Embracing and exploiting annotator emotional subjectivity: an affective rater ensemble model

Lukas Stappen, Lea Schumann, Anton Batliner, Bjorn W. Schuller

Angaben zur Veröffentlichung / Publication details:

Stappen, Lukas, Lea Schumann, Anton Batliner, and Bjorn W. Schuller. 2021. "Embracing and exploiting annotator emotional subjectivity: an affective rater ensemble model." In *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), 28 September 2021 - 01 October 2021, Nara, Japan*, 1–8. Piscataway, NJ: Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/aciiw52867.2021.9666407.

Embracing and Exploiting Annotator Emotional Subjectivity: An Affective Rater Ensemble Model

Lukas Stappen†
University of Augsburg
Augsburg, DE
stappen@ieee.org

Lea Schumann†
University of Augsburg
Augsburg, DE

Anton Batliner
University of Augsburg
Augsburg, DE

Björn W. Schuller
Imperial College London
London, UK

***Abstract*—Automated recognition of continuous emotions in audio-visual data is a growing area of study that aids in understanding human-machine interaction. Training such systems presupposes human annotation of the data. The annotation process, however, is laborious and expensive given that several human ratings are required for every data sample to compensate for the subjectivity of emotion perception. As a consequence, labelled data for emotion recognition are rare and the existing corpora are limited when compared to other state-of-the-art deep learning datasets. In this study, we explore different ways in which existing emotion annotations can be utilised more effectively to exploit available labelled information to the fullest. To reach this objective, we exploit individual raters' opinions by employing an ensemble of rater-specific models, one for each annotator, by that reducing the loss of information which is a byproduct of annotation aggregation; we find that individual models can indeed infer subjective opinions. Furthermore, we explore the fusion of such ensemble predictions using different fusion techniques. Our ensemble model with only two annotators outperforms the regular Arousal baseline on the test set of the MuSe-CaR corpus. While no considerable improvements on valence could be obtained, using all annotators increases the prediction performance of arousal by up to .07 Concordance Correlation Coefficient absolute improvement on test – solely trained on rate-specific models and fused by an attention-enhanced Long-short Term Memory-Recurrent Neural Network.**

***Index Terms*—emotion recognition, annotation optimisation, ensemble models**

I. Introduction

Face-to-face communication has multiple layers alongside with the spoken word. We also react with gesture, postures, and facial expression, conveying emotions which are central to building interpersonal relationships. Nowadays, interaction takes often place via digital channels, both interpersonal and with smart devices. When interacting with machines, the information of these additional layers are lost. However, automatically recognising the expressed emotions, enables to develop products that improve our everyday lives. For example, attention detection when driving a vehicle can prevent accidents and children with autism can easier learn to interact with their surroundings. To this end, models have to be trained on large amounts of (multimodal) data, such as image, audio, or spoken language, to mimic a human-labelled assessment of the emotion being displayed.

†These authors contributed equally.

It is a well-known fact that human emotion perception varies among individuals [1] and is predominantly subjective. In other words, a real 'ground-truth' is simply impossible for systems learning emotions [2]. Furthermore, previous studies have found that human labelling is highly influenced by several factors including environmental distractions, personal bias, and task difficulty [2], [3]. As a result, annotator disagreement is usually higher with dimensional emotion labels than with categorical ones [4].

To incorporate the possible variability in annotator perception and performance, several emotional ratings are usually aggregated to a 'gold standard', and later used as a training target for emotion recognition systems. The emotion gold standard is generally established by around five raters, also referred to as annotators, since agreement increases strongly until this point, while more showed much smaller improvement effects by disproportional increasing costs [5]. It is argued [6] that higher rater disagreement implies that the sample is of less value.

Furthermore, the need for trained raters that possess a cultural understanding of the dataset context and the increased number of ratings necessary, makes the annotation process expensive. As a consequence, the multimodal datasets for time-continuous affective computing are currently either small or scarcely annotated. For example, the Automatic Sentiment Analysis in the Wild (SEWA) database [7] is designed to provide over 33 hours of audio-visual data for emotion research in-the-wild. However, the annotations are only available for 14 % of the data. In this sense, the recently introduced MuSe-CaR dataset [8] compromises of over 40 hours of annotated YouTube videos and is the most extensively annotated multimodal sentiment analysis database [9] for such tasks. However, when compared to large-scale labelled video datasets, e. g. , the YouTube-8M, the aforementioned corpora are still small, considering that such a dataset contains up to 350,000 hours of audio-visual content [10].

