MuSe-Toolbox: The Multimodal Sentiment Analysis Continuous Annotation Fusion and Discrete Class Transformation Toolbox

Lukas Stappen
University of Augsburg
Augsburg, Germany

Lea Schumann
University of Augsburg
Augsburg, Germany

Benjamin Sertolli
University of Augsburg
Augsburg, Germany

Alice Baird
University of Augsburg
Augsburg, Germany

Benjamin Weigell
University of Augsburg
Augsburg, Germany

Erik Cambria
Nanyang Technological University
Singapore

Björn W. Schuller
Imperial College London
London, United Kingdom

ABSTRACT

We introduce the MuSe-Toolbox – a Python-based open-source toolkit for creating a variety of continuous and discrete emotion gold standards. In a single framework, we unify a wide range of fusion methods and propose the novel *Rater Aligned Annotation Weighting* (RAAW), which aligns the annotations in a translation-invariant way before weighting and fusing them based on the inter-rater agreements between the annotations. Furthermore, discrete categories tend to be easier for humans to interpret than continuous signals. With this in mind, the MuSe-Toolbox provides the functionality to run exhaustive searches for meaningful class clusters in the continuous gold standards. To our knowledge, this is the first toolkit that provides a wide selection of state-of-the-art emotional gold standard methods and their transformation to discrete classes. Experimental results indicate that MuSe-Toolbox can provide promising and novel class formations which can be better predicted than hard-coded classes boundaries with minimal human intervention. The implementation[1] is out-of-the-box available with all dependencies using a Docker container[2].

CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**; • **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

Affective Computing; Annotation; Gold-Standard; Smoothing; Emotion classes; Emotion Recognition; Multimodal Sentiment Analysis

[1] https://github.com/anonymous/MuSe-Toolbox
[2] `docker pull musetoolbox/musetoolbox`

ACM ISBN 978-1-4503-8678-4/21/10…$15.00
https://doi.org/10.1145/3475957.3484451

ACM Reference Format:
Lukas Stappen, Lea Schumann, Benjamin Sertolli, Alice Baird, Benjamin Weigell, Erik Cambria, and Björn W. Schuller. 2021. MuSe-Toolbox: The Multimodal Sentiment Analysis Continuous Annotation Fusion and Discrete Class Transformation Toolbox. In *Proceedings of the 2nd Multimodal Sentiment Analysis Challenge (MuSe '21), October 24, 2021, Virtual Event, China.* , 8 pages. https://doi.org/10.1145/3475957.3484451

1 INTRODUCTION

The accelerating pace of digitisation is driving digital interaction into all areas of our daily life, and from the resulting mass of data, a substantial portion can be quantified into human behavioural signals. Learning to recognise emotional cues in interactions e. g., taking place via video, is the purpose of the growing field of Emotion AI. In this process, various modalities, such as body language, voice, text, and facial expression, are examined for patterns that help map the cues to specific emotions. The reference data necessary to learn the mapping is annotated by humans, often as category labels (e. g., happy, sad, surprised, etc.) and continuous annotations. For continuous mapping, behavioural and cognitive scientists assume that the human brain is not divided into hard-wired regions and better represented by dominant primitives (dimensions) whose complex interaction results in a specific emotion (e. g., the dimensional axes of arousal and valence) [33].

The growing demand for emotion technology in various domains led to an increased interest in the annotation of such data. However, the annotation process itself is not trivial to execute, and obtaining meaningful reference data to develop models for automatic pattern recognition is a challenge. One such challenge is the dependency on humans raters. When rating the perceived data (e. g., videos), time-delays in the reaction [27], as well as systematic disagreement due to personal bias and other task-related reasons are well known [2, 4]. To counteract, it is common practice to involve multiple humans in the annotation of the same source and fuse these perceptions. Since emotions are inherently subjective, these fused signals are coined as gold-standard. To date, none of the proposed fusion methods has become a de-facto standard. One reason for this may be that a convenient comparison of the fusion outcomes is hardly feasible. The implementation of the methods is often distributed over many different source bases, coded in different programming languages

and frameworks, or is not publicly available at all. An issue is also the dependency on outdated software (package) dependencies.

Another unresolved problem is the transformation of continuous emotion signals into more general class labels that are easier for humans to interpret. In an empirical approach, Hoffmann et al. [18] mapped discrete emotions into the dimensional emotion space [33]. Similarly, Laurier et al. [23] aimed to cluster emotion tags to find clusters corresponding to the four quadrants of the arousal-valence dimensions. A tool that supports this transformation process by the automatic creation of meaningful classes has not yet been presented in the literature.

With this contribution, we want to tackle both of these issues by proposing an easy-to-use, well-documented toolbox. The input data can be any continuous annotation recorded by an annotation software (e. g., a human-controlled joystick or mouse) or directly from a (physiological) device (e. g., smartwatches). Additionally, the annotations can be easily standardised, smoothed and fused by the most common gold-standard creation techniques, such as *Estimator Weighted Evaluator* (EWE), *DTW Barycenter Averaging* (DBA), and *Generic-Canonical Time Warping* (GCTW). This elegantly makes a comparison of the multiple available fusion methods easily possible, leaving broad flexibility for database creators while allowing reproducibility and exchange over the set of parameters used. To this end, we propose a novel gold-standard method *Rater Aligned Annotation Weighting* (RAAW) to the set of fusion tools, which we derived from methods introduced here and which is inspired by the results we obtained during the work on the toolbox and the limitations of the provided fusion methods. Furthermore, we propose a simple way to extract time-series features from these signals, which may aid the creation of emotional classes from emotion dimensions. The toolbox can be started directly from a Docker container without installing dependencies, and an open-source Github repository is available to the community for further development.

