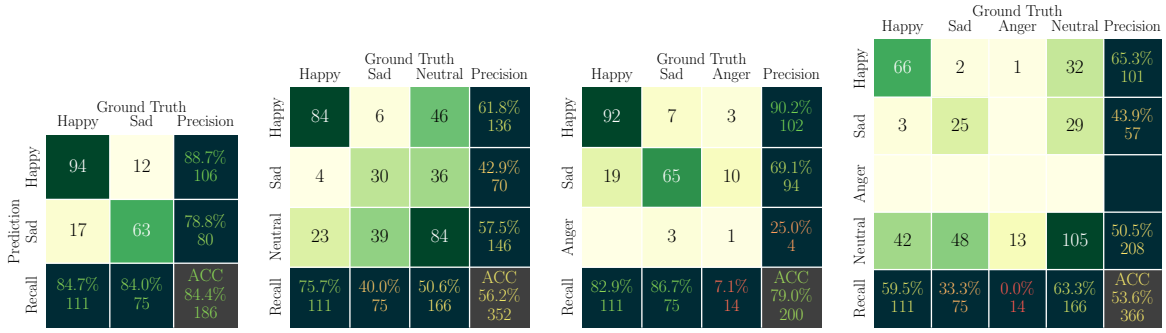

Figure 7: Confusion matrices from results of the FER experiments.

neutral class's detrimental effect on the model's performance. The introduction of the neutral class possibly affects the classifier's sensitivity to identify subtle emotions which make most of the frames in a video.

Unlike the neutral class, the under-represented anger class does not drastically reduce the classifier's accuracy. However, the model performs poorly on anger as it can correctly classify only one out of the fourteen test videos. This is certainly due to the insufficiency of anger in the training data. The model classifies most anger test videos as sad, presumably a relatively closer emotion to anger than happiness.

Both discussed effects from inclusions of neutral and anger classes in Experiment 2 and 3 can be observed in a combined fashion from the results of Experiment 4. The over-represented neutral class hampers the classifier from correctly recognizing even a single test video from the anger class.

7. Summary and Conclusion

This work investigated the ambiguity in human perception of emotions and sentiment in German oral history interviews. Comparing the annotations of three persons using Ekman classes commonly used in emotion recognition revealed substantial differences in human perception. While the annotators in our experiment have a reasonably consistent understanding of the two most common emotions, *happiness* and *sadness*, we found very little correlation for other emotions. Given the ambiguity of the human annotation using predefined emotions classes, we question whether practical learning for machines is even possible.

We further investigated co-occurrence of emotions in the annotation. An annotator survey revealed that Ekman classes were unanimously rated as insufficient for the complexity of multi-layered emotions in oral history interviews. The annotators intuitively combined different emotion classes to describe complex emotions not fitting in the predefined classes. This is reflected in an increased correlation of certain emotion classes, e.g., fear and sadness representing despair or helplessness. Hate was intuitively annotated as a combination of disgust and anger.

We also reported results from initial emotion recognition experiments for facial expressions, speech, and text. A facial emotion recognition system for oral history revealed the system could differentiate happiness and sadness in our interviews. However, adding a neutral class results in the system not being able to differentiate between the subtle emotions and the neutral class. This issue and the combination of emotions described earlier are limited by single-label training. In future work with oral history, multi-label training should be considered to account for these aspects.

So far in our experiments, speech emotion recognition is behind facial emotion recognition. Even differentiating between happiness and sadness based on the voice appears challenging. For sentiment analysis based on raw ASR transcripts, on the other hand, we were able to achieve decent accuracy for our unstructured data. This is also consistent with the human perception, which was highest for sentiment between annotators in our experiments.

In addition to the human ambiguity, other challenges currently limit the application of emotion recognition for oral history. In particular, we identified class imbalance and lack of representative training data as the current primary challenges. The application of pre-trained models, a combination of multiple natural data sets, and fine-tuning of models were essential in our work.

Overall, such indexing technologies for oral history archives seem to be quite limited so far. In oral history interviews, complex, subtle, and multi-layered emotions cannot yet be captured by our systems with the predefined, common classes. Perhaps fundamentally different approaches have to be chosen, e.g., limiting the indexing to recognizing specific patterns in human communication without interpreting them as emotions. However, users need to determine which patterns are relevant for their work in advance for meaningful application in archives. The results and observations of our work can provide initial impetus for this further research, which requires interdisciplinary collaboration between users of such archives and AI researchers.

8. Acknowledgments

The research project is funded by the German Federal Government Commissioner for Culture and Media.

9. Bibliographical References

Akçay, M. B. and Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116:56–76.

Alm, C. O., Roth, D., and Sproat, R. (2005). Emotions from text: Machine learning for text-based emotion prediction. In *Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 579–586. ACL.

Ayadi, M. E., Kamel, M. S., and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587.

Barsoum, E., Zhang, C., Ferrer, C. C., and Zhang, Z. (2016). Training deep networks for facial expression recognition with crowd-sourced label distribution. In *18th ACM International Conference on Multimodal Interaction (ICMI)*, pages 279–283.

Bhowmick, P. K. (2009). Reader perspective emotion analysis in text through ensemble based multi-label classification framework. *Computer and Information Science*, 2(4):64–74.

Dai, D., Wu, Z., Li, R., Wu, X., Jia, J., and Meng, H. (2019). Learning discriminative features from spectrograms using center loss for speech emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7405–7409, May.

Dupré, D., Krumhuber, E. G., Küster, D., and McKeown, G. J. (2020). A performance comparison of eight commercially available automatic classifiers for facial affect recognition. *PLOS ONE*, 15(4):e0231968.