Given that data are sparse, the deliberate loss of annotation information through gold-standard fusion does not seem to be an optimal solution. We argue that even disagreeing annotations may capture an alternative 'true' annotation that could be more fully utilised. Rather than ignoring disagreeing

ratings, we adopt the use of raw annotation signals to train individual, annotator-specific models, each corresponding to one annotator, to predict annotator-specific emotions. Further, we fuse the predictions to test whether the consideration of different rater opinions benefits the predictive performance.

Finding ways to deal with subjectivity in annotations is not new within the machine learning domain. Soft labels are frequently used, for example, to take into account the reviewer subjectivity for the task of classifying whether a conference paper will be accepted or not [11]. Similarly, Fayek et al. attempts to represent an annotator by an individual model using a combination of soft labels and model ensembling [12]. Further examples can be found for machine translation [13] and computer vision [14]. While giving an overview of recent avenues of subjectivity in the field of affective computing, Rizos and Schuller denoted the prospects that lie in this uncertainty [15]. The authors argue that subjectivity can be utilised as an additional and quantifiable information that can create risk-aware systems. Subjectivity also provides an understanding of the confidence in emotion recognition predictions [16], [17]. As far as we are aware, within the context of continuous emotion ratings, individual annotators are yet to be modelled explicitly.

In the context of continuous emotion recognition, considering the annotator subjectivity seems of particular interest due to the complexity of such annotations. Annotators tend to exhibit dynamically varying time-delay in their annotations [18]. Moreover, time-varying annotator disagreement arises systematically because the perceived emotional content can exhibit some degree of inherent ambiguity [19]. Hereby, a spatio-temporal alignment of the individual annotations is important for generating a distortion-reduced gold standard. Different methods have been proposed to come up with a signal which ideally represents a consensus among the raters. Many of them are based on different similarity metrics including correlation coefficients or dynamic time warping distance [7], [20]. Others highlight the individual ratings and assign weights to them based on their agreement (e. g. , Evaluator Weighted Estimator (EWE) [6], [21]). A further challenge is to improve the raw signals ahead of the fusion. Martinez et al. , applied smoothing and further combined them across annotators using a moving median filter with a window size of 500 ms and shift of 1/59 seconds [22]. Ringeval et al. used median filtering with a window width of three samples before creating a single gold standard using EWE [21]. In a 2018 emotion challenge aiming at improving the gold-standard [23], Wang et al. argued that secondary, slight errors in annotations can be removed by a moving average filter [24]. While testing three different filtering techniques to smooth annotation data (Savitzky-Golay filter, moving average filter, and median filter) Thammasan et al. found that the moving average filter is more practical to enhance emotion recognition performance [25].

Guided by previous work, we seek for a better understanding of each annotator's 'emotional value' and the effective utilisation of individual annotations. We intend to answer the following research questions:

1) Can individual subjective ratings be effective in developing emotion recognition systems? If so, how?
2) How can we model subjective emotion and how well do ensembles of rater-specific models perform?
3) What fitting techniques can be used to fuse individual ensemble predictions and how do they perform compared to models trained using the gold standard?

II. Approach: An Affective Rater Ensemble Model

The goal of our approach is to embrace the variability of data, i. e. , the subjectivity that is intrinsic to human emotion annotations. To achieve this, we use an ensemble of models, one for each rater, to model the individual annotator continuous regression targets. Subsequently, we further combine the annotator specific predictions in different ways. While the models used have the same network architecture, they are trained on different targets based on their rater-specific training target. In this section, we initially introduce the two smoothing algorithms, the moving average and Savitzky-Golay filters, that are applied to the (raw) training targets of the rater-specific models. Further, we discuss the emotion recognition model to be employed, and propose multiple fusion methods that can be used to combine rater-specific predictions.

A. Smoothing

Naturally, noise is present in any human-made signal which is recorded at a low sample rate, e. g. , 100 ms [7] or 250 ms [9]. It was found that post-processing steps play a vital role in reducing noise and incidental errors [25]. The fusion of the annotations usually smooths out such errors. Given that the aim is to understand individual ratings as opposed to having them fused, we provide an analysis of the two previously found most successful methods, the moving average and the Savitzky-Golay smoothing using the MuSe-Toolbox [26].

Moving Average Filter: Owing to its simplicity and intelligibility, the moving average filter is one of the most common filters in digital signal processing. It works by taking the average of the input signal of each sliding filter frame.

Savitzky-Golay Filter: The filtered signal is smoothed by passing the data through polynomials of a low degree using a sliding filter, which then evaluates the polynomial for each frame at the central point [27]. What gives it an edge over the moving average is that it does not distort the signal significantly and retains high-frequency signal components [25].

B. Emotion Recognition Model

For every individual annotator-specific model, the exact same model architecture (cf. Figure 1) and feature inputs are utilised. However, they can be differentiated through their training targets, which are the individual annotators' processed ratings. Our model is based on Sun et al. 's architecture

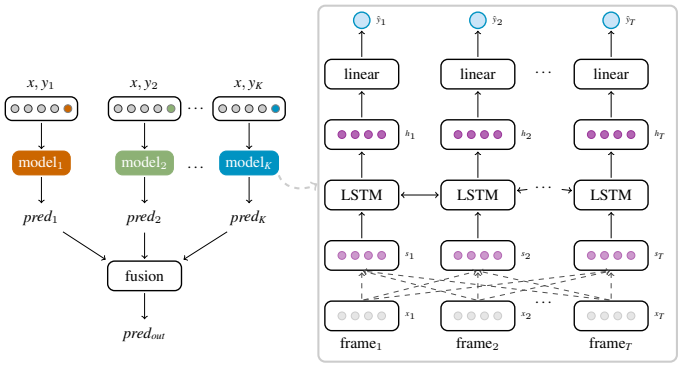

Fig. 1: One network per annotator. *Left:* Shown above is an ensemble of multiple annotators $k \in 1, 2, \ldots, K$. All models receive the same input x during training, and only the targets y_k differ depending on individual annotators. All the predictions $pred_k$ are combined to formulate a final prediction $pred_{out}$. *Right:* The model architecture for each ensemble. A self-attention layer encodes the input sequence x of length T to state s. Next, s is fed into a (bidirectional) RNN layer. The hidden states h_t are then forwarded to a linear layer, outputting $\hat{y}_t$ for each time step t. Right-hand side shows an RNN-LSTM with attention; figure adapted from [28].

[28], other layer types, and data augmentation [8], [29] which were successfully applied to this task. It consists of an Long-short Term Memory (LSTM)-Recurrent Neural Network (RNN) architecture coupled with self-attention for continuous emotion recognition (ATTN+LSTM), which allows for context memorisation [30] over long periods of time [28]. Hereby, it uses two addition modules, namely multi-head self-attention and linear layers. The input is initially encoded by the attention layer. Further, by using a unidirectional or a bidirectional RNN-encoder, the encoded sequence is projected to a context dependent space of hidden states. Finally, a feedforward layer maps the hidden states to the emotion prediction for every time step to predict arousal or valence in a time and value-continuous manner.

C. Ensemble Modelling

After training individual models, the predictions are fused to form a single prediction, which can then be evaluated. There are several methods to do this. The simplest is computing the mean over each time step. However, this approach does not take inter-rater disagreement, captured in the inputs and much likely transferred to the model representations, into consideration. Thus, we consider several fusion techniques. The EWE [6] is commonly used to create gold standard annotations from several ratings r by incorporating the reliability of each rater. It is computed as a weighted mean, where weights are corresponding to the cross-correlation of the annotators' ratings and the mean rating. Based on the correlation coefficient r_k of rater $k \in 1 \ldots K$, the EWE for sample x_n is defined as

$$x_n^{EWE} = \frac{1}{\sum_{k=1}^{K} r_k} \sum_{k=1}^{K} r_k x_{n,k}. \quad (1)$$

Annotators' ratings which are negatively correlated to the mean of others ($r_k <= 0$) are automatically removed by this method. In addition, we seek to learn an automatic, annotator-specific mapping from the individual predictions. To neurally train an individual mapping, we extended the architecture by a dynamically learnt fusion network which either incorporates a bidirectional LSTM-RNN or one with an additional self-attention encoding. For both network approaches, the hidden states are forwarded via a linear output layer.

III. MuSe-CaR Dataset

MuSe-CaR is an extensive, in-the-wild dataset [9] that aids in analysing multimodal sentiment [8], [31]. The idea of the dataset is based on the proliferation of multimodal-based user content on social media platforms; its efficient analysis is vital to establish and improve multimodal sentiment analysis algorithms. It has around 40 hours of English YouTube vehicle review video content and is fully annotated by arousal and valence continuous dimensional emotion annotations. These annotations depict the emotional state of all the individuals featured in the videos and are fused to a gold-standard using the EWE algorithm [6] (cf. Section II-C). To ensure a balance of in-the-wild characteristics, it cautiously considers varying video content from professional, semi-professional, and casual reviewers across all age groups. These video influences range from changing background to face-angles, shot sizes, camera motion, occlusions, ambient noises, microphone types, locations, and colloquialism. The annotations were obtained through a joystick that had a sample recording frequency of 0.25 Hz. The continuous-fashioned labels are within a range between -1000 and 1000 before any normalisation. Taking into account annotator disagreement that is innate to emotional labelling, each video was annotated using a minimum of five annotators. A total of 11 annotators were used throughout the entire process, with all having annotated different videos of either emotion dimension. The annotation is done in two rounds, where after the first round annotators with quantitative and qualitative performance below average were filtered out [9]. For both Arousal and Valence, the rater agreement (CC) is 0.27 and 0.35, respectively, which is close to values initially reported for other emotion datasets [7].

The quality of the annotations depends heavily on each annotator on which our approach relies on. Given that a few annotators conducted all annotation rounds and only few samples have been rated by the others, we choose the annotators with the ids 2, 4, 5, 7, and 8. These annotators exceeded the threshold of 150 videos for every annotated dimension. The new selection, excluding the removed data, has still a fairly equal distribution across the pre-defined partitions.

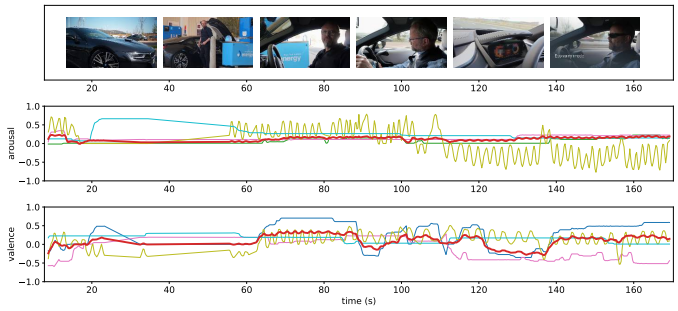

Fig. 2: Annotator-specific ratings for arousal and valence of a sample video. Different annotation styles can be observed. From a qualitative perspective, it can be assumed that systematic differences contain positive variability; while others, e. g. , the olive signals, show a lot of activity, the cyan arousal annotations barley exhibit any change over time. The red signals show the fused annotations.

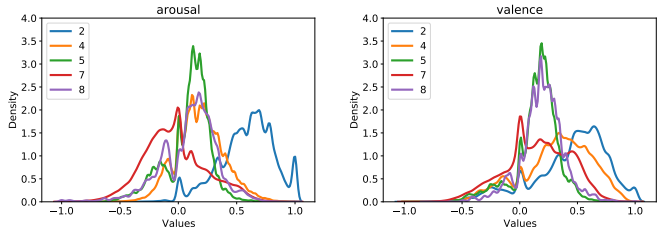

Fig. 3: The estimated density of Arousal and Valence for the annotations of the five annotators. There is a visible difference in value distributions in all the annotators. Annotators 4 and 7 are nearly Gaussian-shaped while annotator 2 is skewed towards the right.

Regardless of the training that individual annotators are given before the annotation process, there is a difference in the understanding of these concepts visible when considering the distinct styles. One notable style feature is the annotator-specific value distribution compared in Figure 3.

IV. Experimental settings and Baseline

This regression task's performance is calculated using the Concordance Correlation Coefficient (CCC) metric [32]. We evaluated the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), VGGFace , FAUs , and FastText features, which were provided and described by length as part of the MuSe-Wild challenge [8] and extracted VGGish embeddings from the raw audio [33] and BERT embeddings from the video transcriptions [34]. The high-level 128-wide VGGish model can be utilised as a feature extractor for deep acoustic representations [33] based on an adjusted VGGNet [35] which was trained for acoustic event detection classification employing a large-scale audio dataset called AudioSet [36]. Another option for well-established word-embeddings can be extracted using the pretrained BERT model, a transformer model [34] trained on English Wikipedia (2.5B words) and BooksCorpus (800M words) in an unsupervised manner. In contrast to the static word-embeddings, these word-vectors are dependent on context; thus, they need to be calculated at run-time.

All models are trained using four NVIDIA Tesla V100 (32 GB) from a DGX Station, which are specifically helpful when dealing with multimodal-based architectures and very long sequential data. Initially, we search for an effective emotion recognition model architecture, which will later be used to model the individual annotators. Based on findings and architectures proposed in [8], [28], we evaluate a large set of hyperparameters, features, and sampling methods for our new data partition. To augment the data, we use the sliding window approach to split the videos' segments into smaller segments of fixed length. As a result, the amount of training data is substantially increased. The experiments are conducted using a window size of 200 with a hop size of 100. Initially, we run an extensive hyperparamter search to optimise our models[1]. The best performing models used BERT input features with hyperparameters *hid_size* $= 128$, *num_layers* $= 2$, and *n_heads* $= 2$ for the prediction of valence with .45 CCC on the development and .58 on the test set. As to be expected from previous work [8], [28], the next best FastText falls .25 behind (valence, test). For arousal prediction, features of the VGGish show that it is most effective to use hyperparameters of *hid_size* $= 32$, *num_layers* $= 2$, and *n_heads* $= 4$ with .50 CCC on development and .40 CCC on the test set, followed by eGeMAPS with .40 CCC on development. We report all following results using these two superior feature sets for each prediction target and their network configurations.

V. Experiments and Results

This section draws on previous findings and evaluates our approach, which proposes an ensemble of rater-specific emotion recognition models. Initially, we attempt to improve the quality of the raw annotation signals and make learning easier by using different smoothing filters. Furthermore, we select the most effective ensembles and fuse their predictions using different approaches. Note that, as explained in the introduction, emotions are inherently subjective and finding a gold standard which accurately reflects something similar to a ground-truth is hard. Due to the lack of better alternatives, we evaluate all annotator-specific models on their very own (raw) annotations, and compare the fusion, independently of our fusion method, to the EWE gold standard fusion.

A. Annotation Signal Prediction

An emotion recognition model is trained for each of the five annotators. The models can be differentiated by their training target which is equal to the raw annotation signal. We analyse whether generalisation can be enhanced by applying a smoothing filter to the ratings. To realise this objective, both the cubic Savitzky-Golay and moving average filters are tested.

[1]We evaluate different hidden feature sizes of the LSTM-RNNs with *hid_size* $\in \{32, 64, 128\}$ and a number of RNN layers *num_layers* $\in \{1, 2, 3\}$. Additionally, we also test different counts of attention heads *n_heads* $\in \{2, 4, 8\}$. For these experiments, we only compare to the gold standard labels using the Adam optimiser with $\alpha \in \{0.001, 0.005, 0.0001\}$.

TABLE I: Rater-specific performance comparing no filter, cubic Savitzky-Golay filter and moving average filter (moving) using different filter frame sizes. The experiments are conducted on each of the chosen annotators and compared to the rater annotation, reporting the CCC for the emotion dimensions arousal and valence for the devel(opment) and test partitions. The best frame sizes are 9 for arousal and 11 for valence. The highest performances for individual raters and across all raters are marked in bold.

		$size_f$	2		4		5		7		8		mean	
			devel	test	devel	test	devel	test	devel	test	devel	test	devel	test
✘	**Arousal**	–	.40	.23	.45	.23	.06	.21	.48	.45	.15	.08	.33	.28
	Valence	–	.42	.42	.40	.52	.17	.25	.13	.04	.15	.08	.25	.26
Savitzky-Golay	**Arousal**	5	.43	.28	.45	.23	.06	.15	.29	.22	**.49**	.47	.34	.27
		7	.41	.25	.45	.16	.10	.20	.29	.30	.48	.40	.35	.26
		9	.43	.27	.45	.22	.10	.21	.31	.30	.47	**.50**	**.35**	**.30**
		11	.44	.23	.45	.23	.06	.21	.27	.27	.48	.45	.34	.28
	Valence	5	.45	.46	.42	.52	.17	.23	.14	.04	.20	.16	.28	.28
		7	**.47**	.43	.42	.52	.15	.24	.14	.04	.19	.00	.27	.25
		9	.46	.37	.41	**.54**	.15	.24	.15	.04	.20	.09	.27	.26
		11	.44	.44	.43	.53	.19	.30	.17	.05	.19	.11	**.28**	**.29**
moving	**Arousal**	3	.40	.23	.45	.23	.06	.21	.27	.27	.48	.45	.33	.28
		5	.40	.23	.46	.21	.14	.20	.27	.27	.48	.45	.35	.27
	Valence	3	.42	.42	.40	.52	.17	.25	.13	.04	.15	.08	.25	.26
		5	.42	.42	.40	.52	.17	.25	.13	.04	.15	.08	.25	.26

The Savitzky-Golay filter is applied by using four different filter frame sizes $size_f \in \{5, 7, 9, 11\}$ and with a value frequency of 4 Hz; a frame size of 5 corresponds to 1.25 seconds and 11 to 2.75 seconds. The convolutional nature of the Savitzky-Golay filter makes it necessary to adopt an odd $size_f$ and set the order of polynomial to less than the filter frame size. From this, we find that arousal $size_f = 9$ generates the most effective recognition performance with a mean CCC over the annotators of .35 and .30 for the development and test partitions, respectively (cf. Table I). Accordingly, valence $size_f = 11$ works effectively resulting in a mean CCC of .28 (devel) and .29 (test).

In the same setting, we apply the moving average filter using frame sizes 3 and 5. While applying this filter, the larger frame sizes were not considered to avoid severe smoothing. The two frame sizes exhibit the exact same valence results. For arousal, however, the average difference on the development partition is .02 CCC and .01 CCC on the test partition (cf. Table I). As such, it can be deduced that $size_f = 5$ is slightly better for arousal, but the difference is negligible.

B. Ensemble Fusion

All the individual predictions from the rater-specific models are combined to measure the final prediction against the gold standard. The performances of the simple mean, EWE, and second-level fusion models (LSTM and ATTN+LSTM) are depicted in Table II. The fusion results are reported with and without annotation smoothing, by using the filter configurations that previously lead to the best generalisation. The arousal models showed better performance when trained with the smoothed training target (cf. Section V-A). Particularly the ATTN+LSTM fusion yields the best results on test improving the baseline by .05 CCC. Contrary, smoothing of the valence signals does not yield better results. This may be attributed to erased information after using the Savitzky-Golay filter with a fairly large frame size of 11. Although the predictive performance seems enhanced by smoothing, this may only be achieved by over-smoothing (simplifying) the rating, however, not by actually increasing the quality of the signal. Here, the non-smoothed targets helped the model to perform slightly better (.01 CCC on test using LSTM fusion) than the smoothed annotations.

TABLE II: Fusion of rater-specific predictions using different ensemble techniques. The annotator models were trained using either raw or smoothed (Savitzky-Golay filter) annotations and the results are reported in CCC scores on devel(opment) and test set. In bold are the highest scores for both arousal and valence.

smooth		mean		EWE		LSTM		ATTN+LSTM	
		devel	test	devel	test	devel	test	devel	test
✘	**Arousal**	.26	.19	.36	NaN	.50	.41	.51	.42
	Valence	.22	.34	.30	.43	**.39**	**.57**	.38	.55
✔	**Arousal**	.25	.19	.19	.23	.50	.47	**.51**	**.47**
	Valence	.25	.36	.32	.43	.39	.56	.37	.53

Additionally, we attempt to combine the predictions of the n most successful rater-specific models, namely $n \in \{2, 3, 4, 5\}$. For this, the ATTN+LSTM and LSTM fusions are used (cf. Table III).

TABLE III: Combining the n best performing rater-specific models. Reported are the CCC scores for devel(opment) and test partitions.

n	Arousal		Valence	
	devel	test	devel	test
2	.32	.42	.38	.53
3	.40	.44	.36	.50
4	.47	.46	.36	.52
5	**.51**	**.47**	**.39**	**.57**
Baseline	.50	.40	.45	.58

For both dimensions, using all five models offers the best results. However, when less rater models are made available, the overall quality of the prediction is of very similar quality. For instance, the top four models of arousal show a test performance only .01 CCC lower than for the fusion of all five. In comparison, the fifth best model for valence improves the overall efficiency by .05 CCC on test compared to fusing only the best four.

Lastly, we compare the performance of our ensemble with that of models that have been solely trained on gold standard annotations. Compared to the baseline, the rater ensemble beats the arousal results by .07 CCC on test, showing .51 CCC on the development partition and .47 CCC on the test partition. For valence, the ensemble falls slightly below the performance of the gold standard model by .01 CCC on test. Given that our experimental setting entails vast data amounts and training multiple models, we only relied on the hyperparameters of the baseline hyperparameter search. Additional performance enhancement can be expected after enhancing both single and overall comparability, but this quantification is beyond the scope of this work.

VI. Discussion

We sought to determine whether subjective ratings can a) effectively build emotion recognition systems, and b) how beneficial they are. We initially established if models can be trained based on individual rater's emotional perception. Our study results suggest that the performance of rater-specific models can indeed be similar to models trained on the gold standard. Annotator 8 had the highest arousal CCC on test with .50 while annotator 4 achieved .54 CCC on test for valence. When compared to the models trained based on the gold standard; these results either fall in range or perform even better. However, this is strongly dependant on the individual annotator, indicated by rater models resulting in CCC scores as low as .10 CCC. Of interest, we found that some annotators such as annotator 8 yield very poor recognition results for valence, yet very strong results for arousal. We assume this might be connected to an annotator's inability to understand one of the dimensions, leading to non-systematic ratings from which models fail to learn underlying patterns. Such patterns might be interesting as an additional evaluation measure in the future when evaluating (and eliminating) raters over several annotation rounds.

Additionally, we assessed which fusion techniques efficiently map the individual annotator ensembles for comparison onto conventional approaches that use gold standard annotations only. In Section V-B, different fusion techniques were tested, including naive ways such as calculating the mean and using the EWE. In addition, we tried to learn the importance of each annotator as well as temporal interactions in the signal dynamically by using an LSTM-RNN by itself and combined with a prior attention encoding. We found that these two models are superior when compared to mean calculation or EWE fusion which provides overall better results as opposed to averaging, but also occasionally fails entirely. The best model, the ATTN+LSTM fusion, achieves a CCC of .47 on the original test set. In fact, the results were nearly .20 better having fewer training data (only the selected, crucial annotators) available when compared to the challenge baselines [8], and matches the MuSe-Wild winner [28]. Both techniques used numerous modalities, while our performance is determined by one modality only for every emotion dimension. For valence, the ensemble of annotators records .57 CCC on test, which is slightly lower when compared to the .60 achieved by the MuSe-Wild winner fusion model. Owing to this slight difference, we omitted conducting an extensive hyperparameter search for our approach. We assume that the valence results can be enhanced.

Within the scope of this study, we were challenged to determine a 'ground-truth' to compare and define the success of our ensemble approaches (cf. Section V-B). Since a real 'ground-truth' is simply impossible for emotions, MuSe-CaR relies on the EWE gold standard, a well-known annotation fusion technique that weighs individual ratings based on inter-rater agreements. While this technique is well established, it does not deal with the systematic differences in annotation styles, in turn leading to a loss of vital information. In contrast, our ensemble of rater-specific models exploit the entire human labelling information available. We are able to measure the performance and success of the model in generalising to a certain degree, by examining the ensemble predictions using the gold standard. Better evaluation grounds for the ensemble predictions can be realised by using different gold standards that have been generated by more advanced and sophisticated aggregation methods. Future research has to show if our claims are still valid under other and new gold-standard fusion methods.

VII. Conclusion

The purpose of this study was to explore the full potential of the subjective information available from raw annotations to predict emotions. By embracing the inherent subjectivity of these annotations, we showed that also rater-specific emotion ensemble models can model human emotions – without loss of any subjective information. Identical concepts have also been tested for emotion recognition systems that use discrete emotions. As far as we are aware, however, we are the first to apply rater-specific ensembles for emotion recognition using the dimensional emotions for both arousal and valence. A key obstacle is to show the full potential of our idea since the success of our ensemble can only be measured against the manufactured ground-truth, which is by nature biased towards a simplification or even loss of information itself and not an objective measure. On the other hand, it is very likely that a less smoothed annotation curve and based on that, emotion modelling, might neither represent the 'ground truth' which might be more ordinal or even categorical – both in the memory of the human generating these emotions and in the one of the observers – and by that, more useful for further processing in applications.

Notwithstanding these basic caveats, we have answered the three research questions that we asked in section I:

1) We found that individual, subjective rater-specific models that use smoothed trained targets can generalise better. This helps explain why the Savitzky-Golay filter performed better in removing incidental errors within the annotation signals.
2) The performance of our rater-specific models is strongly dependent on individual annotators, with some reaching results of .54 for valence on test while others fail to learn generalisable patterns.
3) We have shown that using an ensemble of models, one model for every annotator, enhances the recognition performance most when a neural-base fusion (LSTM, ATTN-LSTM) and not the mean or EWE is used.

Furthermore, we have shown that exploiting individual human annotations by simply fusing individual models is impossible and more advanced architectures are needed for this task.

REFERENCES

[1] H. Hoffmann, A. Scheck, T. Schuster, S. Walter, K. Limbrecht, H. C. Traue, and H. Kessler, "Mapping discrete emotions into the dimensional space: An empirical approach," in *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 3316–3320, IEEE, 2012.

[2] B. M. Booth, K. Mundnich, and S. S. Narayanan, "A novel method for human bias correction of continuous-time annotations," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3091–3095, IEEE, 2018.

[3] Z. Callejas and R. Lopez-Cozar, "Influence of contextual information in emotion annotation for spoken dialogue systems," *Speech Communication*, vol. 50, no. 5, pp. 416–433, 2008.

[4] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100943–100953, 2019.

[5] F. Hönig, A. Batliner, and E. Nöth, "How many labellers revisited–naïves, experts, and real experts," in *Proceedings of Speech and Language Technology in Education*, (Venice), pp. 137–140, 2011.

[6] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, pp. 381–385, IEEE, 2005.

[7] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. W. Schuller, *et al.*, "Sewa db: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[8] L. Stappen, A. Baird, G. Rizos, P. Tzirakis, X. Du, F. Hafner, L. Schumann, A. Mallol-Ragolta, B. W. Schuller, I. Lefter, *et al.*, "Muse 2020 challenge and workshop: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media: Emotional car reviews in-the-wild," in *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, pp. 35–44, 2020.

[9] L. Stappen, A. Baird, L. Schumann, and B. Schuller, "The multimodal sentiment analysis in car reviews (muse-car) dataset: Collection, insights and improvements," *IEEE Transactions on Affective Computing*, pp. 1–16, 06 2021.

[10] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

[11] L. Stappen, G. Rizos, M. Hasan, T. Hain, and B. W. Schuller, "Uncertainty-aware machine support for paper reviewing on the interspeech 2019 submission corpus," *Proceedings of the INTERSPEECH*, pp. 1808–1812, 2020.

[12] H. M. Fayek, M. Lech, and L. Cavedon, "Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels," in *2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 566–570, IEEE, 2016.

[13] T. Cohn and L. Specia, "Modelling annotator bias with multi-task gaussian processes: An application to machine translation quality estimation," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 32–42, ACL, 2013.

[14] M. Guan, V. Gulshan, A. Dai, and G. Hinton, "Who said what: Modeling individual labelers improves classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[15] G. Rizos and B. W. Schuller, "Average jane, where art thou?–recent avenues in efficient machine learning under subjectivity uncertainty," in *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pp. 42–55, Springer, 2020.

[16] J. Deng and B. Schuller, "Confidence measures in speech emotion recognition based on semi-supervised learning," in *Proceedings of the INTERSPEECH*, pp. 2226–2229, ISCA, 2012.

[17] T. Dang, V. Sethu, J. Epps, and E. Ambikairajah, "An investigation of emotion prediction uncertainty using gaussian mixture regression.," in *Proceedings of the INTERSPEECH*, pp. 1248–1252, ISCA, 2017.

[18] M. A. Nicolaou, V. Pavlovic, and M. Pantic, "Dynamic probabilistic cca for analysis of affective behavior and fusion of continuous annotations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1299–1311, 2014.

[19] M. Atcheson, V. Sethu, and J. Epps, "Demonstrating and modelling systematic time-varying annotator disagreement in continuous emotion annotation," in *Proceedings of the INTERSPEECH 2018*, pp. 3668–3672, ISCA, 2018.

[20] F. Zhou and F. De la Torre, "Generalized canonical time warping," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 279–294, 2015.

[21] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "Avec 2017: Real-life depression, and affect recognition workshop and challenge," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pp. 3–9, 2017.

[22] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, "The msp-conversation corpus," *Proceedings of the INTERSPEECH 2020*, pp. 1823–1827, 2020.

[23] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud, *et al.*, "Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pp. 3–13, 2018.

[24] C. Wang, P. Lopes, T. Pun, and G. Chanel, "Towards a better gold standard: Denoising and modelling continuous emotion annotations based on feature agglomeration and outlier regularisation," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pp. 73–81, 2018.

[25] N. Thammasan, K.-i. Fukui, and M. Numao, "An investigation of annotation smoothing for eeg-based continuous music-emotion recognition," in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 003323–003328, IEEE, 2016.

[26] L. Stappen, L. Schumann, B. Sertolli, A. Baird, B. Weigel, E. Cambria, and B. W. Schuller, "Muse-toolbox: The multimodal sentiment analysis continuous annotation fusion and discrete class transformation toolbox," in *Proceedings of the 2nd Multimodal Sentiment Analysis Challenge*, ACM, 2021.

[27] A. Savitzky and M. J. Golay, "Smoothing and differentiation of data by simplified least squares procedures.," *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[28] L. Sun, Z. Lian, J. Tao, B. Liu, and M. Niu, "Multi-modal continuous dimensional emotion recognition using recurrent neural network and self-attention mechanism," in *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*, pp. 27–34, 2020.

[29] L. Stappen, A. Baird, L. Christ, L. Schumann, B. Sertolli, E.-M. Messner, E. Cambria, G. Zhao, and B. W. Schuller, "The muse 2021 multimodal sentiment analysis challenge: Sentiment, emotion, physiological-emotion, and stress," in *Proceedings of the 2nd Multimodal Sentiment Analysis Challenge*, ACM, 2021.

[30] L. Stappen, G. Rizos, and B. Schuller, "X-aware: Context-aware human-environment attention fusion for driver gaze prediction in the wild," in *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI)*, pp. 858–867, IEEE, 2020.

[31] L. Stappen, A. Baird, E. Cambria, and B. W. Schuller, "Sentiment analysis and topic recognition in video transcriptions," *IEEE Intelligent Systems*, vol. 36, pp. 88–95.

[32] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.

[33] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, *et al.*, "Cnn architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 131–135, IEEE, 2017.

[34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186, 2019.

[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[36] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780, IEEE, 2017.