Note, the core focus of this work is emotional annotations. However, all kinds of time-series data are omnipresent in our daily life. Changes in stocks, energy consumption, or weather are all recorded over time and, thus, have natural time-series properties. Predicting these values in time is often challenging and a simplification by fusing them (e. g., energy consumption of several households) transforming sequences into summary classes by clustering (e. g., days in a week) may be beneficial for any of the other applications as well.

2 METHODOLOGY AND SYSTEM OVERVIEW

In the following section, we first describe the methods that underpin the functionality of our toolbox and conclude by placing them in the context of the functionalities in Section 2.6.

2.1 Smoothing of Annotations

As for all fine-grained time-series, short-term errors and distortions can occur in the annotation process. Smoothing digital filters are useful to mitigate these negative noise effects [48, 51]. One common signal processing approach for this is the *Savitzky-Golay filter* (SavGol) which increases the precision of the data points using a low degree polynomial over a moving filter [35]. In our context, this method has the advantage that it still preserves high-frequency characteristics [48]. Also widely applied is the *Moving Average Filter* (MAF). It employs a moving average of a given window to smooth the signal gently. The MAF applied with 4.25 s filter frame (or 17 time steps) is illustrated in Figure 1, alongside a SavGol example, and the raw annotations.

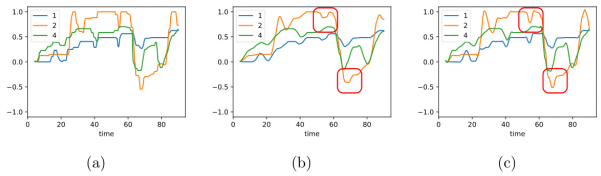


Figure 1: An example of valence annotation signals of three annotators. Figure (a) depicts the raw signals, while the other two figures show the filtered signals of a moving average filter (b), and a cubic Savitzky-Golay filter (c), respectively, with a filter frame size of 17 values (4.25 s). Evidently, the moving average indicates a visibly stronger smoothing effect, when compared to the Savitzky-Golay filter, which preserves signal features more closely.

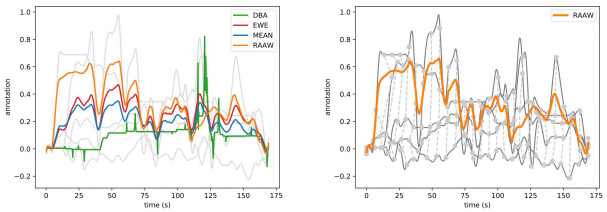


Figure 2: The left side shows all fusion methods in MuSe-Toolbox on a sample annotation (MuSe-CaR database, video id: 100, arousal). The right side is a detailed illustration of the *Rater Aligned Annotation Weighting* (RAAW) alignment, including the warping paths.

2.2 Gold Standard Fusion Methods

A gold-standard method tries to establish a consensus from a group of individual ratings. Some methods are specifically developed for emotion annotations, i. e., EWE , and RAAW , while others are derived from more generic principles of time-series aggregation, i. e., DBA , GCTW . A comparison of all methods is visualised in Figure 2.

2.2.1 Estimator Weighted Evaluator (EWE). The *Estimator Weighted Evaluator* (EWE) is based on the reliability evaluation of the raters [36]. It is essentially a weighted mean of all rater-dependent annotations, sometimes interpreted as the weighted mean of raters' similarity [14, 15]. To compute the weights, the cross-correlations of an annotation to the mean of all other annotations is calculated for each annotation. It can be formally expressed by

$$\hat{x}_n^{EWE} = \frac{1}{\sum_{k=1}^{K} r_k} \sum_{k=1}^{K} r_k \hat{x}_{n,k}, \tag{1}$$

where r_k is the similarity of the k-th annotator to the other time-series. A typical method for calculating similarity between time-series is the Euclidean metric or Pearson coefficients. However, since both do not take sequence order, phase shift, and scaling variance into account, it was replaced by the concordance correlation coefficient (CCC) for similarity calculation:

$$CCC(\hat{\theta}, \theta) = \frac{2 \times COV(\hat{\theta}, \theta)}{\sigma_{\hat{\theta}}^2 + \sigma_{\theta}^2 + (\mu_{\hat{\theta}} - \mu_{\theta})^2} = \frac{2E[(\hat{\theta} - \mu_{\hat{\theta}})(\theta - \mu_{\theta})]}{\sigma_{\hat{\theta}}^2 + \sigma_{\theta}^2 + (\mu_{\hat{\theta}} - \mu_{\theta})^2}. \quad (2)$$

Here, x is the time-series data, θ is a series of n annotations, and $\hat{\theta}$ the reference annotation. This method is broadly applied across different tasks in affective computing [22, 31, 32, 43].

2.2.2 DTW Barycentre Averaging (DBA). Averaging in Dynamic Time Warping (DTW) spaces is widely adopted for similarity-based temporal alignment in the field of machine learning. Similar to the Euclidean metric and CCC, DTW implements a distance metric, adding elastic properties that compute the best global alignment based on a one-to-many mapping of points in two time-series. The *DTW Barycenter Averaging* (DBA) method available in our framework is based on an algorithm originally developed for general time-series barycentre computation to compute the optimal average sequence. A barycentre is a time-series b based on the computation of continuous representative propensities from multiple time series points x. In this particular version, these tendencies are determined by a sub-gradient, majorize-minimize algorithm of d [38] with the advantage of fusing of time-series of varied length. DTW can be expressed as:

$$DTW = \min \sum_i d\,(b, x_i)^2\,. \quad (3)$$

2.2.3 Generic-Canonical Time Warping (GCTW). Another extension of DTW is Canonical Time Warping (CTW) [57], which in addition to DTW integrates Canonical Correlation Analysis [1], a method for extracting shared features from two multi-variate data points. CTW was originally developed with the goal of aligning human motion and multimodal time series more precisely in time [57]. The combination with these two approaches allows a more flexible way of time-warping by adding monotonic functions that can better handle local spatial deformations of the time series. The same authors [56] further extended this approach to *Generic-Canonical Time Warping* (GCTW), which enables a computationally efficient fusion of multiple sequences by reducing the quadratic to linear complexity. Furthermore, the identified features with high correlation are emphasised by weighting.

2.2.4 Rater Aligned Annotation Weighting *(RAAW).* In the context of emotions, we propose a novel method *Rater Aligned Annotation Weighting* (RAAW) for the fusion of dimensional annotations for gold-standard creation. RAAW capitalises on the merits of the underlying alignment technique DTW and the inherent nature of the EWE method. More specifically, DTW is used to align the varying and changing response times of individual annotators over time (cf. Figure 2. This alignment between the fused signal was previously made brute-force by shifting the global or individual annotation by a few seconds and measure the resulting performance. The optimal number of emotion annotators is estimated to be at least three depending on their experience and the difficulty of the task [19]. To perform an alignment in a resource-efficient manner — even for many annotations — we utilise the DTW variant GCTW [56]. Subsequently, the similarity is calculated using the CCC for the individual aligned signals to accommodate the inter-rater agreement (subjectivity). The signals weighted according to this can be completely disregarded when negatively correlated before they are finally merged using EWE [14].

2.3 Emotional Signal Features

Emotion annotations can be seen as a quasi-continuous signal with a high sampling rate [22, 43, 46]. Extracting features from audio-visual and psychological signals is fairly common in intelligent computational analysis [36, 37]. In the context of this work, we extract (time-series) features from an emotional signal segment to summarise the time period in a meaningful way. The resulting representation summarising the segment over time is a vector of the size of the selected features. Starting with the most interpretable features, common statistical measures are extracted [34], such as the standard deviation (std), mean, median and a range of quantiles (q_x). However, these features do not reflect the characteristics of changes over time.

For this reason, the toolkit further offers to extract more complex time-series features namely: relative energy (*relEnergy*) [6], mean absolute change (*MACh*), mean change (*MCh*), mean central approximation of the second derivatives (*MSDC*), relative crossings of a point m (*CrM*) [6], relative number of peaks (*relPeaks*) [6, 28], skewness [8, 10], kurtosis [52], relative longest strike above the mean (*relLSAMe*), relative longest strike below the mean (*relLSBMe*), relative count below mean (*relCBMe*), relative sum of changes (*relSOC*), first and last location of the minimum and maximum (*FLMi*, *LLMi*, *FLMa*, *LLMa*), and percentage of reoccurring data points ($PreDa$). Note that features labelled as "relative" are normalised by the length of a segment, in order to limit the influence of varying segment lengths on the unsupervised clustering.

2.4 Dimension Reduction

Large dimensional feature sets often lead to unintended side effects, such as the curse of dimensionality [49]. However, by reducing or selecting certain dimensions of the available features, these effects can be counteracted. Principal component analysis (PCA) is a well-known dimension reduction method that transforms features into principal components [53]. These components are generated by projecting the original features into a new orthogonal coordinate system. This enables the reduction of the dimensions while preserving most of the data variation. Another method for dimensionality reduction is Self-organising Maps (SOM), a type of unsupervised, shallow neural network that transforms a high-dimensional input space into a low-dimensional output space [21]. Each output neuron competes with the other neurons to represent a particular input pattern, which makes it possible to obtain a comprehended representation of the most relationships in the dataset. SOM can also be used as a clustering or visualization tool, as they are considered to have low susceptibility to outliers and noise [50].

2.5 Clustering

2.5.1 K-means and fuzzy c-means clustering. A common way to differentiate k-means from fuzzy c-means algorithms is how a datapoint belongs to the resulting outcome, which can either be an assignment to exactly one cluster (crisp), or to multiple ones with a certain probability (fuzzy). The most popular fuzzy clustering method is the fuzzy c-means algorithm [3], based on the k-means algorithm [16]. To this end, a fixed number of clusters is defined. The cluster centres are initially set randomly, and the Euclidean distances from them to the data points are calculated. These are assigned to the clusters so that there is a minimal variance increase. By step-wise optimisation (similar to an expectation maximisation (EM) algorithm) of the centres and assignments, the algorithm converges after a few iterations. For the fuzzy version, the degree of overlap between clusters can be specified using the fuzzifier *m* parameter.

2.5.2 Gaussian mixture model . Similar to c-means, a Gaussian Mixture Model (GMM) introduces fuzziness into the clustering process and allows the weak assignment of a single datapoint to several clusters simultaneously. For this purpose, a probabilistic model is generated that attempts to describe all data by Gaussian distributions with different parameters. The optimisation process to find a suitable covariance structure of the data as well as the centres of the latent Gaussian distributions uses the EM algorithm as in k-means .

2.5.3 Agglomerative clustering. Besides the k-means , two other types of crisp clustering are common: agglomerative [20] and density clustering. Agglomerative is a hierarchical clustering technique in which each datapoint starts as its own cluster and is successively merged with the closest datapoint (i. e., cluster) into higher-level clusters. As soon as the distance between two clusters is maximised or the minimum number of clusters is reached, the clustering process is terminated.

2.5.4 Density-Based Spatial Clustering of Applications with Noise (DBSCAN). Density-clustering algorithms such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) have became more popular over the last years [5]. The main difference to other methods is that it also uses the local density of points instead of relying only on distance measures [11]. DBSCAN provides an answer to two common problems in clustering: a) the number of clusters does not have to be specified in advance and b) it automatically detects outliers which are then excluded from the clustering [20, 39]. With other methods, these outliers have to be removed manually after a manual check, otherwise, there is a risk that the clusters would get distorted. The reason for this is that each point must contain at least a minimum number of points in a given radius, called min_samples parameter in the ϵ-neighborhood. However, this simultaneously causes a firm reliance on the defined parameters.

2.5.5 Measures. Clusters are usually evaluated using internal metrics and external assessment. The internal metrics focus on how similar the data points of a cluster are (compactness), and how far the clusters differ from each other (separation) [25]. The Calinski-Harabasz Index (CHI) calculates the weighted average of the sums of squares within and between clusters. Also distance-based is the Silhouette Coefficient (SiC), but it is bounded within an interval of -1 to 1 (1 corresponds to an optimal cluster), allowing for easier comparability between runs and procedures [54]. The Davies-Bouldin Index (DBI) is based on similarity measures and decreases with increasing cluster separability [55]. Specifically for fuzzy c-means , the Fuzzy Partition Coefficient (FPC) can be employed, and measures the separability of fuzzy c-means using Dunn's partition coefficients [9]. Finally, we use the S_Dbw-Index, which is based on intra-cluster variance to measure compactness, where the average density in the area between clusters and the density of clusters is calculated (smaller is better).

2.6 MuSeFuseBox System Overview

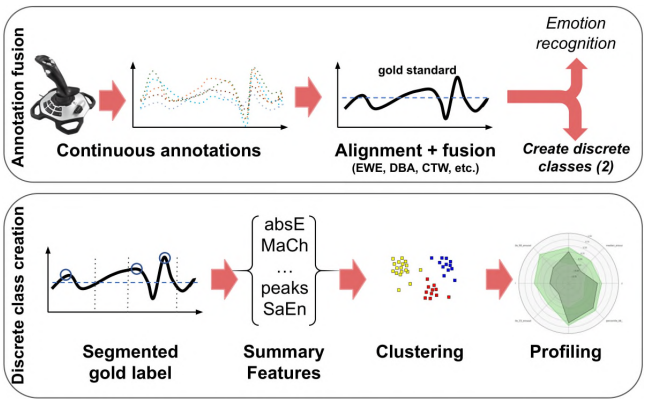


Figure 3: System overview of MuSe-Toolbox

The introduced methodology is integrated into the MuSe-Toolbox as depicted in Figure 3. The upper part shows the annotation fusion process. Given the input of multiple annotations, these can first be smoothed and/or normalised (cf. Section 2.1), which has shown benefits in previous works [26, 31]. The normalisation is either applied on video- or annotator-level. Next, the pre-processed annotations are fused using either DBA , EWE , GCTW , or RAAW (cf. Section 2.1). The lower part represents the creation process of discrete classes from a given signal. All, or a selection of the introduced features from Section 2.3 are extracted from segments of the fused annotation signal. These summary features are either clustered directly by one of the methods described in Section 2.5 or first reduced in dimensionality (cf. Section 2.4) and subsequently clustered. There is an option to either cluster on all data or on the training partition only. For the latter option, the classes of the development and test partitions are predicted based on the resulting clusters from the training set. For internal evaluation, the measures described in Section 2.5.5 are calculated. Since the generated clusters are intended to be used as classification targets, an exclusion of clustering proposals based on a rule-of-thumb can be activated to avoid strong class imbalances. This excludes cluster proposals where one or more clusters are smaller than a factor of the by chance level. For example, the prediction of four classes has a by chance level of 25 %. If the factor is set to 0.5, then, the smallest proposed cluster has to cover at least 12.5 % of the data. Finally, the profiling provides all the information necessary to enable an additional external evaluation by a human. For profiling, we provide a) standard features (mean,

standard deviation, etc.), b) visualisations, such as radar charts of the top distinctive features and scatter plots, and c) correlation of the features within a cluster. Based on these, the resulting clusters can be interpreted, and a name can be given.

2.7 Implementation Details

The MuSe-Toolbox is implemented in Python and relies on several packages, most notably numpy, pandas, scikit-learn, oct2py, and scipy. It can be used as a command line tool (over 50 different settings and configurations are available) or from the Python API. The implementation of DBA is adapted from [12, 29, 30][3] and DTW components are adapted from the Matlab implementation[4] of [56], which we transformed into code of the open-source programming language and environment for octave and access it for our calculations. The code is publicly available on GitHub under the GNU General Public license[5].

3 EXPERIMENTS

To demonstrate the capabilities of the MuSe-Toolbox , we run experiments on the produced gold standards. By doing so, we used them to train models for dimensional affect recognition. To this end, we utilise the MuSe-CaR database [44], used in the 2020 and 2021 Multimodal Sentiment Analysis real-life media Emotion Challenges (MuSe) [41, 43], and several other works [13, 24, 40, 42, 45, 47].

3.1 Continuous Emotion Fusion

In this section, we present the results of several experiments based on outputs from our toolkit to demonstrate its functionality. As explained in the previous sections, gold-standard methods lead to qualitatively different results, meaning that the quantitative results alone are only of limited value.

For our experiments, we build on the MuSe [41, 43], a challenge-series co-located to the ACM Multimedia Conference, which aims to set benchmarks for the prediction of emotions and sentiment with deep learning methods in-the-wild. Since the experimental conditions are predefined and publicly available, this is an ideal test ground. The database utilised for the challenge is called MuSe-CaR , which provides 40 hours of YouTube review videos of human-emotion interactions. Each 250 ms of the video dataset is labelled by at least five annotators, which are used for the following experiments. For more information, we refer the interested reader to the challenge [41] and database paper [44].

We use two of the provided feature sets, VGGish and BERT , from the challenge [41] to predict arousal and valence. VGGish [17], is a 128 dimensional audio feature set pre-trained on an audio dataset including YouTube snippets (AudioSet) with the aid of deep learning methods. These audio samples were differentiated into more than 600 different classes. BERT [7] embeds words in vectors by using transformer networks. Its deep learning architecture is upfront trained on several datasets and training tasks. The embeddings used here is the sum of the last four output layers, which consists of a total of 768 dimensions. Both embeddings were extracted at the same sample rate as the labels. Furthermore, the LSTM-RNN

[3]https://github.com/fpetitjean/DBA, GNU General Public License
[4]https://github.com/zhfe99, free for research use (no licence)
[5]https://github.com/anonymous/MuSe-Toolbox

Table 1: Results comparing with and without pre-smoothing using a savgol filter with a size of 5 on all annotation fusion techniques.

	Arousal				Valence			
	–		smooth		–		smooth	
	Devel.	Test	Devel.	Test	Devel.	Test	Devel.	Test
DBA	.2634	.2615	.2368	.2480	.3580	.4209	.2583	.3638
GCTW	.4809	.3481	.4840	.3502	.4394	.5594	.4503	.5848
EWE	.4410	.2513	.4386	.3210	.4476	.5614	.4454	.5703
RAAW	.4266	.2778	.4225	.3514	.4589	.5493	.4482	.5698
∅	.4030	.2847	.3955	.3177	.4260	.5228	.4006	.5222

baseline model made available by the organisers is utilised and re-trained for 100 epochs with batch size 1024 and learning rate $lr = 0.005$ on the new targets. Further, we run a parameter optimisation for the hidden state dimensionality $h = \{32, 64\}$ for arousal and $h = \{64, 128\}$ to predict valence, as this selection has previously worked well for the MuSe-CaR data, as shown in the 2021 MuSe Challenge baseline publication [41]. As the challenges use the CCC as the competition measure, we use the CCC for evaluation as well as the loss function.

3.1.1 Smoothing. The effect of smoothing can be seen in Figure 1, c) compared to the raw annotations and the filtered signal using the Savitzky-Golay filter. It is apparent that the moving average filter smooths the signal much more than the Savitzky-Golay filter, even to a point at which information from the signal is lost. Hence, we adjust the filter frame-size of the moving average filter to be a smaller value compared to the Savitzky-Golay filter. Following the pre-processing and fusion, the fused signal can further be smoothed using convolutional smoothing. The kernel size of 15 has proven to yield high-quality gold standard annotations whilst reduced signal noise. We further compare the performance of all fusion methods when applying the Savitzky-Golay filter for pre-smoothing in Table 1. In general, it is noticeable that the DBA results are considerably below the level of the other three models. When predicting arousal, the models tend to overfit, while underfitting can be observed for the prediction of valence. This was also found in [24, 43, 47] and is possibly due to the chosen data split, which is speaker-independent, hence leading to imbalances in the label distribution [44]. For arousal, the results without normalisation are slightly stronger on the development set. On the test set, the overfitting gap for EWE and RAAW decreases by at least by .07 CCC with the application of the pre-smoothing filter. For valence, the results without the pre-smoothing filter are also slightly better on the development set, with the exception of GCTW . Pre-smoothing, however, produces atypically low results for DBA, which may indicate the sensitivity of the fusion method. With the other methods, the test result improved moderately.

3.1.2 Normalisation. Across all methods, the maximum deviation on test of the average results is low at .02 CCC for arousal and .04 CCC for valence (cf. Table 2). On an individual level, there are stronger differences, e. g., the results for the fusion of arousal with RAAW differ by more than .05 CCC on the development set and .1 CCC on the test set, with clear advantages for standardisation at the

Table 2: Results comparing different standardisation techniques (no pre-smoothing) on all annotation fusion techniques.

	Arousal						Valence					
	–		per video		per annotator		–		per video		per annotator	
	Devel.	Test	Devel.	Test	Devel.	Test	Devel.	Test	Devel.	Test	Devel.	Test
DBA	.2811	.1993	.3616	.2685	.2634	.2615	.3072	.2868	.3580	.4209	.2800	.3991
GCTW	.4969	.3558	.5175	.3207	.4809	.3481	.4353	.5345	.4256	.5170	.4394	.5594
EWE	.4750	.3563	.4923	.2746	.4410	.2513	.4452	.5551	.4479	.5193	.4476	.5614
RAAW	.4546	.2814	.4898	.3817	.4266	.2778	.4411	.5326	.4430	.5568	.4589	.5493
∅	.4269	.2982	.4653	.3114	.4030	.2847	.4072	.4773	.4186	.5035	.4065	.5173

video level. This is the case for most gold standard procedures in predicting arousal (development set). The results for the prediction of valence are predominantly highest when standardised on the annotator level.

3.2 Emotional Class Extraction

Clustering is by nature an unsupervised machine learning process, and so, human monitoring of the found class clusters ensures they are based on meaningful patterns. The MuSe-Toolbox provides a number of tools for this purpose. After each clustering outcome, detailed profiling is carried out, which contains statistics, e. g., mean and standard deviation, as well as visualisations of the obtained clusters. Figure 4 summarises these: a) shows a correlation between each feature and the cluster classes. This aids identification of influential features. b) offers a visual interpretation of the clustered features through dimension reduction. c) provides an overview of the degrees of influence for individual features in the entire cluster class, ordered by the overall importance (distance from the average value across all classes), while d) shows the statistical (normalised) distribution of a single feature per class.

The outcome of clustering is highly dependent on the dataset, and specifically the distribution of underlying emotional annotations. For this reason, it is difficult to generalise the current findings. In the following, we summarise a few general tendencies that we observe from current experiments.

For this, we run experiments applying k-means , fuzzy c-means , GMM, and agglomerative clustering on MuSe-CaR . For the input features, we select four different feature sets: distribution-based features set_{basic}[6], time-series features as in set_{change}[7], $set_{ext.}$[8], and a very large feature set set_{large}[9]. We further explore the reducing the dimensions before the clustering setting the PCA parameter to {None, 2, 5}, and specify the number of clusters to {3, 5}.

We defined one criterion of a fruitful outcome, i. e., if the cluster measures achieve optimal results (cf. Section 2.5.5 for difficulties). Furthermore, the identification of distinct cluster characteristics and a similar size of the classes may express optimal clusters. The experiments show that the composition of the features has a major influence on achieving the desired results. The features describing the distribution (set_{basic}) achieve slightly better results in terms of clustering measures than the feature set describing changes over time set_{change}. However, the latter seems to capture specific clusters very well, which is expressed by a small set of features (cf. Figure 4c) that stands out strongly from the average characteristics of these across all clusters. Mixing these two feature sets to the $set_{ext.}$ leads to the most evenly distributed class sizes. We recommend experimenting with the two general feature types and compiling your own set of reliable features for a given dataset, depending on your criteria and results obtained.

Regarding the class distribution, in 9 out of 96 setups created, at least one cluster does not cover enough percentage of the total amount of data points to fulfil our class-size-by-chance threshold of 25 %. With an increasing number of clusters (above five), all algorithms tend to split up existing smaller class clusters into even smaller ones, making it more likely to violate the class size rule. This behaviour occurs regardless of the feature set used.

In our feature reduction experiments, brute force was used to determine the best number of components. It showed that almost all clustering metrics except S_dbw perform better when a two-component PCA is used before clustering. However, in terms of the ability to predict the generated class clusters, in our case, five components is the better choice (by chance level vs maximum result). Another decisive aspect in this process is the size (and types) of the feature sets to use for dimension reduction. Prediction results obtained by using this process can be found in [41].

Finally, we find two other high impact aspects noteworthy: the segment length and the data basis for clustering. Regarding the segment length, the time series features (e. g., long strike below mean) are sensitive to the length of the segment compared to the features that only describe the distribution (e. g., quantile). If segments of varying length are given, it is recommended to adjust the length of the segments if possible and to convert the features by length from an absolute to a relative value corresponding to the length of the segment, avoiding the creation of meaningless classes. For the affected features implemented in this toolkit, the normalisation by length is already performed by default.

Depending on the partitioning of the dataset, i. e., the homogeneity between training, development, and test partitions, a clustering algorithm can generate completely different cluster classes. If the tool is used in the sense of an end-to-end process, where first the continuous signals are predicted and then a transformation into classes is automatically performed by a pre-trained clustering model, the exclusive use of the training dataset is advisable to test the method under real conditions. If it is a one-off process where suitable discrete classes are to be found for a given continuous annotation, the extraction can also be carried out on all data.

Of further note, we have found that using DBSCAN[10] for this task is less optimal. First, the class size threshold must be disabled because at least one resulting class does not meet the minimum size (e. g., the noise cluster). Second, the algorithm tends to produce a very low (1-2) or very high number of classes (up to 300).

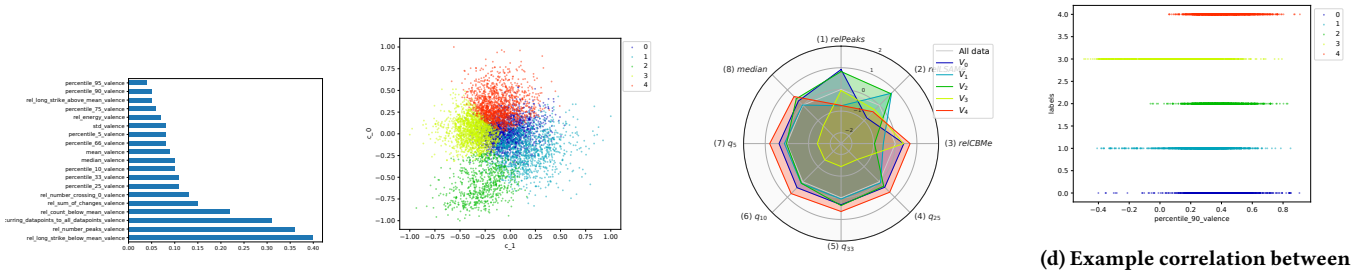
[6] mean, median, std., $q_{\{5,10,25,33,66,75,90,95\}}$

[7] std., rel. energy, rel. sum of changes, rel. number peaks, rel. long strike below mean, rel. long strike above mean, rel. count below mean

[8] $set_{basic} \cup set_{change}$ + rel. number crossing 0, percentage of reoccurring data points to all data points

[9] $set_{ext.}$ + skewness, kurtosis, mean abs. change, mean change, mean second derivative central, and the first and last location of the minimum and maximum, respectively

[10] DBSCAN parameters: $\epsilon = \{0.01, 0.05, 0.1, 0.25\}$; min_samples=3, 5; PCA={None, 2, 5}



(a) Absolute correlation between the labels and all other features.

(b) Cluster classes visualisation after post-PCA.

(c) Distinctive features across all classes.

(d) Example correlation between the labels and the 90th percentile.

Figure 4: Exemplary visualisation capabilities of MuSe-Toolbox for the class extraction process.

4 CONCLUSIONS

In this paper, we introduced the MuSe-Toolbox – a novel annotation toolkit for creating continuous and discrete gold standards. It provides capabilities to compare different fusion strategies of continuous annotations to a gold standard as well as simplify this gold standard to classes by extracting and clustering temporary and local signal characteristics. Hence, we provided a unified way to create regression and classification targets for emotion recognition. Furthermore, we introduced RAAW combining the annotation alignment on every time step and intelligently weighting of the individual annotation. Finally, important configuration parameter were highlighted in our series of experiments to which illustrated the toolkit's capabilities on the MuSe-CaR dataset. In the future, we plan to add further functionality, such as extending the dimension reduction to T-SNE and LDA.

REFERENCES

[1] Theodore W Anderson. 1958. *An introduction to multivariate statistical analysis.* Technical Report.

[2] Mia Atcheson, Vidhyasaharan Sethu, and Julien Epps. 2018. Demonstrating and Modelling Systematic Time-varying Annotator Disagreement in Continuous Emotion Annotation. In *Interspeech*. 3668–3672.

[3] James C Bezdek, Robert Ehrlich, and William Full. 1984. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences* 10, 2-3 (1984), 191–203.

[4] Brandon M Booth, Karel Mundnich, and Shrikanth S Narayanan. 2018. A novel method for human bias correction of continuous-time annotations. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3091–3095.

[5] Ricardo J. G. B. "Campello, Davoud Moulavi, and Joerg Sander. 2013. "Density-Based Clustering Based on Hierarchical Density Estimates. In *"Proceedings of the 17th Pacific-Asia Conference (PAKDD 13)*, Vol. 2. "Springer Berlin Heidelberg", 160–172.

[6] Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W Kempa-Liehr. 2018. Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). *Neurocomputing* 307 (2018), 72–77.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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: Human Language Technologies*. 4171–4186.

[8] David P Doane and Lori E Seward. 2011. Measuring skewness: a forgotten statistic? *Journal of Statistics Education* 19, 2 (2011).

[9] J. C. Dunn†. 1974. Well-Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics* 4, 1 (1974), 95–104. https://doi.org/10.1080/01969727408546059

[10] Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion* 6, 3-4 (1992), 169–200.

[11] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD-96 Proceedings*. AAAI, 226–231.

[12] Germain Forestier, François Petitjean, Hoang Anh Dau, Geoffrey I Webb, and Eamonn Keogh. 2017. Generating synthetic time series to augment sparse datasets. In *Data Mining (ICDM), 2017 IEEE International Conference on*. IEEE, 865–870.

[13] Changzeng Fu, Jiaqi Shi, Chaoran Liu, Carlos Toshinori Ishi, and Hiroshi Ishiguro. 2020. AAEC: An Adversarial Autoencoder-based Classifier for Audio Emotion Recognition. In *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*. 45–51.

[14] Michael Grimm and Kristian Kroschel. 2005. Evaluation of natural emotions using self assessment manikins. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005*. IEEE, 381–385.

[15] Simone Hantke, Erik Marchi, and Björn Schuller. 2016. Introducing the weighted trustability evaluator for crowdsourcing exemplified by speaker likability classification. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2156–2161.

[16] J. A. Hartigan and M. A. Wong. 1979. Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28, 1 (1979), 100–108. https://doi.org/10.2307/2346830

[17] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 131–135.

[18] Holger Hoffmann, Andreas Scheck, Timo Schuster, Steffen Walter, Kerstin Limbrecht, Harald Traue, and Henrik Kessler. 2012. Mapping discrete emotions into the dimensional space: An empirical approach. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 3316–3320. https://doi.org/10.1109/ICSMC.2012.6378303

[19] Florian Hönig, Anton Batliner, Karl Weilhammer, and Elmar Nöth. 2010. How many labellers? Modelling inter-labeller agreement and system performance for the automatic assessment of non-native prosody. In *Second Language Studies: Acquisition, Learning, Education and Technology*.

[20] Manju Kaushik and Bhawana Mathur. 2014. Comparative Study of K-Means and Hierarchical Clustering Techniques. *International Journal of Software and Hardware Research in Engineering* 2 (2014), 93–98.

[21] T. Kohonen. 1990. The self-organizing map. *Proc. IEEE* 78, 9 (1990), 1464–1480. https://doi.org/10.1109/5.58325

[22] Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Antoine Toisoul, Bjoern W Schuller, et al. 2019. SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

[23] Cyril Laurier, Mohamed Sordo, Joan Serra, and Perfecto Herrera. 2009. Music Mood Representations from Social Tags. In *International Society for Music Information Retrieval (ISMIR) Conference*. 381–386.

[24] Ruichen Li, Jinming Zhao, Jingwen Hu, Shuai Guo, and Qin Jin. 2020. Multi-modal Fusion for Video Sentiment Analysis. In *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop*. 19–25.

[25] Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, and Junjie Wu. 2010. Understanding of Internal Clustering Validation Measures. In *IEEE International Conference on Data Mining*. IEEE, 911–916. https://doi.org/10.1109/icdm.2010.35

[26] Luz Martinez-Lucas, Mohammed Abdelwahab, and Carlos Busso. 2020. The MSP-Conversation Corpus. *Proceedings of the INTERSPEECH 2020* (2020), 1823–1827.

[27] Mihalis A Nicolaou, Vladimir Pavlovic, and Maja Pantic. 2014. Dynamic probabilistic cca for analysis of affective behavior and fusion of continuous annotations. *IEEE transactions on pattern analysis and machine intelligence* 36, 7 (2014), 1299–1311.

[28] Girish Palshikar et al. 2009. Simple algorithms for peak detection in time-series. In *Proceedings of the 1st International Conference on Advanced Data Analysis, Business Analytics and Intelligence*, Vol. 122.

[29] François Petitjean, Germain Forestier, Geoffrey I Webb, Ann E Nicholson, Yanping Chen, and Eamonn Keogh. 2014. Dynamic time warping averaging of time series allows faster and more accurate classification. In *Data Mining (ICDM), 2014 IEEE*

International Conference on. IEEE, 470–479.

[30] François Petitjean, Alain Ketterlin, and Pierre Gançarski. 2011. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition* 44, 3 (2011), 678–693.

[31] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. 2017. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge (EmotiW)*. 3–9.

[32] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 1–8.

[33] James A Russell. 1980. A Circumplex Model of Affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161–1178.

[34] Hesam Sagha, Maximilian Schmitt, Filip Povolny, Andreas Giefer, and Björn Schuller. 2017. Predicting the popularity of a talk-show based on its emotional speech content before publication. In *Proceedings 3rd International Workshop on Affective Social Multimedia Computing, Conference of the International Speech Communication Association (INTERSPEECH) Satellite Workshop*. ISCA.

[35] Abraham Savitzky and Marcel JE Golay. 1964. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry* 36, 8 (1964), 1627–1639.

[36] Björn W Schuller. 2013. *Intelligent audio analysis*. Springer.

[37] Björn W Schuller, Anton Batliner, Christian Bergler, Eva-Maria Messner, Antonia Hamilton, Shahin Amiriparian, Alice Baird, Georgios Rizos, Maximilian Schmitt, Lukas Stappen, et al. 2020. The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks. *Proceedings Conference of the International Speech Communication Association (INTERSPEECH)* (2020).

[38] David Schultz and Brijnesh Jain. 2018. Nonsmooth analysis and subgradient methods for averaging in dynamic time warping spaces. *Pattern Recognition* 74 (2018), 340–358.

[39] Narendra Sharma, Aman Bajpai, and Ratnesh Litoriya. 2012. Comparison the various clustering algorithms of weka tools. *International Journal of Emerging Technology and Advanced Engineering* 2, 5 (2012), 73–80.

[40] Lukas Stappen, Alice Baird, Erik Cambria, and Björn W Schuller. 2021. Sentiment analysis and topic recognition in video transcriptions. *IEEE Intelligent Systems* 36, 2 (2021), 88–95.

[41] Lukas Stappen, Alice Baird, Lukas Christ, Lea Schumann, Benjamin Sertolli, Eva-Maria Messner, Erik Cambria, Guoying Zhao, and Björn W Schuller. 2021. The MuSe 2021 Multimodal Sentiment Analysis Challenge: Sentiment, Emotion, Physiological-Emotion, and Stress. In *Proceedings of the 2nd Multimodal Sentiment Analysis Challenge, co-located with the 29th ACM International Conference on Multimedia (ACMMM)*. ACM, Changu, China.

[42] Lukas Stappen, Alice Baird, Michelle Lienhart, Annalena Bätz, and Björn Schuller. 2021. An Estimation of Online Video User Engagement from Features of Continuous Emotions. *arXiv preprint arXiv:2105.01633* (2021).

[43] Lukas Stappen, Alice Baird, Georgios Rizos, Panagiotis Tzirakis, Xinchen Du, Felix Hafner, Lea Schumann, Adria Mallol-Ragolta, Bjoern W. Schuller, Iulia Lefter, Erik Cambria, and Ioannis Kompatsiaris. 2020. 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 (MuSe'20)*. ACM, New York, NY, USA, 35–44.

[44] Lukas Stappen, Alice Baird, Lea Schumann, and Björn Schuller. 2021. The Multimodal Sentiment Analysis in Car Reviews (MuSe-CaR) Dataset: Collection, Insights and Improvements. *IEEE Transactions on Affective Computing (Early Access)* (June 2021). https://doi.org/10.1109/TAFFC.2021.3097002

[45] Lukas Stappen, Gerhard Hagerer, Björn W Schuller, and Georg Groh. 2021. Unsupervised Graph-based Topic Modeling from Video Transcriptions. *arXiv preprint arXiv:2105.01466* (2021).

[46] Lukas Stappen, Björn W Schuller, Iulia Lefter, Erik Cambria, and Ioannis Kompatsiaris. 2020. Summary of MuSe 2020: Multimodal Sentiment Analysis, Emotion-target Engagement and Trustworthiness Detection in Real-life Media. In *28th ACM International Conference on Multimedia (ACMMM)*. ACM.

[47] Licai Sun, Zheng Lian, Jianhua Tao, Bin Liu, and Mingyue Niu. 2020. Multimodal 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 (MuSe), co-located with the 28th ACM International Conference on Multimedia (ACMMM)*. Virtual, 27–34.

[48] Nattapong Thammasan, Ken-ichi Fukui, and Masayuki Numao. 2016. An investigation of annotation smoothing for eeg-based continuous music-emotion recognition. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 003323–003328.

[49] Gerard V Trunk. 1979. A problem of dimensionality: A simple example. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 3 (1979), 306–307.

[50] J. Vesanto and E. Alhoniemi. 2000. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks* 11, 3 (2000), 586–600. https://doi.org/10.1109/72.846731

[51] Chen Wang, Phil Lopes, Thierry Pun, and Guillaume Chanel. 2018. 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*. 73–81.

[52] Peter H Westfall. 2014. Kurtosis as peakedness. *The American Statistician* 68, 3 (2014), 191–195.

[53] Mohammed J. Zaki and Wagner Meira. 2014. *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, New York, NY, 187–191.

[54] Mohammed J. Zaki and Wagner Meira. 2014. *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, New York, NY, 444–445.

[55] Mohammed J. Zaki and Wagner Meira. 2014. *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, New York, NY, 444.

[56] Feng Zhou and Fernando De la Torre. 2015. Generalized canonical time warping. *IEEE transactions on pattern analysis and machine intelligence* 38, 2 (2015), 279–294.

[57] Feng Zhou and Fernando Torre. 2009. Canonical time warping for alignment of human behavior. *Advances in neural information processing systems* 22 (2009), 2286–2294.