Ekman, P., Freisen, W. V., and Ancoli, S. (1980). Facial signs of emotional experience. *Journal of Personality and Social Psychology*, 39(6):1125–1134.

Gera, D. and Balasubramanian, S. (2020). Affect expression behaviour analysis in the wild using spatio-channel attention and complementary context information. *arXiv preprint arXiv:2009.14440*.

Gref, M., Matthiesen, N., Schmidt, C., Behnke, S., and Köhler, J. (2022). Human and automatic speech recognition performance on german oral history interviews. *arXiv:2201.06841 [eess.AS]*.

Gupta, N. K., Gilbert, M., and Fabbrizio, G. D. (2013). Emotion detection in email customer care. *Computational Intelligence*, 29(3):489–505.

Hemmatian, F. and Sohrabi, M. K. (2019). A survey on classification techniques for opinion mining and sentiment analysis. *Artificial Intelligence Review*, 52(3):1495–1545.

Kao, E., Chieh Chun, L., Yang, T.-H., Hsieh, C.-T., and Soo, V.-W. (2009). Towards text-based emotion detection: A survey and possible improvements. In *International Conference on Information Management and Engineering (ICIME)*, pages 70 – 74.

Khanpour, H. and Caragea, C. (2018). Fine-grained emotion detection in health-related online posts. In Ellen Riloff, et al., editors, *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1160–1166. Association for Computational Linguistics.

Kossaifi, J., Tzimiropoulos, G., Todorovic, S., and Pantic, M. (2017). Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36.

Krumhuber, E. G., Küster, D., Namba, S., and Skora, L. (2020). Human and machine validation of 14 databases of dynamic facial expressions. *Behavior Research Methods*, 53(2):686–701.

Kuhnke, F., Rumberg, L., and Ostermann, J. (2020). Two-stream aural-visual affect analysis in the wild. In *15th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 600–605.

Kuo, C.-M., Lai, S.-H., and Sarkis, M. (2018). A compact deep learning model for robust facial expression recognition. In *IEEE conference on computer vision and pattern recognition workshops*, pages 2121–2129.

Li, R., Wu, Z., Jia, J., Zhao, S., and Meng, H. (2019). Dilated residual network with multi-head self-attention for speech emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6675–6679.

Liew, J. S. Y. and Turtle, H. R. (2016). Exploring fine-grained emotion detection in tweets. In *NAACL Student Research Workshop*, pages 73–80. The Association for Computational Linguistics.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 94–101.

Meng, D., Peng, X., Wang, K., and Qiao, Y. (2019). Frame attention networks for facial expression recognition in videos. In *IEEE International Conference on Image Processing (ICIP)*, pages 3866–3870.

Mishne, G. and de Rijke, M. (2006). Capturing global mood levels using blog posts. In *Computational Approaches to Analyzing Weblogs, Papers from the 2006 AAAI Spring Symposium, Technical Report SS-06-03*, pages 145–152.

Neviarouskaya, A., Prendinger, H., and Ishizuka, M. (2007). Textual affect sensing for sociable and expressive online communication. In Ana Paiva, et al., editors, *Affective Computing and Intelligent Interaction, Second International Conference (ACII)*, volume 4738 of *Lecture Notes in Computer Science*, pages 218–229. Springer.

Norambuena, B. K., Lettura, E. F., and Villegas, C. M. (2019). Sentiment analysis and opinion mining ap-

plied to scientific paper reviews. *Intell. Data Anal.*, 23(1):191–214.

Pang, B. and Lee, L. (2007). Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135.

Pessanha, F. and Salah, A. A. (2022). A computational look at oral history archives. *Journal on Computing and Cultural Heritage*, 15(1):1–16.

Rao, Y., Li, Q., Liu, W., Wu, Q., and Quan, X. (2014). Affective topic model for social emotion detection. *Neural Networks*, 58:29–37.

Salah, Z., Al-Ghuwairi, A., Baarah, A. H., Al-Oqaily, A. A., Qadoumi, B., Alhayek, M., and Alhijawi, B. (2019). A systematic review on opinion mining and sentiment analysis in social media. *Int. J. Bus. Inf. Syst.*, 31(4):530–554.

Savchenko, A. V. (2021). Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. *arXiv preprint arXiv:2103.17107*.

Schmidt, C., Stadtschnitzer, M., and Köhler, J. (2016). The Fraunhofer IAIS audio mining system: Current state and future directions. In *12th ITG Conference on Speech Communication*, pages 115–119. VDE / IEEE.

Schoene, A. M. (2020). Hybrid approaches to fine-grained emotion detection in social media data. In *34th Conference on Artificial Intelligence (AAAI)*, pages 13732–13733.

Schuller, B. W. (2018). Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61(5):90–99.

Shannon, C. E. and Weaver, W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press.

Wang, J., Xue, M., Culhane, R., Diao, E., Ding, J., and Tarokh, V. (2020). Speech emotion recognition with dual-sequence lstm architecture. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6474–6478.

Yadollahi, A., Shahraki, A. G., and Zaïane, O. R. (2017). Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys*, 50(2):25:1–25:33.

Zhang, S., Zhang, S., Huang, T., and Gao, W. (2017). Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Transactions on Multimedia*, 20(6):1576–1590.

10. Language Resource References

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., and Weiss, B. (2005). A database of German emotional speech. In *9th European Conference on Speech Communication and Technology (Eurospeech/Interspeech)*.

Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., and Morency, L.-P. (2018). Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL.

Zadeh, A. B., Cao, Y., Hessner, S., Liang, P. P., Poria, S., and Morency, L.-P. (2020). CMU-MOSEAS: A multimodal language dataset for spanish, portuguese, german and french. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL.